International Journal on Advances in Systems and Measurements

2026 vol. 19 nr. 1&2

The *International Journal on Advances in Systems and Measurements* is published by IARIA.
ISSN: 1942-261x
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Responsibility for the contents rests upon the authors and not upon IARIA, nor on IARIA volunteers, staff, or contractors.

IARIA is the owner of the publication and of editorial aspects. IARIA reserves the right to update the content for quality improvements.

Abstracting is permitted with credit to the source. Libraries are permitted to photocopy or print, providing the reference is mentioned and that the resulting material is made available at no cost.

Reference should mention:

International Journal on Advances in Systems and Measurements, issn 1942-261x
vol. 19, no. 1 & 2, year 2026, http://www.iariajournals.org/systems_and_measurements/

The copyright for each included paper belongs to the authors. Republishing of same material, by authors or persons or organizations, is not allowed. Reprint rights can be granted by IARIA or by the authors, and must include proper reference.

Reference to an article in the journal is as follows:

<Author list>, "<Article title>"
International Journal on Advances in Systems and Measurements, issn 1942-261x
vol. 19, no. 1 & 2, year 2026, http://www.iariajournals.org/systems_and_measurements/

IARIA journals are made available for free, proving the appropriate references are made when their content is used.

Sponsored by IARIA
www.iaria.org

Copyright © 2026 IARIA

International Journal on Advances in Systems and Measurements
Volume 19, Number 1 & 2, 2026

Editors-in-Chief

Constantin Paleologu, University "Politehnica" of Bucharest, Romania
Sergey Y. Yurish, IFSA, Spain

Editorial Board

Nebojsa Bacanin, Singidunum University, Serbia
Chaity Banerjee, University of Alabama in Huntsville, USA
Robert Bestak, Czech Technical University in Prague, Czech Republic
Michał Borecki, Warsaw University of Technology, Poland
Vítor Carvalho, 2Ai | School of Technology | IPCA & Algoritmi Research Center | Minho University, Portugal
Paulo E. Cruvinel, Brazilian Corporation for Agricultural Research (Embrapa), Brazil
Miguel Franklin, Federal University of Ceara, Brazil
Mounir Gaidi, University of Sharjah, UAE
Eva Gescheidtova, Brno university of Brno, Czech Republic
Franca Giannini, CNR - Istituto di Matematica Applicata e Tecnologie Informatiche "Enrico Magenes", Italy
Terje Jensen, Telenor, Norway
Wooseong Kim, Gachon University, South Korea
Dragana Krstic, University of Nis, Serbia
Andrew Kusiak, The University of Iowa, USA
Diego Liberati, CNR-IEIIT, Italy
D. Manivannan, University of Kentucky, USA
Stefano Mariani, Politecnico di Milano, Italy
Constantin Paleologu, National University of Science and Technology Politehnica Bucharest, Romania
Paulo Pinto, Universidade Nova de Lisboa, Portugal
R. N. Ponnalagu, BITS Pilani Hyderabad campus, India
Leon Reznik, Rochester Institute of Technology, USA
Gerasimos Rigatos, Unit of Industrial Automation - Industrial Systems Institute, Greece
Claus-Peter Rückemann, Universität Münster / DIMF / Leibniz Universität Hannover, Germany
Subhash Saini, NASA, USA
Adérito Seixas, Escola Superior de Saúde Fernando Pessoa, Porto, Portugal
V. R. Singh, National Physical Laboratory (NPL), New Delhi, India
Miroslav Velev, Aries Design Automation, USA
Manuela Vieira, Instituto Superior de Engenharia de Lisboa (ISEL), Portugal
Xianzhi Wang, University of Technology Sydney, Australia
Kaidi Wu, College of Mechanical Engineering | Yangzhou University, China
Linda Yang, University of Portsmouth, UK
Sergey Y. Yurish, IFSA, Spain
Daniele Zonta, University of Trento / National Research Council, Italy

International Journal on Advances in Systems and Measurements
Volume 19, Numbers 1 & 2, 2026

CONTENTS

pages: 1 - 10
News-based Economic Report Generation for Business Sentiment Index
Kazuhiro Seki, Konan University, Japan

pages: 11 - 20
Visual and Graph-Augmented Evaluation of RAG Systems – Improving Retrieval, Re-Ranking, and Data Quality with the pRAG Framework
Alexander Kress, IU International University of Applied Sciences, Deutschland
Alexander Lawall, IU International University of Applied Sciences, Deutschland
Thomas Zöller, IU International University of Applied Sciences, Deutschland

pages: 21 - 34
From Access to Affirmation: How Technology Exposure and Family Support Shape Confidence in Computer Science
Claudia Hess, IU International University of Applied Sciences, Germany
Sibylle Kunz, IU International University of Applied Sciences, Germany
Cornelia Heinisch, IU International University of Applied Sciences, Germany
Adrienne Steffen, IU International University of Applied Sciences, Germany

pages: 35 - 48
Assessing the Impact of Artificial Intelligence on Job and Task Displacement: A Cross Continental Analysis of Four Economic Sectors
Paul Mupfiga, Midlands State University, Zimbabwe
Dzinaishe Mpini, Midlands State University, Zimbabwe
Maxmillan Giyane, Midlands State University, Zimbabwe

pages: 49 - 60
Estimating the Risk of Failing Physics Courses Through Machine Learning and Monte Carlo Simulation
Isaac Caicedo-Castro, University of Córdoba, Colombia
Rubby Castro-Púche, University of Córdoba, Colombia
Samir Castaño-Rivera, University of Córdoba, Colombia

pages: 61 - 78
Comparative Bias Analysis of Recommendation Algorithms for YouTube Shorts and Long-Form Videos
Nitin Agarwal, COSMOS Research Center, UA-Little Rock; ICSI, University of California - Berkeley, USA

pages: 79 - 88
Metrics, not Myths: The True Price and Value of Productivity in the Era of AI
Sibylle Kunz, IU International University, Germany
Claudia Hess, IU International University, Germany

pages: 89 - 102
Technique Recognition and Arrangement Across Eastern and Western Strings: A Unified CNN Framework for Bowed and Plucked Traditions

Xinyuan Zhu, The Chinese University of Hong Kong, Shenzhen, China
Clement Leung, The Chinese University of Hong Kong, Shenzhen, China

pages: 103 - 116
Modeling and Simulation of Autonomous Mobile Robot Fleets for Automated Baggage Handling in Airports 4.0: A Fuzzy Agent Approach
Alain-Jérôme Fougères, ECAM Louis de Broglie, France
Ouzna Oukacha, ECAM Louis de Broglie, France
Moïse Djoko-Kouam, ECAM Louis de Broglie, France
Egon Ostrosi, UTBM, France

pages: 117 - 129
Target Encoding for Object Orientation Prediction in Smart Manufacturing
Antonio Gambale, School of Computing, Engineering & Intelligent Systems, Ulster University, Northern Ireland
Sonya Coleman, School of Computing, Engineering & Intelligent Systems, Ulster University,, Northern Ireland
Dermot Kerr, School of Computing, Engineering & Intelligent Systems, Ulster University,, Northern Ireland
Philip Vance, School of Computing, Engineering & Intelligent Systems, Ulster University,, Northern Ireland
Emmett Kerr, Dept. of Electronic & Mechanical Engineering, Atlantic Technological University, Ireland
Cornelia Fermüller, Institute for Advanced Computer Studies, University of Maryland,, USA
Yiannis Aloimonos, Institute for Advanced Computer Studies, University of Maryland,, USA

pages: 130 - 146
Representing Human Capabilities in Occupational Databases: An Embedding-Based Audit of O*NET, ESCO and Vocational Standards
Ilona Erika Negro, Alma Mater Europaea, Maribor, Slovenia

pages: 147 - 159
A Modular and Reproducible Framework Based on Plugin Composition for Simulating Cloud Systems
Rubén Luque, University of Oviedo, Spain
Jose Luis Diaz, University of Oviedo, Spain
Joaquín Entrialgo, Polytechnical University of Madrid, Spain
Rubén Usamentiaga, University of Oviedo, Spain

pages: 160 - 177
Memory-Driven Dynamics and Repeated Patterns in Online Toxicity: A Fractal-Fractional SEIQR Modeling Approach
Nitin Agarwal, COSMOS Research Center, UA-Little Rock; ICSI, University of California - Berkeley, USA

pages: 178 - 189
Developing a Framework for Circular Economy Products and Processes within the Construction Industry
Alan Martin Redmond, INCOSE/IEEE, France

pages: 190 - 208
Assessment of Sensor Technologies for Detecting Slurry Emissions in Irish Agriculture
Jacqueline Humphries, Technological University of the Shannon (Midwest), Ireland
Abimbola Ikoyi, Technological University of the Shannon (Midwest), Ireland

pages: 209 - 219
From Backscattering Models to Neural Networks: Simulation-Driven Acoustic Identification of Mesopelagic Organisms
Sander Andre Berg Marx, University of Bergen, Norway

Babak Khodabandeloo, Institute of Marine Research (IMR), Norway
Ketil Malde, Institute of Marine Research (IMR), Norway

pages: 220 - 230
Toward Socially-Aware Explainability for Autonomous Robots
Khatina Sari, Hochschule Bonn-Rhein-Sieg, Germany
Paul Gerhard Plöger, Hochschule Bonn-Rhein-Sieg, Germany
Alex Mitrevski, Hochschule Bonn-Rhein-Sieg, Germany

pages: 231 - 243
Energy Management of Stand-Alone Rural Micro-Grid Application based PV – Batteries–System in Guinea
Mohamed Lamine Touré, GREAH Laboratory, University Le Havre Normandie, France, France
Mamadou-Baïlo Camara, GREAH Laboratory, University Le Havre Normandie, France, France
Fahad Maoulida, GREAH Laboratory, University Le Havre Normandie, France, France
Alireza Payman, GREAH Laboratory, University Le Havre Normandie, France, France
Brayima Dakyo, GREAH Laboratory, University Le Havre Normandie, France, France

pages: 244 - 253
Investigation of Trap-Assisted Tunneling Currents in Capacitive Micromachined Ultrasonic Transducers (CMUT) with Silicon Nitride Membranes
Chukwuebuka Okoro, University of Windsor, Canada
Sazzadur Chowdhury, University of Windsor, Canada

pages: 254 - 274
Adaptive Forward Error Correcting Decoders: A Survey of Architectures and Implementations
Sergei Sawitzki, FH Wedel (University of Applied Sciences), Germany

News-based Economic Report Generation for Business Sentiment Index

Kazuhiro Seki
Faculty of Intelligence and Informatics
8-9-1 Okamoto, Higashinada, Kobe
Hyogo, Japan
e-mail: seki@konan-u.ac.jp

***Abstract*—Business sentiment indices (BSIs) derived from surveys play a crucial role in economic decision-making but suffer from high costs and limited timeliness. While recent work has shown that BSIs can be estimated from news articles more efficiently, such indices lack interpretability, making it difficult to understand the economic factors behind their values. This paper investigates methods for generating concise and interpretable monthly economic reports to explain news-based BSIs by leveraging large language models (LLMs). We compare three methods: direct one-shot report generation, one-shot report generation with topic categorization via prompting, and a topic-sentiment-driven method that integrates neural topic modeling, sentiment aggregation, and multi-stage LLM-based report generation. Experiments on one year of Japanese web news demonstrate that the topic-aware prompting method produces the most distinctive topics but yields less accurate economic impact scores than the topic-modeling method, which reduces API cost but is slower, highlighting trade-offs among report quality, topic diversity, API cost, and processing time.**

Keywords-text summarization; large language models; sentiment analysis; topic analysis.

I. Introduction

Recent advances in natural language processing have enabled the automated generation of economic reports to assist in interpreting business sentiment indices (BSIs) [1]. Business sentiment indices (BSIs) serve as key references for a wide range of economic activities, including monetary-policy decisions, corporate production planning, and investment decisions by institutional and individual investors. In Japan, widely used BSIs include the Current and Future Assessment Diffusion Index of the Economy Watchers Survey [2] by the Japanese government and the Diffusion Index of the Tankan Survey [3] by the Bank of Japan. However, the surveys underlying the computation of these indices require substantial time, labor, and financial resources.

For example, in the case of the Current and Future Assessment Diffusion Index of the Economy Watchers Survey (hereafter, the Economy Watchers DI), regional survey organizations administer questionnaires to 2,050 individuals, such as retail store owners and taxi drivers, who are well positioned to observe local economic conditions. These surveys cover 12 regions of Japan, and the collected regional results are aggregated and analyzed by a central coordinating organization to compute the final index values. Because such surveys take considerable time to conduct and process, the Economy Watchers DI is released only once a month and the Tankan Survey DI only once a quarter, which limits their timeliness.

To address these limitations, Seki et al. [4] developed a framework to estimate a BSI from web news articles automatically collected from the Web. While this framework showed a strong correlation with a traditional survey-based index—specifically, the Economy Watchers DI—and offered significant improvements in timeliness and cost-efficiency, it lacked interpretability. That is, it is unclear why a particular BSI value was obtained.

To bridge this gap, the present study extends the preliminary work by Seki [1]. We investigate methods that leverage sentiment and topic analyses to capture sentiment distribution and topic structure, followed by text summarization to synthesize relevant economic events from news articles. Experiments on news articles collected in 2025 demonstrate promising results, producing concise and easily interpretable summaries of the economic situation for each period.

The rest of this paper is structured as follows: Section II reviews related work. Section III describes our proposed approach, Section IV presents experimental results on real-world news data, Section V introduces the functionality to display economic reports with our BSI nowcasting system, and Section VI concludes with a summary and possible future work.

II. Related Work

This section summarizes related work on forecasting economic indices from textual data and on economic report generation.

There have been a number of attempts to forecast or nowcast economic indices, including BSIs, from textual data [4]–[6]. For example, Seki et al. [4] nowcast BSI using daily Japanese newspaper articles. They used newspaper sentences as input to a regression model to estimate the economic impact score of each sentence. The regression model was a BERT model fine-tuned on the Economy Watchers Survey data in Japan. The economic impact scores were averaged monthly to obtain the BSI for each month. The resulting BSI values showed a strong correlation ($R = 0.937$) with an existing survey-based BSI. Seki [7] also examined the influence of newspaper-specific biases on BSI computation and applied a dynamic factor model [8] to obtain an unbiased BSI using multiple newspapers.

In other languages and countries, Bieri et al. [9] constructed a news-based sentiment indicator from multilingual newspapers (German- and French-language sources) and examined its relationship with existing survey-based indicators. Kalamara et al. [10] demonstrated improvements in forecasting GDP and

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

related macroeconomic variables using UK newspaper texts, while Ashwin et al. [11] showed that sentiment derived from newspaper text can contribute to improved nowcasting of euro area GDP. Furthermore, Barbaglia et al. [12] conducted aspect-based, fine-grained sentiment extraction and reported enhanced predictive performance for multiple macroeconomic variables.

Generating an economic report from multiple news articles can be seen as multi-document summarization (MDS), a topic with extensive literature; thus, the following focuses on economic-report generation and then touches upon topic-guided MDS. Hu et al. [13] proposed a multi-edit Bi-LSTM neural network that first learns an outline for the input news and then generates financial reports from the learned outline. Yan et al. [14] focused on topic accuracy and proposed a variational topic inference model that improves topic faithfulness in report generation by incorporating latent variables. Zhang et al. [15] also proposed a method that improves faithfulness in multi-document summarization by integrating relationships among entities and events extracted from multiple documents into a cross-document information extraction graph and incorporating this structural information into the summarization process. Furthermore, given recent advances in large language models (LLMs), there is growing interest in using LLMs for report generation. For instance, Shukla et al. [16] first identified crucial narrative sections within input financial reports and then provided the extracted sections to LLMs for summarization. Yang et al. [17] conducted a comparative study of three LLMs (GLM-4, Mistral-NeMo, and LLaMA3.1), focusing on their effectiveness in generating automated financial reports. Assis et al. [18] used LLMs to automatically generate market commentary from corporate disclosure documents for the Brazilian financial market.

For MDS, the use of topic modeling has also been widely investigated. Cui and Hu [19] proposed an MDS framework that constructs a heterogeneous graph consisting of word, topic, and document nodes, and integrates latent topics estimated by a neural topic model into summary generation as global semantics that bridge across documents. Their method jointly learns summarization and topic estimation, and introduces an inconsistency loss to encourage alignment between the topic distributions of the input documents and the generated summaries. Dong et al. [20] formulated MDS as a two-stage process: subtopic detection based on content attention in the first stage, followed by a submodular optimization problem in the second stage that simultaneously minimizes inter-sentence distance and maximizes topic diversity. Li et al. [21] proposed topic-guided reinforcement learning, which extracts topic phrases or labels to construct a topic-based reward for preserving topic relevance in MDS, and optimizes the summarization model using the Group Relative Policy Optimization (GRPO) framework. The reward defines topic alignment in terms of both coverage and precision.

The present work differs from the topic-centered text generation studies summarized above in that it aims to generate an economic report guided not only by topics but also by their economic impacts.

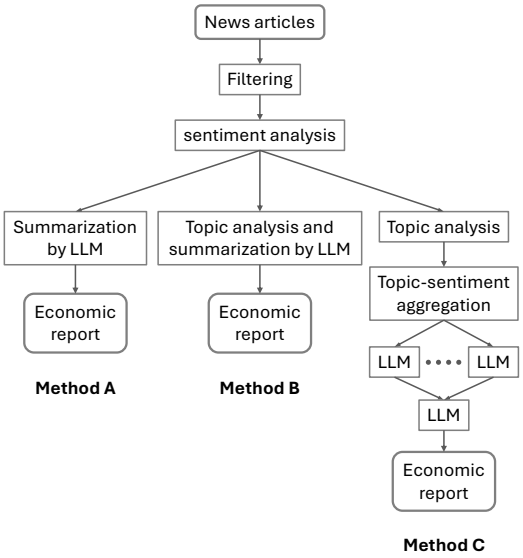


Figure 1. Overview of the three methods.

III. Proposed Approach

A. Overview

We devised three methods to address the limited interpretability of a news-based BSI by generating an economic report from the news articles used to compute the BSI. Figure 1 depicts the processing flows of the three methods. The first method (Method A) generates a report directly from the input news articles using a large language model (LLM). The second method (Method B) also uses an LLM, but first categorizes news articles into representative topics and then summarizes a limited number of major topics using a single detailed prompt. The third method (Method C) employs a separate topic analysis model and aggregates the outputs of topic and sentiment analyses; these outputs are then provided to LLMs to generate topic-wise summaries, which are finally combined by another LLM to produce an economic report.

In the following, we first summarize how BSI is computed, then describe the input data for the three methods, and lastly detail each method.

B. BSI Computation

For BSI computation, we basically follow the framework developed by Seki et al. [4]. The primary difference is that Seki et al. used a fine-tuned BERT model, whereas we adopt ModernBERT [22] as a base model. For completeness, the procedure to compute BSI is summarized as follows:

1) Automatically collect Japanese web news articles from the Sankei Newspaper website [23].
2) Split each news article into sentences based on Japanese periods.
3) Filter out irrelevant news sentences by applying an outlier detection model based on a one-class SVM [24] trained on Economy Watchers Survey data.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

```
The following are headlines of multiple news articles arranged in
chronological order. Based on sentiment analysis, these texts have
been identified as having a significant impact on economic
conditions. Taking a comprehensive view of these texts, generate an
economic report from the perspective of their impact on the economy.

### Output Requirements
- The output must be in JSON format only.
- Do not include any additional explanatory text, extra keys, or
Markdown tags.
- The value of "overall_summary" must be a string written in
Japanese, starting with "Recent events:" and using the formal
declarative style.
- The length of "overall_summary" should be approximately up to 300
Japanese characters.

### Output Format
{
  "overall_summary": "Recent events: ..."
}

### News Texts
{N} lines of news headlines with sentiment scores
...
```

Figure 2. English-translated prompt used for Method A.

4) Apply a sentiment analysis model to each news sentence to predict its business sentiment score. The model is a Japanese ModernBERT model fine-tuned on the Economy Watchers Survey data. If the model output is positive, the economic impact is considered positive, and vice versa.
5) Compute the monthly average of the sentiment scores as the BSI of the month.

C. Input Data

As input for generating an economic report, it would be straightforward to use the news sentences and their sentiment scores computed as in Section III-B. However, our preliminary experiment showed that news sentences were often too fragmented, lacking sufficient context to be analyzed correctly for economic report generation.

To mitigate the problem, we use news articles without splitting them into sentences and compute their sentiment scores. However, as one month of news articles is too long to be fed to an LLM, we use only the headlines with the sentiment scores computed for the corresponding articles as input to the three methods described below.

D. Method A: One-shot economic report generation

Method A relies entirely on an LLM prompt shown in Figure 2 to generate an economic report. The LLM is instructed to generate a report based on the N input news articles (headlines) and their sentiment scores. The N input news headlines are selected based on the absolute sentiment scores of the news articles computed in Section III-B. It should be mentioned that, unlike the other methods, this method does not provide topic information to be presented in our system (see Section V) and is tested only for comparison purposes.

E. Method B: One-shot economic report generation with topic categorization

Method B takes the same input news headlines as Method A but instructs an LLM to categorize input news under k topics and then to generate an economic report based on the most important $k'(< k)$ topics. The prompt is shown in Figure 3.

```
The following are headlines of multiple news articles arranged in
chronological order. Based on sentiment analysis, these texts have
been identified as having a significant impact on economic
conditions.

### Task
You must execute the following steps in the specified order.

1. Internally classify these texts into {k} topics based on their
content (note: these {k} topics themselves must not be output).
2. For each topic, calculate an "impact_score" by simply summing all
numerical values appearing at the end of the sentences belonging to
that topic (negative values indicate negative impact). The sign must
be preserved.
3. Select the top {k'} topics in descending order of the absolute
value of their impact scores.
4. For each selected topic, generate a summary from the perspective
of its impact on economic conditions.
5. Taking an overview of these {k'} topics, generate a single
integrated economic report that reflects their commonalities and
interrelationships. Avoid a mere enumeration of individual topics;
produce a summary that conveys the overall picture.

### Output Requirements
- The output must be in JSON format only.
- Do not include any additional explanatory text, extra keys, or
Markdown tags.
- The value of "overall_summary" must be a string written in
Japanese, starting with "Recent events:" and using a formal
declarative style.
- The length of "overall_summary" should be approximately up to 300
Japanese characters.
- "topics" must be an array with exactly {k'} elements.
- Each topic must be an object with the following keys:
  - "topic_id": a sequential integer starting from 1, ordered by
descending absolute value of impact_score
  - "label": a concise label of approximately 10–15 Japanese
characters representing the topic
  - "summary": a summary of the topic of up to approximately 300
Japanese characters
  - "impact_score": the calculated numerical impact score

### Output Format
{
  "overall_summary": "Recent events:...",
  "topics": [
    {
      "topic_id": 1,
      "label": "...",
      "summary": "...",
      "impact_score": ...
    },
    ...
    {
      "topic_id": {k'},
      "label": "...",
      "summary": "...",
      "impact_score": ...
    }
  ]
}

### News Texts
{N} lines of news headlines with sentiment scores
...
```

Figure 3. English-translated prompt used for Method B.

The intention behind Method B is to mimic the pipeline of Method C, described next, without an external topic model by using a detailed one-shot prompt to see how well it works compared with Method C.

F. Method C: Topic-sentiment-driven economic report generation

Method C integrates topic modeling and sentiment analysis to generate concise, topic-focused economic report using an LLM. The overall process consists of three main stages: topic modeling, topic-level sentiment aggregation, and LLM-based report generation. In the following topic-modeling context, we treat each news headline as a *document*.

First, all documents are tokenized using a custom Japanese tokenizer, which conducts morphological analysis and extracts compound nouns. The tokenized documents are then used as input to a neural topic model. We adopt FASTopic [25], a fast

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

```
The following sentences are news headlines arranged in chronological
order and have been classified into the same topic as a result of
topic analysis. The numerical value at the end of each sentence
indicates the strength of its impact on economic conditions
(negative values indicate a negative impact).

Your task consists of the following two items:
1. Generate a single concise topic label of approximately 10-15
Japanese characters that succinctly represents the overall news.
2. From the perspective of the impact on economic conditions,
produce a summary of the entire news in Japanese, written in a
formal declarative style, with a maximum length of approximately 300
Japanese characters.

Output the result only in the following JSON format, and do not
include any other Markdown tags, explanatory text, preambles, or
additional content.

Output Format:
{
  "label": "topic label",
  "summary": "summary focusing on the impact on economic conditions"
}

### News Texts
{N} lines of news headlines with topic-sentiment scores
...
```

Figure 4. English-translated first prompt to generate a topic summary.

```
The following is a report summarizing recent events that are
considered to have had a strong impact on economic conditions for
each of {k'} topics. Each topic is assigned a label, a summary, a
topic ID, and an impact score (negative values indicate a negative
impact).

### Task
Taking an overview of all the topic-specific reports below, generate
a single economic report that reflects their commonalities and
interrelationships. Do not simply list individual topics; instead,
produce an integrated summary that conveys the overall picture.

### Output Requirements
- The output must be in JSON format only.
- The only key allowed is "summary".
- Do not include any additional explanatory text, line breaks,
comments, extra keys, or Markdown tags.
- The value of "summary" must be a string written in Japanese,
starting with "Recent events:".
- The length of the summary should be approximately up to 300
Japanese characters.

### Output Format
{
  "summary": "Recent events: ..."
}

### Topic-wise Reports (JSON format)
{k'} topic-wise reports follows.
...
```

Figure 5. English-translated second prompt to generate an overall summary.

and stable topic model that reconstructs semantic relations from pretrained embeddings using optimal-transport regularization. FASTopic is reported to achieve higher performance in topic coherence [26] and topic diversity [27] than other models, such as LDA [28] and BERTopic [29]. Given a predefined number of topics k, the model is trained on the input documents using dense document embeddings. As output, the model produces (i) a topic-word distribution $p(\text{word}|\text{topic})$ for each topic and (ii) a document-topic distribution $p(\text{topic}|\text{document})$ representing the degree to which each document is associated with each topic.

Second, document-level sentiment scores are combined with the document-topic distribution to compute topic-level sentiment contributions. Specifically, for each document, its topic membership probabilities $p(t_j|d_i)$ are multiplied by its sentiment score s_i, yielding a topic-sentiment matrix, where i and j denote the indices of a document d and a topic t, respectively. By averaging these values over all documents $d_i \in D$, an aggregated topic-sentiment score $a_j = \frac{1}{|D|}\sum_i s_i \times p(t_j|d_i)$ is obtained for topic t_j, which indicates the overall economic impact of the topic. Topics are then ranked by the absolute magnitude of their aggregated topic-sentiment scores $A = (a_1, \ldots, a_k)$, allowing us to identify topics with the strongest positive or negative economic implications.

For each of the k' highly ranked topics among the k topics, documents are further sorted by the absolute value of their topic-sentiment contribution $s_i \times p(t_j|d_i)$. From this ordering, the top n documents are selected as the most influential textual evidence for that topic t_j, and their timestamps are retained to preserve temporal structure.

Next, an LLM is used to generate summaries for each of the k' economically influential topics. The n documents selected for each topic are arranged in chronological order and provided to the LLM as a single prompt shown in Figure 4.

Each document is accompanied by its topic-sentiment score $s_i \times p(t_j|d_i)$, indicating the strength and direction of its economic impact. The LLM is instructed to (i) generate a concise topic label and (ii) produce a short summary that focuses explicitly on the topic's economic implications.

Finally, the k' outputs of the LLM from the previous step are given to an LLM with a prompt to generate an overall economic report covering the major topics. Figure 5 shows the prompt.

IV. Evaluation

A. Experimental Settings

For development and evaluation of the proposed approach, we crawled Japanese news data from the Sankei Newspaper website from October 2024 to December 2025. Note that the full text of the articles is available only to paid subscribers, and our crawled data include only the headlines and the beginning of each article. The length of the crawled text varies across articles, but it is generally sufficient to understand their contents. The 2024 data (October to December) were used for development, and the 2025 data (January to December) were used for evaluation.

For the input data size, we set the number of news headlines (N) to 1,000, considering the LLM context size for Methods A and B and the burden of the evaluator. Note that the LLM context size is not a big issue for Method C since the first stage (i.e., topic analysis) identifies only relevant news articles for each topic but uses the same N for a fair comparison.

As the FASTopic embedding model for Method C, we used "paraphrase-multilingual-mpnet-base-v2". We set the number of topics for FASTopic (k) to 10, the number of topics presented in the final report (k') to 3, and the number of documents per topic (n) to 100. These parameters were set based on a preliminary experiment on the development data, which were also used to refine the prompts for all three methods. Note that k' can be chosen dynamically by setting a threshold on aggregated topic-sentiment scores A. In this study, however, we

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

fixed $k' = 3$ to directly compare the resulting topics identified by Methods B and C.

As an LLM, we adopted OpenAI's gpt4o-2024-08-06 and set the temperature and nucleus sampling parameters to 0 and 1, respectively, for better reproducibility. We also tested a newer model, gpt-5.2-2025-12-11, which sometimes generated awkward Japanese; therefore, it was not used in this experiment.

All experiments were conducted on a MacBook Pro equipped with an Apple M3 Pro chip (11-core CPU), 18 GB of unified memory, with GPU acceleration enabled via Apple's Metal Performance Shaders (MPS).

B. Evaluation Criteria

We compared the reports generated by the three methods in the following aspects:

- Manual evaluation of the k' identified topics. These aspects were evaluated for only Methods B and C as Method A was not designed to select k' topics.
  - *Economic Significance*: whether the selected topics represent the major economic issues of the month, in terms of including economically significant events, not omitting key topics, and avoiding the selection of minor issues as primary topics.
  - *Distinctiveness*: whether the selected topics are sufficiently distinct from one another, in terms of avoiding substantial content overlap and ensuring that each topic represents a clearly different issue.
  - *Impact Score Validity*: whether the assigned economic impact scores for the three topics are reasonable, in terms of the plausibility of their relative magnitudes within the same month based on the topic descriptions.
- Manual evaluation of the generated economic report.
  - *Factual Accuracy*: whether the generated economic report is factually accurate, in terms of not introducing information unsupported by the source news, correctly describing numerical values and events, and avoiding misleading or incorrect statements.
  - *Economic Adequacy*: whether the generated content is adequate as an economic report, in terms of factual consistency with the input topics and source news and the economic plausibility of the interpretation.
- Automatic evaluation of the k' identified topics. This aspect was not evaluated for Method A for the same reason as above.
  - *Diversity*: whether the k' selected topics are diverse, defined as:

$$Diversity = \frac{1}{{}_nC_2} \sum_{i<j} (1 - \cos(t_i, t_j)) \tag{1}$$

    where t_i and t_j denote the embedding vectors of the topic summaries for topics i and j ($i \leq k'$, $j \leq k'$), and $\cos(\cdot)$ returns the cosine similarity between the two vectors. The embedding vectors were obtained via the OpenAI API using the text-embedding-3-large model.
- API cost in U.S. dollars, based on the number of input and output tokens. At the time of writing, the cost of gpt-4o was \$2.5 and \$10 per 1 million input and output tokens, respectively.
- Processing time in seconds, measured for the OpenAI API calls and, for Method C only, the topic modeling stage.

The five metrics in the manual evaluation were assessed by an expert with a Ph.D. in Economics on a five-point Likert scale, where higher scores indicate better quality. Before assessing the generated contents, the evaluator first read (1) Jiji Press, Ltd.'s website at https://www.jijiphoto.jp/resource/html/ext/news/year/2025/index.html to grasp the major events that occurred in each month and (2) the actual N news headlines provided to the three methods. More detailed instructions are found in Appendix A.

C. Example Outputs

To illustrate how each method worked and how it was assessed, we first present the English-translated economic reports for November 2025 generated by the three methods. The following shows the economic report generated by Method A.

> Incidents and policy changes around the world are having a significant impact on the economy. A robbery in the Philippines, the suspension of SNAP benefits in the United States, and China's request to refrain from overseas travel have negatively affected consumption and the tourism industry. In addition, the shutdown of U.S. government agencies and information leaks caused by cyberattacks have created uncertainty in economic activity. Meanwhile, amid increasing investment in AI and renewable energy, there are also efforts aimed at promoting economic growth.

Although this report is factually accurate as supported by the input news articles, it may not be worth including "a robbery in the Philippines" in the report, since there was only one such news article in the input. Also, it is unclear whether "the suspension of SNAP benefits" has a significant impact on the Japanese economy.

Next, Method B identified $k'(= 3)$ topics as instructed and generated the following topic labels (in italics), summaries, and impact scores in parentheses, followed by the overall report. Note that, to save space, the topic summaries show only the snippets.

1) *U.S. Government Shutdown*: As the shutdown of U.S. government agencies continues, concerns are rising over flight reductions and disruptions to administrative functions. ... (-4.84)
2) *China's Pressure on Japan*: The Chinese government's call for its citizens to refrain from traveling to Japan has dealt a significant blow to Japan's tourism industry and related sectors. ... (-4.65)
3) *Suspension of U.S. Food Assistance*: The U.S. government has indicated its intention to suspend full benefit payments under the Supplemental Nutrition Assistance Program

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

(SNAP), raising concerns about reduced consumption among low-income households. . . . (−1.73)

Amid growing global economic uncertainty, the shutdown of U.S. government agencies and mounting pressure from China on Japan are exerting significant effects. In particular, the suspension of U.S. food assistance and China's request to refrain from travel to Japan have dampened consumption and heightened uncertainty about the economic outlook. In response, countries are exploring economic countermeasures while seeking to strengthen stable supply systems and growth strategies.

Similar to Method A, the output was factually accurate, but it was not evident whether "the suspension of U.S. food assistance" has a substantial impact on the Japanese economy. Consequently, the evaluator judged that the topic was not sufficiently important to be included in the k' major topics.

Lastly, the output of Method C is shown below in the same manner.

1) *International Trade and Economic Policy*: . . . International trade frictions, such as U.S. tariff hikes and China's suspension of imports, may have negative effects on the Japanese economy. . . . (0.476)
2) *Revitalization of Tourism and Transportation*: . . . the redevelopment of tourist destinations, the introduction of new technologies, and the hosting of events are having positive effects on regional economies. . . . (0.435)
3) *Cultural and Economic Trends*: . . . Cultural events and the activities of well-known individuals generally have positive effects, contributing to the revitalization of regional economies and the tourism industry. . . . (0.427)

Developments in international trade, tourism, and culture are exerting multifaceted effects on the Japanese economy. In international trade, concerns persist over U.S.-China trade tensions, while Japanese companies continue to expand overseas. The revitalization of tourism and transportation is stimulating regional economies, and cultural events are supporting the tourism industry. These elements interact with one another, collectively producing complex impacts on Japan's overall economic conditions.

While the report is generally adequate as an economic report, it was not apparent what "Cultural events and the activities of well-known individuals" in the third topic referred to within the source news articles, which resulted in lower scores on the *Economic Significance* and *Factual Accuracy* metrics.

D. Results

Table I shows the results of the manual evaluation of the k' selected topics for Methods B and C, where statistically significant differences at $p < 0.05$ are marked with an asterisk (*) based on Wilcoxon signed-rank tests. Note that we used the Wilcoxon test instead of a t-test because the evaluation scores were measured on an ordinal Likert scale and the normality assumption required for the t-test cannot be guaranteed due to the small sample size (i.e., 12 months).

TABLE I. MANUAL EVALUATION OF THE k' SELECTED TOPICS.

Month	Significance		Distinctiveness		Score Validity	
	B	C	B	C	B	C
1	3	3	5	2	5	5
2	2	3	5	5	3	5
3	3	5	5	5	4	5
4	2	4	5	2	5	5
5	3	3	5	5	4	5
6	4	4	5	5	4	4
7	4	3	5	5	3	3
8	5	5	5	4	5	5
9	5	5	5	5	3	5
10	5	4	5	4	5	5
11	4	4	5	4	5	5
12	5	5	5	5	4	5
Avg.	3.75	4.00	5.00*	4.25	4.17	4.75*

Both Methods B and C performed not particularly well in *Economic Significance*, indicating that some economically significant events were not included. For instance, Sanae Takaichi was designated as the Prime Minister of Japan in October 2025 and more than 10% of the input news articles that month contained her name. Due to her economic and foreign policies, which differ from the previous prime minister's, the event should have had a significant economic impact, but it was not selected as a major topic by either method. More work is needed to identify economically significant topics. As for *Distinctiveness*, Method B worked better than Method C, even though the latter employed a separate topic model for topic identification. To improve Method C, it would be beneficial to incorporate a redundancy measure when choosing k' topics. On the other hand, Method C was rated higher in *Impact Score Validity* than Method B, indicating the strength of the topic-sentiment aggregation.

Then, Table II shows the results of the manual evaluation of the generated reports. All methods showed high *Economic Adequacy*, demonstrating the benefit of an advanced large-scale LLM in logical inference. As for *Factual Accuracy*, while the differences were not statistically significant, Method C showed lower scores on average than the other two. For instance, in the January economic report by Method C, it was stated that "China's market growth was a positive factor". However, the actual input articles contained "deflation concerns", "weak domestic demand", and "economic slowdown" in the Chinese economy, leading the evaluator to judge the statement as factually inaccurate. The particular statement was based on a news article, "China's BYD reports a 41% increase in new vehicle sales to 4.27 million units in 2024, maintaining strong growth." Controlling the influence of each news article on the final report remains an issue to be addressed in future work.

Next, Table III compares Methods B and C using the *Diversity* measure defined in Eq. (1). Consistent with the *Distinctiveness* measure in Table II, Method B performed better than Method C, partly validating the quality of the manual evaluation. In fact, *Distinctiveness* and *Diversity* scores of Method C show a positive correlation ($R = 0.59$). Note that

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE II. MANUAL EVALUATION OF THE GENERATED ECONOMIC REPORTS.

Month	Factual Accuracy			Economic Adequacy		
	A	B	C	A	B	C
1	5	5	2	5	5	4
2	4	4	5	4	5	5
3	5	5	5	5	5	5
4	5	5	5	5	5	5
5	5	5	3	5	5	5
6	5	5	3	5	5	5
7	3	3	3	5	5	5
8	5	5	5	5	5	5
9	5	5	5	5	5	5
10	5	5	5	5	5	5
11	5	5	3	5	5	5
12	5	5	5	5	5	5
Avg.	4.75	4.75	4.08	4.92	5.00	4.92

TABLE III. DIVERSITY OF THE k' SELECTED TOPICS.

Month	B	C
1	0.609	0.490
2	0.537	0.614
3	0.758	0.757
4	0.781	0.356
5	0.773	0.567
6	0.704	0.564
7	0.643	0.559
8	0.669	0.460
9	0.803	0.595
10	0.676	0.349
11	0.545	0.462
12	0.579	0.447
Avg.	0.673	0.518

the correlation coefficient for Method B cannot be calculated, as it has straight fives for *Distinctiveness*.

Table IV compares the API cost for generating the monthly economic reports across the three methods. Method C uses far fewer input tokens (about 33% of Methods A and B) because it first performs topic analysis locally and sends only the news headlines for the k' major topics to the LLM. As a result, the total cost of Method C is approximately 39% of the others on average. The API cost is likely negligible in practice, because only one report is generated per month and the resulting report can be stored and provided to all users of the system. However, we plan to extend the functionality to generate reports tailored to a particular event or topic of interest to individual users [30]. For instance, some users may be interested in the influence of international affairs or U.S. tariff policies on the economy. In such cases, the system would select only news articles relevant to the topic and generate a more focused summary on the fly. The total cost would then scale with the number of users and their specific needs, making cost differences more important.

Finally, Table V shows the processing time for each method, where the processing time for topic analysis and the API call are separately reported for Method C. On average, Method A was the fastest due to its simplicity, whereas Method C took about 7.5 times longer than Method A and 4.5 times longer than Method B. To be used as the on-the-fly report generation illustrated above, the processing time needs to be shortened, particularly for Method C.

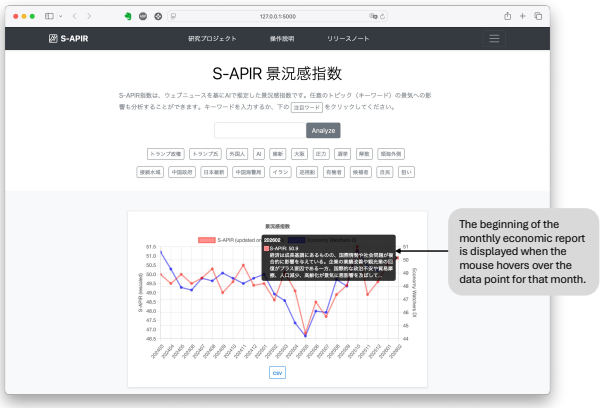


Figure 6. A screenshot of our BSI nowcasting system with a snippet of a monthly economic report.

In summary, all methods generated largely accurate and adequate reports, although Method C scored somewhat lower than the other two in *Factual Accuracy*, with no statistically significant difference. With respect to the identified major topics, Method B identified more distinctive topics than Method C, whereas the economic impact scores were found to be more valid for Method C than for Method B. In terms of API cost, Method C has an advantage over Method B; however, it requires approximately 4.5 times as much processing time. Considering these factors, a hybrid approach that combines Methods B and C may provide the best overall balance. Specifically, we first identify k distinctive topics using an LLM on a smaller subset of news articles, thereby reducing cost and processing time. We then categorize the full set of news articles locally under the identified topics. In this way, topic-level sentiment scores can also be aggregated locally to yield more appropriate estimates of economic impact.

V. IMPLEMENTATION

We designed and implemented the functionality to provide the generated report and the details of the constituent topics for Method B or C in our BSI nowcasting system. Figure 6 shows the chart of BSI values and the beginning of a monthly report for February 2026.

The chart incorporates an on-click event handler, and if the user wishes to look deeper into the month, they click the respective part of the chart and a new Bootstrap modal display appears on top of the page as illustrated in Figure 7. The modal display provides the full text of the generated economic report for the clicked month, the economic impact scores of the major topics as a bar chart, and the labels and summaries of the topics.

VI. CONCLUSIONS AND FUTURE WORK

This paper investigated methods for generating concise yet informative economic reports to enhance the interpretability of a news-based BSI. We proposed and compared the following three methods:

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE IV. COMPARISON OF API COST.

Month	# of input tokens			# of output tokens			Cost (US$)		
	A	B	C	A	B	C	A	B	C
1	39056	39485	13078	134	559	695	0.099	0.104	0.040
2	39210	39639	13248	148	573	695	0.100	0.105	0.040
3	39362	39791	13310	175	527	721	0.100	0.105	0.040
4	39343	39772	13277	162	593	767	0.100	0.105	0.041
5	39323	39752	13524	166	556	728	0.100	0.105	0.041
6	39881	40310	13446	171	574	763	0.101	0.107	0.041
7	39590	40019	13089	162	555	741	0.101	0.106	0.040
8	39784	40213	13260	133	618	738	0.101	0.107	0.041
9	39633	40062	13350	149	541	737	0.101	0.106	0.041
10	39828	40257	13226	155	567	695	0.101	0.106	0.040
11	39522	39951	13305	151	552	675	0.100	0.105	0.040
12	39340	39769	13435	151	522	716	0.100	0.105	0.041
Avg.	39489	39918	13296	155	561	723	0.100	0.105	0.040

TABLE V. COMPARISON OF PROCESSING TIME.

Month	A	B	C		
			Topic	API	Sum
1	5.9	8.3	22.7	25.3	48.0
2	5.5	8.9	20.9	22.1	43.0
3	5.0	8.0	20.4	30.0	50.5
4	6.7	11.0	20.9	28.8	49.7
5	12.6	8.2	19.4	23.8	43.2
6	5.5	8.7	19.5	26.7	46.2
7	4.6	19.7	19.3	26.4	45.7
8	5.7	10.8	20.7	21.3	42.0
9	4.6	10.0	19.9	25.8	45.6
10	6.2	9.4	20.3	21.0	41.3
11	4.8	7.6	19.7	18.9	38.7
12	4.4	7.7	21.1	20.7	41.9
Avg.	6.0	9.8	20.4	24.2	44.6

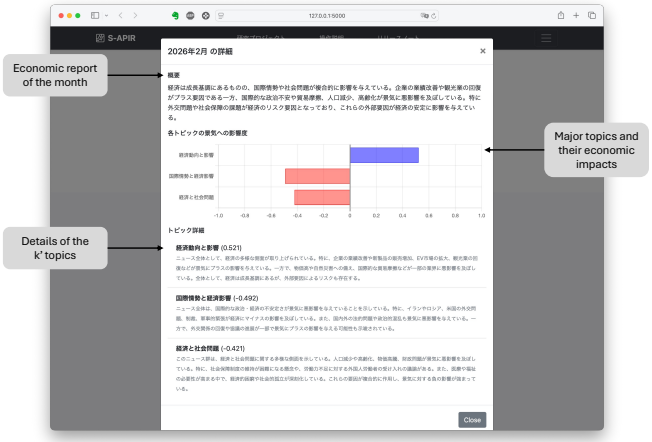


Figure 7. A screenshot of the modal display with a detailed monthly report.

- Method A uses an LLM to summarize input news headlines with high economic impact scores without considering topics.
- Method B also uses an LLM, but with a more detailed prompt to (i) categorize the input into topics and (ii) select important topics to generate an economic report.
- Method C leverages a neural topic model to identify representative, economically influential topics, which are then provided to an LLM (along with supplementary information) to generate an economic report.

We conducted empirical experiments using one year of online news data and examined both the strengths and limitations of the proposed methods. Despite incorporating a separate topic model, Method C was less effective than Method B in identifying distinctive topics. This limitation could potentially be addressed by explicitly enforcing diversity during topic selection. On the other hand, the economic impact scores of the identified topics were found to be more appropriate with Method C than with Method B. Regarding the content of the generated economic reports, the outputs were generally acceptable across all methods, although Method C tended to receive lower scores in factual accuracy. In terms of computational cost, the API cost of Method C was approximately 40% of that of the other methods; however, due to the additional topic-modeling stage, it required about 4.5 times as much processing time as Method B. Combining the strengths of Methods B and C may improve topic identification and the quality of the resulting economic reports. Additionally, we implemented functionality in our BSI nowcasting system to display the generated reports and their constituent topics, which may facilitate the collection of user feedback.

A limitation of this study is that the manual evaluation was conducted by a single expert, which may introduce subjective bias into the assessment. Additionally, the scope of the input data is constrained. Due to the excessive length of full-text news articles, our model relies exclusively on news headlines rather than the full content of the news articles. Future work may include integrating the strengths of Methods B and C, employing more efficient long-context processing techniques to incorporate full texts, and involving multiple evaluators to enhance the robustness and reliability of the evaluation results.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI #22K12280 and MEXT Japan.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Appendix A
Instruction to Evaluator

A. Purpose of the Evaluation

The purpose of this evaluation is to assess the quality of economic reports automatically generated from one month of news articles. The evaluation should be based not on subjective preference, but on the validity of the generated content in light of the provided news articles.

B. Materials to Be Evaluated

Evaluate the reports generated by Methods A, B, and C. These reports are stored in separate files. The news articles used as input for report generation (1,000 articles per month) are also provided.

C. Evaluation Procedure

1) Review the major events of the month using the following sources:
   - Jiji Press: https://www.jijiphoto.jp/resource/html/ext/news/year/2025/index.html
   - Monthly Economic Report by the Cabinet Office: https://www5.cao.go.jp/keizai3/getsurei/getsurei-index.html
2) For each metric described next, assign a score on a five-point scale and record the results in a provided sheet. If a score of 1 or 2 is assigned, briefly state the reason.

D. Evaluation of the Three Selected Topics

This item applies only to Methods B and C. Judge based on the labels, importance scores, and summaries of Topics 1–3. The monthly economic report itself should not be evaluated here.

- Economic Significance
  Evaluation Criteria:
  - Whether major topics that significantly affected the economy during the month are included
  - Whether important topics are omitted
  - Whether minor issues are selected as major topics
  - Refer to the input articles if needed

  Scoring:
  - 5: All three topics are economically important
  - 4: All three topics are economically relevant
  - 3: Difficult to judge
  - 2: Some topics are not economically relevant
  - 1: None of the topics are economically relevant
- Distinctiveness
  Evaluation Criteria:
  - Whether the three topics substantially overlap in content (economic relevance is not required)
  - It is not necessary to check the input articles

  Scoring:
  - 5: All three topics are clearly distinct
  - 4: Two topics are somewhat similar but distinguishable
  - 3: Difficult to judge
  - 2: Two topics are nearly identical
  - 1: All three topics are nearly identical
- Impact Score Validity
  Evaluation Criteria:
  - Whether the relative magnitudes of the three economic impact scores (the numerical values following the topic labels) are reasonable
  - No comparison with other months or methods is required; judge only within the same month
  - It is not necessary to check the input articles; judge based on the topic descriptions

  Scoring:
  - 5: Relative magnitudes are appropriate
  - 4: Generally appropriate
  - 3: Difficult to judge
  - 2: Relative magnitudes appear questionable
  - 1: Relative magnitudes are clearly inappropriate

E. Evaluation of the Generated Economic Report

For this item, evaluate only the "economic report for the month" for each method and do *not* evaluate the labels or summaries of Topics 1–3 here.

- Factual Accuracy
  Evaluation Criteria:
  - Whether information not contained in the news articles has been added
  - Whether numerical values or events are incorrectly described
  - Whether any misleading expressions are present
  - Refer to the input articles if needed

  Scoring:
  - 5: No factual errors are identified
  - 4: Only minor issues are present
  - 3: Some parts are difficult to judge
  - 2: One clear factual error is present
  - 1: Multiple factual errors are present
- Economic Adequacy
  Evaluation Criteria:
  - Whether the interpretation of economic causal relationships is reasonable
  - Whether there are logical gaps
  - Whether there are unclear or unintelligible interpretations
  - Whether the report is natural in an economic context
  - Do not refer to the input articles; evaluate whether the logic of the report itself is coherent. Even if factual errors are present, focus solely on the internal economic reasoning.

  Scoring:
  - 5: Highly plausible
  - 4: Generally plausible
  - 3: Difficult to judge

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

- 2: Contains implausible interpretations
- 1: Not plausible at all

F. Notes for Evaluators

- Do not evaluate based on stylistic preference.
- Evaluate the content rather than writing quality.
- If uncertain, select 3.
- If assigning a score of 1 or 2, or if any notable issue is identified, record it briefly.

References

[1] K. Seki, "How to interpret an economic index? Generating reports with topic sentiment analysis," in *Proceedings of the Thirteenth International Conference on Building and Exploring Web Based Environments (WEB 2025)*, 2025, pp. 9–10.

[2] Cabinet Office, Government of Japan, *Economy watchers survey*, https://www5.cao.go.jp/keizai3/watcher-e/index-e.html, Accessed: Feb. 7, 2026.

[3] Bank of Japan, *Tankan*, https://www.boj.or.jp/en/statistics/tk/index.htm, Accessed: Feb. 7, 2026.

[4] K. Seki, Y. Ikuta, and Y. Matsubayashi, "News-based business sentiment and its properties as an economic index," *Information Processing & Management*, vol. 59, no. 2, 102795, 2022.

[5] A. H. Shapiro, M. Sudhof, and D. J. Wilson, "Measuring news sentiment," *Journal of Econometrics*, vol. 228, no. 2, pp. 221–243, 2022. DOI: https://doi.org/10.1016/j.jeconom.2020.07.053.

[6] Y. Yamamoto and Y. Matsuo, "Sentiment summarization of financial reports by LSTM RNN model with the Japan Economic Watcher Survey Data," in *Proceedings of the 30th JSAI*, In Japanese, 2016.

[7] K. Seki, "Refining sentiment predictions: Obtaining an unbiased business sentiment index from Japanese newspapers," *International Journal of Asian Language Processing*, vol. 33, no. 2, 2350015, 2023.

[8] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*, Second. Oxford University Press, 2012.

[9] M.-C. Bieri, "Assessing economic sentiment with newspaper text indices: Evidence from Switzerland," *Swiss Journal of Economics and Statistics*, vol. 161, 10, 2025. DOI: 10.1186/s41937-025-00139-4.

[10] E. Kalamara, A. Turrell, C. Redl, G. Kapetanios, and S. Kapadia, "Making text count: Economic forecasting using newspaper text," *Journal of Applied Econometrics*, vol. 37, no. 5, pp. 896–919, 2022. DOI: 10.1002/jae.2907.

[11] J. Ashwin, E. Kalamara, and L. Saiz, "Nowcasting Euro area GDP with news sentiment: A tale of two crises," *Journal of Applied Econometrics*, vol. 39, no. 5, pp. 887–905, 2024. DOI: 10.1002/jae.3057.

[12] L. Barbaglia, S. Consoli, and S. Manzan, "Forecasting with economic news," *Journal of Business & Economic Statistics*, vol. 41, no. 3, pp. 708–719, 2023. DOI: 10.1080/07350015.2022.2060988.

[13] W. Hu, X. Zhang, and Y. Ren, "Generating financial reports from macro news via multiple edits neural networks," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, 2020, pp. 667–682. DOI: 10.1007/978-3-030-67664-3_40.

[14] S. Yan, "Disentangled variational topic inference for topic-accurate financial report generation," in *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*, 2022, pp. 18–24. DOI: 10.18653/v1/2022.finnlp-1.3.

[15] Z. Zhang *et al.*, "Enhancing multi-document summarization with cross-document graph-based information extraction," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023, pp. 1696–1707. DOI: 10.18653/v1/2023.eacl-main.124.

[16] N. K. Shukla *et al.*, "Generative AI approach to distributed summarization of financial narratives," in *Proceedings of the 2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 2872–2876.

[17] X. Yang, S. Zang, Y. Ren, D. Peng, and Z. Wen, *Evaluating large language models on financial report summarization: An empirical study*, 2024. arXiv: 2411.06852 [cs.CL].

[18] G. Assis *et al.*, "Language models for automated market commentary from corporate disclosures," in *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF)*, 2025, pp. 727–735. DOI: 10.1145/3768292.3770438.

[19] P. Cui and L. Hu, "Topic-guided abstractive multi-document summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 1463–1472.

[20] L. Dong, M. N. Satpute, W. Wu, and D. Du, "Two-phase multidocument summarization through content-attention-based subtopic detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 6, pp. 1379–1392, 2021.

[21] C. Li, A. Xu, S. Joty, and G. Carenini, "Topic-guided reinforcement learning with LLMs for enhancing multi-document summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 12 395–12 412.

[22] B. Warner *et al.*, "Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference," *arXiv preprint arXiv:2412.13663*, 2024.

[23] The Sankei Shimbun, *The Sankei Shimbun*, https://www.sankei.com, Accessed: Feb. 7, 2026.

[24] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," *The Journal of Machine Learning Research*, vol. 2, pp. 139–154, 2002.

[25] X. Wu, T. T. Nguyen, D. C. Zhang, W. Y. Wang, and A. T. Luu, "FASTopic: Pretrained Transformer is a fast, adaptive, stable, and transferable topic model," in *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[26] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 100–108.

[27] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[29] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022. arXiv: 2203.05794 [cs.CL].

[30] X. Zhang, Q. Wei, Q. Song, and P. Zhang, "TOMDS (topic-oriented multi-document summarization): Enabling personalized customization of multi-document summaries," *Applied Sciences*, vol. 14, no. 5, 1880, 2024.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Visual and Graph-Augmented Evaluation of RAG Systems – Improving Retrieval, Re-Ranking, and Data Quality with the pRAG Framework

Alexander Kreß*, Alexander Lawall*, Thomas Zöller*
*IU International University of Applied Science
Erfurt, Germany
alexander.kress@iu-study.org, alexander.lawall@iu.org, thomas.zoeller@iu.org

***Abstract*—Retrieval-Augmented Generation systems require evaluation methods that expose weaknesses within individual components rather than relying solely on end-to-end metrics. This paper introduces pRAG, a visual and graph-augmented framework that provides interpretable, component-level insights into retrieval, re-ranking, and context selection without using LLM-based judgment. pRAG combines specialized recall metrics, ranking indicators, and timing analysis to identify bottlenecks-most notably in the retrieval stage. We further propose a scalable synthetic dataset pipeline to reduce the need for manually curated ground truth. Experiments show that synthetic question-answer pairs achieve comparable or superior evaluation quality, with human experts preferring them in most cases. The extension of pRAG with knowledge graph relations enables the analysis of contextual dependencies between chunks and supports more informed reasoning. Our results highlight that hybrid retrieval strategies and optimized retrieval sizes significantly improve RAG performance. Overall, pRAG offers an efficient and interpretable methodology for evaluating complex RAG pipelines and supports scalable assessment in specialized domains.**

Keywords-Retrieval-Augmented Generation (RAG); Component-Level Evaluation; Synthetic Data Generation; Knowledge Graph Integration; Retrieval Optimization.

I. Introduction

This work is an extended version of *Retrieval Performance in RAG Systems: A Component-Level Evaluation Framework*, published at GPTMB 2025 [1]. Retrieval-Augmented Generation (RAG) systems are important for improving the factuality and reliability of Large Language Model (LLM) outputs, especially in domain-specific applications. Despite their adoption, evaluating these systems remains challenging, particularly when considering their multi-component nature and varying performance across different use cases [2].

A. Motivation and Problem Statement

The evaluation of RAG systems faces several challenges that current LLM approaches fail to adequately address. While LLMs alone can be evaluated using established benchmarks, RAG systems introduce additional complexities due to their multi-stage architecture spanning document processing, retrieval, re-ranking, and generation components [3][4]. As noted by [3], dynamic data environments further complicate evaluation, as the underlying knowledge sources often change over time.

Current evaluation frameworks typically produce aggregate metrics that mask the performance of individual components, making it difficult to identify specific bottlenecks or optimization opportunities [3][5]. Manual evaluation methods are becoming increasingly inefficient, necessitating automated approaches that can scale with system complexity. Additionally, temporal aspects of RAG performance — such as latency variations across different technical configurations — are rarely incorporated into evaluation methodologies despite their critical importance in real-world applications [3][6].

Established benchmark datasets like HotpotQA [7] and MS MARCO [8] have proven inadequate for evaluating modern RAG systems [9], as they fail to capture the nuanced retrieval and generation scenarios encountered in specialized domains. The availability of ground truth data will become rare in the future [10]. While synthetic dataset generation offers promising alternatives [11], systematic approaches for validating these datasets and incorporating them into holistic evaluation frameworks remain underdeveloped.

B. Research Gap

Despite the proliferation of evaluation methods for RAG systems, major gaps persist in current approaches. Existing frameworks like RAGAS [12] rarely provide granular insights into component-level performance, instead focusing on end-to-end evaluation that obscures the contribution of individual technical elements [3][4]. As [13] observes, the various technical alternatives available at each stage of the RAG pipeline create a complex evaluation space that remains largely unexplored. As [14] points out, this gap cannot be closed by asking LLMs for reasoning.

The role of re-ranking models [15][16] and hybrid retrieval techniques like the combination of embeddings and BM25 [17] in RAG performance is inadequately addressed by current evaluation approaches. Furthermore, while the importance of synthetic datasets for evaluation is increasingly recognized [18], methodologies for generating and validating these datasets remain ad-hoc and not standardized. These gaps demand a comprehensive evaluation framework that addresses both the technical complexity of RAG systems and the practical challenges of meaningful assessment [11].

C. Research Questions

This paper addresses the primary research question: "Which technical concepts are necessary to successfully evaluate RAG systems?". Secondary research questions are investigated to explore this question more comprehensively:

1) "How can we effectively evaluate the retrieval component in RAG systems?"

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

2) "How can synthetic datasets be efficiently generated and validated for RAG evaluation?"
3) "What approaches show promise for evaluating the entire RAG pipeline?"

D. Contributions

This paper makes several contributions to the field of RAG system evaluation:

- The introduction of a methodology for generating and validating synthetic evaluation datasets that can scale efficiently across domains and use cases.
- The development of a visualization approach (pRAG) for assessing retrieval component performance that incorporates both quality metrics and temporal analysis.
- The provision of empirical findings from applying the framework to a real-world RAG system designed for technical documentation question answering.
- The integration of Knowledge Graph logic into the pRAG visualization and evaluation framework

The framework addresses critical gaps in existing evaluation approaches by offering a more granular, component-specific assessment methodology that can adapt to the evolving landscape of RAG system design.

E. Paper Structure

The remainder of this paper is organized as follows: Section II describes the methodology, including the architecture of the evaluation framework, synthetic dataset generation approach, and component-specific assessment techniques. Section III presents the results of applying the framework to a case study RAG system and discusses key findings and implications. Section IV provides a detailed evaluation and discussion of the results, analyzing the role of ground truth, retrieval performance, and the effectiveness of the proposed framework. Finally, Section V concludes the paper and outlines directions for future work.

II. Methodology

A. RAG System Architecture

The RAG system employs a microservice-based architecture designed for scalability and modular development. The system processes user queries through these key components: When a user submits a query via the frontend, the middleware API coordinates the workflow. First, the pre-processing API generates keywords and embeddings from the query for semantic comparison. These are passed to a vector handling API that performs hybrid retrieval, combining BM25 [19], keyword matching, and embedding-based semantic search through paradeDB, a PostgreSQL extension supporting vector operations.

Retrieved contexts and metadata flow back to the middleware API, which forwards them to the pre-processing API where a cross-encoder re-ranker prioritizes the most semantically relevant documents. Finally, these re-ranked contexts together with an initial prompt are provided to a LLM that generates a comprehensive response based on the available information and returns it to the user via the frontend.

B. System Implementation

The system is deployed on a Kubernetes cluster with the frontend developed in React and backend services in Python. For the knowledge base, 50 technical documents from HORSCH machinery manuals using Azure Document Intelligence are processed to convert PDF content into processable text.

For embedding generation, the Hugging Face multi-qa-MiniLM-L6-cos-v1 Sentence Transformer model [20] was implemented, selected for its balance of English language capabilities and computational efficiency. Documents were chunked to match the model's maximum token length and then stored with machine-specific metadata.

We evaluated two cross-encoder models for re-ranking: msmarco-MiniLM-L6-en-dev1 [21] and ms-marco-MiniLM-L-6-v2 [22], which reorder retrieved contexts based on query relevance. For response generation, we utilized ChatGPT-3.5-Turbo-0125 with crafted prompts to ensure responses were relevant, accurate, and focused on HORSCH machinery documentation.

Performance timing was implemented using Python's time module, capturing execution duration for each component to enable system optimization.

C. plot-RAG (pRAG): A Novel Evaluation Framework

1) Motivation and Design: A key contribution of this work is pRAG, a novel visualization and evaluation framework specifically designed to address the lack of quantitative, interpretable evaluation methods for RAG systems. pRAG provides granular insights into the performance of individual RAG components, particularly the critical retrieval and re-ranking stages. This contrasts with current evaluation approaches, which often focus on end-to-end performance or rely on limited metrics like recall and precision, which are susceptible to outliers [23].

2) Visualization Components: The pRAG visualization (see Figure 1) displays multiple dimensions of system performance simultaneously:

- **Context position tracking:** Visualizes where relevant contexts from ground truth appear in both retrieval and re-ranked results (blue numbers).
- **Retrieval method comparison:** Distinguishes between embedding-based and BM25 keyword-based retrievals (y-axis).
- **Ground truth distribution:** Shows distances between relevant contexts in the document corpus (green numbers).
- **Quantitative metrics overlay:** Presents calculated performance metrics alongside visual representations (top right corner).
- **Relevant context count at position:** The number of relevant contexts from ground truth appearing at this specific retrieval position across the entire evaluated data set (numbers in parentheses).

3) Metrics Integration: pRAG calculates and visualizes several critical metrics:

- **Specialized recall metrics:**
  - Recall Emb: Effectiveness of embedding-based retrieval

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

  - Recall BM25: Performance of keyword-based retrieval
  - Recall Full Retrieval: Combined unique contexts retrieval rate
  - Recall Re-ranking from Retrieval: Preservation of relevant contexts after re-ranking
- **Ranking quality:** Normalized Discounted Cumulative Gain (NDCG) calculation highlighting the importance of positioning relevant information earlier in results
- **Retrieval optimization:** Recommended retrieval sizes for both embedding and BM25 components

4) Actionable Insights: The pRAG framework provides actionable insights by visually exposing:

- Which retrieval method (BM25 or embeddings) more effectively captures relevant contexts
- How effectively the re-ranker prioritizes relevant contexts
- Optimal retrieval configuration parameters
- Performance bottlenecks in specific components

This visualization approach enables the identification of system weaknesses without requiring extensive manual analysis, making it particularly valuable for ongoing RAG system development and optimization.

5) Knowledge Graph Integration into pRAG: Furthermore, we supplement the paper with a complementary concept for integrating knowledge graphs into pRAG. Each data point in the pRAG visualization represents one or more chunks contained in the ground truth. If each data point is interpreted as a node for one chunk or as nodes for more chunks, it is possible to draw edges between the data points to symbolize their relationships. In addition, their relationship is made interpretable via several technical components, which which must be interpreted relative to the specific technical components of each RAG implementation.

In pRAG, different chunks can be located at one position. The pRAG plot only shows that a chunk from the ground truth for the question-answer pair occurs there – but not which chunk it is. Since these chunks can now have different knowledge graphs to other chunks, exploration in the Z direction is necessary to display the semantic edges for each node. A 3D plot is not provided here because of the complex visualization.

For this evaluation approach we recommend to use uuids for each chunk row in the database to realize working evaluation without the viable ascii errors of text fields. In order to create a generic pRAG plot with the integration of knowledge graphs, the dataset should be structured in JSON format with the following keys:

JSON Keys for generic pRAG evaluation with knowledge graph combination:

- **user_query**: The query of the user asked to the RAG system
  Datatype: `string`
- **rag_answer**: The answer of the RAG system
  Datatype: `string`
- **ground_truth**: The ground truth of the answer. Can also be created in a synthetic way
  Datatype: `list[string]`
- **contexts**: A list of dicts. The keys in a dict represent the different technical retrieval, re-ranking, or other approaches to fetch chunks for the generation step
  Datatype: `list[dict]`
- **pair_of_contexts**: On the same gate, it is possible to use more than one technical component, like the BM25 and Embedding technique, in the retrieval step.
  Datatype: `list[list]`
- **knowledge_graph**: Displays the relation between edges via nodes as UUIDs
  Datatype: `dict[string, list]`

Figure 5 shows the utilization of the augmented pRAG approach in the analysis of retriever performance. The particular results are further discussed in Section III-E.

D. Synthetic Dataset Generation

For evaluation, both manually curated and synthetically generated question-answer pairs based on three technical manuals for products from the HORSCH portfolio: Avatar 12/40 SD, Joker RX, and Tiger MT were created. These documents were selected based on machine sales volume analysis, indicating likely user query subjects.

For each document, we prepared 50 question-answer pairs with relevant contexts as ground truth. From each set, five pairs were randomly selected as examples for synthetic generation. Using these examples iteratively with different ground truth contexts, we generated 45 synthetic question-answer pairs per document using three different language models: GPT-4o-Mini, Gemini-1.5-Flash, and Nemotron-4-340b-Instruct. Also, two comparison methodologies were implemented:

1) **Absolute comparison:** evaluating curated vs. synthetic datasets based on different contexts
2) **Relative comparison:** generating synthetic data using ground truth from the curated dataset

The concept involved comparing the synthetic data based on the same chunks that were used for the processed datasets. This approach is referred to as a relative comparison. The generator models generated synthetic question-answer pairs based on the chunks selected for the processed datasets. Various discriminant models were then applied to make a selection decision between the synthetic and processed pairs. The combinations of generator and discriminant models are shown in Table II.

In addition to the relative comparison, an absolute comparison was also performed. Here, random chunks were retrieved from the database. The synthetic datasets were then generated and an LLM evaluated the new data independently of the manually generated data.

Quality assessment employed a GAN-like approach, using language models (GPT-4o-Mini, Llama-3-Patronus-Lynx-8B-Instruct [24], and Prometheus-7b-v2.0 [25]) as discriminators to evaluate response quality with Pass/Fail determinations and comparative quality judgments.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

E. Experimental Setup

Multiple experimental configurations are conceptualized to evaluate different aspects of the RAG system and demonstrate the utility of the pRAG framework. For enhanced retrieval configurations, we used a basic setup and changed specific technical components for enhanced setups:

a) Basic Synthetic Data Evaluation (Setup A):

- Generator: ChatGPT-3.5-Turbo-0125
- Retrieval: paradeDB (BM25+Embeddings)
- Re-ranker: Cross-encoder/msmarco-MiniLM-L6-en-de-v1
- Retrieval size: 8 contexts each for BM25 and embeddings
- Re-ranking size: 4 contexts

b) Enhanced Retrieval Configuration (Setup B):

- Generator: ChatGPT-3.5-Turbo-0125
- Metadata: Machine Name
- Re-ranker: Cross-encoder/msmarco-MiniLM-L6-en-de-v1
- Retrieval size: 60 contexts (BM25+Embeddings)
- Re-ranking size: 20 contexts

c) Enhanced Retrieval Configuration (Setup B-1):

- New Re-ranker: Cross-encoder/ms-marco-MiniLM-L-6-v2

d) Enhanced Retrieval Configuration (Setup B-2):

- New Method: HyDE Integration

For additional quantitative evaluation, we implemented the RAGAS framework to assess context precision, answer credibility, relevance, and accuracy. We compared RAGAS with the pRAG framework to substantiate the validity of the pRAG approach. We supplemented the pRAG approach with timing analysis of each system component, capturing minimum, maximum, and median execution times.

The expert evaluation was conducted with domain specialists who assessed question-answer pair quality in a blinded format, comparing synthesized and manually curated responses without knowledge of their origin to eliminate bias. For this experiment, we used the Basic Setup B with three different datasets.

In this comprehensive methodology using the novel pRAG evaluation framework, we aimed to evaluate not only the overall RAG system performance but also the viability of synthetic data for ongoing system improvement. We aimed to address the issue of available ground truth datasets by generating synthetic datasets automatically based on contexts from the database.

III. Results and Discussion

A. Performance Analysis of RAG Components Using pRAG

1) Retrieval Component Performance: The analysis demonstrates that each component contributes differently to the overall performance and can be individually assessed through visualization with pRAG. Figure 1 illustrates the pRAG visualization, where the positions of contexts in the ground truth collection are mapped against their retrieval positions. It shows that several relevant contexts (positions 6, 7, and 8) were missed by BM25 but captured by embedding-based retrieval. Substantial gaps in the diagram, particularly towards the tail of the retrieval list, can support decision-making on whether increasing the retrieval size at the cost of performance should be implemented to identify only a few additional relevant contexts. The automated evaluation of RAG systems with pRAG does not require an LLM as a judge. This makes the evaluation more resource-efficient, considering the substantial computational power required by LLMs.

Since pRAG visualizes the full set of retrieved contexts, the precision metric can be omitted. However, integrating the recall value into the diagram is beneficial to complement the visualization with a quantitative metric. The average values from setups in Section II-E b), c), and d) are presented in Table I.

TABLE I
Recall metrics across enhanced retrieval configuration

Metric	Setup B	Setup B-1	Setup B-2
Recall BM25	0.4314	0.4615	0.3654
Recall Embeddings	0.6863	0.6154	0.3846
Recall Full Retrieval	0.9412	0.9231	0.6923

The pRAG visualization enabled a dedicated evaluation of retrieval techniques, revealing that relevant contexts were retrieved either through BM25 or embedding-based methods and also in both. This points out the importance of evaluating different retrieval strategies individually, as relevant contexts may be identified in one approach but not in another. Further analysis demonstrated that embedding-based retrieval significantly outperformed lexical methods for datasets containing technical terminology. This underlines the necessity of hybrid retrieval approaches, where the combination of strategies ensures a more comprehensive retrieval process and improves overall performance.

2) Optimal Retrieval Size Determination: The experiments demonstrate a relationship between retrieval size and answer quality. With higher retrieval size RAG systems have to handle more irrelevant contexts. Therefore, the optimal retrieval size can be determined using:

$$\text{Retrieval Size} = \frac{\text{Current Retrieval Size} \times \text{Acceptable Response Time}}{\text{Average Response Time}}$$

This formula provides a practical guideline for balancing response time against completeness. As shown in Figure 1, there is no need to put the retrieval size to 60 because the latest relevant contexts were found in positions 21 by embeddings and position 48 by BM25.

The pRAG analysis revealed diminishing returns beyond certain retrieval sizes. For example, in setup B-2 (cf. Figure 1), increasing BM25 retrieval size from 28 to 48 yielded only one additional relevant context, suggesting a practical cut-off point based on efficiency considerations.

B. Synthetic Dataset Evaluation Results

1) Comparative Quality Assessment: To assess the effectiveness of synthetic versus manually prepared datasets, we evaluated both using specialized discriminator models. The results of this evaluation indicate that synthetically generated data achieves comparable or superior performance; specifically,

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

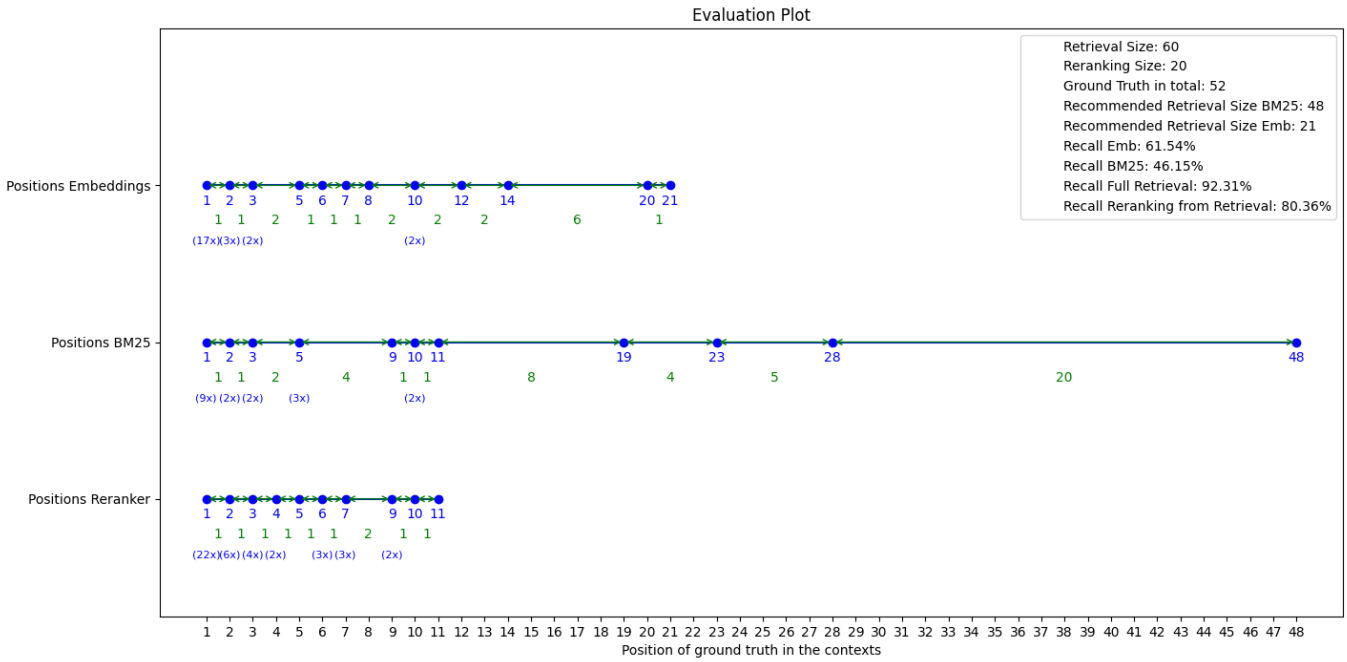


Figure 1. pRAG Visualization Showing Position of Retrieved Contexts Relative to Ground Truth based on Setup B-1.

manually prepared datasets did not offer a notable advantage, and Lynx even performed better on the synthetic data. This confirms that synthetic datasets can provide a similar level of performance to manually prepared ones.

In a direct comparison, the synthetic datasets were rated notably better by the discriminant model than the processed ones when comparing Figures 2 and 3. Additionally, GPT-4o Mini performed the worst, except in cases where it was evaluated directly by GPT-4o Mini. From this, it can be deduced that AI models evaluate their own synthetically generated datasets better and that dedicated AI models developed for synthetic data generation perform best.

Detailed performance results are presented in Figure 4.

TABLE II
GENERATOR-DISCRIMINATOR COMBINATIONS

No.	Generator - Discriminator Combination
1	GPT-4o Mini - GPT-4o Mini
2	Gemini-1.5-Flash - GPT-4o Mini
3	Nemotron-4-340b-Inst. - GPT-4o Mini
4	GPT-4o Mini - Llama-3-Patronus-Lynx-8B-Inst.
5	Gemini-1.5-Flash - Llama-3-Patronus-Lynx-8B-Inst.
6	Nemotron-4-340b-Inst. - Llama-3-Patronus-Lynx-8B-Inst.
7	GPT-4o Mini - Prometheus-7b-v2.0
8	Gemini-1.5-Flash - Prometheus-7b-v2.0
9	Nemotron-4-340b-Inst. - Prometheus-7b-v2.0

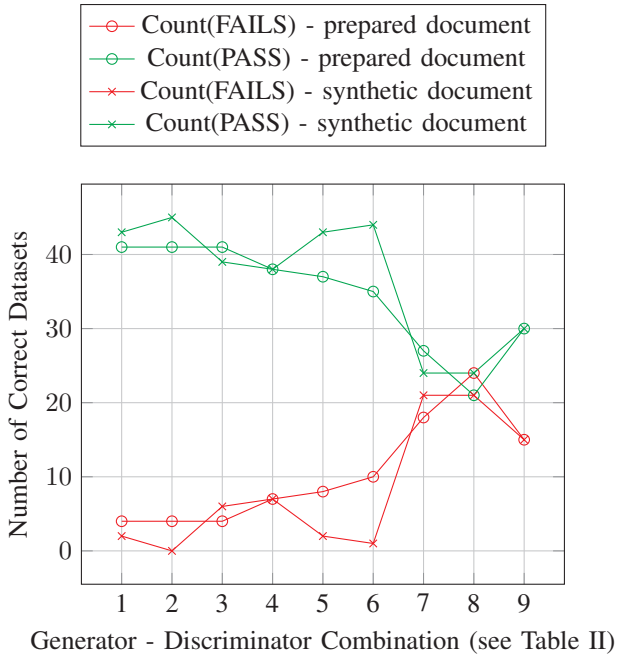


Figure 2. Comparison of Prepared and Synthetic Document for Different Generator-Discriminator Combinations.

2) Human Expert Validation: Human evaluators assessed pairs of question-answer examples from both dataset types. In 68% of comparable cases, experts preferred synthetically generated data, with the remaining 32% showing no clear preference. Table III summarizes these findings. Expert evaluators noted that synthetic datasets showed stronger logical coherence and clearer question formulation. However, they identified a consequential limitation: synthetic datasets generated from tabular data frequently contained factual errors or misinterpretations of numerical relationships, suggesting a specific weakness in

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

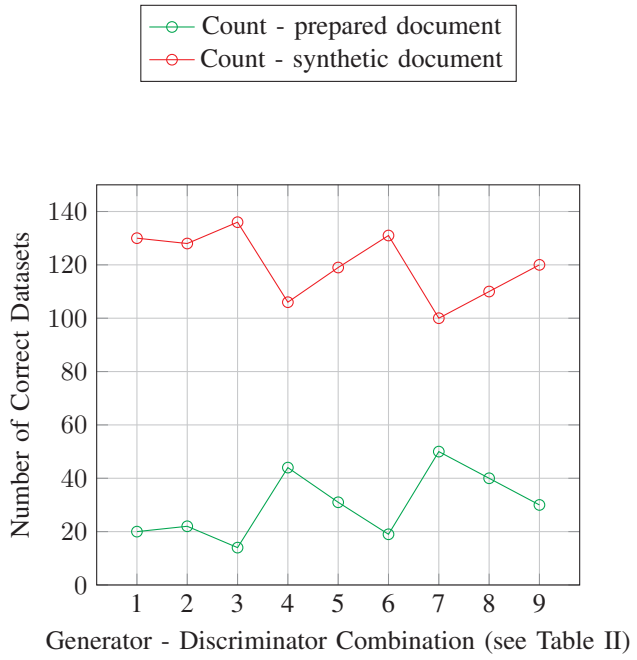


Figure 3. Comparison of Prepared and Synthetic Document for Different Generator-Discriminator Combinations.

current LLM approaches to tabular content. For the evaluation, we used setup B with the three different datasets. The model Nemotron was chosen for its ability to generate better synthetic data [26].

TABLE III
HUMAN EXPERT PREFERENCES IN DATASET EVALUATION

Dataset Pair	Prefer Prepared	Prefer Synthetic	No Preference
Avatar/Nemotron	14	20	16
JokerRX/Nemotron	8	9	31
TigerMT/Nemotron	4	24	22

C. Technical Component Performance Insights

1) Comparative Analysis of Retrieval Enhancements: We evaluated technical enhancements to the base RAG architecture, including re-ranking models and the integration of HyDE (Hypothetical Document Embeddings) [27]. This method decomposes dense retrieval into two distinct tasks: First, it uses an instruction-following language model (like InstructGPT) to generate a hypothetical document in response to a user query. In the second step, an unsupervised contrastively-trained encoder (like Contriever) encodes this hypothetical document into an embedding vector. This vector identifies a neighborhood in the corpus embedding space, from which similar real documents are retrieved based on vector similarity. A detailed component level comparison can be found in Figure 4 while Table IV summarizes these findings.

Contrary to [27], the HyDE approach shows reduced performance despite increased processing time. The pRAG analysis reveals that HyDE's theoretical advantage in generating better query representations did not improve the retrieval of relevant contexts in our test datasets.

TABLE IV
PERFORMANCE COMPARISON OF RETRIEVAL ENHANCEMENT TECHNIQUES

Technique	Recall Re-rank from Retrieval	NDCG	Mean Resp. Time (s)
msmarco-MiniLM-L6-en-de-v1	0.7544	0.54	3.41
ms-marco-MiniLM-L-6-v2	0.8036	0.63	4.08
HyDE Integration	0.7179	0.35	5.43

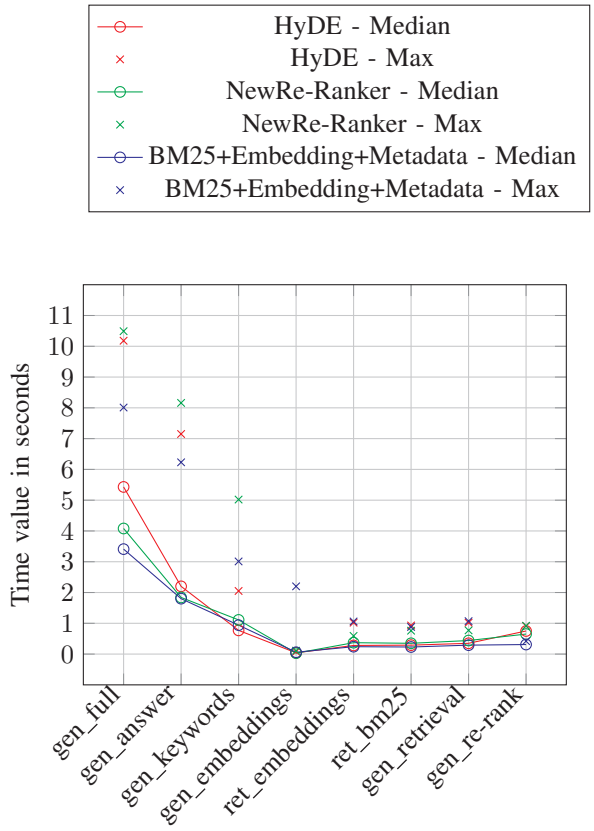


Figure 4. Time Performance across Different Components of the RAG System. Note the Time Advantage of BM25+Embedding compared to HyDE for Comparable Quality Metrics.

Among re-ranking models, ms-marco-MiniLM-L-6-v2 demonstrates the best performance with 80% recall from retrieval but required 20% more processing time than msmarco-MiniLM-L6-en-de-v1. The time-performance analysis shows this tradeoff across various system components.

D. Holistic approach for scaling and evaluating synthetic data

We created data ourselves in our synthetic data generation experiment. This process is not scalable and incurs a lot of human capacity. Rarely does a company invest this huge amount of time to generate prepared datasets. The results of the synthetic data experiments indicate a high usability and intrinsic ability to scale pRAG evaluations. Furthermore, they have the ability to use pRAG at any scale. Further experiments can determine how many synthetic dataset pairs can be generated

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

from a document, and how long the document should be, or how good it has to be to generate question-answer pairs.

Algorithm 1 Synthetic QA Generation and Evaluation Pipeline

Input:
  Big document D
  AI model for QA generation M_{QA}
  Discriminator model M_{disc}
  Retrieval-Augmented Generation system RAG
Output:
  Evaluation metrics comparing synthetic QA vs. RAG
  pRAG visualization
Step 1: Chunk the document
$chunks \leftarrow$ ChunkDocument(D)
Step 2: Generate synthetic QA pairs
for each c in $chunks$ **do**
  $QA_pairs_c \leftarrow M_{QA}(c)$
end for
$QA_pairs \leftarrow$ CombineAll(QA_pairs_c)
Step 3: Filter bad pairs
$Filtered_QA \leftarrow []$
for each (q, a) in QA_pairs **do**
  if $M_{disc}(q, a) =$ good **then**
    Add (q, a) to $Filtered_QA$
  end if
end for
Step 4: Generate RAG answers and save contexts
$RAG_answers \leftarrow []$
for each (q, a) in $Filtered_QA$ **do**
  $(rag_answer, context) \leftarrow RAG(q)$
  Save $context$
  Add $(q, a, rag_answer, context)$ to $RAG_answers$
end for
Step 5: Evaluate
for each $(q, a, rag_answer, context)$ in $RAG_answers$ **do**
  Evaluate rag_answer against a
end for
Return:
  Evaluation metrics e.g NDCG, Recall, ..
  pRAG visualization

Our findings indicate that optimizing the retrieval component provides the greatest improvement in overall system performance. Specifically, the combination of lexical (BM25) and semantic (embedding) retrieval with optimized retrieval size offered the best balance of quality and performance across all test datasets.

E. Knowledge Graph Integration into pRAG Visualization

Within organizations, data is typically stored across a diverse range of systems. It is often the case that no single system contains a comprehensive documentation of a given topic [28]. To obtain such a complete record, the evolution of information maturity must be tracked across these multiple systems.

The development of an individual piece of information can be represented as a directed graph, where nodes correspond to information states and edges represent transitions or transformations between them. For this representation, organizations must have an understanding of how information propagates through their internal processes and how its quality changes throughout this trajectory.

When retrieval encounters directed information, reasoning mechanisms can be applied. This enables dynamic model selection in response to encountering interconnected, directed graphs. In contrast, undirected connections can be interpreted as supplementary relationships and can be incorporated post hoc from the database to support query answering.

Additionally, some connections may point to data points, or “chunks”,that are not part of the established ground truth (shown as orange edges in Figure 5). We propose two potential strategies for handling such cases. First, the chunk can be included in the reasoning process under the assumption that the ground truth may be incomplete. Alternatively, the chunk can be ignored if it is not part of the ground truth. Since ground truth is often unavailable for novel queries, making a definitive distinction between these approaches can be challenging.

When graphs are integrated into pRAG, it is important to recognize that a single data point, or node, may contain multiple chunks. The number of chunks associated with a node is directly dependent on the number of queries incorporated into pRAG for testing.

If a data point has numerous incoming and outgoing connections, it becomes necessary to explicitly indicate which chunks are responsible for establishing these connections. Consequently, pRAG diagrams can become difficult to interpret when applied to large datasets.

However, this complexity can be managed effectively. By assigning weights to individual edges and considering the number of incoming and outgoing connections, nodes can be reordered to enhance clarity. This approach enables a synergistic combination of the advantages of RAG and Knowledge Graph methodologies, optimizing both information retrieval and reasoning capabilities within pRAG.

The starting point of an outgoing connection should always correspond to the last step immediately preceding the generation of a response by the LLM. If additional connections are incorporated into the context prematurely during retrieval, before they are properly retrieved or ranked, these connections may subsequently be deprioritized by a re-ranking model. As a result, the retrieval effort for these connections becomes effectively wasted, increasing inference cost without contributing meaningfully to the generated answer.

IV. Evaluation and Discussion

This section analyzes the empirical and methodological implications of our results, with a particular focus on evaluation reliability, scalability, and system-level insights for RAG architectures.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

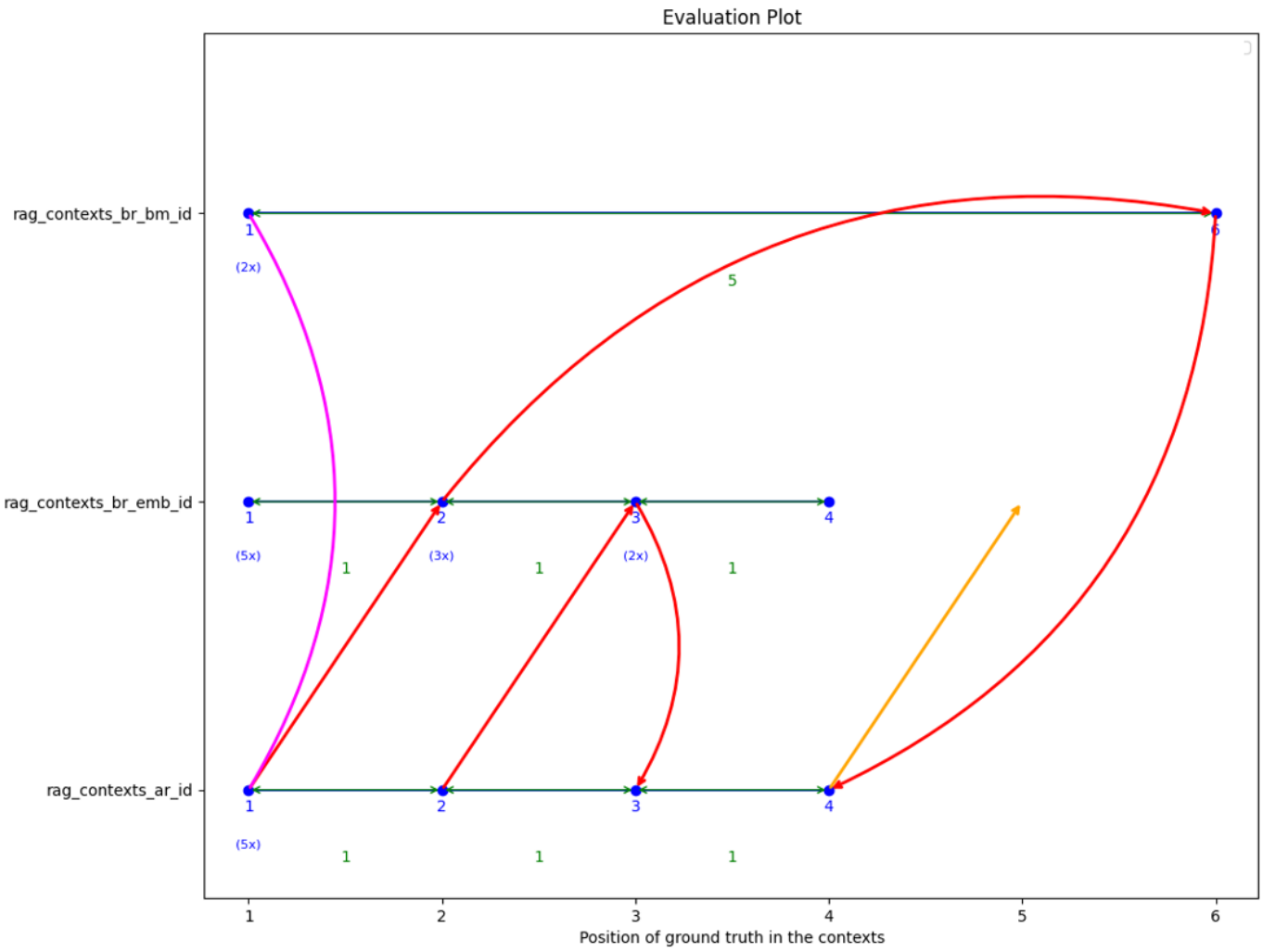


Figure 5. pRAG Visualization Augmented with Knowledge Graph Edges.

A. Role of Ground Truth in RAG Evaluation

A central finding of this work is the decisive role of high-quality ground truth data in evaluating RAG systems. Across all experiments, evaluation outcomes were highly sensitive to the completeness, consistency, and domain alignment of the ground truth. Incomplete or weakly aligned ground truth led to misleading performance signals, particularly in retrieval-heavy configurations.

This highlights a fundamental limitation of many existing evaluation approaches: without reliable reference data, even sophisticated metrics fail to provide meaningful insights. Consequently, evaluation quality is bounded less by metric design and more by the availability of robust ground truth.

B. Effectiveness of pRAG for Component-Level Analysis

The introduction of pRAG enables a shift from aggregated evaluation metrics toward fine-grained, interpretable diagnostics. Our experiments show that pRAG makes it possible to isolate retrieval and re-ranking behaviors, revealing performance bottlenecks that remain hidden in traditional evaluation pipelines.

In particular, pRAG provides:

- Detailed visibility into retrieval depth and ranking quality
- Identification of redundant or low-impact retrieved chunks
- Temporal and structural insights into retrieval dynamics

Compared to LLM-based evaluation methods, pRAG offers greater reproducibility, significantly lower computational cost, and improved transparency. This makes it particularly suitable for iterative system development and benchmarking.

C. Synthetic Data as a Scalable Evaluation Resource

Our results demonstrate that synthetic datasets can serve as a viable substitute for manually curated evaluation data. The generated question-answer pairs achieved high semantic alignment with source contexts and were consistently rated as coherent by both automated discriminators and human evaluators.

This has important implications:

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

- Evaluation pipelines can be scaled without proportional increases in human annotation effort
- Domain-specific datasets can be generated rapidly, enabling continuous evaluation
- Methodological consistency improves due to standardized generation processes

However, limitations remain in handling structured and numerical data, suggesting that future improvements in generation techniques are required.

D. Retrieval as the Dominant Performance Factor

A consistent observation across all configurations is that retrieval quality dominates overall system performance. Variations in retrieval size, ranking strategy, and corpus structure had a significantly greater impact on results than downstream generation differences.

This finding reinforces the importance of:

- Careful tuning of retrieval size
- Balancing completeness against latency constraints
- Avoiding excessive retrieval that introduces noise

Based on our analysis, we derive practical guidelines for selecting retrieval configurations that optimize the trade-off between response quality and computational efficiency.

E. Towards Graph-Enhanced Evaluation

We further explored the integration of knowledge graph concepts into pRAG. Modeling retrieved chunks as nodes and their relationships as edges enables richer representations of information flow within RAG systems.

This perspective provides:

- Improved interpretability of complex retrieval interactions
- Opportunities for graph-based prioritization and reasoning
- Enhanced handling of incomplete or distributed ground truth

While still conceptual, this extension indicates a promising direction for unifying retrieval evaluation with structured knowledge representations.

F. Implications for RAG System Design

Taken together, our findings suggest that effective RAG system development requires:

- Strong emphasis on retrieval optimization
- Scalable and automated evaluation pipelines
- Interpretable tools for diagnosing component-level behavior

The combination of pRAG and synthetic datasets provides a practical foundation for achieving these goals, enabling systematic and data-driven improvement of RAG architectures.

V. Conclusion and Future Work

This work presents a scalable and interpretable framework for evaluating RAG systems by combining pRAG visualization with synthetic dataset generation. Our results demonstrate that reliable evaluation depends fundamentally on high-quality ground truth, and that synthetic data can effectively address this bottleneck at scale.

The proposed pRAG approach enables fine-grained, component-level analysis of retrieval and re-ranking behavior without relying on LLM-based evaluation, improving both transparency and reproducibility. In combination with synthetic datasets, this establishes a practical and resource-efficient methodology for iterative evaluation and optimization of RAG systems.

Our findings further confirm that retrieval is the dominant factor influencing overall system performance, emphasizing the need for careful configuration of retrieval strategies under latency constraints. The integration of graph-based representations offers additional potential for enhancing interpretability and reasoning.

Future work should focus on dynamic evaluation of evolving systems, extensions to multi-modal data, improved synthetic data generation for structured content, and deeper analysis of retrieval–generation interactions. Additionally, graph-aware optimization and advanced visualization techniques remain promising directions for improving both performance and interpretability.

Overall, this framework provides a robust foundation for the systematic development and evaluation of RAG systems in complex, domain-specific environments.

References

[1] A. Kreß, A. Lawall, and T. Zöller, "Retrieval Performance in RAG Systems: A Component-Level Evaluation Framework," in *The Second International Conference on Generative Pre-trained Transformer Models and Beyond (GPTMB) 2025*, 2025, pp. 3–8, [retrieved: January, 2026]. [Online]. Available: https://www.thinkmind.org/library/GPTMB/GPTMB_2025/gptmb_2025_1_20_30012.html

[2] X. Wang *et al.*, "Searching for best practices in retrieval-augmented generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 17 716–17 736.

[3] H. Yu *et al.*, "Evaluation of retrieval-augmented generation: A survey," in *CCF Conference on Big Data*. Springer, 2024, pp. 102–120.

[4] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, 2024, pp. 194–199.

[5] A. Salemi and H. Zamani, "Evaluating retrieval quality in retrieval-augmented generation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2395–2400.

[6] T. Kenneweg, P. Kenneweg, and B. Hammer, "Retrieval augmented generation systems: Automatic dataset creation, evaluation and boolean agent setup," *arXiv preprint arXiv:2403.00820*, 2024, [retrieved: May, 2025].

[7] Z. Yang *et al.*, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," *arXiv preprint arXiv:1809.09600*, 2018, [retrieved: May, 2025].

[8] D. F. Campos *et al.*, "Ms marco: A human generated machine reading comprehension dataset," *ArXiv*, vol. abs/1611.09268, 2016, [retrieved: May, 2025]. [Online]. Available: https://api.semanticscholar.org/CorpusID:1289517

[9] K. Zhu *et al.*, "Rageval: Scenario specific rag evaluation dataset generation framework," *arXiv preprint arXiv:2408.01262*, 2024, [retrieved: May, 2025].

[10] R. Liu *et al.*, "Best practices and lessons learned on synthetic data," *arXiv preprint arXiv:2404.07503*, 2024, [retrieved: May, 2025].

[11] S. Kim *et al.*, "Evaluating language models as synthetic data generators," *arXiv preprint arXiv:2412.03679*, 2024, [retrieved: May, 2025].

[12] S. Es, J. James, L. E. Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158.

[13] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, 2023, [retrieved: May, 2025].

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[14] I. Mirzadeh *et al.*, "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models," *arXiv preprint arXiv:2410.05229*, 2024, [retrieved: May, 2025].

[15] Y. Yu *et al.*, "Rankrag: Unifying context ranking with retrieval-augmented generation in llms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 121 156–121 184, 2024.

[16] G. d. S. P. Moreira *et al.*, "Enhancing q&a text retrieval with ranking models: Benchmarking, fine-tuning and deploying rerankers for rag," *arXiv preprint arXiv:2409.07691*, 2024, [retrieved: May, 2025].

[17] P. Mandikal and R. Mooney, "Sparse meets dense: A hybrid approach to enhance scientific document retrieval," *arXiv preprint arXiv:2401.04055*, 2024, [retrieved: May, 2025].

[18] Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu, and W. Wei, "Machine learning for synthetic data generation: a review," *arXiv preprint arXiv:2302.04062*, 2023, [retrieved: May, 2025].

[19] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[20] Hugging Face, "Model card for multi-qa-minilm-l6-cos-v1," [Online]. Available: https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1, 2021, [retrieved: May, 2025].

[21] ——, "Model card for msmacro-minilm-l6-en-de-v1," [Online]. Available: https://huggingface.co/cross-encoder/msmarco-MiniLM-L6-en-de-v1, 2021, [retrieved: May, 2025].

[22] ——, "Model card for ms-macro-minilm-l-6-v2," [Online]. Available: https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2, 2021, [retrieved: May, 2025].

[23] D. Park and S. Kim, "Probabilistic precision and recall towards reliable evaluation of generative models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 20 099–20 109.

[24] S. S. Ravi, B. Mielczarek, A. Kannappan, D. Kiela, and R. Qian, "Lynx: An open source hallucination evaluation model," *arXiv preprint arXiv:2407.08488*, 2024, [retrieved: May, 2025].

[25] S. Kim *et al.*, "Prometheus 2: An open source language model specialized in evaluating other language models," *arXiv preprint arXiv:2405.01535*, 2024, [retrieved: May, 2025].

[26] B. Adler *et al.*, "Nemotron-4 340b technical report," *arXiv preprint arXiv:2406.11704*, 2024, [retrieved: May, 2025].

[27] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels, 2022," *URL https://arxiv. org/abs/2212.10496*, 2022, [retrieved: May, 2025].

[28] M. Lenzerini and C. Daraio, *Challenges, Approaches and Solutions in Data Integration for Research and Innovation*. Cham: Springer International Publishing, 2019, pp. 397–420. [Online]. Available: https://doi.org/10.1007/978-3-030-02511-3_15

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

From Access to Affirmation: How Technology Exposure and Family Support Shape Confidence in Computer Science

A Study among Young Adults in Germany

Claudia Hess, Sibylle Kunz, Cornelia Heinisch and Adrienne Steffen
IT & Engineering / Management
IU International University of Applied Sciences
Erfurt, Germany
e-mail: claudia.hess@iu.org; sibylle.kunz@iu.org; cornelia.heinisch@iu.org; adrienne.steffen@iu.org

Abstract— **The high demand for computer scientists makes it crucial to understand the career decisions of young adults. This study examines how technology experience shapes young adults' perceived competence in computer science, which is a key factor in career decisions. A novel framework was developed to structure technology experience. This framework guided the quantitative analysis of data from over 800 young people in Germany. Special emphasis was placed on gender differences within this context of digitization adoption. The results show that while a gender gap in technology exposure exists, this disadvantage is compounded by a more significant disparity in the social support they receive at home. Family support emerges as a critical predictor of self-concept, yet it is marked by a stark gender gap: young men receive consistent encouragement for their technical abilities, while young women report significantly less support, often facing neutrality or even doubt. Therefore, fostering a diverse tech pipeline requires a holistic strategy that extends beyond the classroom, engaging with families to dismantle gendered stereotypes and build a supportive foundation for a wider and more diverse talent pool. Taken together, the study contributes a multidimensional framework of technology experience and shows that understanding gender disparities in computer science self-concept requires looking beyond access to the context in which technology use is embedded.**

Keywords: technology experience; family support; computer science self-concept; young adults; gender gap.

I. INTRODUCTION

Past research suggests that technology exposure has a measurable, though modest, influence on career-related beliefs. In particular, greater access to technology and more frequent use have been linked to young adults' confidence in their competence and their perceived ability to succeed in Computer Science (CS) [1]. Such access likely fosters baseline familiarity and comfort with technology. However, the study indicates that frequent exposure is not the sole factor driving the CS self-concept, particularly for young women. A person's self-concept in a particular field such as math, science or CS is typically shaped by past experiences, feedback, the immediate environment, and social comparisons [2], [3]. Crucially, this self-concept acts as a gatekeeper: without it, a person is unlikely to develop a genuine interest or consider a career in that field [2].

Since an individual's history with technology is a pivotal determinant of this self-concept, this study deconstructs that history. While previous research, particularly within school contexts, has examined how mere availability and use of Information and Communication Technology (ICT) affect students' confidence and competence in using technology or its impact on math and science achievement (e.g., [4], [5], [6]), this study takes it a step further. It investigates how the multidimensional nature of the technology experience, ranging from broad quantitative exposure to the **situational context** in which technology exposure occurs (such as in school or at home), relates to their CS self-concept. Building on this framework of multidimensional technology experience, the present study expands the analysis in two primary directions. First, it continues to examine how technology exposure relates to CS self-concept. Second, it introduces an in-depth analysis of one of the most critical situational contexts: the family. Specifically, it investigates how social-environmental feedback within the family—namely, parental encouragement for a CS career and expressed confidence in their child's CS-specific talent—influences a young adult's confidence and career perceptions in the field.

The confidence shaped by these factors is what ultimately drives a young adult's decision to enter and persist in a CS educational path. This educational pipeline is critically important, as the labor market continues to face a widening workforce gap. In recent decades, there has been a high demand for well-trained ICT specialists, and industry has complained about a shortage of skilled workers. Although advances in Artificial Intelligence (AI) raise questions about how automation will impact ICT jobs and the ongoing need for professionals, recent studies [7] predict continued strong demand for professionals skilled in both developing and using ICT and AI systems.

However, the extensive re- and upskilling required to meet this demand will only be successful if a broader demographic develops a genuine interest in CS. In this regard, a large gender gap exists. Despite the importance of ICT roles, women continue to be significantly underrepresented. In 2023, only 19.4% of ICT specialists in the EU were women, compared to 80.6% men [8]. The numbers are similar around the world. According to the Global Gender Gap 2024 report published by the World

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Economic Forum, only 28.2% of the positions in Science, Technology, Engineering, and Mathematics (STEM) worldwide are held by women, in contrast to 47.3% in non-STEM positions [9, p. 49]. This inequality is further amplified by the fact that women are more likely to hold a position with less influence and status, and consequently a lower salary [10]. This gender disparity poses several problems, including the loss of diverse perspectives that are vital for innovation.

To address these disparities, it is necessary to understand young adults' decision-making process underlying career choices. Therefore, the analysis presented in this paper investigates gender disparities in young adults' technology experiences. To this end, data from a survey of more than 800 young adults conducted in Germany in 2024 is analyzed. Understanding these dynamics will help educators, policymakers, and industry leaders encourage a diverse range of young people, especially women, to pursue careers in CS.

The remainder of this paper is organized as follows. Section II presents the theoretical background and derives the research questions. Section III details the methodological approach, including the study design, sample characteristics, and limitations. Section IV presents the quantitative results regarding technology exposure and its relationship to self-concept. Section V analyzes how young adults perceive family support for both their CS-specific talents and a potential career in the field. Section VI discusses the implications of the findings. Finally, Section VII concludes the paper and outlines areas for future research.

II. Theoretical Background

The following sections provide the theoretical basis for the core constructs used in this study. First, key dimensions of technology experience are defined, covering both quantitative aspects such as access and frequency of use, and the contextual and interactional dimensions of engagement with ICT. Subsequently, the construct of the ability self-concept is introduced, which is essential for explaining how specific technology experiences can foster or hinder the development of confidence and interest in CS.

A. Technology Exposure of Young Adults

Exploring the aspect of **technology exposure** requires understanding that it encompasses a wide array of experiences and opportunities through which individuals encounter and interact with ICT. Technology exposure can be conceptualized through a set of behavior- and experience-based indicators that show how much and how often people interact with technology or are exposed to it. These include the amount of time spent using technology, how many times it is used before a person feels comfortable with it, the frequency of technology use, participation in ICT-related trainings, and the context in which technology is used, for instance at work and at home [11]. They also include access-related aspects such as owning ICT devices. The exposure of young adults to technology has been examined in a number of studies from a variety of perspectives. Figure 1 provides an overview of these different perspectives.

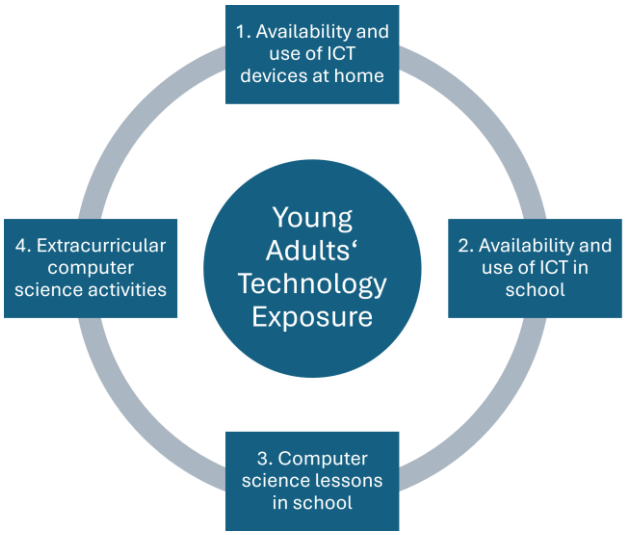


Figure 1. Technology exposure of young adults, Source: own depiction.

Apart from the term "technology exposure", the terms "technology experience" and "technology use" are also commonly used [11], as well as "ICT use" [4].

1) Availability and use of ICT devices at home: Access to digital devices such as computers, tablets, smartphones, game consoles and even Virtual Reality (VR) glasses at home allows young people to explore technology on their own. The presence of ICT devices can encourage experimentation and the development of digital literacy from an early age [12]. Note that, in addition to "classical" ICT devices such as personal computers, laptops, tablets, and smartphones, new devices have entered children's rooms. Programmable robot toys, in particular, offer playful and exploratory entry points into coding and computational thinking [13]. Therefore, engagement with hardware no longer necessarily involves assembling a personal desktop computer; it also includes hands-on activities such as constructing educational robots by connecting wheels, sensors, and actuators, or building simple electronic circuits.

In addition to the availability of ICT devices at home, the type of usage is often differentiated, according to whether it is used for leisure, e.g., the use of social media or gaming, or for schoolwork activities [4]. Analyses of the impact of leisure activities on interest in CS have highlighted gaming in particular. Recent studies confirmed previous research that playing games on computers, game consoles or mobile devices increases interest in CS [14], [15], [16]. In addition, more active engagement with games includes experimenting with mods, level editors, or in-game scripting.

To better understand the context, conditions, and support mechanisms that shape children's ICT use at home, Bonanati and Buhl describe four aspects characterizing a digital home learning environment [17]: stimulation, instruction, interactions, and modeling. Stimulation refers to the availability of and the access to ICT devices at home in general and as children's individual devices. Instruction encompasses the activities and strategies that parents utilize to support their children in using ICT. Interactions describe

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

shared ICT activities in which parents and children use ICT together. One example is searching for information online to plan a family event. Finally, modeling refers to parents' own attitudes toward ICT use, for example, whether they deem it as beneficial for their children to be familiar with ICT.

2) Availability and use of ICT in school: During the Covid-19 pandemic, distance learning was supported by video conferences and digital learning materials such as learning apps. This period accelerated the adoption of ICT in schools and highlighted the importance of digital tools for both teaching and learning. In many education systems, the pandemic, alongside longer-term digitalization efforts, prompted sustained investment in digital infrastructure and the availability of ICT in schools.

Students of all ages are learning basic ICT skills, such as ten-finger typing, internet search, interacting with the user interfaces of generative AI ("AI chatbots"), and using software applications, for instance for creating presentations or using spreadsheet and calculation software, and learning management systems. Sometimes, technology is integrated into different subjects, which helps students recognize the relevance of ICT across diverse fields. In these contexts, ICT may be used both by teachers, for example, to demonstrate concepts or model procedures, and by students themselves in an active and creative way. For example, students use their smartphones in music class to create a short commercial, adding background music that they either record themselves or select from provided audio files.

The extent to which ICT is used as a tool for learning—such as using educational software, engaging in interactive learning experiences via websites and mobile apps, or utilizing online resources—can have a significant impact on students' familiarity with technology [5].

3) CS lessons in school: In many countries, in-school offerings of CS education have been established during the last years. Formal education in CS, including hands-on experiences such as programming or working with robots, can foster students' interest in the field [6]. Such lessons can provide foundational knowledge and inspire confidence in students' abilities to pursue further study or work in this field.

However, the frequency and the level of detail of the CS lessons often varies [18]. In Germany, for example, there are considerable differences between the federal states as the annual "computer science monitor" published by the German Informatics Society reveals. It ranges from a very limited offer of CS lessons in some types of schools or in some grades, sometimes only on a voluntary basis, to several compulsory CS lessons in all types of schools [19]. Moreover, CS curricula differ between the German federal states [19]. To ensure a coherent and high-quality framework for CS education across Germany, the German Informatics Society has developed detailed recommendations and guideline for CS education. An example is the *Educational Standards for the Subject of Computer Science in Lower Secondary Education* published in 2025 [20]. A core consideration is that CS education should not focus on operating applications or devices, but on understanding fundamental concepts of how information and data are represented, processed, and transformed by machines [21]. Rather than merely "doing something with digital media", learners should be enabled to develop digital applications themselves and to creatively modify existing ones. This requires the ability to model problems and to realize solutions through implementation. The goal is to empower learners to not only use ICT, but also to actively shape digital systems.

4) Extracurricular CS activities: Participation in technology-related extracurricular activities, such as coding clubs, robotics teams, or technology competitions (e.g., hackathons) is another context in which teenagers and young adults can experience ICT. These activities often allow students to explore technology in very creative and applied ways [6]. Makerspaces, such as the one in the Amsterdam Public Library [22], illustrate how children can engage creatively with ICT by combining digital technologies with tinkering, for example by using 3D printers and taking part in coding and robotics workshops. Other formats have a more focused nature, with a stronger emphasis on CS and, at times, a competitive format. In Germany, for example, there are several nationwide competitions for school students, the "Bundesweite Informatikwettbewerbe", that aim to encourage engagement with CS and programming [23]. These competitions address different age groups and levels of experience. Not all of them include programming. The Informatik-Biber, for example, consists of short problem-solving tasks that foster computational thinking through everyday scenarios. Besides such competitions, students may also be exposed to online tech communities where people discuss technology trends, share projects, and exchange knowledge. These interactions not only provide informal learning opportunities but also help build a sense of belonging to the broader tech community.

B. Impacts of Young Adult Technology Exposure

The effects of technology exposure on young adults have been analyzed from different perspectives.

1) Impact on students' performance in science and math: In the last fifteen years, many studies used the data from PISA (Programme for International Student Assessment), OECD's Programme for International Student Assessment, which assesses how well 15-year-olds can apply their knowledge and skills in reading, mathematics and science to real-world tasks. In particular, the effects of ICT use in classrooms for teaching and learning purposes as well as ICT availability at home on students' performance in math and science were analyzed for various countries [4], [24], [25], [26], [27]. However, none of these studies analyzed the impact on achievement in CS in particular, which is obvious for studies using PISA data, as PISA has historically only tested students' literacy in reading, mathematics, and science.

The review of a large number of studies in [4] shows mixed results on how ICT availability, use, and engagement affect students' math and science scores. To provide clarity, they compared the results of the four PISA surveys from 2009, 2012, 2015 and 2018. Results indicated that there was

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

no meaningful positive link between student performance in math or science and their use of ICT, whether this usage occurred in school or outside of it, and whether it was tied to specific academic subjects or not [4]. However, students who had a more positive attitude toward ICT, felt confident using it, believed in its usefulness, and used it autonomously, consistently showed better performance in math and science. Similarly, [28] discusses the mixed results of ICT availability on students' academic achievements and examines this relationship in their study, too. In particular, gender differences are investigated. Apart from a direct effect, which was negative for male students and neutral for female students, the study has shown significant positive indirect effects via other variables such as the students' ability to master the use of ICT and their autonomy in using ICT. These aspects are discussed in the next paragraph.

2) Impact on digital literacy: Technology experience also influences comfort and competence with ICT. Starting in 2025, PISA will include a new assessment, the "Learning in the Digital World" test, which evaluates how well students can use computational tools to engage in an iterative process of building knowledge and tackling real-world problems [29]. Pre-tests of this test indicate that students who regularly use ICT across various subjects tend to perform better in solving the tasks [5]. Moreover, students who possess confidence in their ICT skills and show interest in ICT demonstrate higher performance levels on the test. Further studies investigate more specific aspects of students' ability to use ICT tools effectively. For instance, [30] examines how the usage of educational technology affects students' internet self-efficacy. While these studies focus on digital literacy, CS self-concept is about how confident students feel when it comes to engaging with CS tasks.

3) Impact on career decisions in STEM: Research has examined whether and how technology experience increases the sense of familiarity, confidence and interest in pursuing STEM subjects and, consequently, influences an informed career decision regarding a STEM career path. Among the different contexts of technology experience, these empirical studies often focus on STEM school lessons or STEM programs. An example is the impact analysis of a program for students about robotics design and production in the USA, which increased participants' career choice in STEM [31]. While many studies have examined the direct and indirect impact of multiple variables on STEM career choices of young adults (e.g., interest, self-efficacy, influence from family and friends) [18], less attention has been given to the role of ICT availability and use at home or during leisure time in career choice specifically in CS.

C. The Situational Context of Technology Exposure

Analyzing young adults' technology exposure makes clear that it is a multi-faceted experience that happens in various situational contexts. This context can be systematically deconstructed by addressing three fundamental questions: **Where** does the exposure occur? **With whom** is the individual interacting? And **how** are they engaging with technology? These three questions provide the analytical pillars for the framework that underpins our study.

1) Contextual Domains: The first dimension of situational context, **Where**, refers to the distinct Contextual Domains that structure a young adult's life. These domains are more than just physical locations; they are spheres of activity governed by different norms, resources, and expectations, each with a predominant purpose. Based on the literature review in Section II.A, three primary domains for technology exposure can be distinguished: school, home, and extracurricular settings. School provides a formal, structured, and often curriculum-driven context for ICT use. The home, in contrast, is an informal environment where purposes are more varied, ranging from leisure and social connection to self-directed learning and instrumental tasks like homework. Extracurricular domains, such as maker spaces, coding clubs, competitions and hackathons, are typically voluntary, interest-driven, and project-based.

2) Social Actors: The second dimension, **With whom**, focuses on the influential social actors present in each contextual domain. At home, the primary social actors are family members. At school, they are teachers and peers. In extracurricular settings, they are often mentors and like-minded peers who form a community of practice. These mentors shape young adult's experience with technology by providing support, offering feedback (which can be supportive or critical), establishing norms, and acting as role models.

3) Interaction Patterns: By analyzing **how** individuals use ICT devices (see Section II.A), recurring interaction patterns emerge. Regardless of whether the setting is domestic or educational, these patterns can be grouped into the following five distinct clusters of ICT-related activities:

- Acquiring general ICT skills and basic digital literacy (e.g., basic system operation, internet and email use, file management), including the everyday use of digital devices
- Using software applications to accomplish specific tasks (e.g., office tools for word processing, spreadsheets, presentations, editing images or videos)
- Creating and developing computational artefacts, such as programs or websites
- Engaging with hardware systems (e.g., assembling or upgrading computers), including robots and simple electronic kits
- Engaging in digital gaming activities on PCs, game consoles and mobile devices

Table I operationalizes this situational framework. Informed by the literature review in the preceding section, it describes for each contextual domain (home, school and extracurricular context) the involved social actors and the predominant interaction patterns with ICT. The characterization of interaction with the respective social actors is further based on the analysis by [32].

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I. A FRAMEWORK FOR ANALYZING TECHNOLOGY EXPOSURE CONTEXTS (SOURCE: OWN DEPICTION)

Dimension	**Contextual Domains (Where?)**		
	Home Context	***School Context***	***Extracurricular Context***
Social Actors *(With whom?)*	Family (parents, siblings, close relatives), friends (live, online) Interactions are informal and affective: • (+) Informal encouragement, emotional support, and instrumental help • (-) Can be a source of conflict, apathy, or restriction (e.g., over screen time, type of use)	Teachers, peers Interactions are formal and evaluative: • (+) Structured, constructive feedback and formal instructions • (-) Can be punitive, overly comparative, or create social pressure and exclusion among peers	Mentors, experts, like-minded peers Interactions are collaborative and performance-driven • (+) Fosters community through shared interest, mentoring, and specialized collaboration • (-) Can be intensely competitive, leading to gatekeeping, exclusion, or imposter syndrome
Interaction Patterns ***(How?)***	Broad & Interest-Driven: • High prevalence of digital gaming and using software applications for homework/leisure. • Provides unstructured opportunities for self-initiated creation of artefacts (e.g., personal websites), hardware engagement (e.g., PC building), and informal acquisition of skills.	Prescribed & Task-Oriented: • Strong focus on acquiring general ICT skills through formal instruction and using software applications for assigned tasks. • Creation of artefacts is typically limited to specific CS curricula. • Time is limited to the duration of the school lesson. • Unstructured gaming and hardware engagement are generally absent or restricted.	Focused & Project-Based: • Primary emphasis on creating computational artefacts and engaging with hardware systems (e.g., robotics). • Acquiring skills is advanced and goal-oriented. • Digital gaming may be reframed as a creative activity (e.g., game jams).

D. *Framework of Multidimensional Technology Experience*

Building upon these distinct theoretical foundations, this paper develops and puts forward a novel conceptual framework, which we term the **Multidimensional Technology Experience**. Figure 2 presents this framework, which integrates quantitative indicators with contextual and interactional dimensions into a unified model.

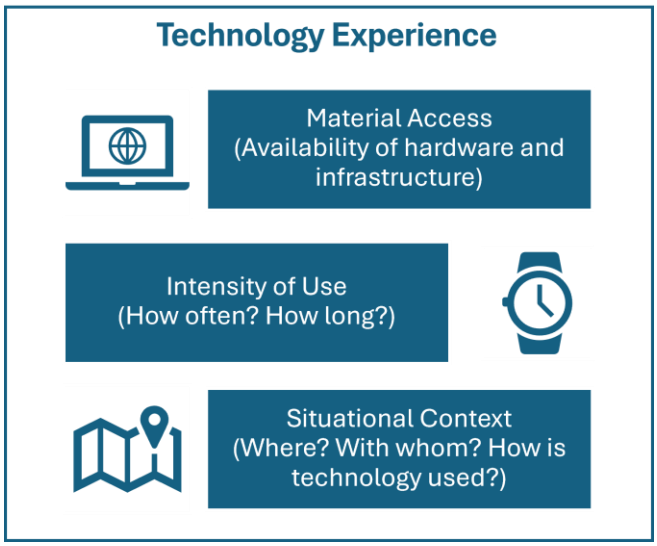


Figure 2. Technology experience, Source: own depiction.

At the quantitative level, 'Material Access' and 'Intensity of Use' serve as the primary indicators of technology exposure, defining the magnitude of the user's contact with ICT. It can be considered as a prerequisite condition for any digital literacy or extended engagement with ICT. This exposure is necessarily situated within specific situational contexts, capturing the socio-spatial setting (e.g., the structured environment of a school vs. the informal environment of the home) and the social nature of the task (individual vs. collaborative).

E. *Ability Self-Concept*

The expectancy-value theory by Eccles and Wiegfield (see e.g., [33]) posits that a person's motivation to engage in a particular task is determined on the one hand by their expectation of success in that task, and on the other hand, on the value they place on it. It helps explain motivational factors that translate into behavioral outcomes, such as selecting a particular study program or career path [2]. The ability self-concept refers to a person's perception of their own competence in a particular domain or area. It is the personal view on the own ability to perform tasks or succeed in activities related to this domain. An example that is intensively studied is the self-concept of math ability, or short math self-concept. It is influenced by previous experiences, feedback from significant others (such as parents, teachers, and peers), the immediate social environment, and social comparisons [2]. For instance, a student's math self-concept would be shaped by their past

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

performance in math (e.g., in tests), the feedback they receive about their math skills, and how they see themselves compared to their classmates in math-related tasks.

1) CS Self-Concept: Just as with math self-concept, the idea of self-concept can be applied to CS, corresponding to the question "How do I assess my own skills in CS?" There are studies that address the self-concept of computer ability (see e.g., [2]), which typically pertains to general computer usage skills, such as using software applications, or internet navigation. However, this idea differs from a CS self-concept, which is concerned with perceptions related to CS as an academic and professional field, including for instance programming, understanding algorithms, or problem-solving. Unlike the large number of studies on math and science self-concept, the CS self-concept seems not well covered.

2) Gender differences in STEM-related self-concepts: Past research often revealed gender differences in STEM-related self-concepts, being already apparent at school. Typically, boys express greater confidence in their success and skills in areas traditionally considered male-dominated, such as mathematics, physical sciences, and sports [2]. Conversely, girls tend to have a higher self-concept of their verbal abilities compared to boys.

3) Influencing Factors on CS Self-Concept: Self-concept is not formed in a vacuum. It is profoundly shaped by social factors such as school interactions, family support, peer influence, the availability of relatable role models to look up to, and personal interest [2]. Regarding the CS self-concept, especially the situational contexts in which young people are exposed to technology are relevant. In structured settings like a classroom, a student's CS self-concept is shaped by feedback, especially from teachers. Grades, teacher comments in class, and competition results establish a system where students are openly judged, which can motivate but also demotivate them [32]. Students inevitably benchmark themselves against their peers, observing who is quicker, more adept, or more celebrated for their abilities in areas like coding, debugging, or robotics.

The home context, however, works more subtly, shaping the value and emotional tone associated with CS. Feedback is not delivered through grades but through daily conversations with family and friends, shared activities, and the implicit encouragement or discouragement of certain pursuits. A parent's enthusiasm can frame a technical challenge as a valuable learning opportunity, while frequent restrictions can signal that such activities are a trivial waste of time.

F. Research Questions

The studies presented in the literature review analyzed the effects of technology experience from various perspectives. However, to the best of our knowledge, they do not examine how the different dimensions of this experience—ranging from the access, intensity of use to the situational context—are associated with interest in and self-concept regarding CS.

To address this, the inquiry begins with the logical foundation: the relationship between technology exposure and CS self-concept. The principle is straightforward: increased access and regular use are expected to build familiarity and technical competence, which in turn should strengthen an individual's CS self-concept, i.e., their belief in their own ability to succeed in the field of CS.

While this provides an essential starting point, it treats exposure primarily as a material phenomenon, often overlooking the crucial situational context in which it occurs. As discussed in Section II, the informal and affective nature of the home environment is profoundly different from the more structured, evaluative contexts of school or extracurriculars.

Building upon this foundational model, it is hypothesized that the social dynamics within the home, specifically, the degree of supportive family interaction with regard to CS, act as a key mediating factor. This proposes that it is not merely access to technology, but the encouragement surrounding its use, that builds a resilient CS self-concept. This integrated line of reasoning is operationalized in the following research questions. Furthermore, special emphasis is placed on differences between young women and young men, as studies in STEM often report significant gender disparities (see Section II.B).

- **RQ1 (The Foundational)**: To what extent is technology exposure in the home context related to young adults' CS self-concept, and are there significant differences between young women and young men?
- **RQ2 (The Contribution of Family Support)**: To what extent is the perceived family support for CS related to young adults' CS self-concept, and are there significant differences between young women and young men?

III. Methodology

The next section presents the methodology, including the study design, the sample composition, and the data analysis strategy. Moreover, it discusses the limitations of the presented approach.

A. Study Design

The data analyzed in this paper were collected in a German-wide study on young adults' career choices with regard to CS. Ethical clearance was obtained from the ethics committee of the authors' institution. In June 2024, an online survey was conducted in collaboration with a reputable market research organization who already had a market panel consisting of young adults in all federal states of Germany. The panel is ISO 20252 certified and adheres to strict ethical standards, including those set by, for example, DGOF, ESOMAR, and ADM, for participation in market research studies. Students under the age of 18 may join the panel only with the written consent of their parents. This form of collaboration was chosen because regulations on whether and how students can be contacted via schools differ among Germany's federal states, significantly limiting direct access to the student population. The study was funded by the authors' higher education institution. The authors declare no conflicts of interest related to this work.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

The online questionnaire contained closed and open-ended questions, taking a total of about 15 minutes to complete. This survey explored a diverse range of topics, including the determinants of young adults' career decisions, their perceptions of CS as a viable professional path, and different factors influencing young adults' career decisions for or against CS. One set of questions focused on technology exposure, specifically the availability and the regular use of ICT devices at home. Another set of questions asked participants how they perceive their family support with regard to CS.

B. Sample Composition

The online survey targeted students and young adults between 15 and 20 years of age who are in the process of choosing a career, or who have just decided on a course of study or occupation. As the aim of the study is to analyze the factors that influence career choice, it is important that participants are in this decision phase. For this reason, participants outside this age group were excluded from the survey. Students under the age of 15 often have not yet decided on a particular profession or field of study. In contrast, young adults over the age of 20 have usually already taken this decision. A total of 822 respondents participated in the online survey. The sample comprised 315 (38.3%) young men, 506 (61.6%) young women, and one person who identified as non-binary (0.1%). Table II summarizes the sample characteristics.

TABLE II. SAMPLE CHARACTERISTICS

Age	Results			
	Female N (%)	***Male N (%)***	***Non-binary N (%)***	***Total N (%)***
15	37 (7.3%)	44 (14.0%)	0	81 (9.9%)
16	60 (11.9%)	53 (16.8%)	0	113 (13.8%)
17	84 (16.6 %)	58 (18.4%)	0	142 (17.3%)
18	133 (26.3%)	75 (23.8%)	0	208 (25.3%)
19	99 (19.6%)	37 (11.7%)	1 (100%)	136 (16.6%)
20	93 (18.4%)	48 (15.2%)	0	141 (17.2%)
Total	**506 (100%)**	**315 (100%)**	**1 (100%)**	**822**

C. Data Analysis Strategy

Descriptive statistics are used to characterize the patterns of technology exposure among young adults, specifically focusing on indicators of exposure (self-reported availability and regular use) within the home context. Subsequently, inferential statistics (independent-samples t-tests) are applied to examine significant differences between young women and men. Finally, correlation and regression analyses are performed to determine the extent to which these exposure metrics are associated with the participants' CS self-concept. The same approach is applied to analyze the perceived family support for CS.

D. Limitations

There are some limitations based on the study design and the data obtained. Although the participants were drawn from all federal states, the data are not fully representative of all German young adults and should not be generalized. Regarding the respondents' age, the young men were slightly younger than the young women. The data set contains only one non-binary person, who was excluded from all gender-specific analyses as no meaningful statistical conclusions can be drawn from a single participant. Moreover, the assessment of technology exposure has certain limitations, notably that it did not quantify the frequency of use of each device in terms of hours per week, nor did it assess the purpose of use like it was done for instance by [27].

IV. ANALYSIS OF TECHNOLOGY EXPOSURE

The quantitative analysis operationalized technology exposure through self-reported availability and regular use of ICT devices at home. It considers both the physical access (Do I have the device?) and functional access (Do I have internet?). The intensity of use is approximated by regular use, which serves as a proxy for sustained engagement with technology.

A. Young Adults' Technology Exposure at Home

The survey asked respondents about the use of IT devices at home, especially a computer/laptop or tablet, the constant availability of Internet access, and the availability of a game console or VR glasses at home. A Likert scale of 1 and 5 was used to scale the responses, ranging from strong disagreement (1) to strong agreement (5).

1) Overview: First of all, an overview of the young adults' exposure to technology is provided. Table III shows the extent of technology exposure of the young adults surveyed. The vast majority of the young adults, 83.4% (sum of the top-2 boxes "agree" and "strongly agree"), have consistent internet access, underscoring prevalence of connectivity. Only 7.2% (sum of "strongly disagree" and "disagree") report limited internet access, indicating a minority facing potential barriers to ICT.

About two-thirds, 66.3% (top-2 boxes), report regularly using a computer or laptop at home, indicating widespread access. A small percentage, 13.9% (top-2 boxes), do not engage in regular use, which may reflect either a lack of necessity or access. Tablets are also common, with 61.5% (top-2 boxes) of respondents engaging regularly. A relatively higher percentage (19.9%) disagrees or strongly disagrees with regular usage compared to computers/laptops (13.9%).

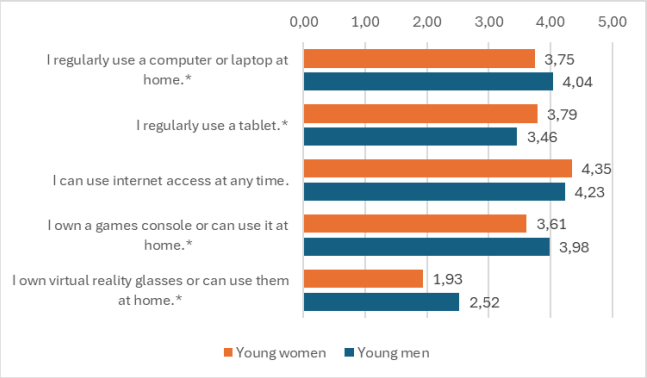
Game consoles are popular, with 66.1% (top-2 boxes) of respondents having access, while 17.8% do not. VR glasses have the least exposure, with only 19.6% reporting access and 67.2% lacking or not using them at home. These low figures reflect either the novelty, cost barrier, or niche market status of VR technology compared to more established devices like computers and tablets.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE III. RESULTS ON TECHNOLOGY EXPOSURE

Technology exposure at home	Answers: Total number = 822 (100%)				
	Strongly disagree N (%)	*Disagree N (%)*	*Neutral N (%)*	*Agree N (%)*	*Strongly agree N (%)*
I regularly use a computer or laptop at home.	35 (4.3%)	79 (9.6%)	163 (19.8%)	236 (28.7%)	309 (37.6%)
I regularly use a tablet.	70 (8.5%)	94 (11.4%)	153 (18.6%)	229 (27.9%)	276 (33.6%)
I can use internet access at any time.	18 (2.2%)	41 (5.0%)	77 (9.4%)	223 (27.1%)	463 (56.3%)
I own a games console or can use it at home.	74 (9.0%)	72 (8.8%)	133 (16.2%)	245 (29.8%)	298 (36.3%)
I own virtual reality glasses or can use them at home.	362 (44.0%)	191 (23.2%)	108 (13.1%)	98 (11.9%)	63 (7.7%)

2) Gender Differences: The next step was analyzing differences in technology availability and accessibility among young women and men. Figure 3 shows the technology exposure reported by both genders for all five statements. Using an independent samples t-test, statistically significant gender differences could be identified. Effect size is measured using Cohen's d and interpreted according to Cohen [34], with values of approximately 0.2, 0.5, and 0.8 indicating small, medium, and large effects, respectively.



t-test, N female: 506, N male: 315, *. p <0 .001

Figure 3. Gender Differences in Technology Exposure.

Regarding the regular use of a computer or laptop, young men have a higher mean score (M_m=4.04, and Standard Deviation SD_m=1.108) compared to young women (M_f=3.75, SD_f=1.164). The difference is statistically significant (t(691)=-3.164, p<0.001) with a small effect size (Cohen's d= -0.256), suggesting a slight gender preference towards computer usage among males. The opposite is the case for tablets, which are used more by young women (M_f=3.79, SD_f=1.204) than young men (M_m=3.46, SD_m=1.369). This difference is also statistically significant (t(602)=-3.518, p<0.001) with a small effect size (Cohen's d= -0.260).

Young men (M_m=3.98, SD_m=1.182) reported higher values with regard to the ownership and use of games consoles than young women (M_f=3.61, SD_f=1.311). This difference is statistically significant (t(717)=-4.219, p<0.001) with a small effect size (Cohen's d=-0.296), indicating a more noticeable gender disparity. With respect to the ownership or use of VR glasses, young men have a substantially higher mean (M_m=2.52, SD_m=1.408) compared to young women (M_f=1.93, SD_f=1.191), indicating greater access or usage among males. This disparity is statistically significant (t(584)=-6.175, p<0.001) with a medium effect size (Cohen's d= -0.461).

Regarding the availability of internet access, there is no statistically significant difference between the genders (M_f=4.35 / M_m=4.23, SD_f=0.915 / SD_m=1.078).

B. Young Adults' CS Self-Concept

To assess the CS self-concept, an established scale based on the expectancy-value theory was used. The items used by [3] to measure expectations of math success were adapted. For instance, the item "I am talented at math" was changed to "I am talented at CS". The same 5-point Likert scale [1-strongly disagree to 5 strongly agree] was used as for the measurement of technology exposure. For the following independent samples t-test, these items are aggregated for each participant to a single score "CS self-concept".

The average CS self-concept in the sample was M=2.96 (SD=0.950). The results show some caution in self-assessing one's competence, with many values falling in the moderate range. Comparing genders, young men have a higher average CS self-concept (M_m=3.35, SD=0.891) than young women (M_f=2.72, SD=0.906). The difference is statistically significant (t(819)=-9.735, p<0.001) with a medium to large effect size (Cohen's d = 0.699).

C. Impact of Technology Exposure On CS Self-Concept

Having analyzed technology exposure and CS self-concept separately, the relationship between these two concepts is explored. Like the CS self-concept, the overall technology exposure scale score for each participant is calculated based on the mean score of the above presented five items measuring various aspects of technology exposure.

1) Correlation: First, it is explored whether there is a positive correlation between technology exposure and CS self-concept. The hypothesis is: The higher a young adult's overall exposure to technology, the higher their CS self-concept. Pearson's Correlation Coefficient was calculated. According to [35], there was a moderate to strong positive correlation between technology exposure and CS self-concept, r(820)=.403, p<0.001. Looking only at young men, a strong positive correlation can be seen with r_m(313)=.542,

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

p<0.001, while there is a moderate positive correlation ($r_f(504)=.292$, p<0.001) for young women.

2) Regression Analysis: Technology Exposure as Predictor of CS Self-Concept: The next step was to examine whether technology exposure predicts the CS self-concept. Overall, a significant regression was found with $R^2 = .163$, $F(1,820) = 159.297$, $p < .001$, corresponding to a moderate effect size according to [34].

For young women, the analysis indicated a significant regression was found with $R^2=.084$, $F(1,504)=47.134$, $p<0.001$, corresponding to a moderate effect (f=0.30). Conversely, for young men, a strong effect (f=0.65) was observed, with $R^2= .294$, $F(1,313) = 130.110$, $p< .001$. These results suggest that technology exposure has varying predictive power on the CS self-concept based on gender. For young women, technology exposure is a minor predictor of their CS self-concept, explaining only 8.4% of the variance. In contrast, it is a substantial predictor for young men. Approximately 30% of their CS self-concept is explained by technology exposure. This means that there are other factors that explain the remaining 70%, which could be factors such as school experiences, family background, or friends.

In summary, to answer RQ1, technology exposure in the home context (availability and regular use) has a moderate relationship with young adults' CS self-concept overall. However, there are profound and significant differences between genders. For young men, technology exposure is a strong predictor, accounting for a substantial 29.4% of the variance in their self-concept. In stark contrast, for young women, the relationship is much weaker, explaining only a modest 8.4% of the variance. This indicates that the home technology environment plays a significantly more influential role in shaping young men CS self-concept than it does for young women.

V. Analysis of Young Adult's Perceived Family Support for CS

After analyzing young adults' technology exposure at home, the focus now shifts to the support that young adults perceive their parents to provide with regard to CS.

A. Young Adults' Perceived Family Support for CS

To measure perceived family support, the survey asked respondents about their parents' beliefs regarding CS as a career and the young adult's personal competence in the subject. The questionnaire employed four items adapted from the scale used by Esch & Grosche [34], which is designed to capture parental encouragement and esteem support. Respondents rated their agreement with statements on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree).

1) Overview: Table IV shows young adults' perception of their parent's beliefs regarding CS and their own competence in this field. The descriptive analysis of the four family support items (N=822) indicates that young adults, on average, perceive their family support as modestly positive. With all mean scores hovering slightly above the neutral midpoint of 3.0 (ranging from M=3.11 to M=3.37 on a 5-point scale), a deeper look at the frequency distributions reveals a more complex and polarized landscape than the means suggest.

A notable contrast emerges between the two items measuring perceived competence. Young adults felt the most support for their general "technical talent", with a majority (50.3%) agreeing with this statement. However, this encouragement did not translate to the specific discipline of CS. In fact, the item measuring perceived "talent in computer science" received the strongest negative feedback, with nearly one-third of respondents (31.9%) actively disagreeing. This gap suggests that while families praise general tech-savviness, they are far less likely to affirm their child's specific abilities within the academic field of CS.

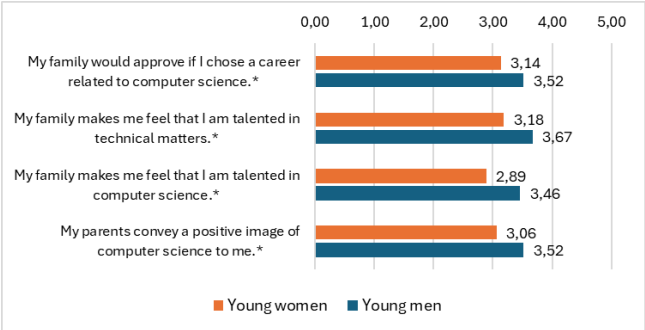
Another key finding is the large number of neutral responses. For questions about career approval and the image of CS, "neutral" was the single most common answer, chosen by nearly one-third of respondents (30.8% and 31.5%). This widespread neutrality suggests that for a large portion of young adults, their family's attitude is not one of support or opposition, but of simple indifference or disengagement. This points to a missed opportunity for parental encouragement.

TABLE IV. Results of the Perceived Family Support for CS

Perceived Family Support for CS	Answers: Total number = 822 (100%)					
	Strongly disagree N (%)	***Disagree N (%)***	***Neutral N (%)***	***Agree N (%)***	***Strongly agree N (%)***	***Average (Standard deviation)***
My family would approve if I chose a career related to computer science.	64 (7.8%)	130 (15.8%)	253 (30.8%)	258 (31.4%)	117 (14.2%)	3,28 (1.13)
My family makes me feel that I am talented in technical matters.	60 (7.3%)	126 (15.3%)	223 (27.1%)	275 (33.5%)	138 (16.8%)	3,37 (1.15)
My family makes me feel that I am talented in computer science.	91 (11.1%)	171 (20.8%)	212 (25.8%)	255 (31.0%)	93 (11.3%)	3,11 (1.19)
My parents convey a positive image of computer science to me.	69 (8.4%)	136 (16.5%)	259 (31.5%)	247 (30.0%)	111 (13.5%)	3,24 (1.34)

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

2) Gender Differences: The data clearly shows that young men and women experience family support for CS in fundamentally different ways (see Figure 4). Across all four categories, young men reported significantly (p<0.001) higher levels of encouragement than young women.



t-test, N female: 506, N male: 315, *. p = 0.000

Figure 4. Gender Differences in Perceived Family Support for CS.

Gender disparities are most pronounced in the area of perceived competence. The most striking difference is in feeling "talented in computer science", where the average score for young women (M_f=2.89, SD_f=1.188) falls below the neutral midpoint. This suggests that young women, on average, feel their families actively doubt their specific abilities in the field, a stark contrast to the clear encouragement young men report (M_m=3.46, SD_m=1.095). The difference is statistically significant (t(819)=-6.924, p=.000) with a medium effect size (Cohen's d= -0.497). A similar, though less severe, pattern appears for general "technical talent". While young women feel modestly supported (M_f=3.18, SD_f=1.143), they receive significantly less praise than their male counterparts (M_m=3.67, SD_m=1.091), and (t(819)=-6.028, p<0.001) with a small to medium effect size (d=-0.433).

This trend extends to broader career support, where young men report consistent, positive reinforcement from their parents for choosing a CS career (M_m=3.52, SD_m=1.057) and a positive image of the field (M_m=3.52, SD_m=1.081). Young women, in contrast, report scores barely above neutral (M_f=3.14, SD_f=1.149, and M_f=3.06, SD_f=1.134, respectively), indicating a lack of active family endorsement for pursuing a CS path. For both items, the differences are statistically significant (t(819)=-4.780, p<0.001 for a CS career, and t(819)=-5.680, p<0.001 for a positive image of CS) with a small effect size (d=-0.343 and d=-0.408).

B. *Impact Of Perceived Family Support for CS On CS Self-Concept*

Now, the relationship between young adults' perceived family support for CS and their CS self-concept is explored. Therefore, the overall perceived family support scale score for each participant is calculated based on the mean score of the above presented four items measuring different aspects of family support.

1) Correlation: A Pearson correlation coefficient was calculated for both young women and young men to evaluate the relationship between perceived family support for CS and CS self-concept. The strength of the relationship is interpreted according to [33]. The hypothesis is that the higher a young adult's perceived family support for CS, the higher their CS self-concept. For young women, a significant strong positive relationship between perceived family support for CS and CS self-concept was found (r_f(504)=0.620, p<0.001). Considering only young men, there was a significant positive relationship, too (r_m(313)=0.657, p<0.001). A direct comparison of the coefficients shows that the strength of the relationship between family support and self-concept is very similar for both genders, though slightly stronger for young men. As hypothesized, the results show that the more family support a young adult perceives, the higher their CS self-concept.

2) Regression Analysis: Family Support as Predictor of CS Self-Concept: Then, regression analysis was used to examine whether perceived family support for CS predicts the CS self-concept. For young women, there was a significant regression F(1,504) = 315.264, p< .001. The R^2 was 0.384, indicating that perceived family support for CS explained 38.4% of the variance in CS self-concept. The effect size is large (f=0.79), interpreted according to [34]. For young men, a significant regression was found F(1,313) = 238.035, p<0.001. The R^2 was 0.432, indicating that perceived family support for CS explained 43.2% of the variance in CS self-concept, corresponding with a large effect size (f=0.87). These results suggest that family support is a powerful predictor of CS self-concept for all young adults. However, the model indicates that family support has slightly more predictive power for young men's self-concept than for young women's.

In summary, to answer RQ2, perceived family support for CS has a strong positive relationship with young adults' CS self-concept overall. The key finding is that this support is a foundational predictor for both genders, explaining a substantial 43.2% of the variance for young men and a similarly large 38.4% for young women. The critical disparity, therefore, is not in the importance of this support, but in the amount of support that is actually perceived. As the analysis in Section V.A showed, young women report receiving significantly less encouragement than young men. The significant gender disparity lies not in the strength of this relationship, but in the unequal distribution of support itself, with young women reporting significantly lower levels of encouragement than young men, depriving them of a crucial resource for building confidence.

VI. Discussion and Implications

This section discusses the key findings from both the quantitative analyses and explores their broader implications.

A. *From Access to a Broader Picture of Technology Experience*

The analysis of the survey data on technology exposure at home (Section IV) confirmed that the widespread access

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

to ICT devices – particularly computers, laptops, tablets, and game consoles – can nurture a general comfort with ICT. A high technology exposure had an overall moderate effect on young adults' CS self-concept. However, the data indicates gender differences. Regarding technology exposure at home, males generally reported higher usage of computers/laptops, game consoles, and VR technology, while females reported higher usage of tablets. The differences, though statistically significant, mostly indicate small effect sizes except for game consoles and VR, which show moderate gender-specific preferences. These results are in line with a German-wide study on the media equipment and media usage of teenagers (12-19 years) in 2024 [36]. In terms of owning devices, boys are more likely to own fixed and portable game consoles and computers compared to girls. In contrast, girls tend to own smartphones, tablets, and e-book readers more frequently than boys [36, pp. 7–8]. The study did not ask for the ownership or usage of VR glasses.

Bitkom e.V. analyzed the usage of consumer technology in Germany in 2024 [37], differentiating between age groups but not gender. In the group of 16 to 29-year-olds, 43% have already used VR glasses. Nearly half of this age group thinks VR glasses will become standard household equipment in the future. In the study presented here, about 20% of young adults reported the availability of a VR glass at home. This discrepancy between usage and availability suggests that while many young adults may have tried VR glasses, possibly at exhibition centers, in museums or event arenas, fewer of them have made the transition to owning this technology. This might be due to factors such as cost, perceived need or little knowledge of possible applications.

The analysis in Section IV showed an overall moderate effect of technology exposure on young adults' CS self-concept. Placing this effect in the context of device availability suggests that the widespread access to ICT devices – particularly computers, laptops, tablets, and game consoles – can nurture a general comfort with ICT. While this baseline familiarity could set a foundational level of confidence, more specific or advanced technology experience might be required to notably impact self-concept. For instance, using a computer for programming or a VR setting for simulations may contribute more significantly than passive use for entertainment.

Beyond differences in access and usage, the study identified a crucial gender disparity in how exposure translates into confidence. While technology exposure serves as a substantial predictor of CS self-concept for young men, its predictive value for young women is comparatively minor. This finding fundamentally underscores the limitations of purely quantitative metrics of technology exposure. Simply having access to ICT devices sets the baseline, but is insufficient to foster a robust belief in young adults' own capabilities with regard to CS. These results validate the theoretical necessity of the multidimensional technology experience framework introduced in Section II, which expands the analytical lens to include the situational context in which technology exposure takes place.

B. When Support Falls Short: Gender Gaps in Family Encouragement for CS

The analysis highlights a critical tension in the formation of young women's CS self-concept. On one hand, the quantitative analysis revealed that young women report a statistically significant lower level of exposure to computers, laptops, and game consoles. On the other hand, our work suggests that bridging this exposure gap is only part of the solution. Simply providing a young woman with a laptop is unlikely to be effective if she believes CS is a "boy's club", lacks female friends who share her interest or relatives working in the field, and feels her questions and contributions in class are overlooked. This example illustrates that the formation of self-concept is influenced by multiple factors, such as family encouragement, peer relationships, school engagement, role models, and one's own interests [2], [3].

It is precisely within this family context that the central finding of this study emerges. The analysis shows that perceived family support is a foundational predictor of CS self-concept for all young adults. Critically, however, the results also demonstrate that this vital resource is unequally distributed: young women report receiving significantly less support than young men, particularly regarding their competence. This means young women are systematically deprived of the very social-emotional support that is most critical for building the confidence needed to persist in the field of CS.

The results may point into a further direction, namely how feedback is given. Literature on this topic highlights the pronounced impact of feedback for young women [38], suggesting that the most empowering feedback often combines two elements: process-related feedback that highlights their effective strategies and effort, and person-related feedback that recognizes their talent and their abilities in STEM [39]. Our results suggest that families may be failing at this second, crucial part. While a parent might offer general encouragement for being "good with technology" (a form of process/effort praise), our data shows a significant drop-off in affirming their daughter's specific talent "in computer science". This lack of direct, identity-affirming validation in the specific domain of CS is likely a key mechanism behind the support deficit we identified.

Interventions should therefore target not only access to technology, but also the communicative practices through which competence in CS is recognized and affirmed. An effective intervention is thus not simply to tell families to "be more supportive". It is to guide them on how to provide this specific, two-pronged feedback. Just as emerging research offers tools to help teachers craft more effective feedback for female students [40], future work could focus on creating accessible resources, such as workshops or informational guides, to help parents recognize and validate their daughters' emerging identities as capable computer scientists.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

VII. Conclusion and Future Work

The study examined the relationship between technology experience and CS self-concept among young adults, with a particular focus on gender disparities. Building on a quantitative analysis in [1], this work provides a complementary quantitative analysis of the perceived family support for CS reported by the same survey cohort. A central contribution of this study is a novel, multidimensional framework for structuring technology experience. It comprises three key dimensions: Material access, intensity of usage and the situational context in which the technology is used.

The findings reveal that while technology exposure is a factor, it is insufficient on its own to shape a strong CS self-concept, particularly for young women. Mere access to technology is not enough. Consider the following example: two users may engage in the exact same activity, such as engaging with hardware, yet experience vastly different levels of cognitive demand and personal agency. For instance, one student might simply watch a video demonstrating the assembly of a robot (passive observation), while another student assembles the hardware and codes the robot's movement (constructive creation). To allow for a more nuanced analysis, future work should extend the multidimensional framework by adding further aspects on how individuals interact with ICT. Such richer forms could be characterized by increased user agency over digital systems, greater involvement in the modification or creation of digital artefacts, or deeper engagement with underlying technological structures such as software, hardware, or computational logic.

Furthermore, the results confirm the importance of a supportive context, underscoring the critical role of the family and their support in fostering confidence and interest in CS. This study demonstrates that family support is a powerful predictor of CS self-concept for everyone, yet it is a resource that young women report receiving far less of. This disparity at home represents a critical, often-overlooked barrier that educational interventions alone cannot overcome. This highlights a critical intervention point: equipping parents with the awareness and language to move beyond gendered praise and specifically affirm their daughters' potential as creators and problem-solvers in the field, thereby building the resilient self-concept needed to succeed. Based on the findings, three primary avenues for future research emerge: intervention studies to test solutions, longitudinal studies to understand development over time, and qualitative studies to explore mechanisms in depth.

First, to directly address the support gap identified in this study, intervention research should focus on designing and testing practical solutions. One promising avenue is the development of parent-child joint CS activities, such as co-designing a simple video game or building a programmable robot. For young women specifically, such an intervention would create a shared, positive experience in which parents can be guided to offer the specific, identity-affirming feedback that our study shows is often missing. Assessing the impact of such an intervention through a pre-post experimental design could provide powerful evidence for a scalable solution that strengthens the entire family support system.

Second, to move beyond correlation and understand the developmental trajectory of CS self-concept, a longitudinal study is a critical next step. Following a cohort of young adults from middle school through their early college years would allow researchers to map how family support, technology experiences, and self-concept interact and influence each other over time. Such a longitudinal design would show how different types of family support (e.g., emotional vs. instrumental) from different caregivers (e.g., mothers vs. fathers) shape career aspirations and academic choices over the long term.

Finally, while our quantitative data shows *that* a support deficit exists, qualitative research is needed to understand the lived experiences *behind* these numbers. A valuable first step in this direction can be taken by analyzing the open-ended data on "critical incidents" already collected in the present study. A first screening has revealed that some incidents describe successfully using ICT, installing software, or solving hardware issues, and highlight how overcoming technical problems can foster a sense of pride and motivate further exploration of CS. This provides a rich inventory of competence-building events. The crucial next step, best explored through in-depth, semi-structured interviews, would be to understand the social context of these very incidents.

By addressing these areas, future work can inform strategies that create a diverse and inclusive CS workforce.

Acknowledgment

The authors gratefully acknowledge the support of their institution, which funded this research as part of a research incubator initiative.

References

[1] C. Hess, S. Kunz, C. Heinisch, and A. Steffen, "Exploring the influence of technology exposure on computer science self-concept: A study among young adults in Germany," in Proceedings of Advances on Societal Digital Transformation, Venice, Italy, 2025, pp. 33–38. Available from: https://www.thinkmind.org/library/DIGITAL/DIGITAL_2025/digital_2025_2_40_10029.html, retrieved: 05, 2026

[2] M. Sáinz and J. Eccles, "Self-concept of computer and math ability: Gender implications across time and within ICT studies," *J. Vocat. Behav.*, vol. 80, no. 2, pp. 486–499, Apr. 2012, doi: 10.1016/j.jvb.2011.08.005.

[3] S. González-Pérez, R. Mateos de Cabo, and M. Sáinz, "Girls in STEM: Is it a female role-model thing?," *Front. Psychol.*, vol. 11, pp. 1–21, Sep. 2020, doi: 10.3389/fpsyg.2020.02204.

[4] M. Courtney, M. Karakus, Z. Ersozlu, and K. Nurumov, "The influence of ICT use and related attitudes on students' math and science performance: multilevel analyses of the last decade's PISA surveys," *Large-Scale Assess. Educ.*, vol. 10, no. 8, pp. 1-26, Aug 2022, doi: 10.1186/s40536-022-00128-6.

[5] OECD, "Technology use at school and students' learning outcomes: exploring the relationship with new data from Germany," *OECD Education Spotlights*, vol. 17, 2022. [Online]. Available from: https://www.oecd.org/content/dam/oecd/en/publications/repor

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

ts/2024/12/technology-use-at-school-and-students-learning-outcomes_4c4f92e6/422db044-en.pdf, retrieved: 05, 2026

[6] C. Chen, J. Rothwell, and P. Maynard-Zhang, "In-school and/or out-of-school computer science learning influence on CS career interests, mediated by having role-models," *Comput. Sci. Educ.*, vol. 34, no. 4, pp. 753–777, Oct. 2024, doi: 10.1080/08993408.2023.2290435.

[7] AI-Enabled ICT Workforce Consortium, "The transformational opportunity of AI on ICT jobs," 2024. [Online]. Available from: https://www.cisco.com/c/dam/m/ai-enabled-ict-workforce-consortium/report.pdf, retrieved: 05, 2025

[8] eurostat, "ICT specialists in employment." [Online]. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=ICT_specialists_in_employment, retrieved: 05, 2025

[9] World Economic Forum, "Global gender gap 2024 - Insight report." [Online]. Available from: https://www3.weforum.org/docs/WEF_GGGR_2024.pdf, retrieved: 01, 2026

[10] E. Young, J. Wajcman, L. Sprejer, and The Alan Turing Institute, "Where are the women? Mapping the gender job gap in AI," Jan. 2021. [Online]. Available from: https://www.turing.ac.uk/sites/default/files/2021-03/where-are-the-women_public-policy_full-report.pdf, retrieved: 01, 2026

[11] M. Salanova and S. Llorens, "Exposición a la Tecnología de la Información y la Comunicación y su Relacion con el Engagement [Exposure to Information and Communication Technology and its Relationship to Work Engagement]," Ciencia & Trabajo, vol. 11, pp. 55-62, no. 32, 2009.

[12] J. Hu and J. Hu, "The influence of age at first regular digital device access on digital reading performance: the mediating effect of cognitive flexibility," Humanit. Soc. Sci. Commun., vol. 11, no. 1, pp. 1–18, Jun. 2024, doi: 10.1057/s41599-024-03292-y.

[13] H. Ehsan, C. Ohland, and M. E. Cardella, "Characterizing child–computer–parent interactions during a computer-based coding game for 5- to 7-year-olds," Education Sciences, vol. 13, no. 2, p. 164, Feb. 2023, doi: 10.3390/educsci13020164.

[14] R. Wu, S. Sunbury, P. Sadler, and G. Sonnert, "Influences of precollege out-of-school time computer science experiences on students' career interest in computer science," ACM Transactions on Computing Education, vol. 26, no. 1, p. 5:1-5:25, Dec. 2025, doi: 10.1145/3770069.

[15] E. K. Peterson, M. Van Noy, S. Scoville, and R. Edwards, "Not 'just a hobby': The influence of early interest and hobbies on community college IT student decision-making," Journal of Advanced Technological Education, vol. 1, no.3, pp.1-12, 2024, doi: 10.5281/ZENODO.10783784.

[16] Z. Shah, C. Chen, G. Sonnert, and P. M. Sadler, "The influences of computer gameplay and social media use on computer science identity and computer science career interests," Telematics and Informatics Reports, vol. 9, p. 100040, Mar. 2023, doi: 10.1016/j.teler.2022.100040.

[17] S. Bonanati and H. M. Buhl, "The digital home learning environment and its relation to children's ICT self-efficacy," Learning Environ Res, vol. 25, no. 2, pp. 485–505, Jul. 2022, doi: 10.1007/s10984-021-09377-8.

[18] M. Oda, Y. Noborimoto, and T. Horita, "International trends in K-12 computer science curricula through comparative analysis: Implications for the primary curricula," *Int. J. Comput. Sci. Educ. Sch.*, vol. 4, no. 4, pp. 1–25, Apr. 2021.

[19] German Informatics Society (GI), "2024/25 - informatik monitor." [Online]. Available from: https://informatik-monitor.de/2024-25, retrieved: 05, 2025

[20] L. Hellmig et al., "Bildungsstandards Informatik für die Sekundarstufe I - Empfehlungen der Gesellschaft für Informatik e.V. [Educational standards for computer science in secondary education – Recommendations from the German Informatics Society (GI)]," 2025, doi: 10.18420/REC2025_052.

[21] R. Schwarz, L. Hellmig, and S. Friedrich, "Informatikunterricht in Deutschland – eine Übersicht [Computer science education in Germany – an overview]," Informatik Spektrum, vol. 44, no. 2, pp. 95–103, Apr. 2021, doi: 10.1007/s00287-021-01349-9.

[22] M. Pijls, T. van Eijck, and B. Bredeweg, "Informal learning in a public library makerspace for youth in the Netherland," in Maker Education Meets Technology Education: Reflections on Good Practices, vol. 19, R. M. Klapwijk, J. Gu, Q. Yang, and M. J. De Vries, Eds., in International Technology Education Studies, vol. 19. , BRILL, 2023, pp. 51–63. doi: 10.1163/9789004681910.

[23] Bundesweite Informatikwettbewerbe (BWINF), "Bundesweite Informatikwettbewerbe." [Online]. Available: https://bwinf.de/, retrieved: 01, 2026.

[24] E. Delen and O. Bulut, "The Relationship between Students' Exposure to Technology and Their Achievement in Science and Math," Turkish Online Journal of Educational Technology - TOJET, vol. 10, no. 3, pp. 311–317, Jul. 2011.

[25] V. Aleksić and D. Politis, "The impact of interactive digital technology exposure on generation Z students learning performance in computer graphics and simulations: A comparative study of Greece and Serbia," Inovacije u nastavi, vol. 36, no. 2, pp. 1–16, 2023, doi: 10.5937/inovacije2302001A.

[26] A. F. Kharuddin, Z. Mustafa, and N. Azid, "Technology exposure, attitude and mathematics achievement: A structural equation model for comparing control and expose group of students," International Journal of Management, Accounting, Governance and Education, vol. 1, no. 1, pp. 3–13, Jun. 2021.

[27] X. Hu, Y. Gong, C. Lai, and F. K. S. Leung, "The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis," *Comput. Educ.*, vol. 125, pp. 1–13, Oct. 2018, doi: 10.1016/j.compedu.2018.05.021.

[28] S. Li, X. Liu, J. Tripp, and Y. Yang, "From ICT availability to student science achievement: mediation effects of ICT psychological need satisfactions and interest across genders," Res. Sci. Technol. Educ., vol. 40, no. 4, pp. 529–548, Oct. 2022, doi: 10.1080/02635143.2020.1830269.

[29] OECD (Organisation for Economic Co-operation and Development), "PISA 2025 Learning in the Digital World Framework (Second Draft)," 2023. [Online], Available from: https://www.oecd.org/en/topics/sub-issues/learning-in-the-digital-world/pisa-2025-learning-in-the-digital-world.html, retrieved: 05, 2025

[30] X. Chen, "The role of the exposure of education technology on learners' internet self-efficacy," Trans. Soc. Sci. Educ. Humanit. Res., vol. 5, pp. 606–617, Apr. 2024, doi: 10.62051/ccwnzw48.

[31] S. R. Yoel and Y. J. Dori, "FIRST high-school students and FIRST Graduates: STEM exposure and career choices," *IEEE Trans. Educ.*, vol. 65, no. 2, pp. 167–176, May 2022, doi: 10.1109/TE.2021.3104268.

[32] A. Steffen et al., "An exploration of critical incidents impacting female students' attitude towards stem subjects," in 6th International Conference on Gender Research (ICGR 2023), Jan. 2023, pp. 215–223. doi: 10.34190/icgr.6.1.994.

[33] J. S. Eccles and A. Wigfield, "In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs," Pers. Soc. Psychol. Bull., vol. 21,

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

no. 3, pp. 215–225, Mar. 1995, doi: 10.1177/0146167295213003.

[34] M. Esch and J. Grosche, "Fiktionale Fernsehprogramme im Berufsfindungsprozess – Ausgewählte Ergebnisse einer bundesweiten Befragung von Jugendlichen [Fictional Television Programs in the Career Choice Process – Selected Results of a Nationwide Survey of Adolescents]," in *MINT und Chancengleichheit in fiktionalen Fernsehformaten,* pp. 16–31, 2011. [Online] . Available from: http://www.mintiff.de/content/0/58/59/137_Broschuere_MINTIFF-Konferenz-2010_deutsch.pdf, retrieved: 05, 2025

[35] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. 1988.

[36] S. Feierabend, T. Rathgeb, Y. Gerigk, and S. Glöckler, "JIM-Studie 2024 - Jugend Information Medien [JIM Study 2024 – Youth, Information, Media]," Medienpädagogischer Forschungsverbund Südwest (mpfs), 2024. [Online]. Available from: https://mpfs.de/app/uploads/2024/11/JIM_2024_PDF_barrierearm.pdf, retrieved: 05, 2025

[37] Bitkom e.V., "Die Zukunft der Consumer Technology – 2024," 2024. [Online]. Available from: https://www.bitkom.org/sites/main/files/2024-09/bitkom-studie-zukunft-der-consumer-technology-2024.pdf, retrieved: 05, 2025

[38] Y. Skipper and P. J. Leman, "The role of feedback in young people's academic choices," International Journal of Science Education, vol. 39, no. 4, pp. 453–467, Mar. 2017, doi: 10.1080/09500693.2017.1294783.

[39] M. Paechter, S. Luttenberger, and B. Ertl, "Distributing feedback wisely to empower girls in STEM," Front. Educ., vol. 5, Jan. 2020, doi: 10.3389/feduc.2020.00141.

[40] S. Kunz and A. Steffen, "Criticize my code, not me: Using AI-generated feedback in computer science teaching," International Conference on Gender Research (ICGR), vol. 7, no. 1, pp. 200–207, Apr. 2024, doi: 10.34190/icgr.7.1.2002.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Assessing the Impact of Artificial Intelligence on Job and Task Displacement: A Cross Continental Analysis of Four Economic Sectors

Paul Mupfiga, Dzinaishe Mpini, Maxmillan Giyane
Department of Computer Science
Midlands State University
Gweru, Zimbabwe
e-mail: {mupfigaps, mpinid, giyanem}@staff.msu.ac.zw

***Abstract*—The rapid advancement of artificial intelligence has ignited extensive debates regarding its impact on the workforce. A widely held assertion is that artificial intelligence replaces specific tasks within jobs rather than entire occupations. This paper critically examines this claim by exploring empirical evidence from case studies of artificial intelligence implementations in the agriculture and healthcare sectors. A stratified purposive sampling method was employed to identify 37 use cases from five continents capturing diverse perspectives from all the different economy classifications. Of these 37 cases only 5 of the cases reviewed involved artificial intelligence tools that can fully replace human jobs, and the majority of them are from high income economies. A significant portion of the analysed cases were not displacing existing jobs but addressing previously unmet needs. In instances were job displacement occurs, there is the opportunity for up-skilling suggesting new career opportunities for the humans. This paper contributes by giving insights from empirical evidence to employment policy makers, employers and educators necessary in preparing the workforce for the future.**

Keywords-artificial intelligence; human-AI collaboration; labor market

I. Introduction

In our earlier work [1], we examined the impact of Artificial Intelligence (AI) in the modern workplace, focusing on the debate surrounding whether AI replaces specific tasks within jobs or entire jobs altogether. The study showed that AI is changing the workplace by replacing tasks rather than whole jobs. However, this effect is not the same across all sectors. Differences in economic conditions and levels of technology adoption shape how AI affects work. The findings only provided evidence of task-level automation in agriculture and healthcare. At the same time, they revealed the need for deeper analysis across sectors of varying manual, or cognitive effort. Furthermore, research is required to better understand how AI branches relates to task and job displacement.

Building on this foundation, the present study extends the original analysis by offering a more comprehensive examination encompassing 4 sectors. It adds empirical evidence with refined analytical frameworks to better capture variations in AI adoption and its socio-economic consequences. The findings of this extended investigation are intended to inform policy-makers, educators, and human resource practitioners by providing clearer evidence on the conditions under which AI augments, transforms, or displaces human labor. To achieve this objective, the study addresses the following research questions:

- To what extent does AI replace tasks within existing jobs compared to entire job roles?
- How do different sectors and skills experience AI-driven task and job displacement?
- How do different classes of AI systems relate to patterns of task and job displacement?
- What are the socio-economic implications of AI-driven task and job displacement for the workforce?

The significance of this research lies in its potential to provide practical and policy-relevant insights into the evolving relationship between AI and the workforce. The study proposes an empirical classification model that distinguishes between task displacement, job displacement, and no displacement outcomes. The research further examines how economic and sectoral contexts mediate the impact of AI-driven automation across different industries and income levels. By doing so, the study aims to develop an evidence-informed framework to support education planning, workforce development, and policy formulation in the context of increasing AI adoption.

In Section II, we present the theoretical framework underpinning our study, exploring prior research related to investigations of the impact of AI on the workforce. Section III details the research methodology which encompasses the research approach, and strategy. It further details the case selection process, as well as the data collection and analysis procedures. Section IV is a presentation of the results of this research followed by a discussion on results with the analysed prior studies. Finally, Section V details the conclusions and proposed future work.

II. Literature Review

In recent years, various studies have been conducted in a bid to analyze perceptions and impact of AI in the workplace. This section reviews such studies guided by a theoretical framework. The identified gaps are put under each theme.

A. Theoretical Framework

This research is guided by two key theories, the Task Based Approach (TBA) and the Job Polarization Theory (JPT). TBA highlights that automation impacts tasks within jobs rather than entire occupations [2]. Guided by TBA, this study evaluates AI's impact at the task level across selected industries, rather than assuming wholesale occupational displacement.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

JPT on the other hand informs that technological advancements including AI, and automation leads to the decline of middle-skill jobs while increasing demand for high-skill and low-skill jobs, thus creating a polarized labor market with a shrinking middle class and fewer middle-class occupations [3]. In this research, JPT provides a structural lens for interpreting sectoral differences in AI adoption and its implications for skill demand distribution. Together, these frameworks inform the study's analytical design, interpretation of findings, and assessment of whether AI primarily replaces tasks or contributes to broader occupational restructuring.

B. Analysis of Related Studies

1) Job Displacement versus Task Displacement: The distinction between job displacement and task displacement is best understood through the TBA theory. It conceptualizes production as a bundle of tasks rather than indivisible jobs. Task displacement occurs when AI automates specific activities within a job [4], while job displacement occurs when sufficient task automation renders an entire occupation economically redundant [5]. From a TBA perspective, technological change alters the composition of tasks within occupations before it eliminates jobs altogether.

Early insights by [6] suggested that AI could perform many human tasks more efficiently. Extending this reasoning, [7] argued that automation displaces certain tasks but simultaneously generates new ones, potentially increasing productivity and labor demand. This effect aligns with TBA's core assumption that technological change restructures task allocation between humans and machines rather than uniformly destroying employment. Similarly, [8] found that highly skilled professionals in fields such as computing, law, and medicine experience augmentation rather than replacement. [8] further indicates non-routine cognitive tasks remain comparatively resilient to displacement.

Empirical studies further reinforce the task-centered nature of AI's impact. [9] estimated that 22% of human tasks could be immediately automated, though without directly linking this to occupational elimination. [10] demonstrated that routine jobs face higher automation risks, while non-routine occupations have expanded, reflecting patterns consistent with JPT. JPT predicts that technology disproportionately displaces middle-skill routine occupations, leading to employment growth at both high- and low-skill ends of the labor market.

Recent contributions deepen this theoretical linkage. [11] synthesized evidence from 2010–2025 and concluded that AI reshapes labor productivity through both automation and augmentation pathways. Their findings show that routine manufacturing and mid-skill cognitive roles are particularly vulnerable, providing empirical support for JPT's "hollowing out" hypothesis of middle-skill occupations. Similarly, [12] identified clerical, administrative, financial, and customer service roles as highly exposed to generative AI.

However, emerging evidence also complicates traditional theoretical expectations. [13] reported significant exposure among occupations in the 80th earnings percentile, challenging the conventional JPT narrative that high-skilled labor is largely insulated from displacement. Furthermore, the observed decline in entry-level employment within high-exposure occupations suggests that AI may be automating foundational tasks that historically enabled skill accumulation and career progression. This indicates that AI-driven task displacement may have dynamic, long-term effects on occupational structures beyond immediate job elimination.

Focusing specifically on generative AI, [14] found displacement effects in certain occupations, particularly within digital freelance markets, yet noted no statistically significant average effects on aggregate employment or wages. This divergence between micro-level task disruption and macro-level employment stability further supports the TBA view that automation initially manifests as task recomposition before translating into measurable job displacement.

Although these studies collectively affirm that AI primarily operates at the task level, several theoretical and empirical gaps remain. First, there is insufficient granular analysis identifying which task clusters within occupations are most susceptible across different industries. Secondly, there is limited longitudinal evidence tracing how cumulative task displacement evolves into structural job displacement remains scarce. Addressing these gaps is essential to empirically determine whether AI reinforces classical job polarisation dynamics or is reshaping the labour market in unseen ways.

2) Sector Specific Impact: The impact of AI varies significantly across sectors. Rather than producing uniform job displacement, AI reshapes sector-specific task bundles, with varying implications for employment structures. [15] observed that regions specialising in technology, healthcare, and creative industries often experience task transformation rather than outright job destruction. Similarly, [16] argued that in healthcare, AI enhances diagnostic accuracy and decision-making, while the interpersonal and ethical complexity of medical work limits full occupational automation.

Recent work provides stronger empirical grounding for these patterns. [17] identified routine-based sectors such as manufacturing and customer service as facing higher displacement risks, whereas knowledge-intensive professions including healthcare, education, and creative industries tend to experience AI as a complementary tool. These findings align with TBA by demonstrating that sectors dominated by routine cognitive or manual tasks are more susceptible to automation at the task level. They also support the JPT, which predicts the erosion of middle-skill routine occupations.

In manufacturing, AI and robotics have produced measurable employment effects. Automated systems now perform assembly, inspection, and packaging with greater precision and efficiency than human workers. Empirically, [18] found that each additional robot per thousand workers reduces the employment-to-population ratio by 0.2 percentage points and wages by 0.42%. This sector exemplifies JPT's prediction of middle-skill displacement, as routine production roles decline while demand increases for high-skill technical maintenance and low-skill service work surrounding automated systems. Supporting this view,

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[19] reported that routine-based occupations in manufacturing face elevated displacement risks, whereas knowledge-intensive roles increasingly adopt AI as a complementary productivity tool.

Healthcare presents a contrasting trajectory. [19] highlighted that AI primarily automates administrative and diagnostic support tasks, while demographic pressures and expanding healthcare demand contribute to net employment growth. This reflects task displacement without proportional job elimination, consistent with TBA. Furthermore, the requirement for empathy, ethical judgment, and contextual reasoning limits full automation, indicating sector-specific resistance to wholesale job displacement.

In finance, AI-driven algorithms increasingly perform data analysis, high-frequency trading, risk modelling, and customer interaction through chatbots. [20] earlier suggested that such automation could reduce demand for routine financial service roles. Sectoral evidence from retail and human resource management further illustrates differentiated impacts. [21] showed that AI accelerates recruitment, personalises training through adaptive learning systems, and enables real-time performance monitoring in the retail sector. While such applications improve efficiency, they also introduce ethical concerns such as algorithmic bias and transparency challenges. These findings demonstrate that in service-oriented sectors, AI often restructures internal task processes rather than directly eliminating occupations.

The public sector similarly exhibits a hybrid pattern of transformation. [22] identified efficiency gains in administrative functions and service delivery, alongside challenges related to trust, ethics, and workforce adaptation. AI automates repetitive administrative tasks while requiring public servants to cultivate higher-order interpersonal and intuitive skills. This pattern again supports the TBA proposition that automation first restructures task composition within occupations.

Collectively, sector-specific evidence suggests that AI adoption does not produce uniform labor market outcomes. Instead, its effects depend on the routine intensity, cognitive complexity, and regulatory environment of each sector. From a JPT perspective, manufacturing and routine service sectors display stronger middle-skill displacement dynamics, whereas healthcare and knowledge-intensive industries show patterns of augmentation and task transformation.

Despite growing sectoral evidence, comprehensive comparative analyses across multiple industries remain limited. Existing studies often focus on single sectors without systematically evaluating task-level exposure across industries. Guided by TBA and JPT, there is a need for empirical research that compares how AI reshapes task structures within occupations across sectors and assesses whether observed task displacement accumulates into broader occupational polarisation. Such comparative analysis is essential to determine whether AI reinforces traditional job polarisation patterns or produces new forms of sectoral restructuring.

3) Socio Economic Impact: [23] provided qualitative insights into employee perceptions on job loss from AI and automation in New Zealand. Their findings revealed sentiments from workers across different professions ranging from perceptions of no threat to potential and real threats to job security. Similarly, [24] highlighted growing concerns regarding the potential for substantial job losses due to AI adoption, despite acknowledging the productivity gains and emergence of new forms of employment that AI technologies may generate.

From the perspective of worker-AI coexistence, [25] conducted a literature survey examining how employees interact with AI systems in the workplace. Their study emphasized the importance of trust in AI systems and the necessity for continuous skill development to enable effective human-AI collaboration.

The economic implications of AI-driven transformation have also been widely explored. For instance, [26] analyzed the economic effects of generative AI tools such as ChatGPT, predicting significant displacement of routine tasks and occupations while also highlighting the potential for increased productivity and the creation of new high-skill job categories. Likewise, [27] estimated that AI technologies could affect a substantial proportion of jobs globally, potentially impacting up to 800 million workers worldwide.

More recent studies provide deeper insights into the mechanisms through which AI reshapes labor markets. For example, [28] investigated the adoption of artificial intelligence in the accounting sector and found that AI-driven automation alters traditional accounting roles by automating repetitive tasks while simultaneously reshaping decision-making processes. The study suggests that while certain clerical and routine accounting tasks may be displaced, the demand for higher-level analytical and strategic roles is likely to increase. These findings align with the TBA, which posits that technologies replace specific tasks rather than entire occupations, leading to a restructuring of job roles rather than complete job elimination.

Similarly, [29] introduced a novel measure termed unemployment risk to quantify technology-driven job loss by analyzing occupation-level unemployment data. Their findings demonstrated that occupations with higher exposure to AI technologies experience greater risks of job separation, reinforcing the importance of understanding how AI affects task composition within occupations rather than employment levels alone.

Supporting this perspective, [30] analyzed global employment trends between 2025 and 2030 and reported that approximately 85 million jobs could be displaced due to AI-driven automation, while around 97 million new roles may emerge, resulting in a net gain of 12 million jobs globally. However, the study also highlighted notable socio-economic disparities, particularly gender disparities in exposure to AI-driven job displacement, with a significantly higher proportion of women occupying occupations susceptible to automation.

While existing studies such as [24] and [25] discuss the broad socio-economic implications of AI adoption, they often overlook how different demographic groups, including age, gender, and education level, are differentially affected by AI-driven transformation. Furthermore, although global projections such as those presented by [27] and [30] provide valuable insights

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

into the overall scale of potential disruption, there remains limited empirical research examining regional disparities in AI-induced labor market changes. In particular, there is a need to investigate how AI adoption affects employment structures in developing economies and rural regions, where labor markets and technological adoption patterns may differ significantly from those in advanced economies.

III. Methodology

This section outlines the research design and procedures employed to investigate the impact of AI on job and task displacement. The methodology is structured to ensure a rigorous and systematic analysis of empirical evidence across diverse sectors and economic contexts. It is grounded in a pragmatic research philosophy, which emphasizes practical problem solving. This approach is particularly suitable for exploring the impacts of AI on job and task displacement across various sectors and geographical locations. A qualitative research approach was chosen to provide in-depth insights into the impacts of AI adoption, focusing on context-specific factors rather than broad generalizations. The content analysis research strategy was used, and this strategy enabled systematic examination and interpretation of textual data on AI use cases across the globe.

A. Case Selection

A purposive and stratified sampling technique was adopted to ensure comprehensive coverage of use cases. Purposive sampling selected four sectors with diverse task structures and varying levels of AI exposure drawing from observations by [31]. The analysed sectors include agriculture, healthcare, finance, and manufacturing. Agriculture and manufacturing represent sectors dominated by routine manual tasks whilst healthcare and finance is characterised by routine cognitive activities. Stratification was conducted by continent and economy type to achieve geographical and economic diversity. Continent classification adhered to [32], encompassing Africa, America, Asia, Europe, and Oceania. Economy classifications followed [33], comprising low-income (LI), lower-middle-income (LIMI), upper-middle-income (UMI) and high-income (HI) economies.

B. Data Collection

Data collection was conducted through desktop research, drawing from academic journals, industry reports, reputable company websites, and blogs. Desktop research was selected due to its broad accessibility to diverse and globally distributed data sources, enabling the study to capture a comprehensive range of implemented AI use cases across sectors and regions. By relying on published academic literature, industry reports, and reputable company websites, we ensured the inclusion of validated and credible data. Peer reviewed articles provide insights on impacts which are verified. Industry reports provide insights into applied observations. Real-world deployments from company websites and reputable blogs provide first hand information of AI deployments. Only implemented AI solutions were considered, excluding conceptual studies, as well as research and development efforts. Each continent was represented for each of the four sectors, ensuring all four economy classifications were covered without repetition of countries.

To ensure a systematic search; under a particular sector, continent and economy classification, the researchers identified the countries which fall in that continent and economy classification and made a search from the Google search engine to identify AI use cases from countries that satisfy the continent and economy classification. The chosen approach ensured representativeness of AI implementations across diverse geographies and economic classifications, minimizing potential bias. Each continent and economy classification was equally represented. The used keywords included "AI powered agent for *category name* use cases in *country name*", as well as *category name* AI agent AND *country*. For example *Healthcare AI agent AND Angola*. Sticking to a systematic search protocol and following these standards ensured a methodological control allowing for replication of results.

The data assessment in this collection prioritized identifying real-world implementations of AI rather than conceptual or research and development (R&D) efforts. The assessment emphasized role classification, guided by the framework outlined in [34], to evaluate the nature of tasks and jobs influenced by AI technologies. Key areas of analysis included the extent to which AI replaced specific tasks within roles versus entire job roles, as well as the socio-economic implications of these changes. The data was analyzed thematically in alignment with the research questions, enabling a structured exploration of task displacement, sectoral impacts, and workforce impacts.

C. Data Analysis

Several critical factors were considered during the analysis. These included the availability of sufficient data to enable sound deductions, the identification of AI agents or classifications, and the specific roles covered by these implementations. Role classifications were examined to determine whether AI-driven technologies led to task or job displacement. Additionally, the study categorized data by economy classification, continent, and sector, ensuring representation from healthcare and agriculture across diverse geographical and economic contexts. Socio-economic impacts, such as the effect on workforce skills, were also assessed, providing a comprehensive understanding of AI's influence. This approach ensured the analysis addressed the complexities of AI adoption within varying global and sectoral environments.

In reporting, findings were presented thematically, guided by the research questions. Themes included job displacement versus task displacement by skill, sector-specific impacts, and socio-economic implications. Analysis considered economy, continent, type of AI, and industry classifications to provide a comprehensive understanding of AI's impact on the workforce.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

IV. Results and Discussion

This study investigated the impact of AI on job and task displacement across four key sectors: agriculture, finance, manufacturing, and healthcare. An attempt was made to obtain at least one use case for each sector across all continents and World Bank economy classifications. However, some regions do not contain countries within certain income classifications, which limited the number of available cases. Additionally, several potential cases were excluded because sufficient information was not available online to clearly identify the AI agent used or the country of implementation. Consequently, the final dataset consisted of 37 AI implementation use cases, summarized in Table I.

The analyzed use cases span all five continents, with Africa (24%, n=11) and Asia (24%, n=9) having the highest representation, followed by America (19%, n=7) and Europe (16%, n=6), with Oceania having comparatively fewer cases (11%, n=4). The uses cases distribution by continent is presented in Figure 1.

From an economic perspective, the cases were distributed across all World Bank income classifications, including low-income economies (8%, n=3), lower-middle-income economies (14%, n=5), upper-middle-income economies (35%, n=13), and high-income economies (43%, n=16). The distribution of use cases by economy type is presented in Figure 2.

Sectoral analysis shows that healthcare accounted for the largest share of use cases (43%, n=16), followed by finance (30%, n=11), manufacturing (14%, n=5), and agriculture (14%, n=5). The distribution of use cases by sector is presented in Figure 3.

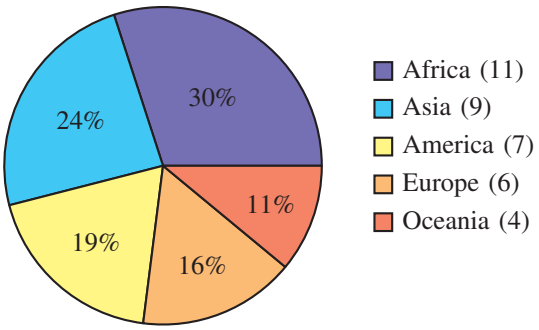


Figure 1. Distribution of AI use cases by continent

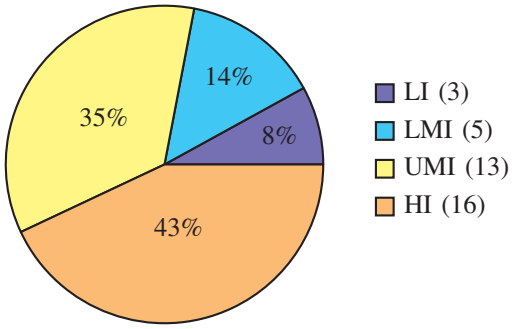


Figure 2. Distribution of AI use cases by economy classification

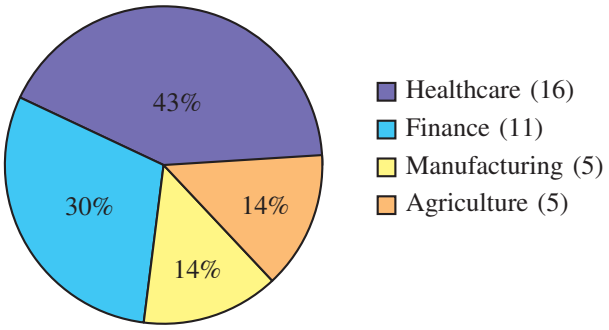


Figure 3. Distribution of AI use cases by sector

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I. SUMMARY OF USE CASES

#	AI Agent	Skill Level	Economy Classification	AI Type	Displacement Type
1	**Radify AI - South Africa**. AI system that diagnose medical images [35][36][37].	4	UMI	Computer Vision	ND
2	**Robin the Robot - Armenia**. Companionship robot providing emotional support to pediatric patients [38][39][40].	2	UMI	Robotics & NLP	ND
3	**FoxTac - Ukraine**. Robotic stretchers for safe medical evacuations in conflict zones [41][42].	2	UMI	Robotics	ND
4	**Montreal's Scale AI - Canada**. AI platforms which optimize hospital operations across Canada [43][44].	4	HI	Expert Systems	ND
5	**Mazor X Stealth Platform - Australia**. Robotic surgery platform that assists spine surgeons [45][46][47].	4	HI	Robotics	ND
6	**MkhulimaGPT - Rwanda**. Smart chatbot which gives farming advice [48][49].	2	LI	NLP	ND
7	**uMudhumeni - Zimbabwe**. Smart chatbot which poses as an agricultural extension worker [50][51].	2	LMI	NLP	ND
8	**Plantix - Bangladesh**. Mobile application for crop disease diagnosis using photos [52][53].	2	LMI	Computer Vision	ND
9	**Anton Tech - Botswana**. AI platform that uses drones to monitor both plant diseases and soil quality [54][55].	2	UMI	Computer Vision	ND
10	**Jeevn AI - Argentina**. Expert system which provides personalized farming advice [56].	2	UMI	Expert Systems	ND
11	**Automated Xray Imaging Device - Sudan**. Screens for tuberculosis (TB) by interpreting X-ray images [57].	4	LI	Robotics	TD
12	**Sophie Bot - Kenya**. AI powered chatbot offering answers to sexual and reproductive health questions [58][59][60].	2	LMI	NLP	TD
13	**Telehealth Learning Robot - Cambodia**. Enhances health education and tele-consultations in low-resource settings [61].	3	LMI	Robotics	TD
14	**LaLuchy Robotina - Mexico**. Alleviate loneliness in patients by enabling virtual communication [62][63][64].	2	UMI	Robotics & NLP	TD
15	**Panafricare Clinic AI Agents - Sychelles**. AI systems to assist in clinical examinations [65].	2	HI	Expert Systems	TD
16	**Medicine Delivery Robot - Singapore**. Voice-activated robot to deliver medications in hospitals [66].	2	HI	Robotics & NLP	TD
17	**NOVISSI AI Targeting System - Togo**. Uses ML to identify poor households to assist in emergencies [67].	4	LI	Machine Learning	TD
18	**Haraka AI Loan App - Zimbabwe**. Uses AI to analyze mobile phone and social media data to assess creditworthiness [68].	4	LMI	Machine Learning	TD
19	**Absa Abby - South Africa**. Virtual assistant that acting as a customer support agent in a bank [69].	3	UMI	NLP	TD
20	**Intelligent Agents for Financial Services - Brazil**. Uses LLMs to automate decision-making in complex digital workflows [70].	4	UMI	NLP	TD
21	**DANA Real-Time Gambling Detection System - Indonesia**. ML system that detects suspicious gambling activity in real time [71].	4	UMI	Machine Learning	TD
22	**Conversational Chatbot for Insurance Operations - Ukraine**. Provides automated insurance support such as claims submission and policy information [72].	3	UMI	NLP	TD
23	**AI-Driven Contact Centre for Tower Insurance - New Zealand**. Provides call transcription, sentiment analysis, and call summarization for the centre [73].	3	HI	NLP	TD
24	**Anaplan Intelligent Role-Based Agents - United States**. Analyses financial data and generate insights for decision-making [74].	4	HI	Generative AI	TD
25	**BlackRock Investment Analysis - United States**. ML models to analyze financial datasets and market signals to support portfolio management [75].	4	HI	Machine Learning	TD
26	**Optasia AI Financial Access Platform - UAE**. AI-driven analytics platform assessing creditworthiness [76].	4	HI	Machine Learning	TD
27	**RPA for Import/Export Operations - UK**. Robotic process automation system to automate contract processing in trade [77].	3	HI	Expert Systems	TD
28	**Yaskawa Robotic Arc Welding System - South Africa**. Automate arc welding by positioning the welding torch and adapting to joint variations [78].	2	UMI	Robotics	TD
29	**Tesla Giga Press - USA**. Produces large single-piece car body components in vehicle manufacturing [79].	2	HI	Robotics	TD
30	**RockyOne Mobile Manipulation Robot - Japan**. Automatically unloads cartons from shipping containers in warehouses [80].	2	HI	Robotics	TD
31	**UR16e Collaborative Robotic Arm - Denmark**. Automates industrial tasks such as packaging, palletizing, and machine tending [81].	2	HI	Robotics	TD
32	**Tesla Grok In-Car AI Assistant - Australia**. Enables drivers to interact with vehicle systems using natural language voice commands [82].	2	HI	NLP	ND
33	**Robotti - Sweden**. Self-driving robot used for various manual farm tasks [83].	3	UMI	Robotics	JD
34	**John Deere CP770 - USA**. Self-driving robot used for various manual tasks in agriculture [84].	2	HI	Robotics	JD
35	**Agrist Robot - Japan**. AI-powered robot that autonomously harvests crops such as cucumbers and tomatoes [85][86].	2	HI	Robotics	JD
36	**YV01 - France**. Autonomous vineyard spraying robot addressing labor shortages and environmental compliance [87].	2	HI	Robotics	JD
37	**Oxin Tractor - New Zealand**. AI-driven autonomous tractor used for plowing, seeding, and weeding tasks [88].	2	HI	Robotics	JD

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A. RQ1: To what extent does AI replace tasks within existing jobs compared to entire job roles?

The findings from the analyzed use cases indicate that AI predominantly replaces specific tasks within occupations rather than entire job roles, supporting the theoretical assumptions of the TBA. Among the 37 analyzed cases, the majority demonstrate task displacement, where AI automates particular functions within a job while leaving the broader occupational role intact. Only a smaller subset of cases shows job displacement JD, and several cases illustrate no displacement ND where AI primarily creates a role which was not being served by humans. This distribution aligns with TBA's central premise that production consists of bundles of tasks rather than indivisible jobs, and technological change first reorganizes these tasks before potentially eliminating occupations entirely [4][5].

Many of the examined use cases demonstrate this task-level automation pattern. In the healthcare sector, for example, systems such as Radify AI [35], which diagnoses medical images, and Plantix [52], which identifies crop diseases using computer vision, automate diagnostic or analytical subtasks traditionally performed by professionals. Similarly, AI-driven contact center systems, banking chatbots such as Absa Abby [69], and fraud detection platforms like the DANA system automate customer service interactions [71], transaction monitoring, or information retrieval tasks. In these cases, the underlying occupations such as healthcare professionals, customer service representatives, or financial analysts are not eliminated. Instead, AI systems assume repetitive or data-intensive subtasks, allowing workers to focus on higher-value cognitive or interpersonal activities. These observations support earlier arguments that AI technologies frequently augment rather than replace skilled professionals, particularly in non-routine cognitive occupations [8].

Empirical evidence from the use cases also reflects the broader coexistence of AI agents and humans described earlier in the literature. According to [7], automation can displace specific tasks while simultaneously generating new complementary activities that increase productivity and labor demand. This dynamic is visible in several implementations, particularly in healthcare and financial services, where AI systems support decision-making, diagnostics, or customer engagement rather than fully replacing professionals. The results therefore reinforce the claim that AI-driven technological change often leads to task recomposition within occupations rather than immediate occupational elimination.

However, the results also reveal instances of full job displacement, primarily within routine physical occupations. Autonomous agricultural machinery such as Robotti [83], John Deere CP770 [84], Agrist Robot [85], YV01 [87], and the Oxin Tractor [88] replaces manual agricultural labor traditionally performed by farm assistants or machine operators. These cases illustrate situations where automation can replace a sufficiently large proportion of an occupation's task bundle, making the job economically redundant. Such findings are consistent with empirical research indicating that routine and repetitive tasks are the most susceptible to automation [10]. They also support predictions from JPT, which suggests that routine middle-skill occupations are more vulnerable to technological displacement.

The distribution of displacement types across sectors further reinforces these theoretical insights. Healthcare and financial services cases largely demonstrate augmentation or task displacement, reflecting the persistence of complex decision-making, ethical judgment, and interpersonal interaction in these professions. In contrast, manufacturing and agriculture show stronger evidence of job-level automation, particularly where tasks are standardized and physically repetitive. These sectoral differences illustrate how the nature of tasks within occupations determines susceptibility to automation, a central argument of the TBA framework.

The findings also provide some insight into emerging debates regarding high-skill exposure to AI, particularly in the context of generative AI. While most high-skill roles in the dataset experienced augmentation rather than displacement, several systems such as financial planning agents and automated analytics platforms demonstrate AI's growing ability to perform complex analytical tasks traditionally reserved for highly trained professionals. This observation aligns with recent studies suggesting that generative AI may expose certain high-skilled occupations to automation risks [13], although the evidence from the present cases suggests that these technologies are currently reconfiguring professional tasks rather than eliminating professional roles.

Overall, the results strongly support the TBA perspective that AI primarily operates at the task level, altering the composition of work within occupations before leading to broader structural labor market changes. The predominance of task displacement across the analyzed cases indicates that AI adoption is currently characterized more by work reconfiguration and augmentation than by widespread job elimination. Nevertheless, the observed cases of job displacement in routine sectors suggest that as automation technologies mature and accumulate across multiple tasks, certain occupations may eventually become fully automated.

B. RQ2: How do different sectors and skills experience AI-driven task and job displacement?

The analyzed use cases demonstrate that the impact of AI varies considerably across sectors, supporting the central premise of the TBA that automation reshapes the composition of tasks within occupations rather than uniformly eliminating jobs. Consistent with previous studies, the results indicate that sectoral characteristics particularly the routine intensity of tasks and the level of cognitive complexity involved play a significant role in determining whether AI results in task displacement, job displacement, or augmentation. This aligns with earlier observations that technology-intensive sectors such as healthcare and knowledge-based services often experience task transformation rather than wholesale job destruction [15]. Similarly, research suggests that AI applications in healthcare frequently enhance diagnostic and decision-making capabilities

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

while leaving the interpersonal and ethical dimensions of medical work largely resistant to automation [16].

C. Healthcare

Healthcare accounts for the largest proportion of analyzed cases and shows the strongest evidence of task-level automation rather than full job displacement. Systems such as Radify AI [36], which assists with medical image diagnosis, and automated X-ray screening devices [57], which analyze tuberculosis scans, automate diagnostic tasks traditionally performed by healthcare professionals. Likewise, conversational health assistants such as Sophie Bot [58] and robotic telehealth [61] platforms automate information delivery, patient monitoring, and administrative support tasks. However, these technologies largely complement rather than replace medical professionals, reflecting the continued need for human judgment, empathy, and contextual decision-making. This pattern supports prior findings that AI in healthcare primarily enhances productivity while leaving core clinical roles intact [19]. In this sector, AI adoption therefore manifests mainly as task displacement and augmentation, consistent with TBA's prediction that complex professional occupations are less susceptible to full automation.

D. Finance

The finance sector also largely exhibits task-level automation. AI-driven systems in banking and financial services perform functions such as credit assessment, fraud detection, customer service automation, and financial data analysis. For instance, Haraka AI Loan App [68] uses machine learning to assess creditworthiness using mobile data, while DANA's real-time detection system identifies suspicious financial transactions [71]. Similarly, conversational agents such as Absa Abby and automated insurance chatbots perform customer service interactions that were previously handled by human agents. These implementations reduce the need for routine clerical and customer-facing tasks but do not fully eliminate the associated occupations. Financial analysts, compliance officers, and customer service representatives continue to play roles in oversight, decision-making, and complex problem resolution. This pattern mirrors earlier predictions that AI would primarily reduce demand for routine financial service activities while transforming rather than eliminating professional roles [20]. The results therefore reinforce the view that in service-oriented sectors, AI typically restructures internal task processes rather than displacing entire occupations.

E. Manufacturing

In contrast, the manufacturing sector demonstrates stronger evidence of automation at the job level, particularly for routine manual roles. Several use cases illustrate the deployment of robotics systems capable of performing highly repetitive industrial tasks with minimal human intervention. Examples include the Yaskawa robotic arc welding system [78], which automates welding operations, the Tesla Giga Press [79], which replaces multiple manual assembly processes with a single automated casting system, and the UR16e collaborative robotic arm [81], which performs packaging and machine-tending tasks. These technologies directly substitute manual labor traditionally performed by assembly workers and machine operators. This pattern aligns with empirical findings that industrial robotics can reduce employment in routine production roles while increasing demand for high-skill technical roles responsible for system maintenance and management [18]. The sector therefore provides strong evidence supporting the JPT, which predicts the displacement of middle-skill routine occupations.

F. Agriculture

The agriculture sector similarly exhibits examples of full job displacement driven by autonomous machinery. Autonomous farming systems such as Robotti [83], John Deere CP770 [84], Agrist Robot [85], YV01 vineyard robots [87], and the Oxin autonomous tractor [88] which automate tasks including planting, harvesting, spraying, and field monitoring. These technologies replace a large proportion of manual agricultural labor tasks traditionally performed by farm assistants and machine operators. Unlike healthcare and finance applications, where AI tends to support professional decision-making, agricultural automation directly replaces physical labor tasks that are repetitive and easily standardized. These findings align with research indicating that sectors dominated by routine manual tasks face higher automation risks [10]. As a result, agriculture demonstrates clearer examples of complete job displacement compared to knowledge-intensive sectors.

G. RQ3: What are the socio-economic implications of AI-driven task and job displacement for the workforce?

The findings from the analyzed use cases suggest that the socio-economic implications of AI-driven task and job displacement are multifaceted, affecting workforce structures, skill requirements, and perceptions of job security. While AI adoption can enhance productivity and efficiency, it also introduces concerns regarding employment stability and the evolving nature of work. These outcomes align with earlier qualitative research indicating that workers exhibit mixed perceptions regarding the risks posed by AI technologies. For instance, [23] reported that employees across various professions perceive AI either as a complementary tool or as a potential threat to job security, depending largely on the nature of their roles and exposure to automation. Similarly, [24] highlighted growing anxieties about the possibility of large-scale job displacement, despite recognizing that AI adoption may also generate new forms of employment.

The analyzed cases provide practical examples of how these socio-economic dynamics unfold across sectors. In many instances, AI systems automate repetitive tasks rather than replacing entire occupations, leading to job transformation rather than immediate job elimination. For example, financial systems such as Absa Abby [69], Haraka AI Loan App [68], and AI-driven financial analytics platforms automate routine tasks including credit scoring, customer support, and financial data analysis. Similarly, healthcare technologies such as Radify AI [35] and automated diagnostic tools assist medical

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

professionals by performing data-intensive diagnostic functions. These cases illustrate how AI reshapes occupational roles by shifting workers away from repetitive activities toward tasks requiring judgment, oversight, and interpersonal interaction. Such patterns are consistent with the TBA, which suggests that technological change restructures the composition of tasks within jobs rather than eliminating occupations entirely.

At the same time, the emergence of fully automated systems in certain sectors highlights the potential for direct job displacement, particularly in occupations dominated by routine manual tasks. Autonomous agricultural machinery such as Robotti [83], Agrist Robot [85], and the Oxin Tractor [88], along with industrial robotics such as the Yaskawa arc welding system [78] and Tesla Giga Press [79], illustrate cases where automation replaces a large portion of human labor in specific roles. These examples reinforce arguments that routine manual occupations are particularly vulnerable to technological displacement. As suggested by [89], such patterns may contribute to labor market polarization, where middle-skill routine jobs decline while demand increases for both high-skill technical roles and low-skill service positions that remain difficult to automate.

Beyond direct employment effects, AI-driven transformation has broader implications for skills development and workforce adaptation. As AI systems assume routine tasks, workers increasingly need to develop competencies that complement automated technologies. Research by [25] emphasizes that successful human–AI collaboration requires workers to cultivate human-centric capabilities such as critical thinking, problem solving, communication, and creativity. The analyzed cases support this observation, as many AI systems function as decision-support tools that require human oversight and interpretation. Consequently, workforce resilience in an AI-driven economy depends heavily on the ability of workers to continuously update their skills and adapt to changing task requirements.

The economic implications of AI adoption further extend to productivity and labor market restructuring. Studies examining generative AI technologies suggest that these tools may significantly increase productivity while simultaneously transforming occupational structures. For instance, [26] predicted that AI could automate a substantial share of routine tasks while also creating new high-skill job categories related to AI development, deployment, and oversight. Similarly, global projections indicate that although AI may displace a large number of jobs, it may also generate new opportunities in emerging technological fields. Estimates presented by [30] suggest that while approximately 85 million jobs may be displaced globally, up to 97 million new roles could emerge, resulting in a net increase in employment.

However, the socio-economic consequences of AI adoption are unlikely to be evenly distributed. Several studies highlight the potential for inequality and demographic disparities in AI-driven displacement. According to [89], automation may disproportionately affect workers in routine occupations, many of which are concentrated among lower-income groups. Furthermore, [30] identified gender disparities in exposure to automation, noting that women are often overrepresented in occupations susceptible to AI-driven task automation. Such disparities suggest that the benefits and costs of AI adoption may be unevenly distributed across different demographic groups.

The present findings also highlight the importance of considering regional differences in AI adoption and labor market impacts. Many of the analyzed cases originate from high-income economies where technological infrastructure and investment in AI development are more advanced. In contrast, developing economies often adopt AI technologies in targeted sectors such as agriculture and financial inclusion, as illustrated by cases like MkhulimaGPT [49], Plantix [52], and the NOVISSI AI targeting system [67]. These applications often aim to improve productivity and service accessibility rather than replace large numbers of workers. Nevertheless, the long-term implications of AI adoption in developing economies remain uncertain, particularly in rural labor markets where employment opportunities may already be limited.

Overall, the socio-economic implications of AI-driven task and job displacement extend beyond immediate employment effects. AI technologies are reshaping workforce structures by transforming occupational tasks, altering skill requirements, and influencing the distribution of economic opportunities. While the analyzed cases suggest that AI adoption currently leads more frequently to task transformation and workforce augmentation than to widespread job elimination, the long-term effects may include structural changes in labor markets, shifts in skill demand, and potential increases in economic inequality. Addressing these challenges requires policies that support workforce reskilling, promote equitable access to technological opportunities, and ensure that the benefits of AI-driven productivity gains are broadly shared across society.

H. RQ4: How do different classes of AI systems relate to patterns of task and job displacement?

The analyzed use cases indicate that the relationship between AI technologies and displacement outcomes varies according to the class of AI system deployed. Different AI technologies automate different types of tasks, and therefore produce distinct patterns of task displacement, job displacement, or augmentation. These findings are consistent with the TBA, which emphasizes that the effects of automation depend on the nature of the tasks being automated rather than the occupation itself. In practice, this means that AI systems designed for analytical reasoning, perception, or physical automation influence the workforce in different ways depending on the task characteristics they target.

I. Robotics

Among the analyzed cases, robotics-based systems are most strongly associated with job displacement, particularly in sectors involving routine physical tasks. Autonomous machinery such as Robotti [83], Agrist Robot [85], John Deere CP770 [84], YV01 vineyard robots [87], and the Oxin autonomous tractor [88] automate agricultural activities including planting, harvesting, spraying, and field monitoring.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Similarly, robotics systems in manufacturing such as the Yaskawa robotic arc welding system [78], Tesla Giga Press [79], and the UR16e collaborative robotic arm [81] perform repetitive industrial operations that were traditionally carried out by human workers. These technologies replace a substantial proportion of the task bundles associated with manual labor occupations, making entire job roles economically redundant. This pattern supports previous research indicating that robotics and industrial automation primarily affect routine manual tasks, which are easier to standardize and automate.

J. Conversational Chatbots

In contrast, natural language processing (NLP) and conversational AI systems tend to produce task displacement rather than full job displacement. Systems such as Absa Abby [69], Sophie Bot [58], MkhulimaGPT [48], and the conversational insurance chatbot automate customer interaction, information retrieval, and advisory functions. These systems reduce the need for human agents to perform routine communication tasks but still require human oversight, escalation handling, and decision-making in complex situations. Consequently, customer service representatives, financial advisors, and extension workers continue to play active roles in service delivery. This pattern aligns with earlier research suggesting that AI technologies in service sectors typically restructure internal workflows rather than fully eliminating occupations.

K. Machine Learning Based Systems

Machine learning systems used for analytics and prediction demonstrate a similar pattern of task-level automation. Financial analytics platforms such as BlackRock's investment analysis tools, Haraka AI Loan App [68], and the Optasia financial access platform automate data analysis [76], credit scoring, and fraud detection. These technologies replace time-consuming analytical tasks previously performed by financial professionals but do not eliminate the broader occupational roles that involve strategic decision-making and regulatory oversight. This observation reinforces the argument that data-driven AI systems augment professional work by improving efficiency and decision accuracy, rather than fully replacing expert labor.

L. Computer Vision

Another important category observed in the dataset is computer vision systems, particularly in healthcare and agriculture. Applications such as Radify AI [36], Plantix [52] , and automated X-ray diagnostic devices perform image-based analysis tasks that previously required specialized human expertise. These systems demonstrate high accuracy in pattern recognition tasks, enabling them to automate diagnostic subtasks within professional workflows. However, the broader occupations of medical professionals and agricultural specialists remain necessary for contextual interpretation, treatment decisions, and field-level interventions. Consequently, computer vision technologies primarily contribute to task augmentation rather than occupational elimination.

M. Expert Systems

The findings also highlight the role of expert systems and generative AI technologies, which automate knowledge-intensive tasks such as financial planning, decision support, and operational optimization. Systems such as Anaplan's intelligent agents [74] and Scale AI healthcare platforms [44] assist professionals by generating recommendations and insights from large datasets. These systems typically function as decision-support tools, requiring human interpretation and oversight. As a result, they primarily lead to task displacement rather than direct job loss, reflecting the continued importance of human judgment in complex decision-making environments.

V. Conclusion and Future Work

This study examined the impact of artificial intelligence on task and job displacement across agriculture, finance, manufacturing, and healthcare sectors using 37 global use cases. RQ1 findings show that, consistent with the TBA, AI most commonly replaces specific tasks within existing jobs rather than entire occupations, with task displacement occurring predominantly in knowledge-intensive sectors such as healthcare and finance. Sectoral differences were also evident in answering RQ2, as manufacturing and agriculture exhibited stronger patterns of job-level automation due to routine physical tasks, whereas healthcare and financial services largely experienced task augmentation and workflow restructuring. The socio-economic implications assessed through RQ3 suggest that AI adoption is likely to reshape skill requirements and labor market structures, increasing demand for higher-order cognitive and interpersonal skills while raising concerns about job polarization and inequality among workers in routine occupations. Finally, the analysis of AI technology classes assessed through RQ4 indicates that robotics systems are more closely associated with job displacement, particularly in manual labor sectors, whereas machine learning, computer vision, NLP, and expert systems primarily lead to task-level automation and human–AI collaboration. Overall, the results highlight that AI-driven labor market transformation is primarily characterized by task recomposition and occupational restructuring rather than widespread job elimination, although certain routine occupations remain vulnerable to full automation as AI technologies continue to evolve.

Based on the findings of this study, several recommendations can be proposed to support policy-makers, educators, and human resource practitioners in responding effectively to AI-driven changes in the labor market. These recommendations are informed by the observed patterns of task displacement, sectoral variation, and the relationship between AI technologies and workforce transformation.

First, policy-makers should prioritize workforce reskilling and lifelong learning initiatives to support workers whose tasks are increasingly automated. Since the findings indicate that AI most commonly replaces specific tasks rather than entire jobs, policies should focus on enabling workers to transition into complementary roles that require human judgment, creativity, and interpersonal skills. Governments should therefore invest

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

in continuous training programs, digital literacy initiatives, and vocational education systems that prepare workers to collaborate with AI technologies rather than compete against them.

Second, education systems should adapt curricula to reflect the evolving skill demands of an AI-driven economy. The results demonstrate that routine tasks both cognitive and manual are more susceptible to automation, while non-routine analytical, creative, and interpersonal tasks remain more resilient. Educational institutions should therefore emphasize interdisciplinary learning that combines technical knowledge with human-centered competencies such as critical thinking, problem solving, ethical reasoning, and communication. In addition, integrating AI literacy into school and university curricula will help prepare future workers to effectively interact with AI systems across different industries.

Third, human resource practitioners should redesign job roles to leverage human–AI collaboration. Rather than viewing AI solely as a tool for labor substitution, organizations should adopt strategies that integrate AI systems into existing workflows in ways that augment employee capabilities. This may involve redefining job roles so that AI handles routine data processing tasks while employees focus on strategic decision-making, innovation, and customer engagement. HR departments should also implement structured reskilling programs to ensure that employees can transition into new roles created by AI-enabled transformation.

Fourth, sector-specific workforce strategies should be developed, particularly for sectors where job displacement risks are higher. The findings indicate that sectors such as manufacturing and agriculture, where tasks are highly routine and standardized, are more likely to experience full job displacement through robotics and automation. In these sectors, targeted labor transition programs such as retraining for technical maintenance roles, digital agriculture management, or advanced manufacturing operations can help mitigate the negative employment effects of automation.

Fifth, inclusive AI adoption policies should be implemented to reduce socio-economic inequalities that may arise from automation. Since routine occupations are often concentrated among lower-income workers, AI-driven displacement could exacerbate existing labor market inequalities if adequate support mechanisms are not established. Governments and organizations should therefore implement policies that support equitable access to training, promote fair AI deployment practices, and ensure that productivity gains from AI adoption are shared across the workforce.

Finally, future labor market monitoring and research should be strengthened. As AI technologies continue to evolve, it will be essential for policy-makers and organizations to track emerging patterns of task and job displacement across sectors and regions. Establishing labor market observatories, collecting task-level employment data, and supporting interdisciplinary research on AI and employment will help ensure that workforce policies remain responsive to technological developments.

Despite providing valuable insights into the relationship between artificial intelligence and labor displacement, this study has several limitations that should be acknowledged. Addressing these limitations presents important opportunities for future research.

First, the number of analyzed use cases was limited. Although the study initially aimed to collect approximately 80 use cases across sectors, continents, and income classifications, only 37 cases met the inclusion criteria due to limited publicly available information about implemented AI systems and their deployment contexts. In many instances, online sources did not provide sufficient details regarding the specific AI technology used or the country of implementation. Consequently, while the dataset spans all continents and income classifications, the relatively small sample size may limit the generalizability of the findings.

Second, the study relies primarily on secondary data obtained from publicly available case studies and online reports. Such sources may emphasize successful implementations of AI technologies and may not fully capture unsuccessful deployments or negative employment outcomes. This may introduce publication bias, where visible implementations receive greater attention than cases where AI adoption did not achieve expected results or led to unintended consequences. Future research could address this limitation by incorporating primary data sources, such as surveys, interviews with organizations adopting AI, or labor market datasets that track occupational changes over time.

Third, the analysis focused primarily on qualitative classification of displacement types namely task displacement, job displacement, or no displacement. While this approach provides valuable conceptual insights, it does not quantify the magnitude of displacement effects. For example, it does not measure the proportion of tasks within a job that have been automated or the number of workers affected by each AI implementation. Future studies could integrate quantitative approaches, such as task-level automation metrics, employment statistics, or econometric analyses to estimate the scale of AI-driven displacement across industries.

In light of these limitations, several avenues for future research emerge. Future studies should aim to expand the dataset of AI use cases, incorporate quantitative labor market data, and conduct comparative analyses across sectors and regions. Additionally, research focusing on longitudinal labor market changes and demographic impacts of AI adoption would provide deeper insights into the evolving relationship between artificial intelligence and employment. Such efforts would contribute to a more comprehensive understanding of how AI reshapes the global workforce and inform policies aimed at promoting equitable and sustainable technological transformation.

References

[1] P. Mupfiga, D. Mpini, and M. Giyane, "Assessing the impact of artificial intelligence on job and task displacement: Evidence from the agriculture and healthcare sectors", in *Proceedings of DIGITAL 2025: Advances on Societal Digital Transformation*, Venice, Italy, 2025, pp. 1–9.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[2] D. H. Autor, F. Levy, and R. J. Murnane, "The skill content of recent technological change: An empirical exploration", *The Quarterly Journal of Economics*, vol. 118, no. 4, pp. 1279–1333, 2003.

[3] D. H. Autor, "Why are there still so many jobs? The history and future of workplace automation", *Journal of Economic Perspectives*, vol. 29, no. 3, pp. 3–30, 2015.

[4] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?", *Technological Forecasting and Social Change*, vol. 114, pp. 254–280, 2017.

[5] J. Bessen, "Automation and jobs: When technology boosts employment", *Economic Policy*, vol. 34, no. 100, pp. 589–626, 2019.

[6] N. J. Nilsson, "Artificial intelligence, employment, and income", *AI Magazine*, vol. 5, no. 2, pp. 5–5, 1984.

[7] D. Acemoglu and P. Restrepo, "Artificial intelligence, automation, and work", in *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2018, pp. 197–236.

[8] Q. Wang, X. Liu, and K. Huang, "Displaced or Augmented? How does Artificial Intelligence Affect Our Jobs: Evidence from LinkedIn.", in *Fortieth International Conference on Information Systems*, 2019, pp. 1–10.

[9] R. Gruetzemacher, D. Paradice, and K. B. Lee, "Forecasting extreme labor displacement: A survey of AI practitioners", *Technological Forecasting and Social Change*, vol. 161, p. 120 323, 2020.

[10] R. Tiwari, "The impact of AI and machine learning on job displacement and employment opportunities", *International Journal of Engineering Technologies and Management Research*, vol. 7, no. 1, pp. 1–8, 2023.

[11] A. Olatunbosun, D. A. Ojo, N. A. Eboh, A. Ukwa, and T. S. Jimoh, "How Artificial Intelligence Is Reshaping Labor Productivity: A Narrative Review of Theoretical Perspectives and Emerging Empirical Evidence", *Journal of Economics, Business, and Commerce*, vol. 2, no. 2, pp. 414–425, 2025.

[12] S. Pokhrel, S. Poudel, S. R. Banjade, S. Ganesan, and C. Goonawardane, "Generative AI and Job Vulnerability: A Global Review", *Recent Research Reviews Journal*, vol. 4, no. 2, pp. 281–305, 2025.

[13] K. Pandey, "Artificial Intelligence and the Evolving Labor Market: A Comprehensive Review and Policy Roadmap", *Journal of Computer Science and Technology Studies*, vol. 7, no. 10, pp. 306–315, 2025.

[14] E. Carbonara and E. Santarelli, "The Labor Market Impact of Generative AI: A Critical Survey of the Empirical Literature", *Rivista Trimestrale di Scienza dell'Amministrazione*, vol. 70, no. 4, pp. 1–30, 2025.

[15] T. Berger and B. Frey, *Digitalisation, jobs and convergence in Europe: Strategies for closing the skills gap*. Oxford Martin School Oxford, 2016, vol. 50.

[16] F. Jiang et al., "Artificial intelligence in healthcare: past, present and future", *Stroke and Vascular Neurology*, vol. 2, no. 4, 2017.

[17] S. Chhibber, S. R. Rajkumar, and S. Dassanayake, "Will Artificial Intelligence Reshape the Global Workforce by 2030? A Cross-Sectoral Analysis of Job Displacement and Transformation", *Blockchain, Artificial Intelligence, and Future Research*, vol. 1, no. 1, pp. 35–51, 2025.

[18] D. Acemoglu and P. Restrepo, "Robots and jobs: Evidence from US labor markets", *Journal of Political Economy*, vol. 128, no. 6, pp. 2188–2244, 2020.

[19] P. Muthukrishnan, *Artificial Intelligence and the Future of Work: A Comprehensive Analysis of Job Transformation and Workforce Adaptation by 2030*, Artificial Intelligence and the Future of Work, 2025.

[20] M. Arntz, T. Gregory, and U. Zierahn, "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis", OECD Publishing, Paris, Tech. Rep. 189, 2016. DOI: 10.1787/5jlz9h56dvq7-en.

[21] I. Rizvi, C. Bose, and N. Tripathi, "Transforming Education: Adaptive Learning, AI, and Online Platforms for Personalization", in *Technology for Societal Transformation: Exploring the Intersection of Information Technology and Societal Development*, Singapore: Springer Nature Singapore, 2025, pp. 45–62.

[22] L. J. Tveita, A. Kotsialos, and P. Vassilakopoulou, "Human-AI Collaboration for Case Management: Exemplars, Required Capabilities and Transition Management", *Procedia Computer Science*, vol. 256, pp. 166–173, 2025.

[23] D. Brougham and J. Haar, "Smart technology, artificial intelligence, robotics, and algorithms (STARA): Employees' perceptions of our future workplace", *Journal of Management & Organization*, vol. 24, no. 2, pp. 239–257, 2018.

[24] A. Ansari and A. Ansari, "Consequences of AI Induced Job Displacement", *International Journal of Business, Analytics, and Technology*, vol. 2, no. 1, pp. 4–19, 2024.

[25] A. Zirar, S. I. Ali, and N. Islam, "Worker and workplace Artificial Intelligence (AI) coexistence: Emerging themes and research agenda", *Technovation*, vol. 124, p. 102 747, 2023.

[26] A. Zarifhonarvar, "Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence", *Journal of Electronic Business & Digital Economics*, vol. 3, no. 2, pp. 100–116, 2024.

[27] A. S. George, "Bracing for the Coming AI Tsunami: Preparing the Global Workforce for Displacement and Inequality in the Age of Accelerating Technological Advancement", *Partners Universal Innovative Research Publication*, vol. 2, no. 3, pp. 126–146, 2024. DOI: 10.5281/zenodo.11668571.

[28] A. Rawashdeh, "The consequences of artificial intelligence: an investigation into the impact of AI on job displacement in accounting", *Journal of Science and Technology Policy Management*, vol. 16, no. 3, pp. 506–535, 2025. DOI: 10.1108/JSTPM-02-2023-0030.

[29] M. R. Frank, Y. Y. Ahn, and E. Moro, "AI exposure predicts unemployment risk: A new approach to technology-driven job loss", *PNAS Nexus*, vol. 4, no. 4, pgaf107, 2025.

[30] J. Nartey, "AI Job Displacement Analysis (2025–2030)", *SSRN Electronic Journal*, 2025. DOI: 10.2139/ssrn.5316265.

[31] K. McElheran et al., "AI adoption in America: Who, what, and where", *Journal of Economics and Management Strategy*, vol. 33, no. 2, pp. 375–415, 2024.

[32] P. Markert, A. Adhikari, and S. Das, "A transcontinental analysis of account remediation protocols of popular websites", *arXiv preprint arXiv:2302.01401*, 2023.

[33] World Bank, "World Bank Country and Lending Groups", 2024, Accessed: May 16, 2026. [Online]. Available: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.

[34] International Labour Organisation, "International Standard Classification of Occupations (ISCO)", 2024, Accessed: May 16, 2026. [Online]. Available: https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/.

[35] EnvisionIT, "Raddify for Mammograms", 2024, Accessed: May 16, 2026. [Online]. Available: https://www.radify.ai/for-mammograms/.

[36] A. Njanja, "South Africa's Envisionit Deep AI gets $1.65M to expand access to medical imaging", 2023, Accessed: May 16, 2026. [Online]. Available: https://techcrunch.com/2023/02/23/south-africas-envisionit-deep-ai-gets-1-65m-to-expand-access-to-medical-imaging/.

[37] Daba Finance, "South Africa's Envisionit Deep AI Partners with Pharma Giant Bayer", 2024, Accessed: May 16, 2026. [Online]. Available: https://dabafinance.com/en/news/south-africas-envisionit-deep-ai-partners-with-pharma-giant-bayer.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[38] J. Kart, "Robin The Robot Comforts Kids In Hospitals, Can Help With Covid-19", 2020, Accessed: May 16, 2026. [Online]. Available: https://www.forbes.com/sites/jeffkart/2020/06/17/robin-the-robot-comforts-kids-in-hospitals-can-help-with-covid-19/.

[39] R. Melkumyan, "The Making of Robin: A Robotic Friend for Pediatric Patients", 2024, Accessed: May 20, 2026. [Online]. Available: https://evnreport.com/creative-tech/the-making-of-robin/.

[40] J. Mahakian, "Investigating the use of an autonomous robot assistant toimprove the wellbeing of institutionalized older adults inArmenia", in *Alzheimer and Dementia*, 2022.

[41] R. Manuel, "Ukraine OKs Robotic Stretchers for Military Distribution", 2024, Accessed: May 20, 2026. [Online]. Available: https://thedefensepost.com/2024/09/10/ukraine-military-robotic-stretchers/.

[42] Ukrainska Pravda, "Medical evacuation robots tested in Ukraine", 2024, Accessed: May 20, 2026. [Online]. Available: https://www.pravda.com.ua/eng/news/2024/02/29/7444397/.

[43] M. Brady, J. Coco, and K. McDonald, "Implementing Robots to Support Nursing Home Staff and Socialize with Residents", 2022, Accessed: May 16, 2026. [Online]. Available: https://digital.wpi.edu/downloads/tb09j879b.

[44] D. Daniel, "With Montreal's Scale AI, hospitals across Canada are testing new AI applications", 2024, Accessed: May 16, 2026. [Online]. Available: https://www.canhealth.com/2024/11/01/with-montreals-scale-ai-hospitals-across-canada-are-testing-new-ai-applications/.

[45] Medtronic, "MAZOR - Robotic Guidance System for Spinal Surgery", 2024, Accessed: May 20, 2026. [Online]. Available: https://www.medtronic.com/ca-en/healthcare-professionals/products/spinal-orthopaedic/spine-robotics/mazor-x-stealth-edition.html.

[46] Warringal Private Hospital, "Mazor X Stealth Edition", 2024, Accessed: May 20, 2026. [Online]. Available: https://www.warringalprivate.com.au/Our-Services/Mazor-X-Stealth-Edition.

[47] A. Gamal et al., "Clinical applications of robotic surgery platforms: a comprehensive review", *Journal of Robotic Surgery*, vol. 18, no. 1, 2024.

[48] MkulimaGPT, "About MkulimaGPT", 2024, Accessed: May 20, 2026. [Online]. Available: https://mkulimagpt.com/about-1.

[49] Africa News, "Rwanda: an AI platform for small holder farmers", 2024, Accessed: May 20, 2026. [Online]. Available: https://www.africanews.com/2023/09/28/rwanda-an-ai-platform-for-small-holder-farmers/.

[50] Farm Hut, "Discover Our Wholesome Features", 2024, Accessed: May 19, 2026. [Online]. Available: https://www.farmhutafrica.com/features.

[51] V. Makuvaro et al., "Constraints to crop production and adaptation strategies of smallholder farmers in semi-arid Central and Western Zimbabwe", *African Crop Science Journal*, 2017.

[52] P. Kendra, "Agriculture and Climate Change: Call "Plantix — Your Crop Doctor" for Help", 2022, Accessed: May 19, 2026. [Online]. Available: https://agrilinks.org/post/agriculture-and-climate-change-call-plantix-your-crop-doctor-help.

[53] I. Samal, T. K. Bhoi, A. K. Pradhan, and D. K. Mahnta, "Plantix App: A Success story of artificial intelligence in plant protection", *Van Sangyan*, 2023.

[54] B. Kanana, "Botswana Startup Pioneers AI and Drone Use in Agriculture for Sustainable Farming", 2024, Accessed: May 19, 2026. [Online]. Available: https://www.cryptopolitan.com/botswana-pioneers-drone-use-in-agriculture/.

[55] Anton Tech, "Project Anton, powered by Anton Tech", 2024, Accessed: May 19, 2026. [Online]. Available: https://www.anton.co.bw/.

[56] Fermonaut, "Revolutionizing Agriculture: Introducing JEEVN AI by Farmonaut Technologies", 2024, Accessed: May 19, 2026. [Online]. Available: https://farmonaut.com/precision-farming/revolutionizing-agriculture-jeevn-ai-your-personal-farm-advisor-2/.

[57] United Nations Development Programme, "Harnessing artificial intelligence for health", 2024, Accessed: May 20, 2026. [Online]. Available: https://www.undp.org/stories/harnessing-artificial-intelligence-health.

[58] Artificial Intelligence for Development, "Building Responsible AI For Global Health", 2024, Accessed: May 20, 2026. [Online]. Available: https://ai4d.ai/project/building-responsible-ai-for-global-health/.

[59] SophieBot, "SophieBot - Artificial Intelligence to answer your questions on sexual reproductive health, try her out today", 2024, Accessed: May 20, 2026. [Online]. Available: https://www.sophiebot.ai/.

[60] Villgro Africa, "AI Innovator Spotlight: Sophie Bot", 2024, Accessed: May 20, 2026. [Online]. Available: https://villgroafrica.org/ai-innovator-spotlight-sophie-bot/.

[61] Health Volunteers Overseas, "Robot Enriches Telehealth Learning in Cambodia", 2021, Accessed: May 20, 2026. [Online]. Available: https://hvousa.org/blog/robot-enriches-telehealth-learning-in-cambodia/.

[62] Medical Press, "Robot eases loneliness of Mexican virus patients", 2020, Accessed: May 16, 2026. [Online]. Available: https://medicalxpress.com/news/2020-08-robot-eases-loneliness-mexican-virus.html.

[63] A. France-Presse, "Robot eases loneliness of Mexican virus patients", 2020, Accessed: May 16, 2026. [Online]. Available: https://www.abs-cbn.com/classified-odd/08/30/20/robot-eases-loneliness-of-mexican-virus-patients.

[64] International Social Security Association, ""LaLuchy Robotina", the humanoid robot: Robotic therapist programme for mental health", 2020, Accessed: May 16, 2026. [Online]. Available: https://www.issa.int/gp/242717.

[65] Sychelles Nation, "The health trends of 2022", 2021, Accessed: May 16, 2026. [Online]. Available: https://www.nation.sc/articles/11915/the-health-trends-of-2022.

[66] C. Lok, S. Seah, and J. A. Baker, "Robots that deliver medication, voice-activated devices: Hospitals tap technology to improve patient care", 2023, Accessed: May 16, 2026. [Online]. Available: https://www.channelnewsasia.com/singapore/tech-ai-healthcare-ttsh-smart-ward-cgh-ae-emergency-robots-3634801.

[67] World Bank, *AI Targeting for Social Protection: The NOVISSI Program in Togo*, https://documents1.worldbank.org/curated/en/099751009222330502/pdf/IDU0e83f857301ff1047bf082710a8d21ddf42c3.pdf, Accessed: 2026-05-14, 2022.

[68] AlgorithmX Lab, *MyBucks Using AI Offers Loans via Social Media*, https://algorithmxlab.com/blog/mybucks-using-ai-offers-loans-via-social-media/, Accessed: 2026-05-14, 2024.

[69] Absa Group, *Absa Abby Virtual Assistant*, https://www.absa.co.za/self-service/absa-abby/, Accessed: 2026-05-14, 2024.

[70] Nubank, *Exploring the Future with Intelligent Agents*, https://building.nubank.com/exploring-the-future-with-intelligent-agents/, Accessed: 2026-05-14, 2024.

[71] Asian Banking and Finance, *DANA Cuts Gambling with Real-Time Detection*, https://asianbankingandfinance.net/exclusive/dana-cuts-gambling-real-time-detection, Accessed: 2026-05-14, 2024.

[72] Diceus, *Chatbot Development for Insurance Company Kniazha*, https://diceus.com/case-studies/chatbot-development-for-insurance-company-kniazha/, Accessed: 2026-05-14, 2023.

[73] Deloitte, *Transforming Contact Centre Experience with AI*, https://www.deloitte.com/nz/en/services/consulting/case-

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

studies/transforming-contact-centre-experience-with-ai.html, Accessed: 2026-05-14, 2024.

[74] Anaplan, *Anaplan Intelligent Role-Based Agents*, https://www.anaplan.com/platform/intelligent-role-based-agents/, Accessed: 2026-05-14, 2024.

[75] BlackRock, *AI Investing Insights*, https://www.blackrock.com/us/individual/insights/ai-investing, Accessed: 2026-05-14, 2024.

[76] Optasia, *Optasia AI Financial Access Platform*, https://optasia.com/, Accessed: 2026-05-14, 2024.

[77] Diceus, *RPA for Import Export Operations*, https://diceus.com/case-studies/rpa-for-import-export-operations/, Accessed: 2026-05-14, 2023.

[78] Yaskawa, *Arc Welding Robot Applications*, https://www.yaskawa.za.com/applications/application/arc-welding_a10876, Accessed: 2026-05-14, 2024.

[79] Electrek, *Tesla Giga Press Producing Single Piece Rear Body*, https://electrek.co/2021/02/04/tesla-video-giga-press-in-action-producing-giant-single-piece-rear-body/, Accessed: 2026-05-14, 2021.

[80] XYZ Robotics, *RockyOne Mobile Manipulation Robot for Container Unloading*, https://www.xyzrobotics.com/news/unloading-office-supplies-sanwa-2025, Accessed: 2026-05-14, 2025.

[81] IEEE Spectrum, *Universal Robots Introduces Its Strongest Robotic Arm Yet*, https://spectrum.ieee.org/universal-robots-introduces-its-strongest-robotic-arm-yet, Accessed: 2026-05-14, 2023.

[82] Not a Tesla App, *Tesla Launches Grok in Australia and New Zealand*, https://www.notateslaapp.com/news/3684/tesla-launches-grok-in-australia-new-zealand-with-update-2026261, Accessed: 2026-05-14, 2026.

[83] Lantmännen, "Lantmännen in unique robot project as part of the Farming of the Future", 2021, Accessed: May 16, 2026. [Online]. Available: https://www.lantmannen.com/about-lantmannen/newsroom/press-releases/2021/lantmannen-in-unique-robot-project-for-the-farming-of-the-future/.

[84] John Deere, "CP770 Cotton Picker", 2024, Accessed: May 16, 2026. [Online]. Available: https://www.deere.com/en/harvesting/cotton/cp770-cotton-picker/.

[85] Agrist, "Robot", 2024, Accessed: May 16, 2026. [Online]. Available: https://agrist.com/products/robot.

[86] K. Funaki, "FOCUS: Use of AI-based robots on the rise in Japanese agriculture", 2023, Accessed: May 16, 2026. [Online]. Available: https://english.kyodonews.net/news/2023/11/1b953d6db421-focus-use-of-ai-based-robots-on-the-rise-in-japanese-agriculture.html.

[87] Ymedia, "The Yanmar spraying robot – making work in the vineyard safer and more sustainable", 2024, Accessed: May 16, 2026. [Online]. Available: https://www.yanmar.com/global/about/ymedia/article/spraying-robot.html.

[88] Oxin by Smart Machine, "Meet Oxin, the autonomous and multi-tasking robotic tractor", 2024, Accessed: May 16, 2026. [Online]. Available: https://www.oxin.nz/.

[89] K. Ali, "The Automation Paradox: A Critical Review of Technological Innovation, Labor Displacement, and Social Inequality", *IOARP Journal of Social Sciences and Humanities Research*, vol. 4, no. 1, 2025.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Estimating the Risk of Failing Physics Courses Through Machine Learning and Monte Carlo Simulation

Isaac Caicedo-Castro*†‡ Rubby Castro-Púche†§ Samir Castaño-Rivera*‡
*Socrates Research Team
†Research Team: Development, Education, and Healthcare
‡Department of Systems and Telecommunications Engineering
§Department of Social Sciences
University of Córdoba
Carrera 6 No. 76-305, 230002, Montería, Colombia
e-mail: {isacaic | rubycastro | sacastano}@correo.unicordoba.edu.co

***Abstract*—This research is conducted within the Systems Engineering undergraduate program at the University of Córdoba (Colombia) with the aim of estimating the risk of failing physics courses, which are widely regarded as particularly challenging for students. At this university, each semester is divided into three sessions, each contributing equally to the final course grade. Our objective is to estimate the probability of course failure based on student performance in the early sessions. To this end, we compiled a dataset containing session grades and final outcomes for students enrolled in Physics I, II, and III during the 2024 and 2025 academic years. A logistic regression model was employed to estimate the probability that a student fails a course given their grades in the first two sessions. Using the predictive function obtained from the fitted model, we subsequently conducted a Monte Carlo simulation to estimate both absolute and relative risks of course failure across a broader range of academic scenarios. The results suggest that poor performance in early sessions is strongly associated with an increased probability and absolute risk of course failure, particularly in Physics I and Physics III. In several scenarios, the estimated risk of failure approaches critical levels when students fail multiple early assessments. These findings highlight the importance of early academic performance as a predictor of final outcomes and provide evidence that might support instructors and academic administrators in designing early interventions aimed at reducing course failure and improving student success.**

***Keywords*-machine learning; logistic regression; education; Monte Carlo method; absolute and relative risk.**

I. Introduction

Nowadays, academic success is a major concern for universities worldwide. Consequently, identifying effective strategies and conducting research to predict the risk of academic failure has become an active area of study within social computing, particularly through educational data mining approaches. This research builds upon a previous study [1], extending it through the use of machine learning methods and a larger dataset.

Educational data mining methods are commonly utilised to predict student dropout, delayed graduation [2][3][4], and the likelihood of course failure or withdrawal [5][6][7][8][9][10][11][12][13][14].

Our research is centred on the academic context of the University of Córdoba in Colombia, where each academic semester is split into three sessions, each lasting six weeks and contributing equally to the final grade. The final course grade is calculated as the mean of the student's grades across the three sessions. Within each session, no single assessment exceeds 40% of the session grade, meaning that each student undergoes at least nine evaluations during a semester.

This structure is designed to achieve several pedagogical goals: reducing the pressure of final exams, diversifying assessment strategies, encouraging consistent study habits, enabling continuous monitoring of learning progress, facilitating early interventions, and providing timely support to students. This approach is supported by several educational theories and instructional strategies, including constructivism, formative assessment, multiple intelligences theory, cognitive load theory, active learning, and outcome-based education.

According to constructivist theory, students build their understanding through interaction with their environment. Frequent evaluations help instructors monitor this evolving understanding and adjust teaching strategies accordingly.

Formative assessment emphasises the use of ongoing evaluations throughout the instructional period to monitor learning, identify challenges, and guide teaching. This method offers continuous feedback to both students and instructors, aligning well with the university's evaluation strategy.

The theory of multiple intelligences posits that students possess diverse talents and learning preferences. A variety of assessments throughout the semester provides a more inclusive way to evaluate these varied strengths.

Cognitive load theory suggests that students learn more effectively when information is presented in manageable segments. Multiple evaluations distributed over time align with this principle by reducing cognitive overload.

Active learning promotes student engagement through problem-solving, discussions, and hands-on activities. Multiple evaluations throughout the semester may reinforce this approach by encouraging students to actively engage with the thorough material.

At the University of Córdoba, Outcome-Based Education (OBE) is the foundational approach. It emphasises clearly defined learning outcomes and assessments aligned with those outcomes. Dividing the semester into multiple sessions allows for a more granular alignment of evaluations with specific goals.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

The university's OBE model is integrated with the Structure of the Observed Learning Outcomes (SOLO) taxonomy, which categorises learning into five levels:

1) Prestructural (0.0–2.0): The student has not yet grasped the key concepts.
2) Unistructural (2.1–2.9): The student understands a single aspect of the task.
3) Multistructural (3.0–3.7): The student understands several aspects, but without integration.
4) Relational (3.8–4.5): The student can integrate multiple aspects meaningfully.
5) Extended Abstract (4.6–5.0): The student demonstrates deep understanding and applies concepts to new contexts.

To pass an evaluation, a student must achieve at least the multistructural level, corresponding to a grade above 3.0.

By structuring learning outcomes around SOLO taxonomy principles, the curriculum offers a coherent and progressively challenging learning experience. This structure emphasises the development of deeper understanding as students progress. However, despite this pedagogical framework, physics courses remain particularly challenging for systems engineering students, who often struggle to reach the relational or extended abstract levels. For example, during 2025 and 2024, the average final grades for Physics I, II, and III were 3.42, 3.55, and 3.73, respectively, suggesting limited integration of concepts or application to real-world contexts.

In an effort to mitigate course failure and student dropout, the university considers students to be at risk of failing a course if they fail either of the first two sessions. This institutional criterion motivates the following research questions:

- What is the risk of failing a physics course if a student fails the first session?
- What is the risk if a student fails the second session?
- What is the risk if a student fails the first and the second sessions?

These questions were previously addressed in [1] by simulating a wide range of grade scenarios across the three sessions using the Monte Carlo method. In contrast, in our research, a predictive model is fitted using a larger dataset to estimate the probability of failing a course given a student's grades in the first two sessions. We then employ the Monte Carlo method to simulate a wide range of grade scenarios across these sessions in order to analyse the research questions posed above.

The Monte Carlo method has been applied in related educational contexts, for instance, to evaluate curriculum effectiveness [15] and to estimate students' motivation when learning scientific computing [16][17]. As far as we know, no prior research has addressed these questions using the methodological approach adopted in this study.

The results of our simulations suggest the following findings:

- For Physics I, approximately 62 out of 100 students are likely to fail the course. Among students who failed the first session, about 81 out of 100 are at risk of failing the course; this risk increases to 83 out of 100 for those who failed the second session, and to 99 out of 100 for those who failed both sessions.
- For Physics II, approximately 59 out of 100 students are likely to fail the course. Among students who failed the first session, about 72 out of 100 are at risk of failing; this increases to 85 out of 100 for those who failed the second session, and to 99 out of 100 for those who failed both sessions.
- For Physics III, approximately 52 out of 100 students are likely to fail the course. Among students who failed the first session, about 71 out of 100 are at risk of failing; this increases to 73 out of 100 for those who failed the second session, and to 95 out of 100 for those who failed both sessions.

These insights contribute to implement early intervention strategies and improve academic support in physics courses.

Finally, the rest of this paper is outlined as follows: in Section II, we present the research and simulation methodology adopted in this research, while we present and discuss the results in Section III. The article concludes in Section IV.

II. Research Methodology

We adopted a quantitative approach, collecting the session and final grades of 611 students enrolled in physics courses at the University of Córdoba between 2024 and 2025. Specifically, 280 students were enrolled in Physics I, 193 in Physics II, and 138 in Physics III. The relatively small dataset size reflects the recent implementation of the previously described Outcome-Based Education (OBE) framework at the institution.

Let $\mathcal{D}$ denote the dataset used in this research, defined as follows: $\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathcal{X} \subset \mathbb{R}^D \wedge y_i \in \mathcal{Y} = \{0, 1\}; \forall i = 1, \ldots, M\}$, where the grades of the ith student in the first two sessions are represented by the vector x_i, which is a two-dimensional real-valued vector. Thus, $D = 2$ because there are only two independent variables corresponding to the grades obtained in these sessions. The target or dependent variable y_i, might take one of two possible values: $y_i = 1$ if the ith student failed the course, and $y_i = 0$ otherwise. Furthermore, M is the size of the dataset.

We preprocessed the independent variables by standardising each component for later practical convenience. Thus, for a given ith student's grade in the jth session, the component x_{ij} is yielded as follows:

$$\hat{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{1}$$

where $\bar{x}_j$ and s_j denote the mean and standard deviation of the grades obtained in the jth session, respectively.

Given the dataset $\mathcal{D}$, our goal is to estimate the probability that the ith student fails the course given the grades obtained in the first two sessions, that is, $P(y_i = 1 \mid x_i)$.

To accomplish this goal, we employ logistic regression, a supervised machine learning method. In this model, the logistic function $\sigma : \mathbb{R} \to [0, 1]$ applied to a linear combination of the model weights and the independent variables to compute $P(y_i = 1 \mid x_i)$ as follows:

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

$$P(y_i = 1 \mid x_i) = \sigma(w^T \hat{x}_i + w_0) = \sigma(w_2 \hat{x}_{i2} + w_1 \hat{x}_{i1} + w_0) \tag{2}$$

where the components of the D-dimensional vector $w \in \mathbb{R}^D$ are the weights associated with the independent variables (D = 2), and $w_0 \in \mathbb{R}$ is the intercept of the prediction function. The logistic function is defined as follows:

$$\sigma(u) = \frac{1}{1 + \exp(-u)} \tag{3}$$

The values of the weights and intercept are estimated by finding w and w_0 that minimise the objective function, namely the regularised cross-entropy error $\mathcal{L}$ with L_2 regularisation:

$$\begin{aligned} \min_{w_0, w} \mathcal{L}(w_0, w) = -\sum_{i=1}^{M} \{ y_i \ln[\sigma(w^T \hat{x}_i + w_0)] + \dots \\ \dots + (1 - y_i) \ln[1 - \sigma(w^T \hat{x}_i + w_0)] \} + \frac{\lambda}{2} ||w||^2 \end{aligned} \tag{4}$$

where $\lambda \in \mathbb{R}$ is the regularisation parameter used to shrink the vector w and prevent overfitting.

The model weights may be estimated using gradient-descent–based optimisation algorithms, which rely on the partial derivatives of the objective function $\mathcal{L}$. We standardised the independent variables to speed up the convergence of the gradient-descent-based optimisation process. In particular, the partial derivatives are as follows:

$$\frac{\partial \mathcal{L}(w_0, w)}{\partial w_j} = \sum_{i=1}^{M} [\sigma(w^T \hat{x}_i + w_0) - y_i] \hat{x}_{ij} + \lambda w_j \tag{5}$$

and

$$\frac{\partial \mathcal{L}(w_0, w)}{\partial w_0} = \sum_{i=1}^{M} [\sigma(w^T \hat{x}_i + w_0) - y_i] \tag{6}$$

A broader description of logistic regression is provided in [18]. We used the implementation available in the machine learning library known Scikit-learn [19] and employed its default optimisation algorithm, namely L-BFGS [20], to minimise the above-mentioned objective function.

The probability threshold used for classification in [19] is 0.5. Accordingly, for a new input vector x_i corresponding to the ith student, whose data were not used to fit the model, the predicted label is $\hat{y}_i = 1$ (failure) if $P(\hat{y}_i = 1 \mid x_i) \geq 0.5$; otherwise, $\hat{y}_i = 0$.

On the other had, to select the regularisation parameter and evaluate the performance of the logistic regression model, we adopted K-fold Cross-Validation. In this procedure, the dataset $\mathcal{D}$ is divided into K subsets, i.e., $\mathcal{D}_1$, $\mathcal{D}_2$, ..., $\mathcal{D}_K$, and the model is evaluated K times. At each iteration, $K - 1$ subsets are used to fit the model, whereas the remaining subset is employed for testing. More formally, in the kth evaluation the model is fitted using $\cup_{l=1; l \neq k}^{K} \mathcal{D}_l$ and tested on the subset $\mathcal{D}_k$.

We chose $K = 10$ for the models fitted to the datasets corresponding to Physics I and Physics II, while $K = 3$ was selected for Physics III due to the small number of positive examples. To address the class imbalance in the dataset, we used the stratified K-fold cross-validation procedure implemented in Scikit-learn. The dataset partitions used in each experimental setting are presented in Tables I to III.

TABLE I. 10-FOLD CROSS-VALIDATION SETTING FOR THE DATASET CORRESPONDING TO THE STUDENTS WHO ENROLLED PHYSICS I COURSE

Fold	Number of Examples for Fitting the Model			Number of Examples for Testing		
	Positive	Negative	Total	Positive	Negative	Total
1	68	184	252	8	20	28
2	68	184	252	8	20	28
3	68	184	252	8	20	28
4	68	184	252	8	20	28
5	68	184	252	8	20	28
6	68	184	252	8	20	28
7	69	183	252	7	21	28
8	69	183	252	7	21	28
9	69	183	252	7	21	28
10	69	183	252	7	21	28

TABLE II. 10-FOLD CROSS-VALIDATION SETTING FOR THE DATASET CORRESPONDING TO THE STUDENTS WHO ENROLLED PHYSICS II COURSE

Fold	Number of Examples for Fitting the Model			Number of Examples for Testing		
	Positive	Negative	Total	Positive	Negative	Total
1	29	144	173	3	17	20
2	28	145	173	4	16	20
3	28	145	173	4	16	20
4	29	145	174	3	16	19
5	29	145	174	3	16	19
6	29	145	174	3	16	19
7	29	145	174	3	16	19
8	29	145	174	3	16	19
9	29	145	174	3	16	19
10	29	145	174	3	16	19

TABLE III. 3-FOLD CROSS-VALIDATION SETTING FOR THE DATASET CORRESPONDING TO THE STUDENTS WHO ENROLLED PHYSICS III COURSE

Fold	Number of Examples for Fitting the Model			Number of Examples for Testing		
	Positive	Negative	Total	Positive	Negative	Total
1	2	90	92	1	45	46
2	2	90	92	1	45	46
3	2	90	92	1	45	46

The evaluation aims to measure the accuracy, precision, recall, and harmonic mean or F_1-score of the classification model. Moreover, the regularisation parameter is tuned by selecting the value that minimises the error (defined as $1 - accuracy$) using the elbow rule. Since the dataset is imbalanced in terms of classes, we analysed the Receiver Operating Characteristic (ROC) curve and calculated the area under this curve.

Once w and w_0 are estimated, we employed the Monte Carlo simulation method [21] to explore the space of possible

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

student performance outcomes during the first two sessions. Instead of relying solely on the limited number of observed cases, we reconstructed the grade distribution using random numbers drawn from a uniform distribution over the interval [0, 5]. This allowed us to simulate a large number of plausible grade combinations and estimate the associated risks under uncertainty.

In essence, the Monte Carlo simulation serves as a data-informed method for approximating risk in underrepresented or unobserved configurations, enabling generalisation beyond the empirical observations while remaining grounded in the observed statistical characteristics of the data.

This method may be used to estimate the probability that a student fails a physics course as follows:

$$P(\hat{y}_i = 1) = \int_{\mathcal{X}} P(\hat{y}_i = 1 \mid x_i) P(x_i) dx_i \tag{7}$$

where $\hat{y}_i = 1$ if the model predicts that a new ith student fails the course. Since this integral cannot be solved analytically, the numerical method known as Monte Carlo simulation can approximate its value by generating an $N \times 2$-dimensional matrix $X \in [0,5]^{N\times 2}$, whose components $X_{ij} \sim \mathcal{U}(0,5)$ are independently drawn from a uniform distribution over the interval $[0,5]$.

Then, by the law of large numbers, the integral above is approximated by computing the expected value of $P(\hat{y}_i = 1 \mid X_{i,:})$ (here $X_{i,:}$ is the ith row of X) as follows:

$$P(\hat{y}_i = 1) \approx \mathbb{E}_{X_{i,:}\sim\mathcal{U}[0,5]^2}[P(\hat{y}_i = 1 \mid X_{i,:})] \tag{8}$$

or

$$P(\hat{y}_i = 1) \approx \frac{1}{N}\sum_{i=1}^{N} P(\hat{y}_i = 1 \mid X_{i,:}) \tag{9}$$

The grades are sampled from a uniform distribution because our goal is not to reproduce the empirical distribution of grades observed in the dataset. Instead, the purpose of the simulation is to systematically explore the full domain of admissible grade values when evaluating the risk predicted by the logistic regression model. Under this assumption, all combinations of grades in the first two sessions are treated as equally plausible a priori, preventing the simulation from being biased toward patterns that might be overrepresented in the observed data.

Assuming a uniform distribution is consistent with the goal of risk analysis under uncertainty, which focuses on understanding how the predictive model behaves across the full range of possible inputs rather than only within the empirical distribution of the data used to fit the logistic regression model. Consequently, the Monte Carlo method enables the approximation of expected failure probabilities while allowing the analysis of scenarios that may be rare or even absent in the available data.

Thus, the probability that the ith student fails a physics course given that they failed the jth session is denoted as $P(\hat{y}_i = 1 \mid x_{ij} < 3)$, where $x_{ij} < 3$ indicates that the ith student failed the jth session because the minimum grade required to pass the course is 3.

The *Absolute Risk (AR)* of failing the course given failure in session jth is defined as:

$$AR(\hat{y}_i = 1 \mid x_{ij} < 3) = \int_{\mathcal{X}} \frac{P(\hat{y}_i = 1, x_{ij} < 3)}{P(x_{ij} < 3)} dx_i \tag{10}$$

Similarly, the absolute risk of failing the course given that the student did *not* fail session jth is:

$$AR(\hat{y}_i = 1 \mid x_{ij} \geq 3) = \int_{\mathcal{X}} \frac{P(\hat{y}_i = 1, x_{ij} \geq 3)}{P(x_{ij} \geq 3)} dx_i \tag{11}$$

whereas, the *Relative Risk (RR)* is defined as the ratio of these two quantities:

$$RR(\hat{y}_i = 1 \mid x_{ij} < 3) = \frac{AR(\hat{y}_i = 1 \mid x_{ij} < 3)}{AR(\hat{y}_i = 1 \mid x_{ij} \geq 3)} \tag{12}$$

Using the Monte Carlo method, the absolute risk $AR(\hat{y}_i = 1 \mid x_{ij} < 3)$ is approximated as:

$$AR(\hat{y}_i = 1 \mid x_{ij} < 3) \approx \frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1 \wedge X_{ij} < 3)}{\sum_{i=1}^{N} \mathbf{1}(X_{ij} < 3)} \tag{13}$$

Herein we assume that $\hat{y}_i = 1$ if $P(\hat{y}_i = 1 \mid X_{i,:}) \geq 0.5$, and $\mathbf{1}(u) = 1$ if the condition u is true, and 0 otherwise. Similarly, the absolute risk for students who did *not* fail session j is:

$$AR(\hat{y}_i = 1 \mid x_{ij} \geq 3) \approx \frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1 \wedge X_{ij} \geq 3)}{\sum_{i=1}^{N} \mathbf{1}(X_{ij} \geq 3)} \tag{14}$$

Finally, the relative risk is calculated as:

$$RR(\hat{y}_i = 1 \mid x_{ij} < 3) \approx \frac{\frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1 \wedge X_{ij} < 3)}{\sum_{i=1}^{N} \mathbf{1}(X_{ij} < 3)}}{\frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1 \wedge X_{ij} \geq 3)}{\sum_{i=1}^{N} \mathbf{1}(X_{ij} \geq 3)}} \tag{15}$$

This simulation-based approach enables us to estimate the conditional risks associated with failing individual sessions and provides a probabilistic understanding of academic outcomes based on partial performance.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

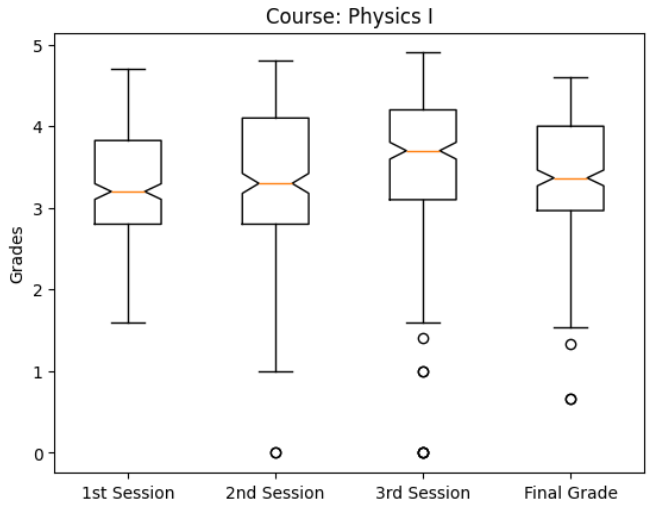


Figure 1. Grades of the students enrolled in Physics I during 2024 and 2025

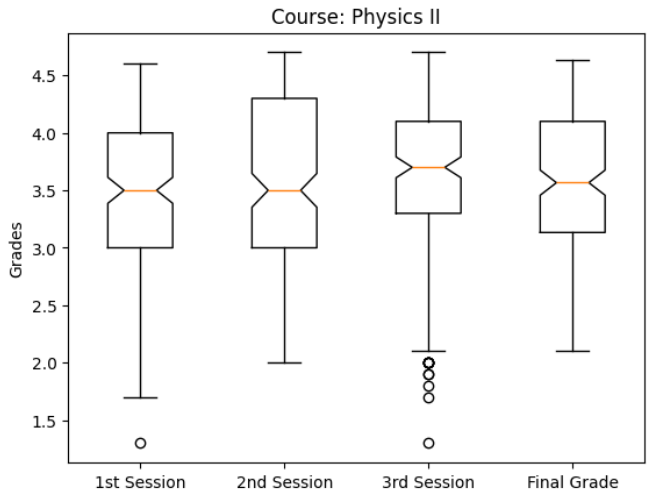


Figure 2. Grades of the students enrolled in Physics II during 2024 and 2025

The simulation was implemented in Python using the NumPy and Matplotlib libraries. The anonymised dataset and corresponding source code are available upon request from the first author.

III. The Research Results and Discussion

Based on the collected dataset, the mean grades for the first, second, and third sessions of the Physics I course were 3.29, 3.36, and 3.59, respectively, with corresponding standard deviations of 0.73, 0.81, and 0.85. As shown in Figure 1, the box plot of the final grade indicates that most students passed the course, while exceptionally high grades were relatively uncommon. Furthermore, a general trend of improved performance can be observed in the final session.

Similarly, the mean grades for Physics II were 3.51, 3.56, and 3.56 for the first, second, and third sessions, respectively, with standard deviations of 0.69, 0.72, and 0.70. Figure 2 shows a performance pattern comparable to that observed in Physics I, where exceptionally high performances were relatively infrequent.

Additionally, the dataset shows that the mean grades for the first, second, and third sessions in the Physics III course were 3.68, 3.66, and 3.86, respectively, with standard deviations of 0.66, 0.60, and 0.49. Figure 3 illustrates a pattern similar to that of the previous two physics courses, where students tend to gradually improve their grades across sessions.

TABLE IV. 10-Fold Cross-Validation Results Corresponding to the Students who Enrolled Physics I Course

Fold	**Accuracy** (%)	**Precision** (%)	**Recall** (%)	**F_1 score** (%)
1	100.00	100.00	100.00	100.00
2	92.86	100.00	100.00	100.00
3	96.43	88.89	100.00	94.12
4	89.29	72.73	100.00	84.21
5	67.86	40.00	25.00	30.77
6	92.86	100.00	75.00	85.71
7	92.86	85.71	85.71	85.71
8	89.29	100.00	57.14	72.73
9	92.86	100.00	71.43	83.33
10	89.29	75.00	85.71	80.00
Mean	90.36	86.23	77.50	80.23

On the other hand, concerning the machine learning method employed in this research, the results of the 10-fold cross-validation for the logistic regression models of Physics I and Physics II are presented in Tables IV and V, respectively. Table VI reports the results of the 3-fold cross-validation for the Physics III model.

The results reveal that all models achieve high accuracy despite the class imbalance in the datasets. Moreover, recall exceeds 70% for the models corresponding to the Physics I and Physics II courses, and reaches at least 66% for the model corresponding to the Physics III course. In addition, the area under the Receiver Operating Characteristic (ROC) curve is above 0.95 for all models (see Figures 4 to 6). These results indicate that the models are effective at predicting whether a student is likely to fail a given physics course based on their grades in the first two sessions.

These results are consistent with the confusion matrices presented in Tables VII to IX. In particular, the model correctly identified 69 out of 76 positive examples for the Physics I course (see Table VII), 26 out of 32 positive examples for the Physics II course (see Table VIII), and 2 out of 3 positive examples for the Physics III course (see Table IX). The results for the latter case are aligned with those reported for the second fold in Table VI, where the precision, recall, and F_1 score were equal to zero.

Figures 10 to 12 visualise the classification of all examples in the dataset. In particular, Figure 12 shows that only one positive example lies above the decision boundary (a false negative), which might correspond to an exceptional case.

Once the logistic regression models were fitted for each course, Monte Carlo simulations were performed as described in the previous section. Thus, the results of the numerical simulations indicate that the probabilities of failing Physics I, II, and III are 61.62%, 58.61%, and 51.82%, respectively (see

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE V. 10-FOLD CROSS-VALIDATION RESULTS CORRESPONDING TO THE STUDENTS WHO ENROLLED PHYSICS II COURSE

Fold	Accuracy (%)	Precision (%)	Recall (%)	F_1 score (%)
1	100.00	100.00	100.00	100.00
2	95.00	100.00	75.00	85.71
3	95.00	80.00	100.00	88.89
4	94.74	75.00	100.00	85.71
5	94.74	100.00	66.67	80.00
6	89.47	100.00	33.33	50.00
7	84.21	50.00	33.33	40.00
8	100.00	100.00	100.00	100.00
9	94.74	100.00	66.67	80.00
10	89.47	100.00	33.33	50.00
Mean	93.74	90.50	70.83	76.03

TABLE VI. 10-FOLD CROSS-VALIDATION RESULTS CORRESPONDING TO THE STUDENTS WHO ENROLLED PHYSICS II COURSE

Fold	Accuracy (%)	Precision (%)	Recall (%)	F_1 score (%)
1	100.00	100.00	100.00	100.00
2	97.83	0.00	0.00	0.00
3	100.00	100.00	100.00	100.00
Mean	99.28	66.67	66.67	66.67

TABLE VII. CONFUSION MATRIX ILLUSTRATING THE PERFORMANCE OF LOGISTIC REGRESSION MODEL CORRESPONDING TO THE COURSE OF PHYSICS I.

	Forecasted Class		
True Class	$\hat{y}_i = 0$	$\hat{y}_i = 1$	**Total**
$y_i = 0$	194	10	204
$y_i = 1$	17	59	76
Total	211	69	280

TABLE VIII. CONFUSION MATRIX ILLUSTRATING THE PERFORMANCE OF LOGISTIC REGRESSION MODEL CORRESPONDING TO THE COURSE OF PHYSICS II.

	Forecasted Class		
True Class	$\hat{y}_i = 0$	$\hat{y}_i = 1$	**Total**
$y_i = 0$	158	3	161
$y_i = 1$	9	23	32
Total	167	26	193

TABLE IX. CONFUSION MATRIX ILLUSTRATING THE PERFORMANCE OF LOGISTIC REGRESSION MODEL CORRESPONDING TO THE COURSE OF PHYSICS III.

	Forecasted Class		
True Class	$\hat{y}_i = 0$	$\hat{y}_i = 1$	**Total**
$y_i = 0$	135	0	135
$y_i = 1$	1	2	3
Total	136	2	138

TABLE X. PROBABILITY OF FAILING A PHYSICS COURSE OBTAINED FROM THE MONTE CARLO SIMULATION RESULTS.

Course	Probability of Failing $P(\hat{y}_i = 1)$	Standard Error	95% CI
Physics 1	61.73%	9.3×10^{-4}	[61.60%, 61.86%]
Physics 2	58.44%	6.3×10^{-4}	[58.31%, 58.56]
Physics 3	51.82%	5.0×10^{-4}	[51.72%, 51.92%]

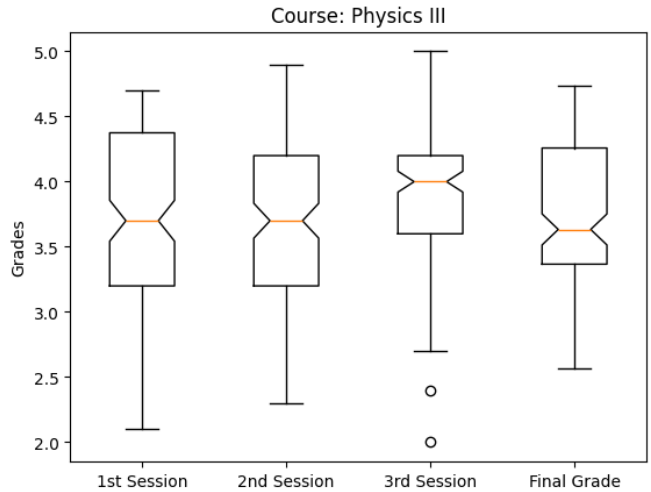


Figure 3. Grades of the students enrolled in Physics III during 2024 and 2025

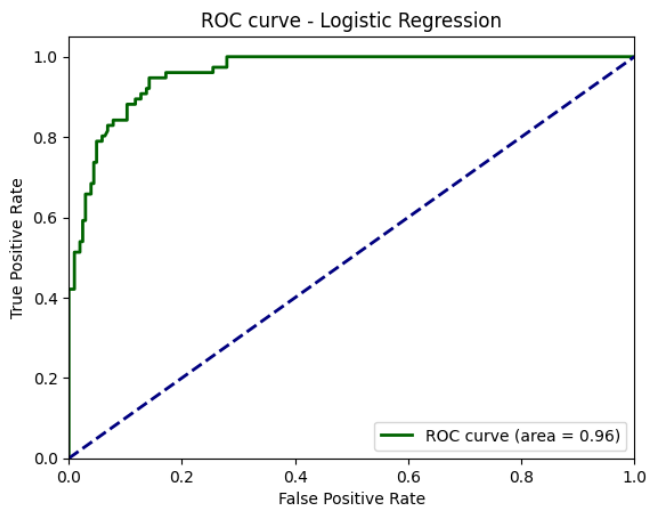


Figure 4. Receiver Operating Characteristic (ROC) curve for the logistic regression model corresponding to the course of Physics I

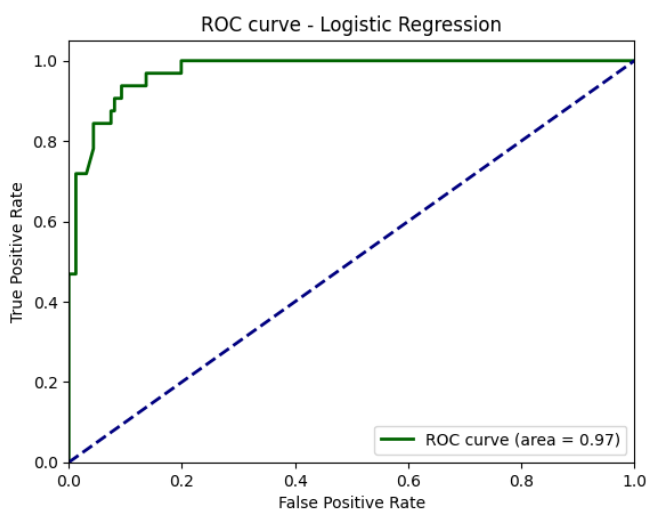


Figure 5. Receiver Operating Characteristic (ROC) curve for the logistic regression model corresponding to the course of Physics II

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

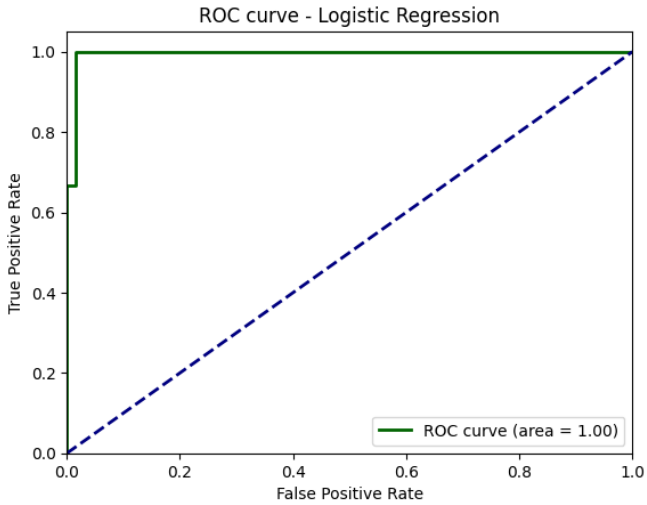


Figure 6. Receiver Operating Characteristic (ROC) curve for the logistic regression model corresponding to the course of Physics III

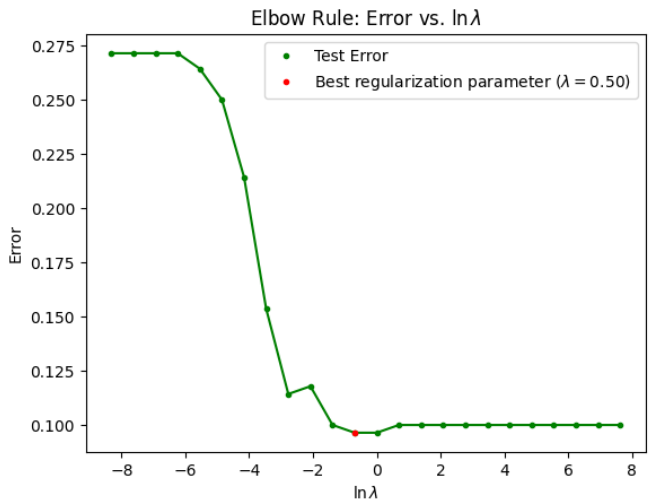


Figure 7. Elbow rule used to select the regularisation parameter for the logistic regression model corresponding to the course of Physics I

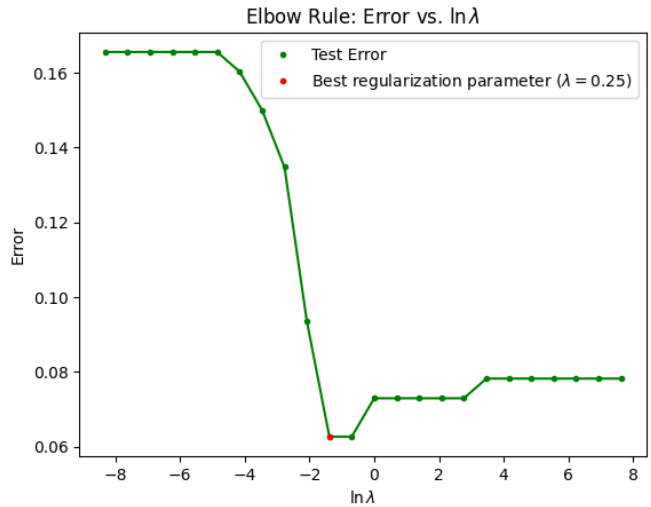


Figure 8. Elbow rule used to select the regularisation parameter for the logistic regression model corresponding to the course of Physics II

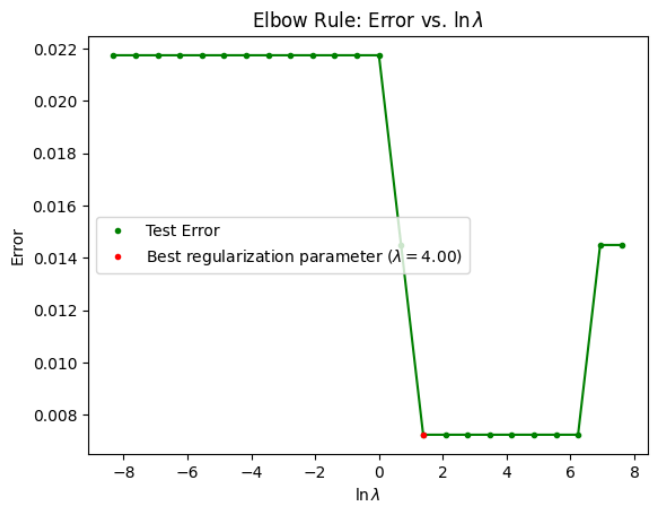


Figure 9. Elbow rule used to select the regularisation parameter for the logistic regression model corresponding to the course of Physics III

Table X). Figures 13–15 illustrate the convergence behaviour of the Monte Carlo simulations towards these estimates.

Figures 16–18 show the histograms of the failure probability obtained from simulated inputs to the logistic regression model. Instead of a bell-shaped pattern, the histograms exhibit a U-shaped distribution, reflecting the bimodal tendency typically observed in well-separated classification problems, where most randomly generated inputs yield probabilities close to either 0 or 1. Under these conditions, the simulated outcomes follow a Bernoulli distribution, and the sample mean of the simulations provides a Monte Carlo estimate of $P(y_i = 1)$, interpreted here as the probability that a student fails the physics course.

Additionally, Table XI summaries the relative and absolute risk estimates derived from the Monte Carlo simulation. The corresponding relative risks for each course are illustrated using forest plots in Figures 19–21. A similar pattern is observed

TABLE XI. ABSOLUTE AND RELATIVE RISK BY SESSION AND COURSE.

Course	Session(s) Failed	Absolute Risk (%)	Relative Risk (%)	95% CI (RR)
Physics I	S1	80.89	2.32	[2.31, 2.34]
	S2	82.82	2.61	[2.59, 2.63]
	S1 and S2	99.37	2.86	[2.84, 2.88]
Physics II	S1	71.94	1.85	[1.83, 1.85]
	S2	85.17	4.47	[4.42, 4.51]
	S1 and S2	98.75	2.54	[2.52, 2.55]
Physics III	S1	70.63	2.97	[2.95, 2.98]
	S2	73.23	3.67	[3.64, 3.69]
	S1 and S2	95.54	4.01	[3.98, 4.04]

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

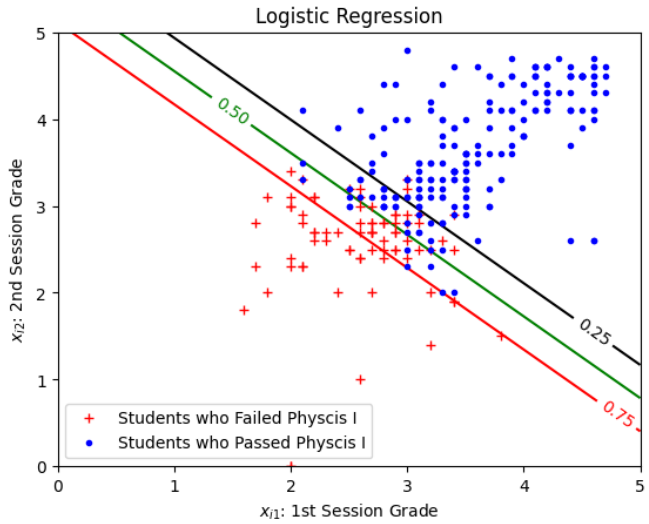


Figure 10. Classification of students by the logistic regression model for the Physics I course. Regions above the black line correspond to failure probabilities below 25%, while regions below the red line correspond to probabilities above 75%. The green line indicates the decision boundary, i.e., $P(\hat{y}_i \mid x_i) = 50\%$

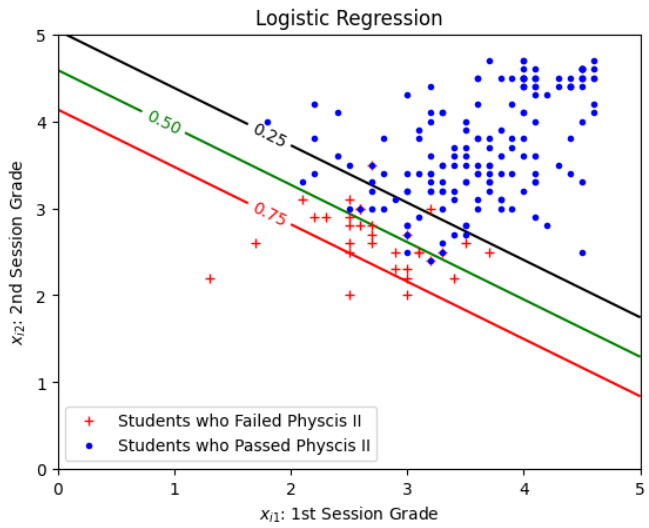


Figure 11. Classification of students by the logistic regression model for the Physics II course. Regions above the black line correspond to failure probabilities below 25%, while regions below the red line correspond to probabilities above 75%. The green line indicates the decision boundary, i.e., $P(\hat{y}_i \mid x_i) = 50\%$

for Physics I and Physics III. When comparing the risk of failing the course given that a student fails either of the first two sessions ($RR(\hat{y}_i \mid x_{i1} < 3) = 2.32$ and $RR(\hat{y}_i \mid x_{i2} < 3) = 2.61$ for Physics I, whereas $RR(\hat{y}i \mid x_{i1} < 3) = 2.97$ and $RR(\hat{y}_i \mid x_{i2} < 3) = 3.67$ for Physics III), failing both sessions is associated with a substantially higher risk of failing Physics I ($RR(\hat{y}_i \mid x_{i1} < 3, x_{i2} < 3) = 2.86$) and Physics III ($RR(\hat{y}_i \mid x_{i1} < 3, x_{i2} < 3) = 4.01$).

Interestingly, for Physics II (see Figure 20), failing the second session alone ($RR(\hat{y}_i \mid x_{i2} < 3) = 4.47$) is associated

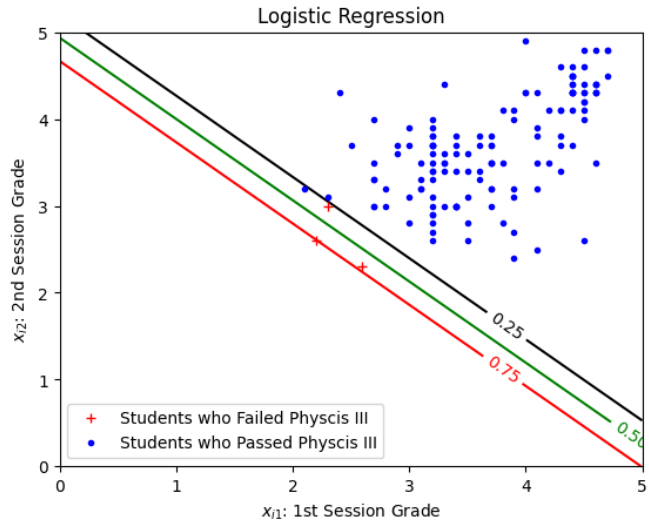


Figure 12. Classification of students by the logistic regression model for the Physics III course. Regions above the black line correspond to failure probabilities below 25%, while regions below the red line correspond to probabilities above 75%. The green line indicates the decision boundary, i.e., $P(\hat{y}_i \mid x_i) = 50\%$

with a higher relative risk than failing either the first session alone ($RR(\hat{y}_i \mid x_{i1} < 3) = 1.85$) or both sessions combined ($RR(\hat{y}_i \mid x_{i1} < 3, x_{i2} < 3) = 2.54$). This pattern suggests that interventions targeting students who fail the first two sessions might have mitigated their risk of ultimately failing the course.

The absolute risk of course failure among students exposed to session failures versus those unexposed is compared in Table XII. The simulation results suggest that the probability of failing Physics I is 80.89% for students who fail the first session. This corresponds to a risk difference of 46.09 percentage points (95% CI: [45.81%, 46.36%] – CI stands for Confidence Interval –) relative to students who pass the first session, whose absolute risk of failure is 34.80%. This difference is statistically significant, indicating a meaningful association between failing the first session and ultimately failing Physics I. Furthermore, the relative risk is 2.32, suggesting that students who fail the first session are at least twice as likely to fail the course as those who do not (see Figure 19).

TABLE XII. COMPARISON OF ABSOLUTE RISK (AR) OF COURSE FAILURE BETWEEN STUDENTS EXPOSED AND UNEXPOSED TO FAILING PREVIOUS SESSIONS, WITH CORRESPONDING RISK DIFFERENCES (RD)

Course	Session(s) Failed	AR (%) exposed	AR (%) unexposed	RD (%)	95% CI (RD)
Physics I	S1	80.89	34.80	46.09†	[45.81, 46.36]
	S2	82.82	31.75	51.07†	[50.80, 51.34]
	S1 and S2	99.37	34.80	64.57†	[64.34, 64.81]
Physics II	S1	71.94	38.94	32.99†	[32.70, 33.29]
	S2	85.17	19.07	66.10†	[65.86, 66.34]
	S1 and S2	98.75	38.94	59.81†	[59.57, 60.05]
Physics III	S1	70.63	23.82	46.81†	[46.62, 47.01]
	S2	73.23	19.98	53.26†	[53.07, 53.44]
	S1 and S2	95.54	23.82	71.72†	[71.56, 71.89]

† (p-value < 0.05)

The absolute risk of failing the course increases to 82.82%

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

among students who fail the second session. This corresponds to a risk difference of 51.07 percentage points (95% CI: [50.80%, 51.34%]) compared with students who pass the second session, whose absolute risk of failure is 31.75%. The relative risk is 2.61, indicating that students who fail the second session are 2.61 times as likely to fail the course as those who pass it (see Figure 19).

When students fail both the first and second sessions of Physics I, the absolute risk of failing the course increases to 99.37%. This corresponds to a risk difference of 64.57 percentage points (95% CI: [64.34%, 64.81%]) compared with students who do not fail the first two sessions, whose absolute risk of failure is 34.80%. The relative risk is 2.86, further highlighting the substantially increased likelihood of course failure under these conditions (see Figure 19).

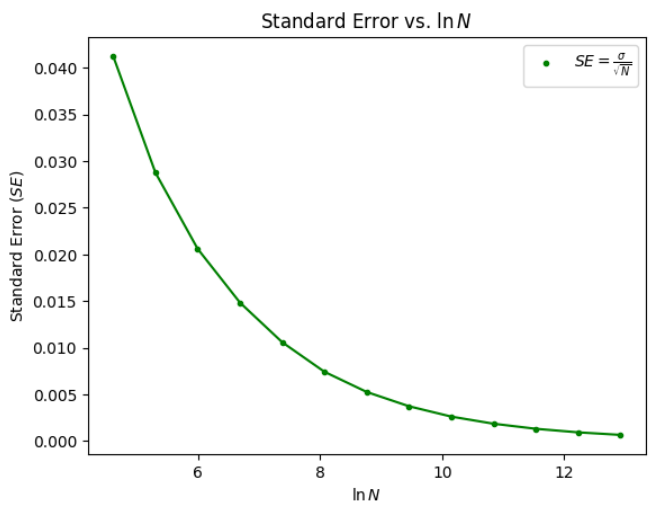


Figure 13. The simulation converges to the expected probability $P(\hat{y}_i = 1) = 61.73\%$ in the Physics I course as $N = 409,600$, with a standard error of 6.6×10^{-4}. The result lies within the 95% confidence interval of [61.61%, 61.86%].

For Physics II, the absolute risk of failing the course among students who fail the first session is 71.94%. This corresponds to a risk difference of 32.99 percentage points (95% CI: [32.70%, 33.29%]) compared with students who pass the first session, whose absolute risk of failure is 38.94%. The relative risk is 1.85, indicating that students who fail the first session are almost twice as likely to fail Physics II (see Figure 20).

The absolute risk of failing the course among students who fail the second session increases up to 85.17%. This corresponds to a risk difference of 66.10 percentage points (95% CI: [65.86%, 66.34%]) compared with students who pass the second session, whose absolute risk of failure is 19.07%. The relative risk is 4.47, indicating a substantially higher likelihood of failing the course under this condition than the previous one (see Figure 20).

Failing both sessions in Physics II rises the absolute risk of failing the course until 98.75%. This corresponds to a risk difference of 59.81 percentage points (95% CI: [59.57%, 60.05%]) compared with students who pass both sessions,

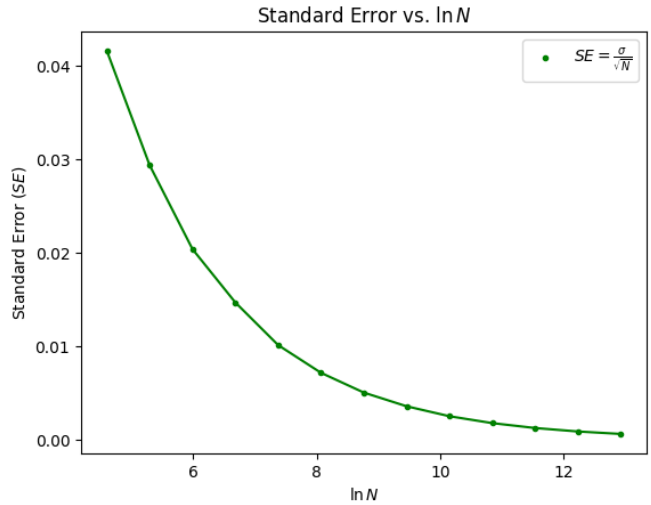


Figure 14. The simulation converges to the expected probability $P(\hat{y}_i = 1) = 58.44\%$ in the Physics II course as $N = 409,600$, with a standard error of 6.3×10^{-4}. The result lies within the 95% confidence interval of [58.34%, 58.56%].

whose absolute risk of failure is 38.94%. However, the relative risk decreases to 2.54, suggesting that students who fail the first two sessions are more than twice as likely to fail Physics II (see Figure 20).

For Physics III, the absolute risk of failing the course among students who fail the first session is 70.63%. This corresponds to a risk difference of 46.81 percentage points (95% CI: [46.62%, 47.01%]) compared with students who pass the first session, whose absolute risk of failure is 23.82%. The relative risk is 2.97, suggesting that students who fail the first session are nearly three times as likely to fail the course (see Figure 21).

For students who fail the second session, the absolute risk of course failure increases to 73.23%, compared with 19.98% among those who pass that session. This corresponds to a risk difference of 53.26 percentage points (95% CI: [53.07%, 53.44%]). The corresponding relative risk of 3.67 suggests that students who fail the second session are far more than three times as likely to fail the course (see Figure 21).

Furthermore, for students who fail the first two sessions, the absolute risk of failing Physics III increases to 95.54%, compared with 23.82% among students who pass both sessions. This corresponds to a risk difference of 71.72 percentage points (95% CI: [71.56%, 71.89%]). The relative risk of 4.01 suggests that students who fail both sessions are about four times as likely to fail the course (see Figure 21).

The pattern of absolute risk in all courses increases across sessions, whereas the relative risk does not follow this pattern in the case of Physics II. Despite the pattern observed in absolute risk, the risk differences follow the same pattern as the relative risk.

Besides, although the dataset contains few instances of students failing physics courses, especially Physics III, the Monte Carlo simulations suggest a noteworthy risk of failure.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

In particular, Physics II requires further analysis to understand the pattern of lower failure risk among students who failed in the previous two sessions.

Several threats to validity should be considered when interpreting the results. First, the simulation model does not account for potential confounding variables such as prior preparation, study habits, or instructor differences, which may influence course performance. Second, the evaluation structure analysed is specific to the University of Córdoba, limiting the generalisability of the findings to other institutions or disciplines. Finally, because observed cases of course failure, particularly in Physics III, are relatively scarce in the dataset, the risk estimates rely heavily on the assumptions of the Monte Carlo simulation model.

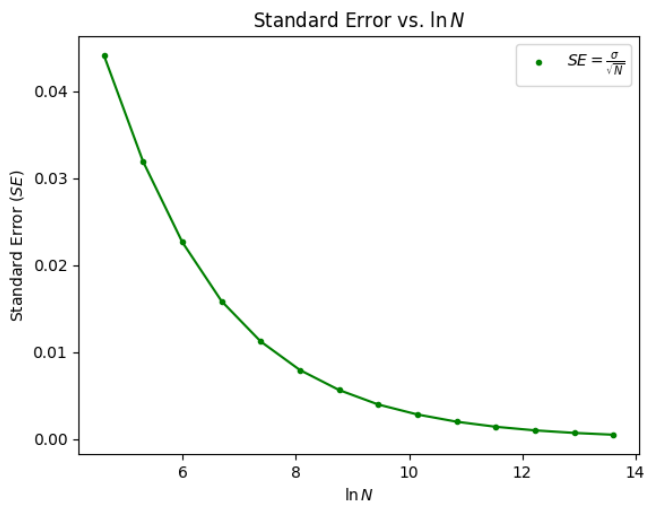


Figure 15. The simulation converges to the expected probability $P(\hat{y}_i = 1) = 51.82\%$ in the Physics II course as $N = 819,200$, with a standard error of 5×10^{-4}. The result lies within the 95% confidence interval of [51.72%, 51.92%].

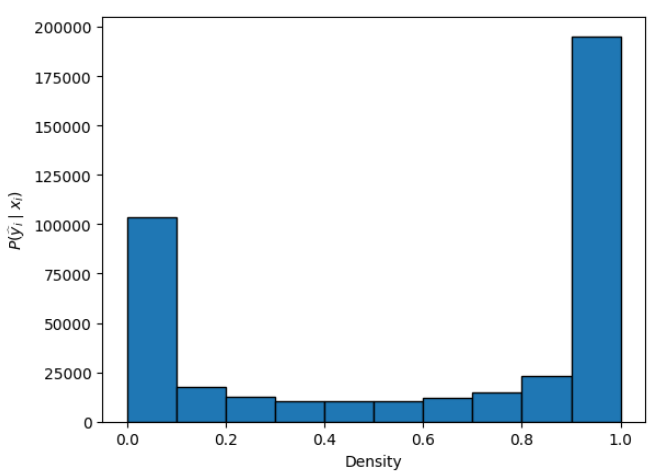


Figure 16. Distribution of probabilities $P(\hat{y}_i = 1 \mid x_i)$ obtained from the simulation for the Physics I course.

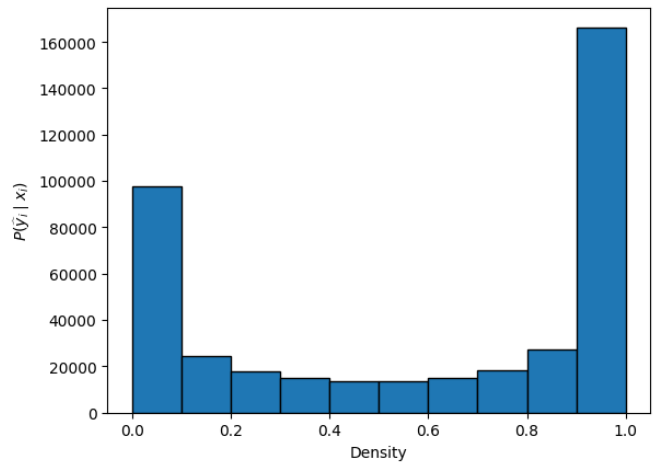


Figure 17. Distribution of probabilities $P(\hat{y}_i = 1 \mid x_i)$ obtained from the simulation for the Physics II course.

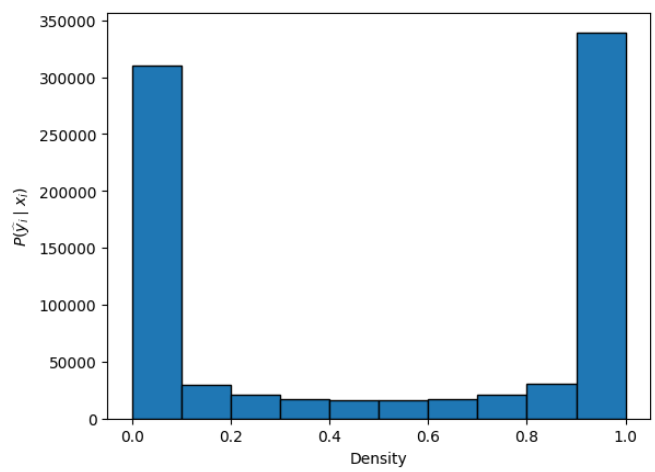


Figure 18. Distribution of probabilities $P(\hat{y}_i = 1 \mid x_i)$ obtained from the simulation for the Physics III course.

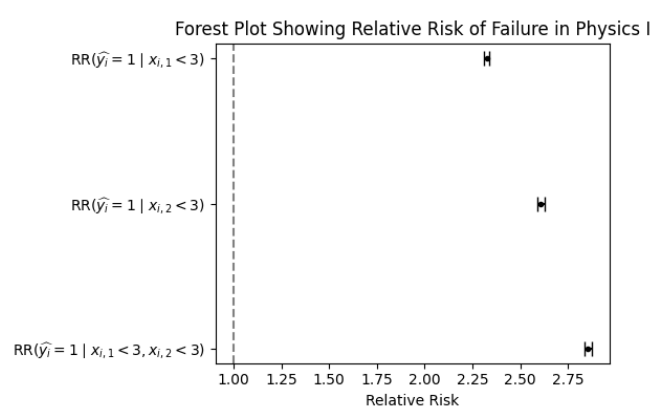


Figure 19. Forest plot showing the relative risk (RR) of failing the Physics I course. $RR(\hat{y}_i \mid x_{i1} < 3) = 2.32$, with a 95% confidence interval of [2.31, 2.34]; $RR(\hat{y}_i \mid x_{i2} < 3) = 2.61$, with a 95% confidence interval of [2.56, 2.63]; and $RR(\hat{y}_i \mid x_{i1} < 3, x_{i2} < 3) = 2.86$, with a 95% confidence interval of [2.84, 2.88]. In all cases, the Wald test p-value is less than 0.05.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

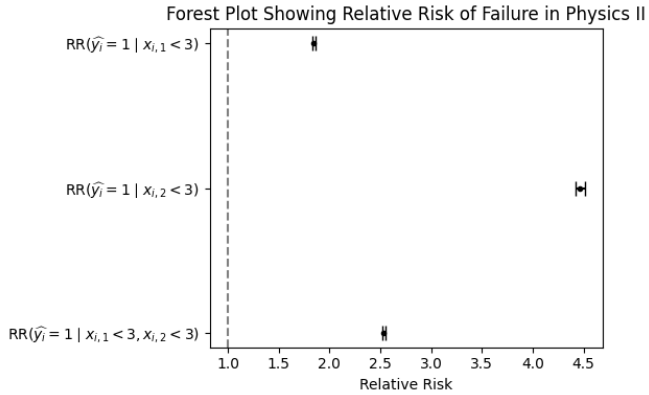


Figure 20. Forest plot showing the relative risk (RR) of failing the Physics II course. $RR(\hat{y}_i \mid x_{i1} < 3) = 1.85$, with a 95% confidence interval of [1.83, 1.85]; $RR(\hat{y}_i \mid x_{i2} < 3) = 4.47$, with a 95% confidence interval of [4.42, 4.51]; and $RR(\hat{y}_i \mid x_{i1} < 3, x_{i2} < 3) = 2.54$, with a 95% confidence interval of [2.52, 2.55]. In all cases, the Wald test p-value is less than 0.05.

IV. Conclusion and Outlook

In this research, we combined machine learning and Monte Carlo simulation to analyse the risk of failing physics courses in the Systems Engineering bachelor's program at the University of Córdoba. This approach was adopted to address the limited number of observed course failures, which makes direct empirical estimation of risk difficult.

The results demonstrate that early-session performance is a strong predictor of course failure, both at the level of individual students and at the population level. The estimated risks might therefore be employed to inform institutional policies aimed at preventing course failure and improving student retention.

The analysis also reveals a progressive increase in absolute risk as students fail additional sessions. In several scenarios, the probability of failing the course reaches alarming levels, approaching approximately 95–99%.

These findings suggest that the grading scheme may strongly amplify the consequences of early session failures. This effect might be associated with factors such as student discouragement, reduced engagement, and diminished resilience when facing subsequent academic challenges [22][23]. More broadly, performance in the first or second session is strongly associated with final course outcomes. Beyond mere statistical correlation, early academic difficulties might signal underlying motivational or behavioural challenges, making them valuable triggers for early intervention and academic support strategies. Further research is needed to design and evaluate targeted measures that could help students recover from early setbacks and improve their overall academic trajectory.

Possible interventions include tutoring programs, additional problem-solving sessions, and timely formative feedback implemented shortly after the first assessment to mitigate the risk of course failure.

As directions for future research, we propose the following:

- Collect additional data in order to apply the proposed methodology to other courses and broaden the scope of academic risk analysis.
- Extend the simulation framework to incorporate the specific coursework and evaluation structure of each session, with the aim of estimating risk more accurately.
- Adapt the simulation to consider non-uniform weighting of sessions when calculating final grades, in order to explore grading schemes that may reduce the risk of course failure.
- Incorporate additional data and evaluate alternative machine learning methods to improve the accuracy of the risk estimations.
- Consider additional explanatory variables to further strengthen both the predictive model and the simulation framework.

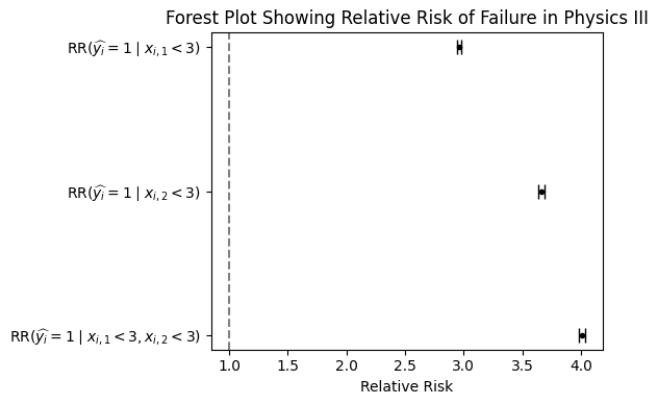


Figure 21. Forest plot showing the relative risk (RR) of failing the Physics III course. $RR(\hat{y}_i \mid x_{i1} < 3) = 2.97$, with a 95% confidence interval of [2.95, 2.98]; $RR(\hat{y}_i \mid x_{i2} < 3) = 3.67$, with a 95% confidence interval of [3.64, 3.69]; and $RR(\hat{y}_i \mid x_{i1} < 3, x_{i2} < 3) = 4.01$, with a 95% confidence interval of [3.98, 4.04]. In all cases, the Wald test p-value is less than 0.05.

Acknowledgement

Caicedo-Castro thanks the Lord Jesus Christ for blessing this project. The authors thank the Universidad de Córdoba, Colombia, for supporting this research. They also thank the Universidad de Córdoba for providing them with the data for conducting this research. Finally, the authors thank the anonymous referees for their comments that contributed to improve the quality of this paper.

References

[1] I. Caicedo-Castro, R. Castro-Púche and S. Castaño-Rivera, 'Estimating the Risk of Failing Physics Courses through the Monte Carlo Simulation', in *GPTMB 2025, The Second International Conference on Generative Pre-trained Transformer Models and Beyond*, International Academy, Research, and Industry Association, 2025, pp. 15–22.

[2] D. E. M. da Silva, E. J. S. Pires, A. Reis, P. B. de Moura Oliveira and J. Barroso, 'Forecasting Students Dropout: A UTAD University Study', *Future Internet*, vol. 14, no. 3, pp. 1–14, 2022.

[3] I. Caicedo-Castro, O. Velez-Langs, M. Macea-Anaya, S. Castaño-Rivera and R. Catro-Púche, 'Early Risk Detection of Bachelor's Student Withdrawal or Long-Term Retention', in *The 2022 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications*, International Academy, Research, and Industry Association, 2022, pp. 76–84.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[4] S. Zihan, S.-H. Sung, D.-M. Park and B.-K. Park, 'All-Year Dropout Prediction Modeling and Analysis for University Students', *Applied Sciences*, vol. 13, p. 1143, Jan. 2023. DOI: 10.3390/app13021143.

[5] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis and V. Loumos, 'Dropout Prediction in E-Learning Courses through the Combination of Machine Learning Techniques', *Computers and Education*, vol. 53, no. 3, pp. 950–965, 2009, ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2009.05.010.

[6] J. Kabathova and M. Drlik, 'Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques', *Applied Sciences*, vol. 11, p. 3130, Apr. 2021. DOI: 10.3390/app11073130.

[7] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka and P. C. Nshimyumukiza, 'Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization', *Computers and Education: Artificial Intelligence*, vol. 3, p. 100 066, 2022, ISSN: 2666-920X. DOI: https://doi.org/10.1016/j.caeai.2022.100066.

[8] V. Čotić Poturić, I. Dražić and S. Čandrlić, 'Identification of Predictive Factors for Student Failure in STEM Oriented Course', in *ICERI2022 Proceedings*, ser. 15th annual International Conference of Education, Research and Innovation, Seville, Spain: IATED, 2022, pp. 5831–5837, ISBN: 978-84-09-45476-1. DOI: 10.21125/iceri.2022.1441.

[9] V. Čotić Poturić, A. Bašić-Šiško and I. Lulić, 'Artificial Neural Network Model for Forecasting Student Failure in Math Courses', in *ICERI2022 Proceedings*, ser. 15th annual International Conference of Education, Research and Innovation, Seville, Spain: IATED, 2022, pp. 5872–5878, ISBN: 978-84-09-45476-1. DOI: 10.21125/iceri.2022.1448.

[10] I. Caicedo-Castro, M. Macea-Anaya and S. Rivera-Castaño, 'Early Forecasting of At-Risk Students of Failing or Dropping Out of a Bachelor's Course Given Their Academic History - The Case Study of Numerical Methods', in *PATTERNS 2023: The Fifteenth International Conference on Pervasive Patterns and Applications*, ser. International Conferences on Pervasive Patterns and Applications, Nice, France: IARIA: International Academy, Research, and Industry Association, 2023, pp. 40–51, ISBN: 978-1-68558-049-0.

[11] I. Caicedo-Castro, M. Macea-Anaya and S. Castaño-Rivera, 'Early Risk Detection of Bachelor's Student Withdrawal or Long-Term Retention', in *The 2023 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications*, International Academy, Research, and Industry Association, 2023, pp. 177–187.

[12] I. Caicedo-Castro, 'Course Prophet: A System for Predicting Course Failures with Machine Learning: A Numerical Methods Case Study', *Sustainability*, vol. 15, no. 18, 2023, 13950. DOI: 10.3390/su151813950.

[13] I. Caicedo-Castro, 'Quantum Course Prophet: Quantum Machine Learning for Predicting Course Failures: A Case Study on Numerical Methods', in *Learning and Collaboration Technologies*, P. Zaphiris and A. Ioannou, Eds., Cham: Springer Nature Switzerland, 2024, pp. 220–240, ISBN: 978-3-031-61691-4. DOI: 10.1007/978-3-031-61691-4_15.

[14] I. Caicedo-Castro, 'An Empirical Study of Machine Learning for Course Failure Prediction: A Case Study in Numerical Methods', *International Journal on Advances in Intelligent Systems*, vol. 17, no. 1 and 2, pp. 25–37, 2024.

[15] D. Torres, J. Crichigno and C. Sanchez, 'Assessing Curriculum Efficiency Through Monte Carlo Simulation', *Journal of College Student Retention*, vol. 22, no. 4, pp. 597–610, 2021. DOI: 10.1177/1521025118776618.

[16] I. Caicedo-Castro, O. Vélez-Langs and R. Castro-Púche, 'Using the Monte Carlo Method to Estimate Student Motivation in Scientific Computing', in *Patterns 2025, The Seventeenth International Conferences on Pervasive Patterns and Applications*, International Academy, Research, and Industry Association, 2025, pp. 15–22.

[17] I. Caicedo-Castro, R. Castro-Púche and O. Vélez-Langs, 'Identifying Factors that Increase the Risk of Demotivation in Scientific Computing Courses Using Monte Carlo Methods', *International Journal on Advances in Systems and Measurements*, vol. 18, no. 3 and 4, pp. 162–171, 2025.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, ISBN: 0387310738.

[19] F. Pedregosa et al., 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] D. C. Liu and J. Nocedal, 'On the limited memory BFGS method for large scale optimization', *Mathematical Programming*, vol. 45, no. 1, pp. 503–528, 1989.

[21] N. Metropolis and S. Ulam, 'The Monte Carlo Method', *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 1949, ISSN: 01621459, 1537274X.

[22] V. Tinto, *Leaving College: Rethinking the Causes and Cures of Student Attrition*, 2nd. Chicago, IL: University of Chicago Press, 1994, ISBN: 9780226804521.

[23] M. Yorke, 'Retention, Persistence and Success in On-Campus Higher Education, and their Enhancement in Open and Distance Learning', *Open Learning: The Journal of Open, Distance and e-Learning*, vol. 19, no. 1, pp. 19–32, 2004. DOI: 10.1080/0268051042000177827.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Comparative Bias Analysis of Recommendation Algorithms for YouTube Shorts and Long-Form Videos

Nitin Agarwal*†
*COSMOS Research Center, University of Arkansas at Little Rock, Arkansas, USA
†International Computer Science Institute, University of California, Berkeley, California, USA
e-mail: nxagarwal@ualr.edu

***Abstract*—The growing popularity of short-form video content, such as YouTube Shorts, has transformed user engagement on digital platforms, raising critical questions about the role of recommendation algorithms in shaping user experiences. These algorithms significantly influence content consumption, yet concerns about biases, echo chambers, and content diversity persist. This study develops an efficient data collection framework to analyze YouTube's recommendation algorithms for both short-form and long-form videos, employing parallel computing and advanced scraping techniques to overcome limitations of YouTube's API. The analysis uncovers distinct behavioral patterns in recommendation algorithms across the two formats, with short-form videos showing a more immediate shift toward engaging yet less diverse content compared to long-form videos. Furthermore, a novel investigation into biases in politically sensitive topics, such as the South China Sea dispute, highlights the role of these algorithms in shaping narratives and amplifying specific viewpoints. By providing actionable insights for designing equitable and transparent recommendation systems, this research underscores the importance of responsible AI practices in the evolving digital media landscape.**

***Keywords*—YouTube; recommendation collection; parallel computing; optimization; recommendation bias**

I. Introduction

Recommendation algorithms have become a cornerstone of digital platforms, transforming the way users discover and interact with content. By offering personalized suggestions, these systems help mitigate the problem of information overload, enabling users to navigate vast content repositories effectively. However, their influence extends beyond utility, shaping user engagement, content consumption patterns, and even societal behaviors. As noted in [1] and [2], recommender systems must balance their utility with fairness and transparency, ensuring equitable access to diverse content while avoiding the amplification of biases.

YouTube, one of the world's most influential video platforms, exemplifies the power and challenges of recommendation algorithms. Historically, YouTube has been synonymous with long-form video content, offering creators opportunities for monetization through advertisements, sponsorships, and thematic partnerships. However, the emergence of YouTube Shorts has disrupted this paradigm, catering to the modern digital audience's preference for bite-sized, fast-paced content. As highlighted in [3], the rise of YouTube Shorts has significantly influenced viewer behavior, leading to a decline in engagement with long-form videos. This shift underscores the need to understand how recommendation algorithms operate across these two distinct formats and the biases that may arise as a result.

The differences between short-form and long-form content extend beyond duration. Long-form videos often provide deeper narrative and thematic exploration, fostering meaningful user engagement and discourse. In contrast, short-form videos prioritize instant gratification, often relying on visually engaging elements to capture attention. This division presents unique challenges for recommendation algorithms, as their optimization strategies may differ for each format. For instance, while long-form videos may emphasize sustained engagement metrics, short-form content often prioritizes click-through rates and immediate interactions, as discussed in [4].

The investigation of biases within these algorithms is critical. Biases can arise from multiple sources, including user interaction data, algorithmic design, and platform policies. As described in [5], the integration of advanced technologies, such as large language models, offers new opportunities to address these biases by improving the interpretability and fairness of recommendations. Additionally, [3] highlights the potential for recommendation systems to inadvertently prioritize certain content types, creating echo chambers that limit exposure to diverse perspectives. Understanding these dynamics is essential to ensure that algorithms serve users equitably and transparently.

Efficient data collection is a fundamental step in studying these algorithms. Despite the availability of APIs, they often fall short in providing comprehensive recommendation data, particularly for platforms like YouTube, where recommendation chains are essential for analysis. This challenge necessitates alternative methodologies, as noted in [2], where the need for robust and scalable data collection techniques is emphasized. By enabling the large-scale analysis of recommendation patterns, such methods provide a foundation for examining the biases and behaviors embedded within these systems.

The South China Sea (SCS) adds a unique dimension to this study. This region has long been a geopolitical hotspot, with overlapping territorial claims by multiple nations and significant strategic and economic importance [6][7]. The SCS's

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

rich natural resources and critical trade routes, which account for nearly a third of global maritime trade, make it a focal point of international relations [7][8]. China's strategic use of information and digital platforms, including YouTube, to influence perceptions about its claims in the SCS underscores the importance of studying recommendation biases in this context. As explored in [8], the dissemination of content related to territorial disputes can shape narratives and sway public opinion, raising ethical questions about the role of recommendation algorithms in amplifying such content.

This research investigates biases in YouTube's recommendation algorithms through the lens of the SCS dispute, analyzing both short-form and long-form content. Understanding how the algorithms prioritize content related to politically charged topics, such as the SCS, is crucial for addressing concerns about algorithmic fairness and misinformation. Additionally, the efficient collection of recommendation data is critical to ensure the scalability and reliability of such analyses, especially given the limitations of YouTube's API. To address these challenges, this study poses the following research questions:

- RQ1: How can YouTube recommendations for short-form and long-form videos be collected efficiently without relying on API support?
- RQ2: What optimizations are necessary to enable large-scale data collection while maintaining accuracy and efficiency?
- RQ3: How does the recommendation algorithm exhibit biases in prioritizing user engagement, emotional responses, and toxicity within short-form and long-form videos?
- RQ4: How do the recommendation algorithms for short-form and long-form videos compare in their overall behavior and outcomes?

By addressing these questions, this research contributes to a deeper understanding of recommendation systems, particularly their ethical implications and biases. Unlike prior studies, this work uniquely integrates a comparative analysis of short-form and long-form video algorithms, shedding light on their distinct behaviors and outcomes. Furthermore, it provides actionable insights for designing equitable and transparent recommendation algorithms, offering a framework for mitigating biases and enhancing fairness. In doing so, this study not only advances the field of recommender systems but also underscores the critical importance of responsible AI practices in shaping the future of digital content platforms. The research flow chart outlining the methodology and key processes can be observed in Figure 1.

The remainder of this paper is organized as follows. Section II reviews the relevant literature. Section III presents techniques for efficient data collection for recommendation analysis. Section IV describes the research methodology and analytical approaches used to examine recommendation algorithms using the collected data. Section V presents the results and findings. Finally, Section VI concludes the study and outlines directions for future research.

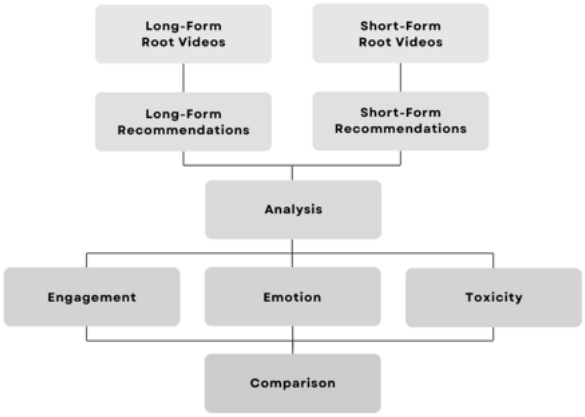


Fig. 1: Research Methodology

II. Literature Review

Understanding YouTube's recommendation algorithms has been a subject of significant interest due to their impact on content consumption and user engagement. Previous studies, such as [9], have highlighted limitations in the YouTube API, particularly its inability to directly extract recommendation data. This limitation has necessitated alternative methodologies for data collection. However, while methods for extracting recommendation data have been developed, detailed processes for efficient and accurate data collection remain underexplored.

Studies have increasingly focused on analyzing biases within YouTube's recommendation algorithms. For instance, [10][11] employed a multi-method approach to evaluate biases across recommendation depths, integrating network analysis, emotion modeling, and topic modeling to uncover systematic patterns in the algorithm's behavior. Similarly, [12] demonstrated the tendency of YouTube Shorts' recommendations to shift from serious, topical content to entertainment-focused themes, highlighting the algorithm's impact on user engagement and content diversity. Such findings underscore the need for more nuanced studies of algorithmic biases in both regular videos and Shorts.

Recent research has also examined YouTube Shorts as a distinct format. For example, [13] explored biases within the recommendations of short-form videos, revealing a strong emphasis on entertainment content. This observation aligns with [14], which analyzed engagement metrics and found that while Shorts achieved higher view counts, they elicited fewer comments per view, suggesting limited meaningful discourse compared to regular videos. Furthermore, [15] highlighted the role of symbolic content in reinforcing cohesive but narrow recommendation networks, limiting content diversity. Recommender systems have also been accused of leading users to increasingly homogeneous content, resulting in filter bubbles and echo chambers [16][17] where users are isolated from diversified content and viewpoints [18][19][20][21].

Metrics for evaluating recommendation algorithms have evolved alongside these studies. Emotional and moral dimen-

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

sions have been identified as critical evaluative metrics in studies such as [22]. Meanwhile, [23] utilized toxicity as a key metric for analyzing textual content in video metadata. Engagement metrics, including views, likes, and comments, were emphasized by [24] and [25], providing insights into user interaction patterns. Additionally, [26] analyzed thematic differences between regular videos and Shorts, noting that regular videos often encompass a broader range of topics, including science, education, and politics, whereas Shorts predominantly cater to entertainment-driven content.

Multimedia elements play a crucial role in shaping user engagement, emotional responses, and the biases inherent in YouTube's recommendation system. [27] demonstrated the emotional impact of color patterns in YouTube videos, revealing strong correlations between color combinations and viewer sentiment. These findings are particularly relevant when comparing YouTube Shorts and long-form videos, as the shorter format often relies heavily on visually striking elements, such as vibrant thumbnails, to capture attention quickly. Similarly, [28] examined the role of multimedia in mobilizing social movements, showcasing how visual content fosters connectivity and collective action. This dynamic highlights the potential for biases in recommendations to amplify certain narratives or content types based on engagement metrics. Furthermore, [29] explored the interplay between textual, visual, and auditory elements, showing how these modalities interact to evoke user emotions. These findings provide valuable insights into how both Shorts and long-form videos are differentially prioritized in YouTube's recommendation system, potentially reinforcing biases and shaping user behavior.

Despite these advancements, gaps remain in the comparative analysis of recommendation algorithms for regular YouTube videos and Shorts. Existing studies, such as [14], have highlighted the differences in engagement metrics and thematic focus, but the implications of these differences on content diversity and user behavior remain underexplored. Furthermore, while studies like [15] and [30] have provided insights into the biases introduced by symbolic and emotional content, their impact on the broader recommendation ecosystem requires further investigation.

This study seeks to bridge existing gaps by offering a comprehensive analysis of YouTube's recommendation dynamics across both short-form (YouTube Shorts) and long-form videos. By addressing challenges in efficient recommendation data collection, this work provides a robust framework for studying algorithmic biases. Integrating key metrics such as engagement, emotion, and thematic focus, the study aims to deepen our understanding of how these biases influence content accessibility, user interaction, and diversity across video formats. Furthermore, this research sets itself apart by not only analyzing biases in individual formats but also drawing comparisons between Shorts and long-form videos, providing a holistic view of the recommendation ecosystem. Ultimately, this work contributes novel insights into the ethical implications of recommendation algorithms, offering a valuable foundation for future studies and practical strategies for mitigating biases in digital media.

III. Efficient Data Collection for Recommendations

Effective data collection is fundamental to understanding the extent of bias in YouTube's recommendation algorithm. There are many theories about the severity of the bias in YouTube's recommendation algorithm; however, these theories cannot be tested without collecting recommendation data. Since the official YouTube Data API does not support the gathering of recommendation data, a more nuanced data collection method must be developed. Our approach was to simulate a user browsing YouTube using the Python library Selenium. Selenium is a library with a wide range of automated website interactions, which was initially designed for website testing but is perfect for any application that requires the simulation of user behavior.

In the following subsections, we describe the process of using Selenium to collect recommendation data from both YouTube short-form and long-form videos. Later on, we will go through the challenges we faced with both of the video formats, but first, we mention the measures we took to ensure an unbiased data collection for both formats.

The recommendation algorithm can be influenced by a plethora of factors. This forces us to eliminate as many of those factors as possible. The clearest step is to not log in to any account when collecting data. Furthermore, the history of the browser must be empty, since the algorithm can be influenced by past watched videos. Finally, the cache and cookies are cleared to further prevent the recommendation algorithm from being influenced by external factors. One factor, however, that was impossible to control for was the IP location; the internet connection has to be initiated from a certain location, and for this study we chose to connect from the United States.

Finally, to collect root video IDs, which represent the initial videos for gathering recommendations, several methods can be employed. For example, keyword searches can be performed using the YouTube Data API, scrapers such as APIFY can be utilized, or video IDs can be obtained from certain channels with same approaches. The method chosen depends on the study's requirements. Once the root video IDs are collected, the recommendation data collection process begins.

In this study, we have used keyword search. As shown in Table I, our selected keywords for the South China Sea dispute study encapsulate the conflict's key aspects: legal rulings ("Permanent Court", "Arbitration", "UNCLOS"), geopolitical tensions ("China + Philippines", "sovereignty"), and economic interests ("economic cooperation", "natural resources"). These terms are essential to examine the intricate blend of legal, political, and economic factors in the dispute, particularly focusing on China's territorial claims and the responses of neighboring states like the Philippines. The keywords enable a comprehensive analysis, aligning with the diverse perspectives and complexities discussed in our literature review.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I: South China Sea Dispute Related Keywords

Keywords
"Permanent Court" + Arbitration + South China Sea, China + Philippines + "South China Sea" + sovereignty, "South China Sea" + UNCLOS + 2016 ruling, Philippines + "West Philippine Sea" + China + sovereignty, China + "nine-dash line" + historical claims, South China Sea + maritime " justice" + international law, China + Philippines + "economic cooperation" + trade, Philippines + "Scarborough Shoal" + territorial dispute, "West Philippine Sea" + environmental concerns + coral reefs, China + "Maritime Silk Road" + regional influence, South China Sea + ASEAN + regional stability, Philippines + China + "mutual benefit" + diplomacy, China + "artificial islands" + military presence, South China Sea + "international arbitration" + dispute resolution, China + Philippines + "joint exploration" + resources.

A. YouTube Short-Form Video Recommendation Collection

Before the YouTube Shorts recommendation scraping process begins, the initial videos for which we want to collect recommendations, referred to as root videos, are imported into the program from a predefined CSV file containing their video IDs. The code can be modified to import data from an Excel or TXT file as well. Once the root videos are imported, the scraping process begins.

The process for short-form videos started with initializing the web driver using Selenium and opening the browser and as well as opening the root videos in the browser. The opening of the browser was included in the main scraping loop to ensure that the program opened a new browser with a clear cache and history for each root video. Eliminating possible bias-inducing factors like cache and history guaranteed that any bias seen in the recommended videos was purely from YouTube's recommendation algorithm.

Using the link format, https://www.youtube.com/shorts/*VIDEO_ID*, the root videos are opened. Once the root short-form video was opened, YouTube automatically loaded in recommendations under the opened video. Normally, the user would scroll down to start watching the next video recommended by YouTube. In this case, the 'user' is the program, and scrolling down can be implemented in Selenium in a few different ways.

As represented in Figure 2, the recommendation collection for short-form YouTube videos involves evenly distributing the root videos into multiple parallel processes that scroll through the videos as they collect the URLs of the recommended videos. This methodology also includes a fail safe to ensure the scroll input was registered by the browser by checking if the URL has in fact changed, meaning a new video is being played.

1) Scrolling using Selenium: To collect recommendations for short-form videos, it was necessary to simulate scrolling behavior similar to that of a real user. One method uses the buttons on the bottom right that the user can click to scroll up or down. To implement this, first the button element needs to be identified and then clicked. This is not ideal since large social media platforms occasionally change the tags and classes of elements, and there are more robust methods of scrolling down. One alternative uses the scroll wheel input using Selenium. This method works regardless of any tag

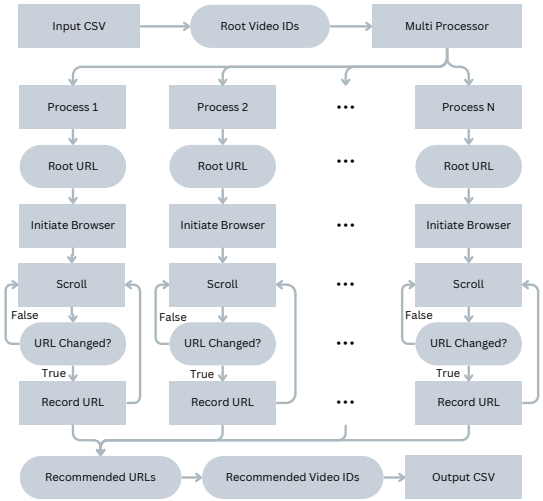


Fig. 2: Flow Chart of YouTube Shorts Scraping

or class change YouTube might implement on any part of the website. However, using the scroll wheel introduces a new problem stemming from the fact that to scroll down, the program needs a specified amount to scroll down. If selenium's scroll down function is given too low a value, YouTube will not scroll to the next video, and if the function is given too high a value, YouTube will skip one or more recommended videos. The final and best alternative uses the arrow keys or the page up and down keys. Selenium still requires an element to attach the "send keys" function to. The most stable and robust element to do this with is the "html" element that encompasses the whole page and cannot be removed. Using one of these key pairs makes the program precise in its scrolling and adaptable against changes YouTube might implement in its HTML. For simplicity, the down arrow key was used in Selenium in this way.

2) Wait Time Optimization: Once the page is scrolled and a new video appears on the screen, the next step is to capture the video ID of the newly displayed content. The simplest and most robust way to do this is to use the URL of the page which looks like in this format "https://www.youtube.com/shorts/*VIDEO_ID*". When the program scrolls down, and the next recommended video appears. The URL is updated to contain the video ID of the new video. This makes it easy to collect the ID of the new video. Furthermore, the wait time when scrolling between two consecutive videos is important. This can be done by remembering the old URL or video ID, and continuously checking the URL to detect the new URL as soon as it has changed. This checking mechanism replaces any arbitrary wait time that would have otherwise been implemented.

To improve on this basic recommendation scraping method and to simulate an actual user interaction in a more accurate fashion, a wait time was introduced to allow the video to play a certain amount. This was done because a user generally does not watch a short-form video for a few milliseconds and immediately jump to the next video. Users generally

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

watch a video for either its entire duration or a significant fraction of the video. Furthermore, our hypothesis was that recommendations would be affected and altered depending on the user's watch time (in seconds), since Shorts videos' recommendations are not immediately visible like regular videos' are. Therefore, this is also important to consider while collecting recommendations to ensure accurate recommendations. To simulate this, a wait function was implemented between collecting the video ID and scrolling to the next video. This wait time can be defined as needed, and for our case the wait times of 3 seconds, 15 seconds, and 60 seconds were chosen.

3) Multiprocessing: Introducing wait times drastically increases the collection time. For example, collecting recommendations for 2,000 short-form videos with a 10 recommendation depth and 60 seconds wait time would theoretically take 13.9 days to complete without any optimization. Therefore, this straightforward method was not a viable solution to collecting any meaningful amount of data in any reasonable amount of time. Hence, parallel processing was implemented using Python's "multiprocessing" library to decrease the total duration of recommendation collection. This theoretically decreases the time it collects the data by a magnitude of the number of processes used. This was accomplished by first splitting the root videos for recommendation scraping and then starting each on a separate process. Each browser instance was run in a separate process to enable parallel execution. Multiple browser instances are opened using each multiprocessing pool, as illustrated in Figure 3, which provides an example of the YouTube Shorts recommendation collection process. The number of processes used depends on the capabilities of the computer used. This has the potential to increase the speed of recommendation collection by ten- or twenty-fold. However, there is one more optimization that can be done to get the most amount of recommendation data in the shortest amount of time.

4) Asynchronous Scrolling: Within the individual data collection process, the computational load of scrolling to the next recommended video is exceedingly higher than that of waiting and watching the video. Traditionally, if 10 processes are being used to collect several recommendations, all of the processes start at the same time. This results in a synchronization of the scroll timings for all the processes. This is problematic because it causes a spike in the computational demand every 60 seconds or however long the wait time is defined. A better way to approach this is to distribute the scrolling timings of the processes evenly throughout the wait time. This creates a much smoother computational load with many small spikes within the defined wait time. Hence, the maximum computational load with a given number of processes is decreased, and the number of processes allowed is increased, without crashes or glitches, for a given system.

B. YouTube Long-Form Video Recommendation Collection

The first step in collecting recommendations for long-form YouTube videos was to open the root video in YouTube with this URL format https://www.youtube.com/watch?v= *VIDEO_ID*. Before collecting the recommendations, it was critical that the program ensured the page had completely loaded in, in order to avoid errors; however, it was equally important not to wait an excessive amount of time due to the exponential consequences. This was ensured using a three-step process. Firstly, within the Selenium WebDriver class, there is an option variable called *implicitly_wait*. Setting this option to a set number of seconds caused Selenium to wait that set amount of time before returning with an error if the element being called had not been loaded yet. Furthermore, the program was paused while searching for the title element, one of the main components of the YouTube video page, and continued only after the title element had been loaded in. The final step was an additional one-second pause for good measure and to allow other less essential components of the page to load.

As depicted in Figure 4 the recommendation collection for long-form videos starts off, similar to short-form recommendations, by separating the root videos evenly between N number of processes. Then the top five recommendations for each root video is collected. This process is looped for a predefined number of desired depth, Y.

1) Collection of URL Tags: Once the page had fully loaded in, the recommendations could be scraped. On YouTube, the recommendations are listed in a column on the right side of the page. Each recommended video is represented by a video thumbnail block and a description block that houses the title, the channel, number of views, and time passed since the video was uploaded. Block examples has been showed in below and in Figure 5:

Description Block Example:

```
<a class="yt-simple-endpoint style-scope
ytd-compact-video-renderer"
rel="nofollow" href="/watch?v=uWWVNq5GHp4">
... </a>
```

Thumbnail Block Example:

```
<a id="thumbnail" class="yt-simple-endpoint
inline-block style-scope ytd-thumbnail"
aria-hidden="true" tabindex="-1"
rel="nofollow" href="/watch?v=uWWVNq5GHp4">
... </a>
```

Both block elements have the video ID of the recommended video they represent. However, the thumbnail element has an ID attached to the tag named "thumbnail" whereas the description element only has a class. This class could be used to collect the recommended videos but this same class is currently used for the title and description of ads and video playlists as well. Furthermore, the class defined consists of character and numbers which means it is subject to change in the future. Therefore, using the ID of a tag for locating it is significantly more reliable than using the class identifier. This is why the thumbnail block was chosen as the focus for collecting the recommended videos.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

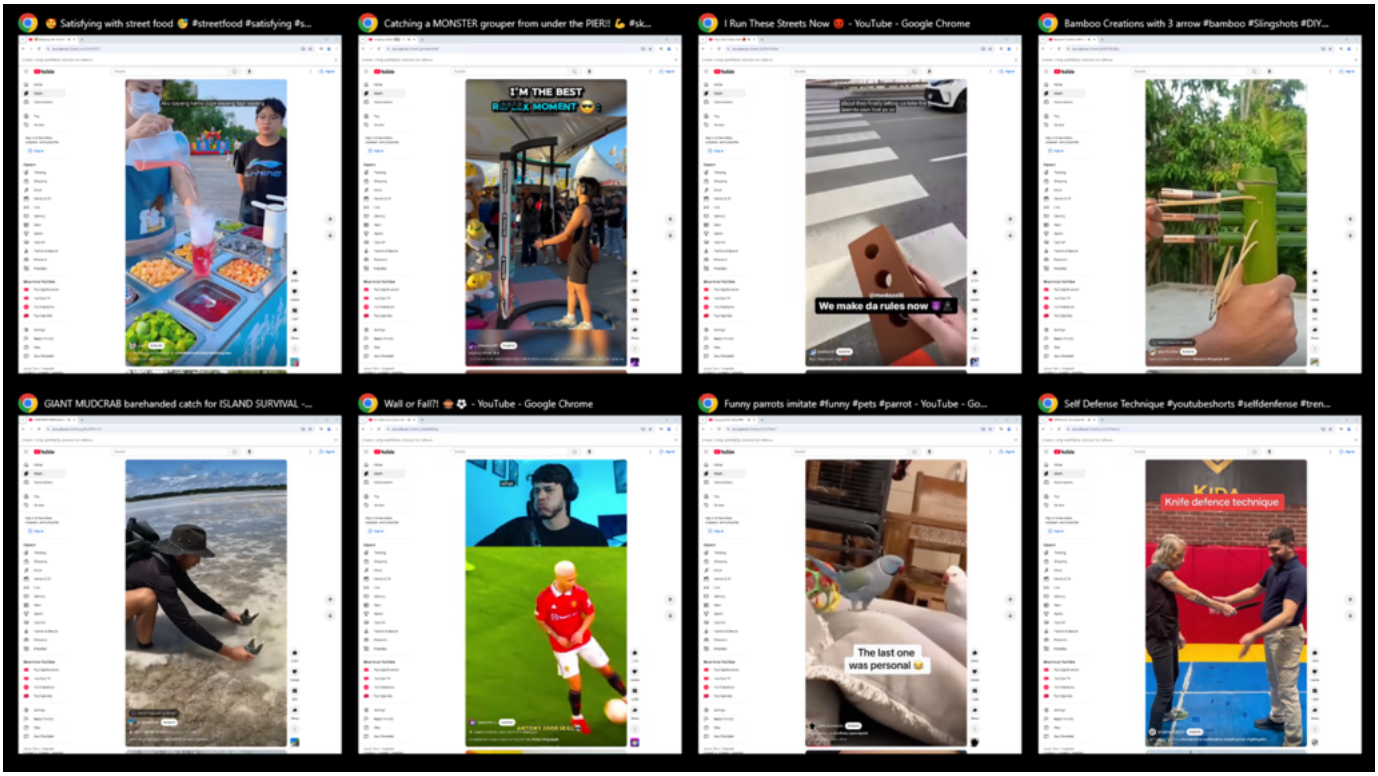


Fig. 3: YouTube Shorts Collection with Multiprocessing

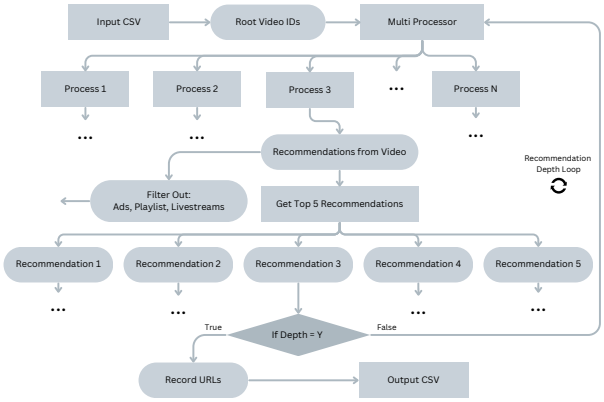


Fig. 4: Flow Chart of YouTube Long-form Video Scraping

For optimization, we aimed to investigate the scrolling functionality of browsers as well, in order to see whether scrolling had an effect on speed optimization. For this we did two experiments, one being synchronous and the other asynchronous. In synchronous, the browsers scrolled at the same time, causing a spike on CPU usage. In the asynchronous, the scrolling happened at different times, since we waited intentionally during scrolls so that they did not scroll at the same time.

The next step was to filter the elements and reduce the size of the recommendation video IDs list to the required length, which, as mentioned above, was recommended to be set as five. There were three possible non-video elements that could be present in the recommendation's column: ads, video playlists, and live streams.

2) Filtering: Using the thumbnail element already dodged the ads, since the div section for advertisements does not contain a block with an ID equal to the thumbnail. Furthermore, the video playlists were filtered by the fact that the link format is different, which was another reason thumbnail elements were focused on. Regular videos, by comparison, have a hyperlink tag attached with a string that starts with "https://www.youtube.com/watch?v=". Filtering out the live streams was a more intricate problem, because the HTML for live streams is very similar to the HTML for regular videos. There were a few indicators in terms of the HTML to differentiate the two. The most practical and robust indicator was the 'is-live-video' property on the parent element of a thumbnail.

Live Video Tag Example:

```
<ytd-thumbnail use-hovered-property=""
class="style-scope ytd-compact-video-
renderer" size="medium" is-live-video=""
loaded=""> ... </ytd-thumbnail>
```

Checking for that property and then filtering out the live videos produced a remaining list of video links for the initial recommendations loaded, typically with a length between 10 and 15. Note that if the number of recommendations to be collected per video exceeds that range, it is highly

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

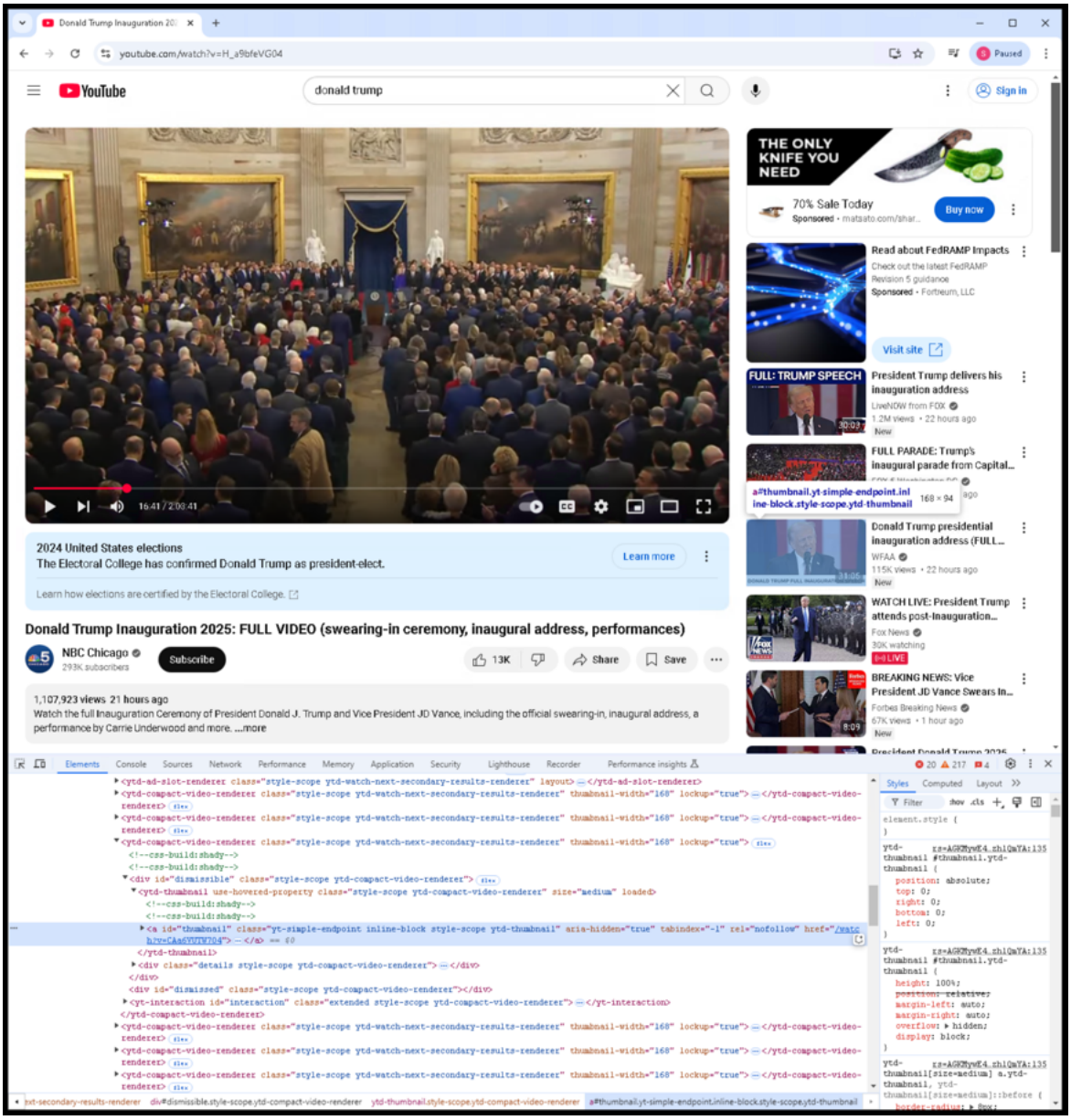


Fig. 5: Engagement Metrics of Long-Form Videos

recommended to implement a scrolling function after giving the page time to load and before collecting the recommendations, to ensure YouTube adds more videos to the recommended column. The list is further simplified by extracting the video IDs from the video URLs by removing the initial "https://www.youtube.com/watch?v=" from the strings. Finally, the list is reduced to the desired size, with a default of 5, which is safely in the no need to scroll range, if not specified by the user.

This method above allowed the program to scrape the recommendations for YouTube videos in a robust and customizable manner. Sometimes, the YouTube page had an error loading, and the page would crash. In this situation, the only solution was to refresh the page or to restart the scraping process. Therefore, the scraping function above was put into a loop that called this function repeatedly until the returned string of recommended video IDs was the required length. This was further put into a loop to allow the scraper to collect recommendations to a desired depth. The program scraped the first depth, then the second, and so on. This created a logistical problem since each depth had an exponentially greater number of videos to collect than the previous depth. This exponential growth of demand increased the time it would take to collect the required data to days, sometimes weeks. This was not an acceptable situation and had to be optimized.

3) Multiprocessing: To address the challenge of exponential growth in recommendation collection, we implemented multiprocessing. This approach significantly reduced the time

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

required for data scraping by enabling the program to run multiple processes concurrently. By distributing the workload across multiple CPU cores, multiprocessing provided an efficient solution to handle the vast number of recommendations generated as we recursively explored recommendations of recommendations.

We utilized Python's multiprocessing library to achieve this parallelization. Specifically, the Pool class was employed to manage the processes, allowing us to execute tasks in parallel and aggregate their results efficiently. Each process handled a subset of root videos and their subsequent recommendation chains. For each process, a new browser instance was launched using Selenium to ensure isolation and independent data collection, preventing interference between processes. The results from all processes were dynamically collected and combined into a single output.

The aggregated data was then structured into a Pandas DataFrame for seamless post-processing and analysis. This DataFrame not only facilitated the organization of the collected data but also allowed for straightforward export to a CSV file. The use of multiprocessing not only mitigated the computational bottleneck but also ensured scalability for handling the extensive recommendation network, making the data collection process more robust and time-efficient.

IV. Analysis

In this section, we overview the methods used to analyze the recommendation algorithms using the data collected. As mentioned previously, the recommendation algorithms for both short- and long-form videos have an immense impact on the content that users are exposed to. Furthermore, these algorithms may have tendencies to show one type of content over another. This is called recommendation bias. This bias in the recommendation algorithm might prefer to show the user one video or a handful of videos from a sea of seemingly infinite content depending on a verity of variables. Hence, to study and understand the recommendation algorithms better, it is required to analyze the recommendations across multiple metrics. The metrics chosen to be analyzed in this research are engagement, emotion, and toxicity.

A. Engagement

Analyzing engagement metrics is essential for understanding biases in video recommendation systems. Metrics like views, likes, and comments serve as indicators of a video's popularity and audience interaction. These metrics may reveal whether recommendation algorithms disproportionately favor popular content, potentially creating a bias in the recommendations.

Our analysis was based on the hypothesis that recommendation algorithms might favor videos with higher engagement metrics as users delve deeper into content recommendations. This could create a feedback loop where popular videos gain even more visibility, overshadowing less-viewed content, regardless of its relevance or quality. For this study, we identified popular videos as those with high view counts, substantial likes, and significant comment activity.

To investigate this, we tracked the engagement metrics of videos across different recommendation depths. This method enabled us to identify patterns in how the algorithm prioritized content based on engagement. If videos with higher engagement consistently dominated recommendations, it would suggest a bias toward popular content.

Understanding how engagement metrics influence recommendations is vital, as such biases may limit the visibility of newer or highly relevant content. This could narrow the range of ideas and perspectives presented to viewers, reducing content diversity. By analyzing these metrics, we aim to uncover potential biases and gain a deeper understanding of the factors shaping recommendation systems on digital platforms.

B. Emotion

The emotional dynamics of social media content can heavily influence user engagement and can make or break a recommendation algorithm. The content being consumed by a user can have a noteworthy impact on the emotional state of the user. Depending on their emotions, the users may choose to engage with the content, engage with different content on the platform, or even leave the platform altogether. Therefore, the recommendation algorithm is intrinsically incentivized to recommend users the type of content that will keep them emotionally invested. Although measuring the user's average emotional state after watching a piece of content is not logistically possible, the metadata of the videos can be used in combination with emotional analysis models to detect the emotional conveyance of the video.

The emotional expressions of the content can be measured by looking at a few qualities of the metadata extracted from videos. The first and the most obvious quality of a video to look at with regards to its emotional tendencies is its title. However, the amount of information that can be drawn from a video's title is clearly limited. Therefore, the next piece of data to look at is the description, in which the author of a video puts the video's general idea or summary. This can be a much more comprehensive text-based representation in regards to the emotionality of the video itself. Furthermore, the transcription of a video's audio can be used to apply to the same emotional analysis to understand the video's emotional contents. Finally, the comments under the video can be collected and used in a similar analysis to demonstrate the emotion inspired in the users by the content.

All of the video data mentioned can be pulled using the YouTube Data API except for the transcripts. For collecting the transcriptions of YouTube videos, there have been a few successful methodological approaches such as described in [31] and [32]. The transcript collection in this research mostly used the YouTube Transcript API, since this API had transcripts available for most of the videos. However, for the videos that did not have transcripts available, the video's audio was downloaded and transformed into text using OpenAI's

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Whisper model. These methods were inspired by and followed according to [33].

After the data for the text-based attributes of the video had been extracted, the next and final step was to use an emotional analysis model to detect what emotions were present in the various data collected. There were a few current models we considered that could be viable for emotion detection. The first option was to use a GPT-based model such as GPT4. Although GPT4 is known for being competent in a wide range of scenarios as a 'jack of all trades,' it is not specialized in emotional analysis as shown in [34], hence, there were more effective and specialized tools to conduct this analysis. The BERT model was another option that was considered. BERT would've been a good option to analyze the emotion in the text collected if a more efficient variant did not exist, namely RoBERTa as shown in [35]. RoBERTa has also been further improved by DistilRoBERTa.

Finally, in this analysis we decided to use a refined version of DistilRoBERTa, downloaded from Hugging Face, specifically trained on identifying a collection of emotions: anger, disgust, fear, joy, neutral, sadness, and surprise. Using this model we were able to analyze the pieces of data that we extracted from the videos and uncover the underlying emotional levels of the videos recommended by the algorithm.

C. Toxicity

In exploring biases in video recommendations, measuring toxicity is just as important as assessing emotional content. Toxicity refers to content that is rude, disrespectful, or unreasonable, and it can significantly impact how people interact with online media. Toxic elements in videos may subtly influence viewer preferences and behaviors, potentially affecting how recommendation algorithms function. Including toxicity as a factor in our analysis was therefore essential for gaining a comprehensive understanding of the biases that shape user engagement and content consumption patterns.

To assess toxicity in video content, we used the Detoxify model developed by Hanu (2020). Detoxify is a cutting-edge machine learning tool designed to detect and quantify toxic behavior in text. For this study, we selected the "unbiased" version of the model, which minimizes biases related to factors such as gender, race, or ideology. Built on a transformer-based architecture like BERT, the unbiased Detoxify model is effective at context-sensitive analysis and has been trained on a diverse dataset, allowing it to identify and score a range of toxic behaviors, including insults, threats, and hate speech.

We applied the Detoxify model to various textual elements of videos, such as titles, descriptions, transcriptions, and user comments. This process generated toxicity scores for each element, providing a quantifiable measure of toxic content within recommended videos. These scores allowed us to systematically evaluate the prevalence and severity of toxicity in video recommendations and to examine its potential influence on viewer behavior and preferences. This analysis was critical for understanding how toxic content affects the users' experience and the biases in recommendation algorithm.

V. Results

This results section includes both the improvement metrics for the optimizations implemented in the recommendation data collection and the analysis conducted on the recommendations themselves.

A. Data Collection Optimizations

The data collection, as mentioned in Section III, was separated into two parts: the collection of short-form videos and long-form videos. The process of collecting these two type of videos was almost completely different. The optimizations implemented were also different and resulted in different time gains for varying reasons. Therefore, the results of these data collection optimizations were also separated in the two sections below.

1) Collecting Short-Form YouTube Recommendations: The optimizations for the data collection of YouTube Shorts were theoretically decreasing, but we decided to test and measure exactly how much time was saved with these improvements. This data collection tool was developed for large scale analyses with thousands of videos. To test the time saved, we would have to run the data collection without the improvements implemented. However, if we had run a realistic data collection scenario such as with 100 root videos with 50 depths of recommendations and a 60-second watch time, it would take months for the data to be collected. Hence, we decided to test the optimizations with a much smaller data sample of 100 root videos and 5 depths of recommendations with a 60-second watch time. This ensured that the measurements would not take an unreasonable amount of time.

The collection of short-form videos on YouTube involved a number of different optimizations that affected the results. Some of the methodological optimizations mentioned in Section III made sure time was not wasted between videos and ensured the collection of accurate data. These accuracy improvements cannot be empirically measured and are not included in the following results. However, the optimizations that increase the speed of the data collection such as using multiprocessing and asynchronous scrolling had an impact that could be measured.

Looking at Table II, it is immediately evident that using a higher number of processes has a direct correlation to the data collection speed, reducing the time it takes for the collection of the data.

Furthermore, we can see that, initially, the data collection with synchronized scrolling was finishing faster than the data collection with asynchronized scrolling. This was due to the initial wait time to offset the processes to ensure that the scrolling was initiated evenly. This optimization was to prevent the computer hardware from experiencing computational load spikes so for low processes, it was not expected to affect the data collection times that much. However, when looking at higher number of processes, there was a clear improvement between the two. The improvement may seem insignificant with this small a sample of data collection, but with large scale data collection requests, it becomes more significant.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE II: Data Collection Times for Short-Form YouTube Videos

Number of Processes	Time [Asynchronized]	Time [Synchronized]
1	32,594	32,594
5	6,764	6,723
10	3,417	3,371
15	2,707	2,699
20	2,046	2,037
25	1,406	1,407
30	1,387	1,388
35	1,379	**1,383**
40	**1,374**	1,397

The initial value representing the time that it took to collect the data with a single process was 32,594 seconds. This was done with a single run for both columns since there was no additional processes to synchronize the scrolling to. Furthermore, we can see that the time it took to collect the data was higher that the theorized time. The time it should theoretically take is the number of root videos times the depth of the recommendations collected times the wait time which is calculated as 100 x 5 x 60 = 30,000. The extra two and a half thousand seconds come from the fact that the opening and closing of the browser instances take a small amount of time. Therefore, for the experiments proceeding, we expected the actual time to be at least about 7% higher than the theorized time.

2) Collecting Long-Form YouTube Recommendations: In this section, we analyze the data collection optimizations made for the collection of recommendations for long-form YouTube videos. Contrary to the YouTube Shorts recommendation collection, the collection of regular YouTube video recommendations was improved more with data accuracy increase optimizations than with time decrease optimizations. Therefore, the results below show only the time improvements made by implementing multiprocessing.

Similar to the testing and evaluation process for collecting short-form videos, a small test sample was utilized to measure improvements in this context. Specifically, the test sample consisted of 50 root videos, with five recommendations collected per video and a recommendation depth of three. The recommendations for regular videos exhibited exponential growth. For instance, starting with a single video, the top five recommendations were collected. Subsequently, the top five recommendations for each of those recommended videos were also gathered, and this process continued recursively. To ensure the feasibility of data collection within a manageable time-frame, it was crucial to limit the recommendation depth in the test sample.

Furthermore, the accuracy of the data collection tool for long-form YouTube videos was also optimized and this was done by eliminating the collection of video playlists, live-streams, advertisements, and anything that was not a regular YouTube video recommended to the user by the recommendation algorithm. Similarly, these improvements con not be measured empirically, hence will not show in Table III.

Looking at Table III, it is immediately evident that the decrease in the time taken for the data collection was almost proportional to the number of processes used. There could be some errors or some micro-glitches during the loading of the page that caused slight variances but the general trend was that using more processes resulted in faster data collection. This statement, however, starts to break down when increasing the number of processes, which pushes the limits of the pages loading. As mentioned in Section III, the data was collected as soon as the page is loaded, the bottleneck most likely stemmed from the network connection and, hence, the number of pages that can be loaded in a given amount of time. When that limit was reached, increasing the number of processes resulted in diminishing returns.

TABLE III: Data Collection Times for Long-Form YouTube Videos

Number of Processes	Time (seconds)
1	6404.8
5	1323.2
10	723.9
15	540.7
20	480.4
25	454.0
30	476.3
35	472.5
40	**471.9**

B. Engagement Analysis

The first metric analyzed in the recommendations for YouTube's algorithm was the engagement the users had with the video content. There were three categories of engagements, and these categories were the same across short-form and long-form content. The first one was the viewership. The view count, as the name suggests, represented the number of times the video was viewed by any user; to clarify, repeat views by the same user were also included. The second type of engagement was video likes, where users click a thumbs up icon to show that they like and enjoy the content they are watching. The third, final way a user can engage with videos on YouTube, whether they are short form or long form, is by posting a comment under the video. These comments can be seen by the account that posted the video and any other users who are watching the video and have navigated to the comments section. This type of engagement shows that the users find the video interesting and would like to add some information or voice their own opinions on the subject.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Before diving into the results of the analysis, it is important to note that the number of views will always be the highest measured metric on a video. Generally, they are followed second by the number of likes that the video received and finally by the number of comments left under the video. Since the discrepancies of views, likes, and comments on videos can vary drastically, we decided to use a logarithmic scale to make the visualizations more readable. Furthermore, using a logarithmic scale allowed us to put all three engagement measuring metrics into one visualization, making it easier to compare the trends in each side by side.

When examining Figure 6, the first thing to point out is that there was a general increasing trend in all three metrics. This increase in engagement was gradual and tapered off slightly going from depth 2 to 3. This clearly shows that YouTube's recommendation algorithm for long-form videos recommends and pushes users towards videos of higher engagement.

There is another trend that increased gradually as the depths increased, and that was the variance in the data. In the first depth, the data points were grouped close to each other and with little variance. However, as the depths increased, the data became more variant, and the box plots' minimum and maximum lines show this. This occurrence was largely caused by the fact that the higher depths have exponentially more videos than the previous ones. Hence, it is expected that a depth with more videos will have higher variance in these engagement metrics. This is evident especially when looking at some of the outliers in depth 4 that reach more than a billion views.

Another thing to note about the results in Figure 6 is the fact that the minimum values of the comments were always zero. This happens when the authors of the videos choose the option to disable the comments. Especially in political topics, this is a common occurrence.

When comparing Figure 7 with Figure 6, we can observe that both show an increase in the engagement metrics as the depths increase. However, the stark difference between the two sets of results is the sharp increase in engagement in the recommended YouTube Shorts videos in comparison to the regular videos. This shows that the YouTube Shorts recommendation algorithm more heavily prefers to recommend videos with high engagement compared the algorithm for recommending regular videos.

Another thing to note is that the depth in Figure 7 went up to 50, while the depth in Figure 6 went only up to 5 depths. This was a decision based on the logistical difficulty in collecting more depths with exponentially increasing recommendations. Nevertheless, even looking at the first 5 depths in Figure 7, it becomes clear that there was a much more sudden change in the amount of engagements in the videos that were being recommended.

Comparing the shift in the engagement metrics throughout the depths between the various wait times, it can be observed that different patterns emerged in each and that these patterns were caused by different reasons. Initially, the 3-second wait time communicates to the algorithm that the user did not find the content engaging and skipped the short-form video without watching it. This causes some spikes and variability in the engagement metrics as the algorithm tries to find something that the user will be interested in. This is most apparent when looking at the mean view line in Figure 7(a). There were small spikes that were random and did not follow any pattern.

When looking at the engagement values of the 15-second wait time (Figure 7(b)), however, most of those random spikes disappear, and it can be observed that the line is a bit smoother. The largest difference between the engagement trend out of these three wait times shows up when analyzing the 60-second wait time engagement metrics (Figure 7(c)). There seems to be a significant dip in engagement in exactly one in every five videos. When we observed this phenomenon occurring, we looked into the data more closely and realized that when the algorithm detects the user will watch anything they are recommended, YouTube's algorithm starts to recommend advertisements to the user to watch every fifth video. This was then confirmed by further manual testing, and it was verified that when watching YouTube Shorts, if the user fully watches every video, they will be periodically be shown ads, with this periodicity most commonly set at five.

C. Emotion Analysis

The engagement metrics of a video and its emotional response can be intertwined within the recommendation preference of the algorithm. It is important to measure both to see if there are independent correlations between these metrics. In this section, we take a look at the results for the emotional analysis for both short-form and long-form videos. Furthermore, we highlight the differences, if any, between the emotional metrics for the recommendations given by YouTube's short-form video recommendation algorithm and long-form recommendation algorithm.

Before drawing conclusions, it is important to note that the emotional assessment resulted in a score between zero and one for each emotion. This was the same for both the analysis of short-form and of long-form videos. However, the emotional analysis for short-form videos was conducted on the metadata of videos, which included titles and descriptions. However, the long-form videos had extra data, such as the transcription and comments. We collected that data and ran it through the emotional analysis to see if there were any differences. Furthermore, when we tried to collect the transcripts for short-form videos manually, we saw that most of the short-form content that was collected had music only and no speech to transcribe and analyze. Since collecting and analyzing the transcripts of the few short-form videos that had speech would have caused skewed results, the emotional analysis of short-form videos were completed once using a combination of the title and the description. The YouTube Shorts emotionality analysis was split into the 3-second, 15-second, and 60-second wait times, identical to the split done in the analysis for engagement metrics.

In Figure 8, the emotions detected in the four different pieces of data extracted from the videos are shown as they

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

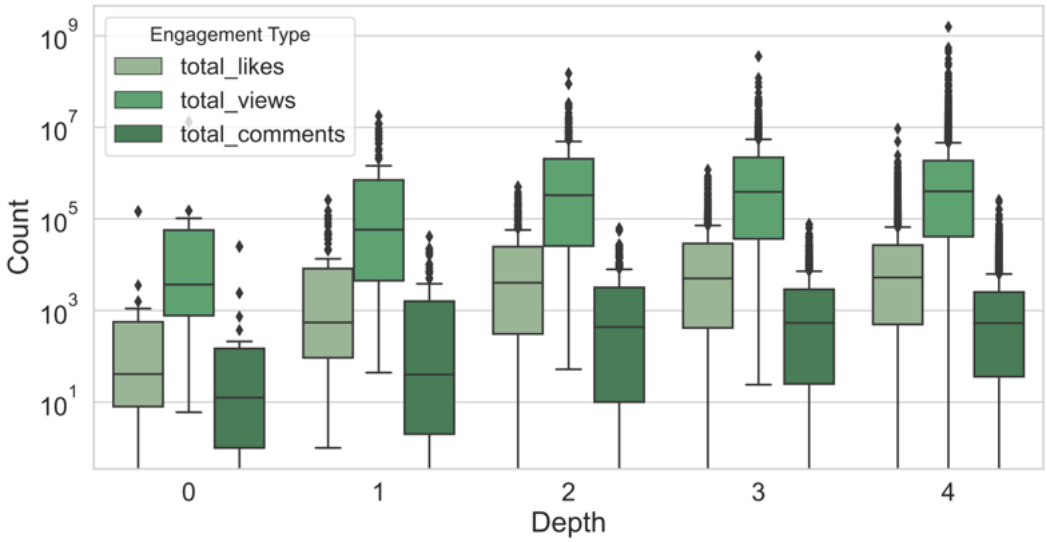


Fig. 6: Engagement Metrics of Long-Form Videos

change across the depths of recommendations. It is important to note that there was an emotion labeled as neutral that represents the absence of any clear emotion in the text provided to the model. The neutral emotion shows a clear increase when looking at the title, the description, and the transcription. Even when looking at the comments, there was only a slight decrease. This shows that the recommendation algorithm prefers to show content with increasingly neutral emotion and with a comments section with decreasing neutral emotion. Hence, although the algorithm recommends videos with lower amounts of clearly identifiable emotion, the emotional response from the users in the comments section seems to increase as the depths of the recommendation increases.

Next, the most obviously consistent trend in emotional shifts through the depths was the decrease in the anger emotion in all four pieces of video text data extracted. This is by far the clearest piece of inference that can be drawn from Figure 8. Another emotional trend that was consistent throughout the sub-figures was the increase in surprise. These observations together are evidence that the algorithm actively tries to decrease the topics that incite anger in users and to recommend content that surprises users and keeps them engaged. This could be problematic if it means that the algorithm is suppressing critical information about current events that may rightfully cause the public to get angry, such as the escalation in military tension in the south-east region of China near Taiwan.

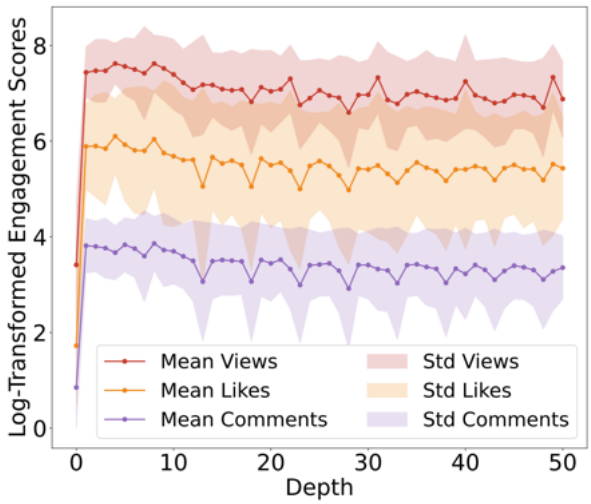
Observing some of the other emotional trends, there is no clear inference that can be drawn from the trend in fear since there is a drastic difference between how the trends through the depths for the four different pieces of data. However, if the anomalies are ignored, especially looking at the title and the comments, there is a clear decrease in fear. Similarly, it is difficult to draw any conclusions about disgust or sadness since they both started low, ended low, and had no clear trends in the data. The emotion of joy had more subtle trends. When looking at the overall data extracted from the videos themselves, there was a gradual increase in joy. However, joy had a significant and clear increase when looking at the comments. Therefore, the recommendation algorithm prefers to recommend videos that incited joy in users.

Now, looking at the emotional shifts in the recommendations for short-form content in Figure 9, it is immediately evident that the emotional trends happened in a more sudden fashion, starting from the first recommendation. There was an increase in joy in all watch times, similar to the recommendations for long-form videos. Contrary to the emotional shifts in long-form videos, there was a decrease in neutral emotion as joy increased. This was most likely caused by the fact that not only did joy increase a lot faster compared to regular videos but that the magnitude of this increase was also significantly higher, almost by a factor of five.

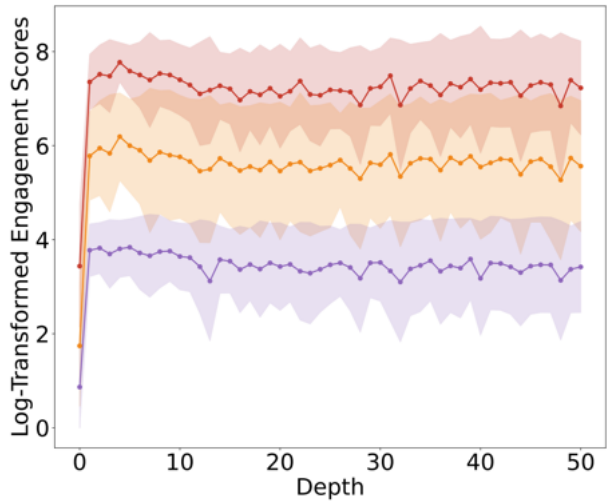
Another similarity between the recommendation algorithms for regular YouTube videos versus YouTube Shorts was that they both reduce anger over time. Looking at all three subfigures in Figure 9, anger has a clear decrease from its starting point around 0.3, and it later on tapered off right under 0.2. The emotional trends of sadness and fear were inconsistent and insignificant, hence, there was no clear conclusion that can be drawn from their results.

One anomaly to point out for the emotional trends of the recommendations for YouTube Shorts was the spikes found in the 60-second watch time results. This, similar to the spikes in engagement values, was caused by the ads shown to users when their watch times were consistently high. When an advertisement was shown, anger spiked down due to the non-angering nature of advertisements, and this caused a resulting spike increase in neutral emotion.

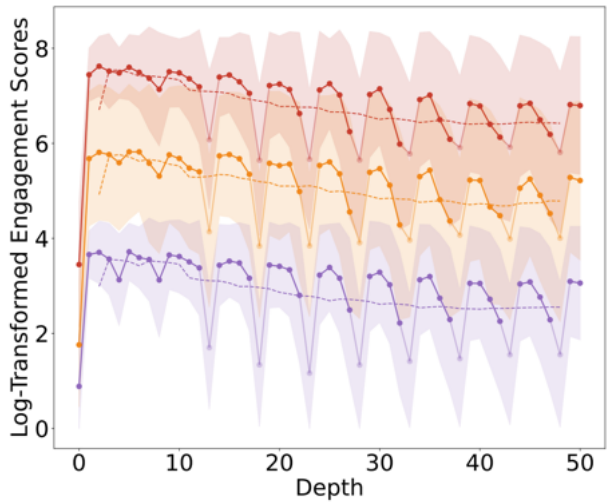
2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) 3-Second Watch Time



(b) 15-Second Watch Time



(c) 60-Second Watch Time

Fig. 7: Engagement Metrics of Short-Form Videos

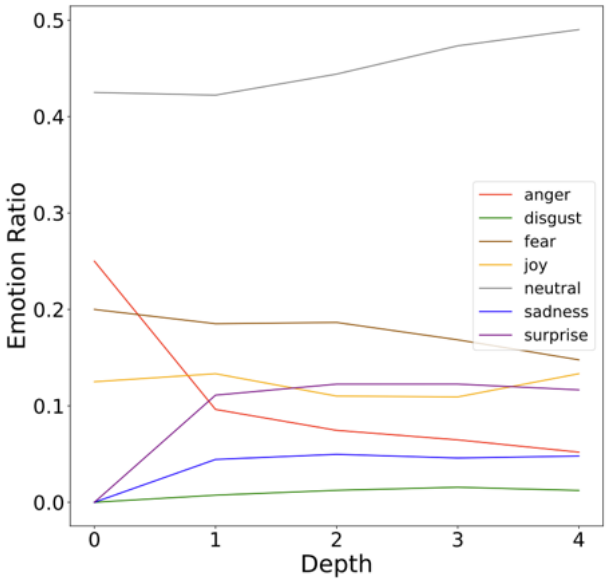
D. Toxicity Analysis

Toxicity analysis, although it might sound similar to emotionality analysis, focuses primarily on the aspects of social media that constitute harassment, divisiveness, or other forms of serious negativity. This type of behavior did not, and will generally not, show distinctly in categorical emotionality analysis. Furthermore, toxicity can draw attention, be infectious, and therefore be interconnected with the results of engagement. Our hypothesis going into this analysis was that the algorithm would prefer to recommend increasingly toxic content to the users. In this section, we have outlined the toxicity analysis results and drawn decisive conclusions from them.

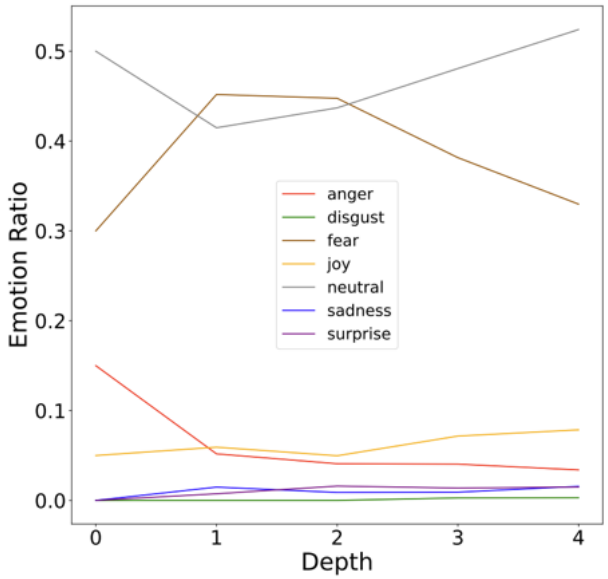
After analyzing the results of the toxicity levels across the long-form recommendation depths in Figure 10, it is immediately noticeable that the mean toxicity of all the videos—both the root videos and the recommended ones—was so low that it seems to be zero. However, when looking at the mean toxicity of the more toxic videos, there was a clear increase in the toxicity in the data extracted from the videos themselves (i.e., the title, description, and transcription). On the other hand, there was a consistently high amount of toxicity found in the video's comments, regardless of the depth of recommendation. These results show that the comments of YouTube videos were almost always toxic. Furthermore, the recommendation algorithm, as hypothesized, preferred to show increasingly toxic videos, regardless of its possible effect on the users, in order to maximize engagement.

Looking at the results of the toxicity analysis in the YouTube Shorts recommendations in Figure 11, it is immediately evident that, when considering all three charts as representing

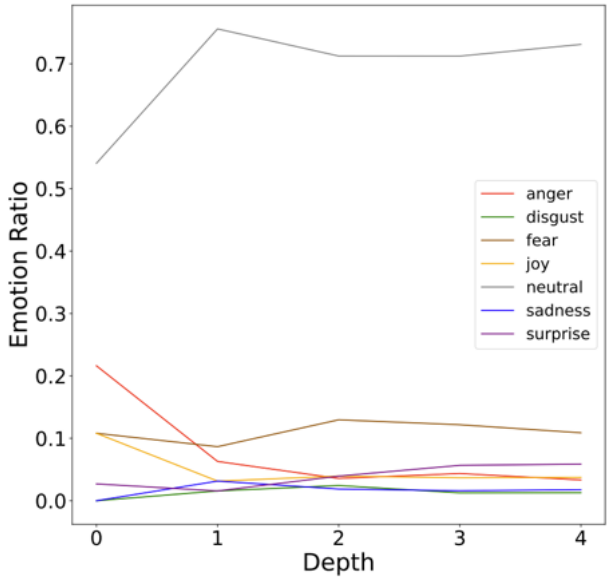
2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



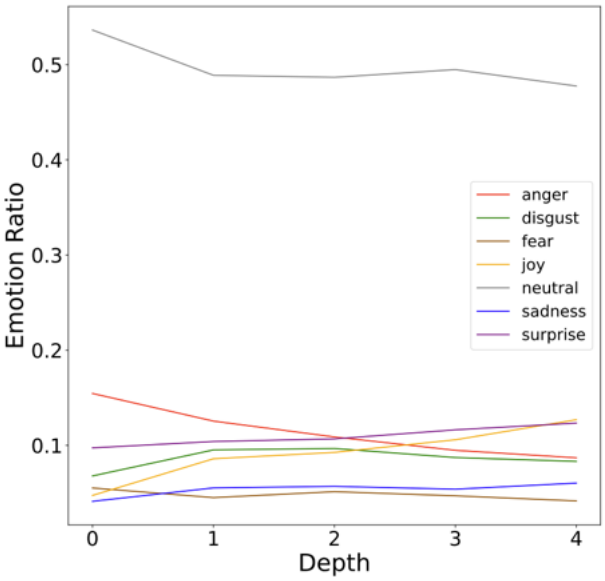
(a) Emotions in the Title



(b) Emotions in the Description



(c) Emotions in the Transcription



(d) Emotions in the Comments

Fig. 8: Emotional Analysis on Long-Form Videos

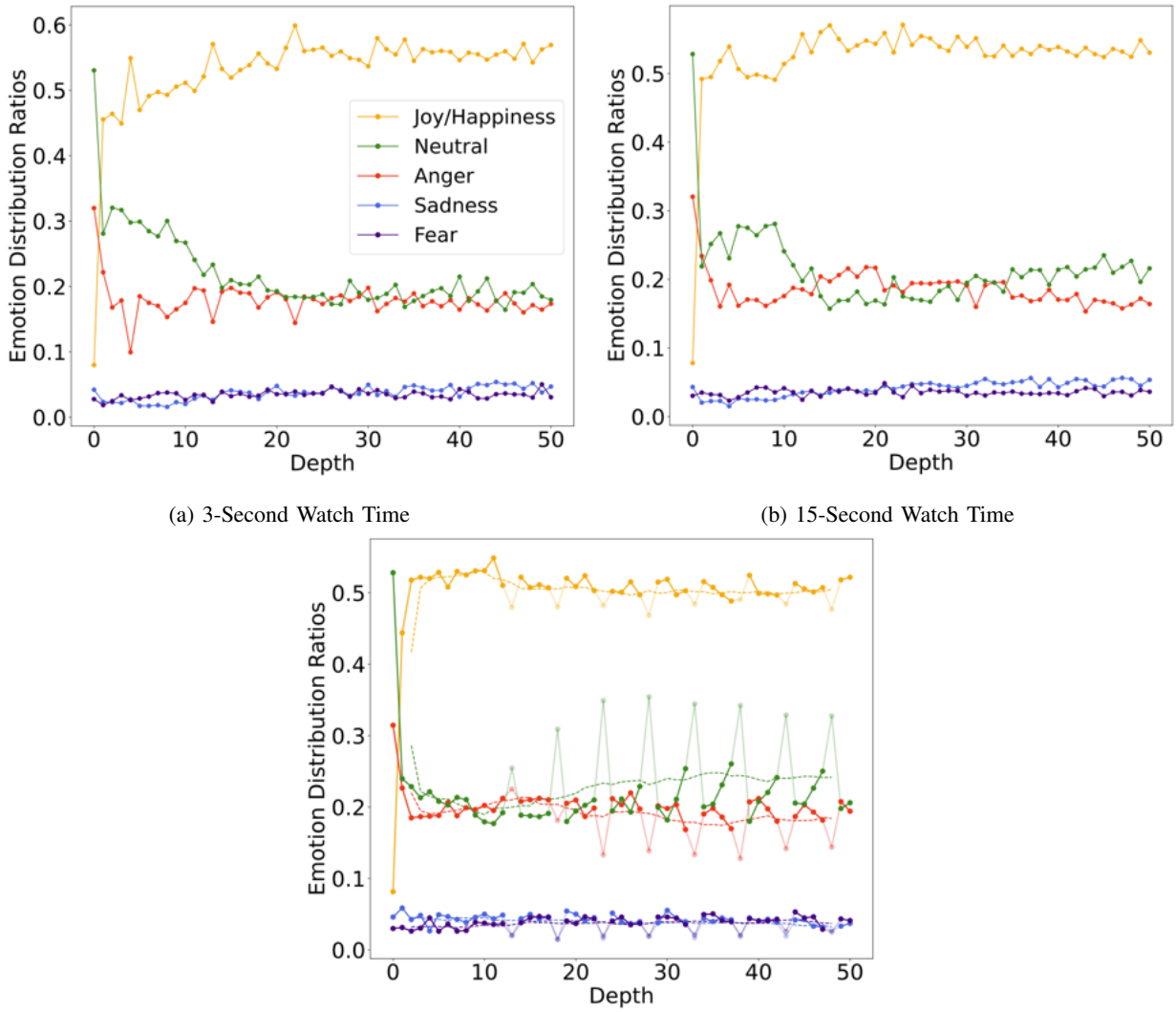
the three different wait times, there was a consistent increase in toxicity throughout the recommendation depths. This, however, is the mean toxicity. When taking into consideration the scale of the results, by paying close attention to the values in the y-axis, it becomes clear that the average toxicity of all of the videos was close to zero. Since these charts are more closely zoomed in, the increase becomes more noticeable. This is similarly caused by the fact that there are outliers in the recommendations with much higher than average toxicity. This shows the willingness of the algorithm to recommend toxic videos to increase user engagement.

Furthermore, focusing on the sub-figure Figure 11c, we saw that the advertisement spikes in the previous two results of the previous analyses shows up here as well. It can be inferred from these results that the toxicity of the advertisements was significantly lower than the videos the recommendation algorithm would otherwise recommend.

VI. Conclusion and Discussion

In this research paper, we have successfully showed that collecting recommendation data on both regular YouTube videos and YouTube Shorts was possible. Not only did we develop a method that made this possible, but we improved upon it to make it practically viable enough to collect data for generating decisive results and analysis. This shows that,

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) 3-Second Watch Time

(b) 15-Second Watch Time

(c) 60-Second Watch Time

Fig. 9: Emotional Analysis on Short-Form Videos

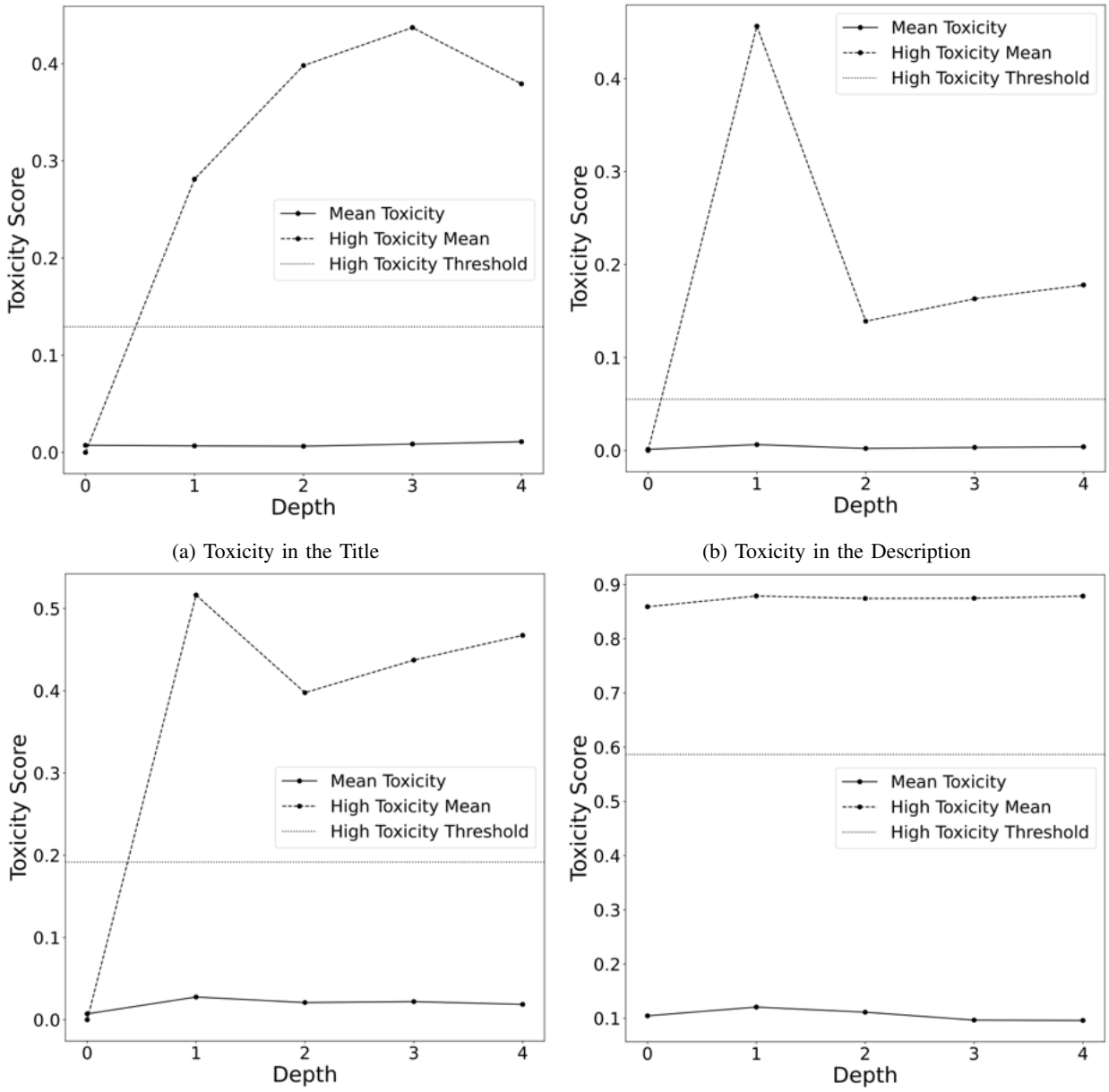
even if there are some roadblocks in the path, it is possible to devise unconventional solutions to overcome those challenges.

Furthermore, the timed results of multiprocessing applications, when compared to the base data collection methods for both long- and short-form content, shows that multiprocessing can be used to improve the processing time of light computational tasks by almost a factor equaling the number of processes used. The results further show that these improvements taper off at a certain point, and that that taper off point can be further improved upon by implementing clever solutions to reduce the computational demand of individual processes.

This research also successfully compares the recommendation algorithms of regular YouTube videos and YouTube Shorts. The overarching conclusion of the analysis is that both of the recommendation algorithms push the users toward more engaging videos. Moreover, both algorithms preferred less angry and more joyful videos in their recommendations. They both also tended to recommend content with higher toxicity, seemingly to increase engagement. Additionally, the recommendation algorithm for YouTube Shorts recommended higher engaging videos much earlier and stronger than the algorithm for regular YouTube videos did. This faster deviation to high engagement videos—combined with the fact that recommendation in short-form content is a much stronger factor in deciding what users are exposed to—makes the recommendation algorithm for short-form content much more dangerous and biased.

There were many successful conclusions drawn from the data, and there were also a few lessons learned. One of these lessons was to more carefully filter out advertisements when collecting recommendation data in the future. Nevertheless, this oversight helped show the engagement, emotion, and toxicity levels affected by YouTube Shorts advertisements. It

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) Toxicity in the Title

(b) Toxicity in the Description

(c) Toxicity in the Transcription

(d) Toxicity in the Comments

Fig. 10: Toxicity Analysis on Long-Form Videos

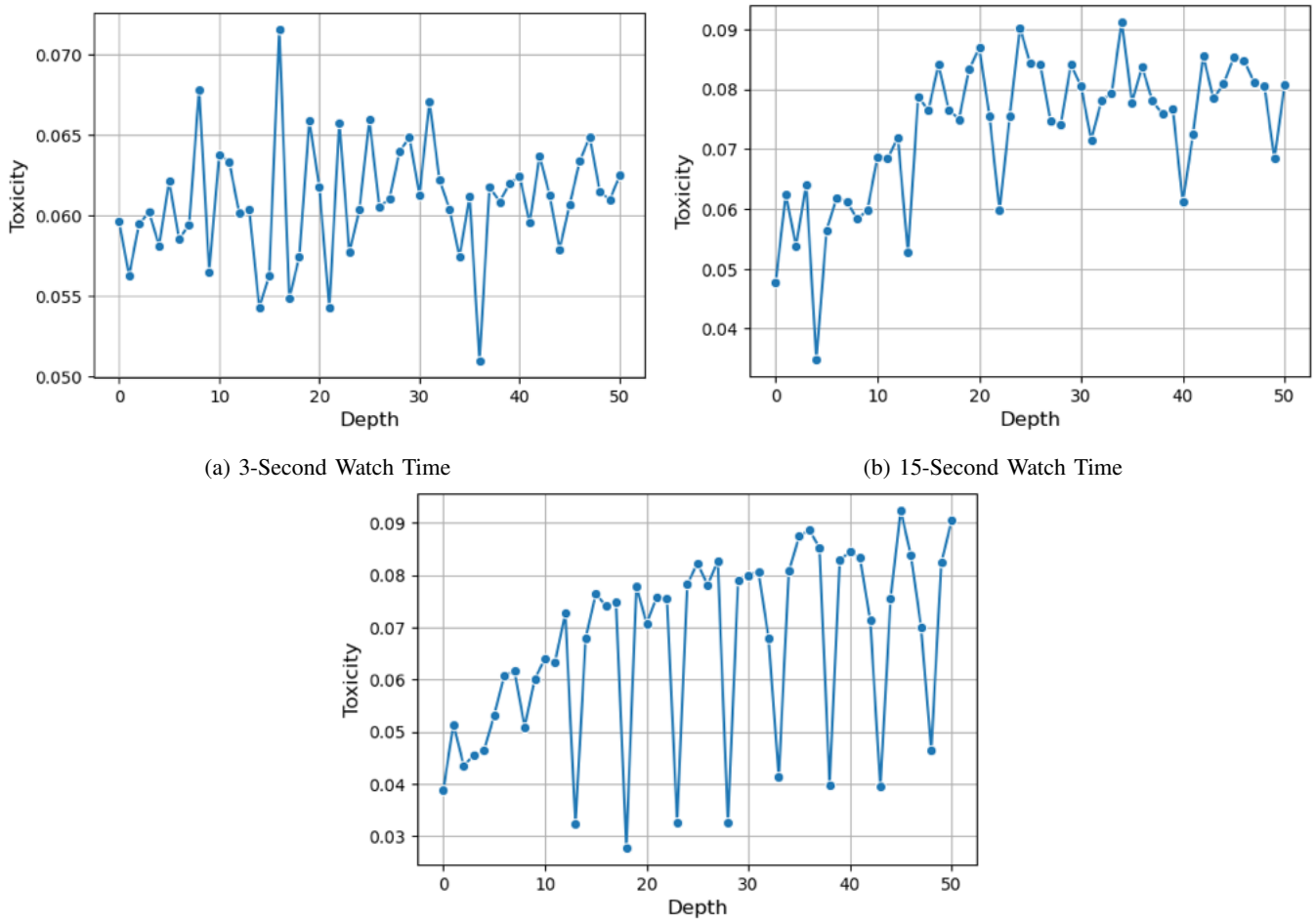
also showed us that advertisements show up periodically and, more specifically, once every five videos. This observation is helpful for future researchers attempting to collect YouTube Shorts data. They can avoid collecting advertisements unless that is a part of the research experiment. For our study, once the advertisements were discovered, the results were analyzed by factoring out the advertisements' engagement, emotion, and toxicity, thereby removing the risk of result contamination by advertisements.

After conducting this study, it was also noticed that reacting with the content, such as liking, disliking, or commenting, could affect the recommendations. Hence, this research will be extended in the future by reacting positively with relevant content and reacting negatively with irrelevant content to measure if the algorithm picks up on user engagement. This study could additionally be extended by performing similar data collection and analysis methods on different platforms with short-form content to see if the recommendation algorithm in those platforms also behaves in a similar fashion. Future research would also consider collecting data from different parts of the world to evaluate global vs. regional bias patterns.

Acknowledgements

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) 3-Second Watch Time

(b) 15-Second Watch Time

(c) 60-Second Watch Time

Fig. 11: Toxicity Analysis on Short-Form Videos

Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078, W911NF-25-1-0147), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency, the Australian Department of Defense Strategic Policy Grants Program, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment, and the Donaghey Foundation at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

[1] N. Agarwal, "From Metadata to Meaning: GPT-4 Reveals Bias Trends in YouTube," in *Proc. 1st IARIA Int. Conf. AI-based Media Innovation (AIMEDIA)*, Venice, Italy, Jul. 2025.

[2] Y. Deldjoo, D. Jannach, A. Bellogin, A. Difonzo, and D. Zanzonelli, "Fairness in recommender systems: research landscape and future directions," *User Modeling and User-Adapted Interaction*, vol. 34, no. 1, pp. 59–108, 2024.

[3] P. T. Rajendran, K. Creusy, and V. Garnes, "Shorts on the rise: Assessing the effects of youtube shorts on long-form video content," *arXiv preprint arXiv:2402.18208*, 2024.

[4] S. Wang, X. Zhang, Y. Wang, and F. Ricci, "Trustworthy recommender systems," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 4, pp. 1–20, 2024.

[5] Z. Zhao, W. Fan, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang *et al.*, "Recommender systems in the era of large language models (llms)," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[6] A. Chubb, "PRC Assertiveness in the South China Sea: Measuring Continuity and Change, 1970–2015," *International Security*, vol. 45, no. 3, pp. 79–121, 01 2021. [Online]. Available: https://doi.org/10.1162/isec_a_00400

[7] C. E. Macaraig and A. J. Fenton, "Analyzing the causes and effects of the South China Sea dispute," *The Journal of Territorial and Maritime Studies*, vol. 8, no. 2, pp. 42–58, 2021.

[8] M. T. Fravel, "China's strategy in the South China Sea," *Contemporary Southeast Asia*, pp. 292–319, 2011.

[9] R. Zhou, S. Khemmarat, and L. Gao, "The impact of youtube recommendation system on video views," in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, 2010, pp. 404–410.

[10] M. C. Cakmak, O. Okeke, U. Onyepunuka, B. Spann, and N. Agarwal, "Analyzing bias in recommender systems: A comprehensive evaluation of youtube's recommendation algorithm," in *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '23. New York, NY, USA:

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Association for Computing Machinery, 2024, p. 753–760. [Online]. Available: https://doi.org/10.1145/3625007.3627300

[11] E. Alp, B. Gergin, Y. A. Eraslan, M. C. Çakmak, and R. Alhajj, *Covid-19 and Vaccine Tweet Analysis*. Cham: Springer International Publishing, 2022, pp. 213–229. [Online]. Available: https://doi.org/10.1007/978-3-031-08242-9_9

[12] M. C. Cakmak and N. Agarwal, "Unpacking algorithmic bias in youtube shorts by analyzing thumbnails," in *Proceedings of the 58th Hawaii International Conference on System Sciences*, 2025, p. 2498. [Online]. Available: https://hdl.handle.net/10125/109144

[13] M. C. Cakmak, N. Agarwal, S. Dagtas, and D. Poudel, "Unveiling bias in youtube shorts: Analyzing thumbnail recommendations and topic dynamics," in *Social, Cultural, and Behavioral Modeling*, R. Thomson, A. Hariharan, S. Renshaw, S. Al-khateeb, A. Burger, P. Park, and A. Pyke, Eds. Cham: Springer Nature Switzerland, 2024, pp. 205–215.

[14] C. Violot, T. Elmas, I. Bilogrevic, and M. Humbert, "Shorts vs. regular videos on youtube: A comparative analysis of user engagement and content creation trends," in *Proceedings of the 16th ACM Web Science Conference*, 2024, pp. 213–223.

[15] M. C. Cakmak, N. Agarwal, D. Poudel, and S. Bhattacharya, "Examining the Impact of Symbolic Content on YouTube's Recommendation System," in *Proceedings of the 13th International Conference on Complex Networks and Their Applications*. Springer, 2024, pp. 421–432.

[16] B. Kitchens, S. L. Johnson, and P. Gray, "Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption." *MIS Quarterly*, vol. 44, no. 4, 2020.

[17] M. I. Gurung, M. M. I. Bhuiyan, A. Al-Taweel, and N. Agarwal, "Decoding youtube's recommendation system: A comparative study of metadata and gpt-4 extracted narratives," in *Companion Proceedings of the ACM Web Conference 2024*, ser. WWW '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1468–1472. [Online]. Available: https://doi.org/10.1145/3589335.3651913

[18] K. Hosanagar, D. Fleder, D. Lee, and A. Buja, "Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation," *Management Science*, vol. 60, no. 4, pp. 805–823, 2014.

[19] J. Treviranus and S. Hockema, "The value of the unpopular: Counteracting the popularity echo-chamber on the web," in *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*, 2009, pp. 603–608.

[20] C. Ballard, I. Goldstein, P. Mehta, G. Smothers, K. Take, V. Zhong, R. Greenstadt, T. Lauinger, and D. McCoy, "Conspiracy brokers: Understanding the monetization of youtube conspiracy theories," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2707–2718. [Online]. Available: https://doi.org/10.1145/3485447.3512142

[21] A. J. B. Chaney, B. M. Stewart, and B. E. Engelhardt, "How algorithmic confounding in recommendation systems increases homogeneity and decreases utility," in *Proceedings of the 12th ACM Conference on Recommender Systems*, ser. RecSys '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 224–232. [Online]. Available: https://doi.org/10.1145/3240323.3240370

[22] M. C. Cakmak, O. Okeke, U. Onyepunuka, B. Spann, and N. Agarwal, "Investigating Bias in YouTube Recommendations: Emotion, Morality, and Network Dynamics in China-Uyghur Content," in *International Conference on Complex Networks and Their Applications*. Springer, 2023, pp. 351–362.

[23] V. Su and N. Thakur, "Covid-19 on youtube: A data-driven analysis of sentiment, toxicity, and content recommendations," *arXiv preprint arXiv:2412.17180*, 2024.

[24] M. L. Khan, "Social media engagement: What motivates user participation and consumption on youtube?" *Computers in Human Behavior*, vol. 66, pp. 236–247, 2017.

[25] W. Hoiles, A. Aprem, and V. Krishnamurthy, "Engagement and popularity dynamics of youtube videos and sensitivity to meta-data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, pp. 1426–1437, 2017.

[26] D. Poudel, M. C. Cakmak, and N. Agarwal, "Beyond the click: How youtube thumbnails shape user interaction and algorithmic recommendations," in *The 16th International Conference on Advances in Social Networks Analysis and Mining (ASONAM), accepted for presentation*, 2024.

[27] M. C. Cakmak, M. Shaik, and N. Agarwal, "Emotion assessment of youtube videos using color theory," in *Proceedings of the 9th International Conference on Multimedia and Image Processing (ICMIP 2024)*, ser. ICMIP '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 6–14. [Online]. Available: https://doi.org/10.1145/3665026.3665028

[28] M. Shaik, M. C. Cakmak, B. Spann, and N. Agarwal, "Characterizing multimedia adoption and its role on mobilization in social movements," in *57th Hawaii International Conference on System Sciences, HICSS 2024, Hilton Hawaiian Village Waikiki Beach Resort, Hawaii, USA, January 3-6, 2024*, T. X. Bui, Ed. ScholarSpace, 2024, pp. 146–155. [Online]. Available: https://hdl.handle.net/10125/106393

[29] N. Yousefi, M. C. Cakmak, and N. Agarwal, "Examining multimodel emotion assessment and resonance with audience on youtube," in *Proceedings of the 2024 9th International Conference on Multimedia and Image Processing (ICMIP)*, 2024, pp. 85–93.

[30] O. Okeke, M. C. Cakmak, B. Spann, and N. Agarwal, "Examining Content and Emotion Bias in YouTube's Recommendation Algorithm," in *The Proceedings of the Ninth International Conference on Human and Social Analytics (HUSO 2023)*. Barcelona, Spain: Copyright (c) IARIA, 2023, Mar. 2023, pp. 15–20. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=huso_2023_1_40_80032

[31] M. C. Cakmak, O. Okeke, B. Spann, and N. Agarwal, "Adopting parallel processing for rapid generation of transcripts in multimedia-rich online information environment," in *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2023, pp. 832–837.

[32] M. C. Cakmak and N. Agarwal, "High-speed transcript collection on multimedia platforms: Advancing social media research through parallel processing," in *2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2024, accepted for presentation.

[33] M. C. Cakmak, N. Agarwal, and R. Oni, "The bias beneath: analyzing drift in youtube's algorithmic recommendations," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 171, 2024.

[34] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz *et al.*, "Chatgpt: Jack of all trades, master of none," *Information Fusion*, vol. 99, p. 101861, 2023.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach. arxiv [preprint](2019)," *arXiv preprint arXiv:1907.11692*, 2019.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Metrics, not Myths: The True Price and Value of Productivity in the Era of AI

A systematic framework to determine AI-induced productivity gains with regard to the "AI paradox"

Sibylle Kunz, Claudia Hess
IT & Engineering
IU International University of Applied Sciences
Erfurt, Germany
e-mail: sibylle.kunz@iu.org; claudia.hess@iu.org

Abstract— **Many organizations tout productivity gains when deploying Analytical and Generative Artificial Intelligence (AI), yet few offer structured methods to quantify real value. This article presents a framework to classify use cases by time and cost savings or quality improvements, and links productivity to profitability for calculating measurable outcomes. Beyond headline gains, the article uncovers hidden costs—such as the need to validate, verify, and re-work AI outputs—that can erode expected benefits and impact true efficiency. A comprehensive Total-Cost-of-Ownership (TCO) approach addresses not only acquisition and operation of AI systems, but also the ongoing labor associated with refining AI-generated results. Comparing both overt and covert costs against claimed benefits, the framework enables more accurate estimations of AI's impact. Demonstrated with examples from diverse industries, the article provides practical guidance for more transparent assessment of AI projects and their real-world returns.**

Keywords-artificial intelligence; productivity; TCO; hidden costs; AI paradox.

I. INTRODUCTION

With the increasing use of Artificial Intelligence (AI), the number of publications and press releases proclaiming the economic success of this technology is also rising, often without concrete evidence or verifiable figures. In many cases, only percentage improvements are estimated. This makes it difficult to compare use cases and identify best practices, as Kunz and Hess [1] explain. To address this issue, they propose a framework for assessing the benefits and costs associated with implementing AI solutions.

Productivity gains resulting from analytical AI, such as in medical diagnosis or fraud detection, have been discussed for several years now. AI can solve complex problems that would otherwise need much more time, resources or would be intractable due to the sheer volume of data involved. Since the introduction of Open AI's Large Language Model (LLM) GPT3.5 to the general public in November 2022 and the emergence of many competing models, the number of news articles and publications that postulate even higher expectations concerning the use of Generative AI (GenAI) in the form of LLMs has multiplied.

For example, a McKinsey study conducted in 2023 [2] proclaimed an estimated world-wide productivity gain induced by GenAI of 2.6 to 4.4 billion dollars and a rise in working productivity of 0.1% to 0.6% per year. According to this study, branches benefitting the most from GenAI will be finance, high-tech, media and bioscience. 75% of this potential can be found in the field of customer service, sales and distribution, software development and research & development. The study evaluated 850 different jobs and 2.100 different job tasks. The study also distinguished between three clusters of task types: Physical work (foreseeable or unforeseeable) with productivity gains of 70% rsp. 34% (slightly more than using analytic AI only), data collection (79%, 65% by analytical AI only) and data management (92%, 75% by analytical AI only) and decision making and collaboration (management, stakeholder communication and interaction, knowledge application) with an increase of 50-55% in productivity – almost three times as much as with analytical AI only. That the amount of productivity gains seems correlated with the level of education comes as a surprise – employees with lower qualifications saw less possibilities to raise their productivity.

There are further national and international studies that aim to quantify the economic effects of AI and its impact on functions and job profiles. In 2024, a study conducted by ifo among German companies [3] revealed that almost 84% of them expected productivity gains for the national economy within five years, estimating an average increase of 12%. But 70% of all managers predicted productivity gains of averagely 8% for their own company, thereby estimating lower values for their company than nationwide.

Hammermann et al. [4] report that 45% of all employees in the 815 companies included in their study, who have used AI in their daily work between 2022 and 2024, claim productivity gains in their own job. On the other hand, 15% of the employees using AI stated the opposite.

Demary et al. [5] arrive at a moderate assessment: their study conducted for the Institute of the German Economy (IW) estimates a rise of the gross domestic product by 0.9% that can be derived from AI-usage between 2025 and 2030, so it does not act as a strong growth driver. For the decade starting in 2030, an increase of 1.2% is predicted. Demary et al. see AI as complementary to human work. They also raise the question whether rises in productivity can be accomplished by AI alone or only if AI is flanked by other digitalization technologies like robotics, software, internet access etc.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

The saving of processing time is often cited as a reason for productivity gains. Looking at these ambitious expectations, AI seems to be a must-have for companies to stay competitive. However, AI solutions and their integration in a company's way of working often come with considerable costs and sometimes ethical implications. Before making decisions on investment, organizations must be able to assess the expected productivity gains and construct a fact-based calculation to make sure that AI-investments will be amortized and do not cause any uncontrollable risks, especially with regard to reductions in the workforce as a consequence of productivity gains. It seems insufficient to just ask employees about their general personal impression of the extent to which productivity has increased when using (generative) AI, especially when it is unclear whether this can be attributed to the use of AI alone.

According to the 29th CEO survey of PriceWaterhouse Coopers [6], only 12% of all enterprises using AI manage to realize both cost savings and rising revenues at the same time. PWC interviewed 4,454 CEOs in 95 countries. Many CEOs could not see direct measurable results from their AI activities. 38% of all US companies reports lower costs, 51% of all Chinese enterprises said they had increased their revenue. In Germany, 16% saw lower costs, 11% stated revenue increases. But all in all, more than half of the respondents saw no identifiable benefits at all.

Meta, Facebook's parent company, goes one step further: a new performance tracking tool called "Checkpoint" is used to link employee performance reviews to AI tool usage at work [7]. The management wants staff to integrate artificial intelligence into daily tasks, aiming for greater efficiency and innovation. Guidelines for using AI across different functions have been circulated, and employees are expected to automate routines, support decision-making, and enhance content creation via these tools. Performance evaluations will partially reflect each employee's AI adoption, marking a cultural shift within the company and promising significant bonuses for successful AI adoption.

All of these results are in line with the so-called "AI Productivity Paradox" highlighted by Brynjolfsson et al. [8], which observes that despite widespread adoption of AI technologies, measurable productivity improvements often lag behind investments and expectations. Brynjolfsson and co-authors emphasize that organizations frequently struggle to translate AI capabilities into tangible business gains, partly due to challenges in integration, organizational change, and the need for complementary innovations. As the PWC survey shows, the majority of enterprises are yet to realize clear benefits from AI, reflecting this paradox—the promise of AI is substantial, but the practical gains remain elusive for most firms. This affirms Brynjolfsson's "Productivity J-Curve" [9], stating that productivity does not immediately rise. Instead, it may first stall or even decline before increasing later. This is because early investments in general purpose technologies generate intangible assets (like new business processes, workforce training, and organizational redesign) that are not captured in conventional productivity metrics.

To reduce the ambiguity of statements about savings and revenue increases, a framework is needed which allows empirical AI use cases to be classified, their outcomes measured and compared. Therefore, the research question is: what kind of values can be measured and in which dimensions?

This article compares measurement dimensions and reference values for different use cases and develops a systematic framework for calculating realistic economic effects of AI usage. Instead of calculating each new AI business case from scratch, it might be helpful to have standard categories, benchmarks and a set of indicators that can be measured.

To provide a solid foundation, the concepts of productivity and profitability are clarified and differentiated in Section II. Section III shows the dimensions for measuring AI-induced productivity and profitability gains related to time, cost and quality. In Section IV, some examples and case study projects are presented to illustrate which productivity gains have already been proven and how they have been calculated. Section V describes a Total Cost of Ownership (TCO)-based approach to calculate investments needed to plan, train, implement and run an AI system. Section VI focuses on the "hidden costs of AI", resulting in an enhancement to the TCO approach and discussing the "Junior Paradox". Section VII sums up both sides of the equation: cost structure versus profitability to sketch a framework showing how to develop benchmarks in the future. Finally, Section VIII draws a conclusion and offers suggestions for future work.

II. Productivity and Profitability

Due to Thommen et al. [10], productivity is defined as the quantitative relationship ratio between output and input of the production process (1):

$$\text{Productivity} = \frac{\text{Work Outcome}}{\text{Input Quantity of production factors}} \quad (1)$$

Since it is impossible to measure productivity for a whole enterprise at once, partial productivities are calculated related to work hours, machine running times or area sizes. For example, an accountant could execute 20 bookings per hour, a punching machine could produce 1.000 parts per hour, or a toy shop could achieve a turnover of 1.000 € per square meter. If the usage of AI is somehow beneficial to the company, these partial productivities can be expected to rise.

This effect can be direct or indirect: An increase in productivity per area means better utilization of limited resources and thus a direct increase in sales. Conceivable areas of application include stationary retail or agricultural production. Conversely more output from a machine or worker per unit of time means - assuming investment or labor costs remain the same (considered here as fixed costs) - that less machine running time or labor is now required for the same result, indirectly reducing costs.

Productivity is an output-oriented concept measured in physical units like pieces or amounts. Since productivity figures can usually also be expressed in monetary units, the

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

concept of profitability (or economic viability) can be used as a substitute for productivity in the planned framework.

Profitability can be defined in different ways. The following equation assumes a direct relationship to productivity via (2):

$$\text{Profitability} = \frac{\text{Productivity x Revenue per Unit} - \text{Total Expenditures}}{\text{Revenue}} * 100\% \quad (2)$$

Alternatively, the concepts of the net profit ratio (after deduction of all taxes) (3)

$$\text{Net Profit Ratio (\%)} = \frac{\text{Net Income}}{\text{Revenue}} \quad (3)$$

or the gross margin ratio (after deduction of all direct costs) (4)

$$\text{Gross Margin Ratio (\%)} = \frac{\text{Gross Profit}}{\text{Net Revenue}} \quad (4)$$

can be used.

The central question is: to what extent can the use of AI reliably contribute to the productivity gains predicted by recent studies, and which investments in AI technologies are necessary to leverage this potential? Although there are currently many studies that attempt to quantify the expectations of AI use in this regard, it is difficult to verify whether these expectations will actually materialize. There is a lack of clear assignability and classification of measurement methods. The following section will therefore examine which dimensions and characteristics are suitable for operationalization.

III. Dimensions for Measuring AI-Induced Productivity and Profitability Gains

AI can be classified into different fields based on its purpose, underlying methodologies, and applications. For our framework, we will distinguish between Analytical AI, Generative AI, and Reactive AI.

- Analytical AI focuses on data-driven decision-making, pattern recognition, and predictive analytics. It processes structured and unstructured data, extracts insights, and assists in optimization and forecasting without autonomously generating new creative content. It can be used in the fields of medical diagnosis, fraud detection, predictive maintenance or algorithmic trading as well as for natural language recognition, sentiment analysis, etc.
- Generative AI is designed to create new, synthetic content, such as text, images, music, or videos, by learning patterns from existing data. It uses advanced models like Generative Adversarial Networks (GANs) that allow the creation of videos and Transformer-based architectures like GPT4.x and other Large Language Models for natural language generation.
- Reactive AI operates based on real-time inputs and predefined rules, without memory or learning from past experiences at runtime. It was mainly used for fast-response, rule-based systems like chess-playing systems, such as “Deep Blue”, or older AI-powered, but rule-based chatbots.

Today, especially the methods in Analytical and Generative AI can be used to leverage potentials for productivity - and therefore - profitability gains. There are three dimensions to be improved: time, cost and quality. Table I shows the relation between these dimensions and whether Analytical (A) or Generative (G) AI is used. It lists examples for measures that can be used to calculate AI-related profitability increases.

TABLE I. Measuring Dimensions for AI Productivity

Dimen-sion	Context/Scenarios	AI class
Time	Planning time Project planning, Product planning, Logistic optimization (e.g., airports, freight forwarders, harbors, railways)	A+G
	Design time Product design, Service design, Individualization of consumer goods, Developing protein structures for pharmaceutical or chemical applications (e.g., AlphaFold), Developing recipes	A+G
	Production time of physical goods (time needed to produce one piece or unit)	A
	Production time of immaterial artefacts Creation of text, audio, video, e.g., in journalism, marketing, consulting, arts, Creation of program code	G
	Testing time Creation of test cases (e.g., software testing), Automatic test execution and evaluation	A+G
	Delivery time Demand forecasting, Route optimization	A
	Support time Analyze service requests by natural language recognition, Speech-to-Text, Answer customer requests, Analyze customer feedback	G
Cost	Material usage: production factors like raw materials, supplies and energy	A
	Waste, Offcut: raw materials	A
	Required space: inventory optimization	A
Quality	Quality control and error detection (checking products, processes, or results for deviations, errors, anomalies, and risks, e.g., automatic anomaly detection in production processes, medical diagnosis, e.g., skin cancer, tumor detection in X-rays or MRIs)	A+G
	Correction and quality improvement (adjustment or optimization of existing artifacts to meet defined quality standards, e.g., proofreading of texts or translations)	A+G
	Quality of results and performance (AI elevates the quality of the end result or human performance beyond what was originally possible, e.g., easier readability, improved accessibility of a website, personalized content)	A+G

With regard to GenAI, it needs to be pointed out that time and cost might tend in a different direction than quality, i.e., gains in the first two dimensions might cause reductions in

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

the third. Consider this example: It is often stated that GenAI helps save a lot of time in producing text and illustrations. Employees working with text, such as journalists or marketers, can produce results in a shorter time or reduce the effort necessary to check translations or edit articles. And this means that one person can create a larger amount of output (i.e., text or graphic elements) in a given time, therefore getting more things done for the same salary. But in this case, productivity gains are more difficult to estimate, since the quality or originality of results is also important for the artefacts. Just speeding things up might lead to counterproductive effects in the long run. This effect is discussed in more detail in Section VI.

IV. Case Study Examples for Successfully Measuring Productivity Gains

The two following tables show some examples for case studies in the fields of Analytical and Generative AI that documented concrete absolute or relative values for productivity gains. In most cases, percentages or absolute values for savings were mentioned, but none of them listed any data on the cost side as described in Section V. Table II lists projects using mainly analytical AI.

TABLE II. Case Studies Using Analytical AI to Increase Profitability or Save Costs.

Company and branch	**What was measured?**	**Scope**	**Relative or absolute change in productivity / reported savings**
Salling Group, Energy consulting [11]	Cost: Energy consumption in supermarket buildings via smart meters or data from energy providers	AI system analyses weather data and energy consumption and optimizes usage of device during closing hours	700 supermarkets were evaluated, savings in the millions are reported
Municipality of Holstebro, Denmark [12]	Cost: Energy consumption in community buildings	AI system analyses weather data and energy consumption and optimizes usage of device during closing hours	Savings: 1 million DKK = ca. 146.000 $
SWMS Systemtechnik Ingenieur-gesellschaft mbH [13]	Cost: Automated production of composite materials using a printing robot arm: usage of 3D printing materials	Reduction of printing errors through AI-supported monitoring: image-based object recognition and segmentation.	Material savings of 1/3 (estimated), Savings in energy for robot and cooling of printed artefacts
Orthopedical insoles [14]	Cost: Material usage in 3D printing instead of insole construction using blanks	AI system calculates ideal form to realize material saving	Material savings: >70% in plastic materials, up to 60% in energy
FRAPORT AG, Aviation [15]	Time: staff are assigned to ground handling based on qualification and availability	AI system IDA simulates and optimizes staff planning	(no data yet, project in beta-status)

Table III shows some examples of projects using GenAI. Productivity gains here often mean that standard tasks can be automated, so that the employee gets more time for more difficult tasks.

TABLE III. Case Studies Using Generative AI to Increase Profitability or Quality or Save Costs.

Company & Branch	**Scope**	**What was measured?**	**Change in productivity / reported savings**
heise (IT magazine publisher, online portal provider) [16], [17]	System heiseIO (based on LLM and a process-oriented approach with predefined prompts)	a) Time: Annotation time of 14.000 pictures in a CMS b) Time: Production time of a newsletter	a) 5-15min human-based annotation time saved per picture, total cost of 300 € Daily “Botti” newsletter can be produced 12 minutes faster, saving a total of 1,5 person days per moth
Fieldcode GmbH [18]	System Fieldcode (based on LLM) planning of field service assignments	Time: Ticket diagnosis: analysis of field service requests to solve service problems remotely instead of sending a service technician, optimize “first fix rate” Cost: avoid unnecessary order of spare parts, avoid unnecessary travel cost, fuel etc.	Up to 50% of spare parts could be saved, Rise of First fix rate (no exact number given)
Klarna [19]	System Kiki (based on LLM) used	Time, Quality: The system answers up to	Contracts can be generated in 10 minutes

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

	for internal knowledge management	2.000 questions from employees per day and is used by 85% of all staff	instead of 1 hour Chatbot performs 2,3 million chats with customers (equals work of 700 employees) Estimated profit increase: 40 million $
Boston Consulting Group [20]	General usage of AI	Time, Quality	12% more tasks accomplished, 25% savings in time 40% increase in quality

The two tables show that the case studies can easily be categorized in terms of the dimensions mentioned in Section III. Table IV shows an overview of all case studies mapped to these dimensions.

TABLE IV. ASSIGNMENT OF CASE STUDIES TO MEASURING DIMENSIONS AND AI CLASS

Case Study	**Dimension**	**AI class**
Salling Group	Cost	A
Municipality of Holstebro	Cost	A
SWMS Systemtechnik Ingenieur-gesellschaft mbH	Cost	A
Orthopedical insoles	Cost	A
Fraport	Time	A
Heise	Time	G
Fieldcode GmbH	Time, Cost	G
Klarna	Time, Quality	G
Consulting	Time, Quality	G

The next section discusses how different AI-system investments can be made comparable using a lifecycle.

V. COST OF AI-USAGE: A TOTAL COST OF OWNERSHIP-APPROACH FOR AI-SYSTEMS

Before productivity gains can be realized, AI-based systems need to be developed, trained and fine-tuned. To get an overall picture of the total cost of ownership (TCO) of an AI system, the following cost components need to be considered. They can be organized along the life cycle of an AI system, using a TCO-like approach with lifecycle phases and corresponding tasks as shown in Fig. 1.

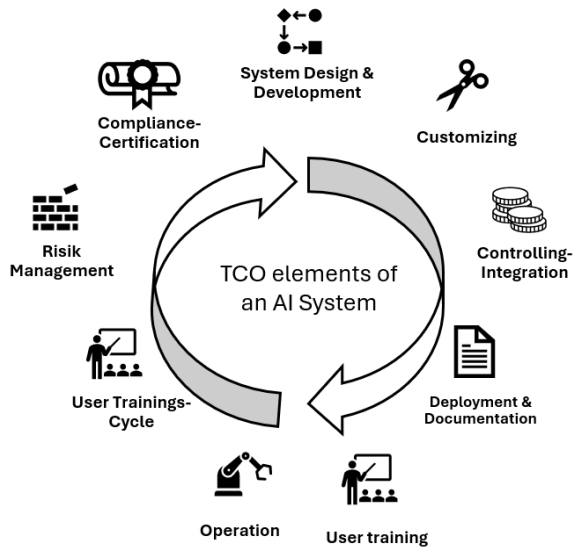


Figure 1. TCO Elements of an AI System

The following cost components need to be evaluated:

A. System Design and Development

All AI systems need to be modelled and trained. This can either be done using supervised or unsupervised learning, reinforced and/or deep learning. Developing these systems requires a large amount of storage and compute resources like GPUs or other AI-chips. Training data needs to be collected or artificially generated, and the data needs to be cleaned and consolidated. Interfaces to control and monitor the settings are needed and the system may have to be integrated into existing processes. This results in a significant fix-cost block before installing the final system. In the case of prefabricated AI models, this cost block will be covered by later subscription fees.

The cost determinants in this phase are especially the programming time, data engineering time (calculated via salary), on-premises hardware or cloud cost for CPU/GPU time, and energy cost.

B. Customizing

Especially in the case of GenAI, the resulting models need to be finetuned to perform specific tasks or to improve their performance in a particular domain. Guard rails need to be developed to ensure responsible and ethical AI use.

The cost determinants are salary (programmers and domain experts), acquisition of domain-specific data sets, and computing resources.

C. Integration into Controlling

Key Performance Indicators (KPIs) and performance metrics need to be developed in order to monitor the system and to assess its performance. This is a prerequisite to ensure that the AI system meets the desired outcomes

The main cost determinant is the time required to choose and to agree upon KPIs and appropriate metrics with AI engineers, domain experts and executives being involved.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

D. Deployment and Documentation

The AI system needs to be deployed within the existing infrastructure, which involves technical setup, integration with systems and platforms already in use, and testing. Documentation facilitates maintenance, troubleshooting and further development.

The cost determinant is the time spent by the technical team.

E. User training

Users need to be initially trained in how to use the AI systems (e.g., required by the EU AI act). The training ensures employees understand AI opportunities, risks, and legal compliance requirements. The curriculum also depends on what tasks the employees are assigned to, e.g., whether they work in the IT department, in a dedicated AI team, human resources etc.

The cost determinants are salary, course fees, course material, travelling and accommodation costs.

F. Operation

Each request to the running system causes a certain amount of costs for inferencing, i.e., producing a solution or an answer. These costs are often covered using subscription fees like user licenses for LLMs on a monthly or per-token basis. For critical systems, human workers need to be kept in the loop to meet ethical requirements.

The cost determinants are subscription fees (fix costs on monthly/annual basis), token consumption (variable cost), salary.

G. User Training Cycle

After certain intervals, these training courses need to be repeated to keep users up-to-date and refresh their knowledge.

The cost determinants are salary, course fees, course material costs, travel and accommodation costs.

H. Risk Management

Installing AI systems in certain processes might also require additional insurance, e.g., to protect against potential damage caused by AI system failures or monitoring frameworks to assess and mitigate associated risks.

The cost determinants are salary and insurance premiums.

I. Certification for Compliance

Certifications might be required regarding compliance with legal and regulatory standards, which can vary by industry and region (e.g., the AI Act by the European Union). They must be renewed in prescribed intervals. Renewal processes typically involve reassessment and auditing of systems to confirm that they still meet the necessary requirements.

The cost determinants are certification fees on an annual or long-term basis and salaries for the internal and external experts involved in the certification.

This structure can be used to calculate a concrete AI project, resulting in the estimated total costs of the system. It can further be used to calculate the time needed to amortize.

But there is another aspect not directly related to the application lifecycle, but to the quality of AI-based output, which is discussed in the following section.

VI. THE "HIDDEN COST" OF AI

If increases in productivity or profitability have been documented and proven using this framework, a second consideration is needed. Does the output quality match the original expectations? The effort required to ensure this quality introduces a crucial, often underestimated, or even overlooked layer of human work.

A systematic **AI Output Rework** process is the critical quality gate that transforms a promising AI draft into a trustworthy and valuable business asset. This section presents a formal framework for this human-led validation, outlining the essential activities needed to mitigate risks and ensure that all AI outputs meet the highest standards of quality and integrity.

A. The AI Productivity Paradox Revisited

AI output rework is especially important for GenAI-generated content like documentation, communication with customers, suppliers and stakeholders or program code. This quality inspection and the potential rework effort require time and, as a result, increase staff costs.

A recent study conducted and published in January 2026 by Workday [21] showed that 77% of all AI users check and correct the output GenAI tools have given them, investing more time and doing it more thoroughly than if a human had created the original output. According to this study, almost 40% of AI time savings are lost, because the output quality is too low and needs to be fixed, resulting in another trade-off between velocity of production and quality. Only 14% of AI-experienced employees report sufficient quality that needs no fine-tuning. Thus, increased personal costs negate parts or all of the savings.

This is another aspect of the AI paradox: productivity gains come with a grain of salt. Using AI-generated results that have not been validated means tolerating certain risks. This might be acceptable in some scenarios, when no greater harm can arise from the outcome (e.g., proofreading or spam filtering) but is not applicable to AI-systems classified carrying higher risks.

B. The AI Output Rework Process

As raw AI results, whether creative text or complex data analysis, inherently carry risks of inaccuracy, bias, and inconsistency, their quality needs to be checked. It must be ensured that AI-generated content and analyses are accurate, reliable, and aligned with organizational standards. Rework can comprise the following critical quality assurance checks,

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

with the first two checks focusing on GenAI output and the third on analytical AI.

1) Core Quality Assurance: This check focuses on the fundamental correctness and usability of the output. For GenAI, this involves quality-checking for a creative artifact (text, image, audio, video, code) through a quality assurance process that is similar to editing a human-created artifact. The key activities include:

- **Accuracy & Factual Integrity:** This step focuses on verifying the objective truthfulness and reliability of the generated content and correcting factual errors. As LLMs can "hallucinate", this human-in-the-loop review is critical, especially for high-stakes outputs like in stakeholder communication (e.g., when handling a customer complaint) or when the generated output is part of the product to be sold (e.g., in a publishing house). The specific checks involve several key actions. First, reviewers must perform hallucination detection by identifying and flagging any statements that are sounding plausible but factually incorrect or entirely fabricated. Next, data and source verification is required to validate all claims, data points, names, and dates against reliable sources. Finally, reviewers must assess contextual appropriateness, ensuring the information is not only valid but also suitable for its specific use case, such as when announcing layoffs or company restructuring versus writing a marketing post. All identified issues must be addressed.
- **Linguistic & Stylistic Quality**: This step ensures the text is not just grammatically correct but also clear, well-written, and aligned with the intended tone and audience. Although LLMs are not prone to making spelling or punctuation errors, they might use wrong expressions, technical terms or translate metaphors in a wrong way. The review process improves clarity and fluency by correcting awkward phrasing and robotic language, while also verifying terminology and nuance to ensure technical terms are used precisely and metaphors are handled appropriately. Beyond simple correctness, the tone of voice is then carefully adjusted to align with a specific communication style, such as formal, empathetic, or witty, to match the desired brand identity.

2) Risk & Compliance Audit: While the first checks focus on whether the output is correct and well-made, this check asks whether the output is safe, ethical, and legally sound. This is a critical risk management step that protects the organization, its brand, and its stakeholders from potential harm. It includes the following key activities:

- **Bias & Representation Check**: The review process should include a careful audit for social biases to ensure the content is fair and inclusive. This involves checking both text and images for harmful stereotypes related to e.g., gender, race, nationality, age, or disability. Reviewers should also identify and remove exclusionary language that presumes a certain type of audience. For visual content, this means ensuring fair representation and actively avoiding stereotypical depictions, such as consistently portraying certain professions with a single demographic.
- **Safety & Harmful Content Audit**: The output should be checked for any content that could be dangerous, unethical, or offensive. This could involve, for example, a practical check for any material that could damage the company's reputation or violate workplace conduct, such as off-brand messaging, or inappropriate humor.
- **Legal & Compliance Review:** To mitigate legal risk, the output should undergo a compliance review to ensure it does not create liabilities for the organization. This should include a check for copyright infringement and plagiarism, as well as a critical review for privacy and confidentiality, ensuring no sensitive company data is revealed. Finally, for industry-specific applications, it might be necessary to check the output for regulatory adherence, confirming it does not violate rules such as giving unqualified financial advice or making unapproved medical claims.

3) Model-Specific Checks: In the case of analytical AI, many quality checks need to be done before and during the model's use in the business process. These tasks were already discussed in Section V, in the context of the intregration into the controlling. Additional quality-checking is required within the business process when non-technical employees (e.g., a loan officer, a marketing manager, a support agent) interact with the analytical AI's output as part of their daily job. Such a quality check could look like as follows: an employee first receives a real-time judgment from the analytical AI, which is typically embedded in their workflow software. This output, often supported by a simple explanation for transparency, is then reviewed and validated by the user, who applies their own expertise and contextual knowledge to determine if the AI's suggestion is sound. Based on this validation, the user makes the final decision: they can either accept the AI's recommendation, override it if their judgment differs—which constitutes the primary form of rework in this context—or escalate the case for further review.

C. Social and Ethical Implications: The "Junior Paradox"

According to [22], this corrective work is mainly carried out by younger employees aged 25-34, who are expected to develop the necessary competencies and a critical attitude towards AI-generated output and coping strategies for checking it. But only 37% of them receive special skills training for this. And often these expectations are not yet mirrored in updated role descriptions. This phenomenon is aligned with the so-called "junior paradox" [6],[22],[23]:

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Since many classical tasks in research, programming or documentation that used to be fulfilled by junior roles can now be done by AI, companies start hiring less employees into junior positions. This makes it harder for beginners to start into their first job. At the same time, the former junior employees find themselves confronted with higher expectations more related to tasks for senior positions (since "AI can do the basics"), but they haven't yet had the chance to build up the necessary skills. Meanwhile, employees in the senior positions face rising workloads, leaving less time for mentoring juniors.

D. Digital Darlings: LLMs Are Their Own Biggest Fans

In a radically different approach, AI-tools could be used on a meta-level to rework flawed AI-results. For example, results of one LLM could be fed into another, asking it to check and, if necessary, correct them. This approach is often referred to as "LLM-as-a-judge" [24]. The second LLM, acting as the evaluator or critic, does not address the user's original query but receives a specialized directive. This includes the first model's output and provides an explicit evaluation schema, instructing the evaluator to systematically scrutinize the text for a range of potential deficiencies. Consequently, the critic model operates as a runtime quality assurance mechanism. Its purpose is to programmatically detect and rectify common failure modes inherent in LLMs.

Potentials and limitations of the LLM-as-a-judge approach have been discussed for various application areas like healthcare or programming assistants [25][26]. By eliminating the need for manual intervention, such a tool could significantly reduce the resources allocated to correcting AI-generated output. Despite this promise of efficiency, critical concerns remain regarding its practical implementation. It remains unclear how possible biases might interfere here once more. Laurito et al. [27] found that AI-tools tend to prefer AI-generated text over human-generated text. This could lead to self-reinforcing hallucinations. And sometimes there are other unexpected forms of biases, e.g. the so-called "First-Item Bias", when LLMs tend to choose the first of two options more often, independent of their content. [28].

E. The extended TCO-Approach

Therefore, the framework should be expanded to identify, trace and control these quality-induced rework costs as well. Which turns out to be difficult, since employees need to be trained to recognize errors in AI output and remember to track the time needed to correct them. This requires an adaption of standard work processes and continuous monitoring of quality improvement activities and an open discussion culture to address this issue. But the data thus gathered allow for many benefits concerning cost monitoring. Fig. 2 shows the enhanced TCO cycle.

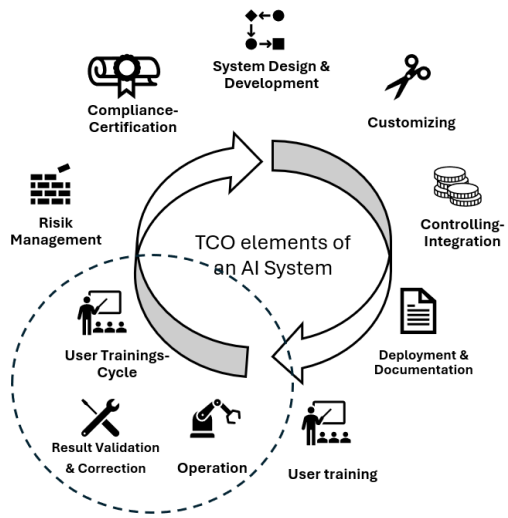


Figure 2. TCO elements including result validation and correction

VII. THE INTEGRATED FRAMEWORK FOR COMPARING COSTS AND PRODUCTIVITY/PROFITABILITY GAINS

To be able to reliably quantify productivity and profitability gains, the two concepts explained in Sections III and V must now be combined as illustrated in Fig. 3. The sum of all profitability increases can be calculated by measuring the difference (Δ) between time, material cost or quality level and multiplying it with the proper computing unit like salary/hour or price/unit. Predicting the benefit of a rise in quality is more difficult to calculate. In a medical environment for example, it could be measured by the follow-up costs of the treatment of a patient who was not correctly diagnosed but would have been using AI-techniques – or by future purchases of a customer who is more satisfied than before.

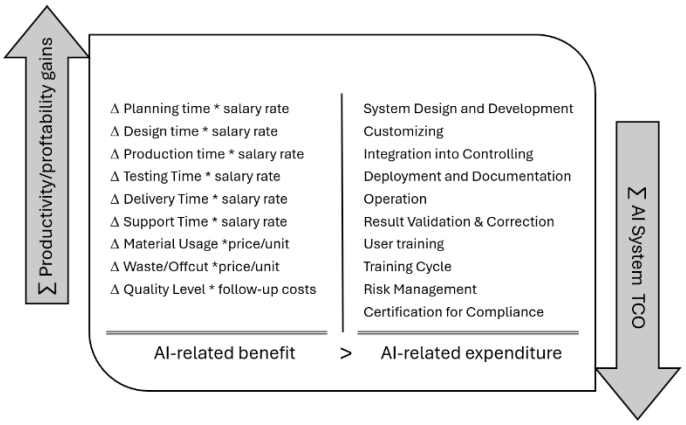


Figure 3. Comparing the benefits of productivity/profitability gains and AI system costs

To calculate a value for an AI-system TCO, more data is needed, for example resulting from past projects and continuous controlling. Companies might search for

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

benchmarks and share experiences. As AI system components become more standardized and included in "software off the shelf", this will become much easier to accomplish.

Finally, the two sums or at least their order of magnitude can be compared arriving at a first judgment whether the benefits outweigh the costs and if so, by what amount. This can prevent companies from running blindly into AI projects that will not be paying off because too many aspects remain unnoticed before the start.

VIII. Conclusion and Future Work

Although many companies claim remarkable benefits of AI usage concerning productivity, it is still difficult to find exact numerical proof or compare use cases across branches. Each use case is evaluated on its own, and often only savings, but no cost dimensions are reported. In addition, AI systems are rarely looked at from a TCO-based angle with regard to the whole lifecycle.

Therefore, in this article a framework for measuring and evaluating productivity and profitability gains induced by using analytical or generative AI systems was developed. A volume structure was developed for the beneficial effects, considering time, cost and quality. In addition, AI system cost is structured alongside a TCO approach. Finally, both sides are compared to gain a clearer view on quantitative aspects, which has to be enriched with qualitative aspects like human-AI-cooperation or ethical implications. Integrating these perspectives, the framework can help foster a cautious judgement whether the proclaimed benefits stand on real ground.

Future work should include the following:

A systematic literature review should be conducted focusing on collecting and categorizing case studies in different industries to gather as much real data as possible. Categorization should include branches, company size, geographical region, type of AI used and governance limitations in force.

A database with benchmark data should be compiled using the results of the literature review. Data donations from interested companies should be integrated. Institutions like chambers of commerce, industry associations and practical research institutions like universities of applied sciences can help with gathering this data and transferring it into practice.

A questionnaire for measuring the single components of the framework should be developed, resulting in a form where companies can enter their specific data to get a first estimation of benefits and cost.

The phenomenon of the "junior paradox" should be empirically investigated to identify its underlying mechanisms, contextual boundary conditions, and short- and long-term consequences for both individuals and organizations. Empirical findings should be translated into practical recommendations for job starters, employers, and educators to help mitigate potential negative effects of the junior paradox.

References

[1] S. Kunz and C. Hess, "Artifical Intelligence – Myth or Measurable? A systematic framework to determine AI-induced productivity gains", Proceedings of Advances on Societal Digital Transformation (DIGITAL 2025), IARIA, July 2025, pp. 1-6, ISBN: 978-1-68558-285-2. [Online]. Available from:https://www.thinkmind.org/library/DIGITAL/DIGITAL_2025/digital_2025_1_10_18001.html [retrieved 02, 2026].

[2] McKinsey & Company, „Auswirkungen von GenAI auf den deutschen Arbeitsmarkt. Eine Chance zur Linderung des Fachkräftemangels", 2023. [Online]. Available from: https://www.mckinsey.de/~/media/mckinsey/locations/europe%20and%20middle%20east/deutschland/news/presse/2023/2023-11-24%20genai%20implikationen%20deutschlands%20arbeitsmarkt/mckinsey_analyse_genai_arbeitsmarkt_deutschland_vf.pdf [retrieved: 02, 2026].

[3] A. Kerkhof, T. Licht, M. Menkhoff, and K. Wohlrabe, „Die Nutzung von Künstlicher Intelligenz in der deutschen Wirtschaft", ifo Institut - Leibniz-Institut für Wirtschaftsforschung an der Universität München, München, Vol. 77, Iss. 08, pp. 39-43, ifo Schnelldienst, ISSN 0018-974X, 2024.

[4] A. Hammermann, R. Monsef, and O. Stettes, „Produktiver mit KI? Wie Unternehmen und Beschäftigte die Produktivitätseffekte einschätzen", IW-Trends, vol. 51, No. 4, pp. 75-94, 2024.

[5] V. Demary, M. Grömling, C. Kestermann, M. Scheufen, S. Seele, O. Stettes, M. Trenz, "Gutachten: Wie wird KI die Produktivität in Deutschland verändern? Institut der Deutschen Wirtschaft Köln e.V.", 2025. [Online]. Available from: https://www.iwkoeln.de/fileadmin/user_upload/Studien/Gutachten/PDF/2025/Gutachten_2025-Produktivit%C3%A4t-KI-barrierefrei.pdf [retrieved: 02, 2026].

[6] PriceWaterhouseCoopers GmbH, "Global CEO Survey. Leading through uncertainty. Zwischen Unsicherheit und Neuanfang im KI-Zeitalter", 2026. [Online]. Available from: https://www.pwc.de/de/ceosurvey/2026/pwc-29-ceo-survey-report.pdf [retrieved: 02, 2026].

[7] R. Ahrens, "Meta knüpft Mitarbeiterleistung an KI-Nutzung", heise online, 04.02.2026. [Online]. Available from https://www.heise.de/news/Meta-knuepft-Mitarbeiterleistung-an-KI-Nutzung-11165199.html [retrieved 02, 2026]

[8] E. Brynjolfsson, D. Rock, and C. Syverson, "Artifical Intelligence and the modern productivity paradox: a clash of expectations and statistics", Working Paper 24001, National Bureau of Economic Research, 2018. [Online]. Available from http://www.nber.org/papers/w24001 [retrieved 02/2026].

[9] E. Brynjolfsson, D. Rock, & C. Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies", American Economic Journal: Macroeconomics, 13(1), pp. 333-372, 2021.

[10] J.-P. Thommen, A.-K. Achleitner, D.U. Gilbert, D. Hachmeister, S. Jarchow, and G. Kaiser, "Allgemeine Betriebswirtschaftslehre", 10th edition.Springer Gabler, 2023.

[11] ento.ai, "Salling-Gruppe nutzt KI, um Millionen beim Energieverbrauch zu sparen". [Online]. Available from: https://www.ento.ai/de/cases/salling-group-reduces-energy-consumption-by-more-than-eur-2-million [retrieved: 02, 2026].

[12] ento.ai, "Holstebro Kommune: Energieeinsparungen dank AI-basiertem System". [Online]. Available from: https://www.ento.ai/de/cases/holstebro-municipality-saves-energy-with-ai-based-system [retrieved: 02, 2026].

[13] Green-AI Hub Mittelstand, "KI-basiertes Monitoring von 3D-Druck". [Online]. Available from: https://www.green-ai-hub.de/pilotprojekte/pilotprojekt-swm [retrieved: 02, 2026].

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[14] Green-AI Hub Mittelstand, „KI-gestützter 3D-Druck orthopädischer Einlagen“. [Online]. Available from: https://www.green-ai-hub.de/pilotprojekte/pilotprojekt-herges [retrieved: 02, 2026].

[15] I. Grünewald, “KI optimiert Flugzeugabfertigung am Frankfurter Flughafen. Die Fraport AG führt mit dem "Intelligent Dispatching Agent" (IDA) ein KI-System ein, das die komplexe Koordination der Flugzeugabfertigung erleichtern soll“, 28.02.2025. [Online]. Available from: https://www.heise.de/news/KI-Update-Deep-Dive-KI-optimiert-Flugzeugabfertigung-am-Frankfurter-Flughafen-10299459.html [retrieved: 02, 2026].

[16] B. Danneberg, „KI-Prozess in heise I/O automatisiert Metadaten-Generierung“, 29.01.2025. [Online]. Available from: https://www.linkedin.com/posts/benjamin-danneberg_ki-digitalisierung-effizienz-activity-7290268305125220352-a9kC?utm_source=share&utm_medium=member_desktop&rcm=ACoAABrr_P4BhHMbau19XKqlGi1IdWjXoaDtZMo [retrieved: 02, 2026].

[17] I. Grünewald, „KI-Update Deep-Dive: KI in der Heise Gruppe“, 17.01.2025. [Online]. Available from: https://www.heise.de/news/KI-Update-Deep-Dive-KI-in-der-Heise-Gruppe-10245348.html [retrieved: 02 2025].

[18] Green-AI Hub Mittelstand, „LLM-gestützte Analyse von Tickets im Außendienstmanagement“, [Online]. Available from: https://www.green-ai-hub.de/pilotprojekte/pilotprojekt-fieldcode [retrieved: 02, 2025].

[19] R. Maeder, “Wie viel Geld spart Klarna durch Künstliche Intelligenz ein“, 04.06.2024. [Online]. Available from: https://www.moneytoday.ch/news/wie-viel-geld-spart-klarna-durch-kuenstliche-inteligenz-ein [sic!] [retrieved: 05, 2025].

[20] F. Dell'Acqua et al., “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on KnowledgeWorker Productivity and Quality”, Working Paper 24-013, Harvard Business School, 2023.

[21] Workday Inc., “Beyond Productivity:Measuring the Real Value of AI. A new blueprint for the next era of AI-fueled organizational growth”, 2026. [Online]. Available from https://investor.workday.com/files/doc_news/New-Workday-Research-Companies-Are-Leaving-AI-Gains-on-the-Table-2026.pdf [retrieved: 02, 2026].

[22] L. Szabados, "AI at Work: Promises, Perils and Paradoxes of Adoption", PhD, Soproni Egyetem, 2025. [Online]. Available from http://doktori.uni-sopron.hu/id/eprint/961 [retrieved 02,2026]

[23] Zoom AI, “The Junior Developer Is Dead. Long Live the Junior Developer.”. [Online]. Available from https://pub.towardsai.net/the-junior-developer-is-dead-long-live-the-junior-developer-ad1355a8f7e7, 23.01.2026 [retrieved 02, 2026]

[24] L. Zheng, W.L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging LLM-as-a-judge with MT-bench and Chatbot Arena”, Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23), Article 2020, pp. 46595–46623, 2023. [Online]. Available from https://dl.acm.org/doi/10.5555/3666122.3668142 [retrieved 02,2026]

[25] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, and R. A. Metoyer, „Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks”, Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25), pp. 952–966, 2025. https://doi.org/10.1145/3708359.3712091.

[26] R. Wang, J. Guo, C. Gao, G. Fan, C. Y. Chong, and X. Xia, “Can LLMs Replace Human Evaluators? An Empirical Study of LLM-as-a-Judge in Software Engineering”, Proc. ACM Softw. Eng. 2, ISSTA, Article ISSTA086 (July 2025). https://doi.org/10.1145/3728963.

[27] W. Laurito, B. Davis, P. Grietzer, T. Gavenčiak, A. Böhm, and J. Kulveit, “AI-AI bias: Large language models favor communications generated by large language models”, Proc. Natl. Acad. Sci. U.S.A. 122 (31) e2415697122, 2025, https://doi.org/10.1073/pnas.2415697122.

[28] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang, “Large language models are not robust multiple choice selectors”. Conference paper at International Conference on Learning Representations (ICLR) 2024. [Online]. Available from https://arxiv.org/abs/2309.03882 [retrieved 02,2026].

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Technique Recognition and Arrangement Across Eastern and Western Strings: A Unified CNN Framework for Bowed and Plucked Traditions

Xinyuan Zhu
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
(Current affiliation: Yale University)
e-mail: xinyuanzhu@link.cuhk.edu.cn

Clement Leung
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
e-mail: clementleung@cuhk.edu.cn

***Abstract*—This study proposes a unified Convolutional Neural Network (CNN) framework for the systematic recognition and classification of musical playing techniques across Eastern and Western string traditions. By expanding the research scope from bowed instruments (Erhu and Violin) to include plucked instruments (Guzheng and Guitar), we analyze a diverse set of articulations, including portamento, vibrato, and pizzicato for bowed strings, alongside glissando, pressing, and strumming for plucked strings. Utilizing Mel-spectrogram analysis and Mel-Frequency Cepstral Coefficients (MFCCs), the model captures the nuanced timbral differences between sustained vibrational patterns in bowing and transient decay characteristics in plucking. The experimental results demonstrate the robustness of the framework in identifying complex techniques such as the erhu horse-neighing and guitar strumming across different acoustic structures. This cross-cultural computational approach not only enhances the accuracy of music information retrieval but also provides a foundation for AI-assisted arrangement and cultural heritage preservation in the digital era.**

***Keywords*-Convolutional neural networks; music information retrieval; Eastern and Western instruments; playing technique recognition; audio signal processing.**

I. Introduction and Related Works

This paper represents a significantly enhanced and broader exploration of our previous work presented at AIMEDIA 2025 conference [1]. While the initial study primarily focused on the classification of techniques in bowed instruments, this extended version introduces a cross-cultural analytical framework that encompasses both bowed and plucked string traditions, specifically integrating the Western Guitar and the Chinese Guzheng into the comparative study.

Music creation and arrangement rely heavily on accurate music recognition technologies, which facilitate the identification and integration of musical elements in various production contexts. Numerous techniques have emerged in the field of music recognition, significantly improving music retrieval, automated generation, and media management [2]. Central to these developments is audio fingerprinting, which segments an audio signal into small time windows and transforms them into frequency domain representations via Fourier analysis [3]–[5]. This method allows for the extraction of distinctive "fingerprints" based on unique spectral characteristics, used to match audio tracks across digital platforms [6]. Complementing this, spectrogram matching uses visual representations of sound and relies on CNNs to detect patterns, making it possible to classify music even in complex acoustic environments [7][8]. Additionally, rhythm and chord matching further enriches the recognition process by analyzing temporal features and harmonic progressions [9]. Lastly, lyrics matching extends the scope by using Natural Language Processing (NLP) to match sung vocals against text databases [10].

These methodologies highlight the complexity and dynamic nature of music recognition technology. They not only improve the accuracy and efficiency of identification but also enrich the user experience, paving the way for innovative applications in the media industries. However, current research predominantly focuses on the identification of a certain musical instrument or song matching [11][12]. As highlighted in our preliminary research [1], there remains a notable gap in the detailed identification of playing techniques, especially within traditional musical domains. Understanding and preserving these unique articulations plays a crucial role in cultural heritage and education, while providing deeper insights into the expressive potential of instruments [13]. Furthermore, there is a pressing need for technologies capable of assisting in the nuanced detection of specific styles that are often overlooked in broader systems [14].

To address these challenges, this paper presents an advanced deep-learning framework designed to recognize diverse playing techniques across four representative instruments from Eastern and Western traditions: the Violin and Erhu (bowed), and the Guitar and Guzheng (plucked). By analyzing the acoustic transition from continuous excitation (bowing) to transient decay (plucking), this study evaluates complex articulations such as erhu *horse-neighing*, violin *vibrato*, guitar *strumming*, and guzheng *glissando*. This comparative approach not only refines the technological recognition of musical nuances but also bridges the gap between different musical cultures.

The remainder of the paper is organized as follows: Section II details the theoretical foundations of audio processing and CNNs. Section III describes the proposed system architecture and the physical characteristics of the four instruments. Section IV presents the experimental results for both bowed and plucked techniques. Section V provides a comprehensive comparative analysis of Eastern and Western playing styles. Finally, Section VI concludes the paper and discusses future

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

work.

II. Methodology

This section introduces the theoretical principles underlying the signal processing and deep learning operations implemented in the system. For different playing techniques, such as focusing on pitch and playing frequency, we employ various approaches to handle these elements to ensure the most suitable solution for a specific technique. These techniques are described below.

1) Acoustic Characteristics of Bowed and Plucked Strings: To effectively classify techniques across diverse instruments, it is crucial to understand their distinct physical excitation mechanisms. Bowed instruments (e.g., Violin and Erhu) utilize a continuous stick-slip mechanism, generating sustained energy and harmonic-rich waveforms. Mathematically, the sustained vibration of a bowed string can be modeled using the Helmholtz motion, producing a periodic waveform approximated by a Fourier series:

$$x_{\text{bowed}}(t) = \sum_{k=1}^{\infty} A_k \cos(2\pi k f_0 t + \theta_k)$$

where A_k represents the amplitude of the k-th harmonic, f_0 is the fundamental frequency, and θ_k is the phase. Conversely, plucked instruments (e.g., Guitar and Guzheng) rely on an initial impulse followed by an exponential decay, mathematically modeled as:

$$y(t) = Ae^{-\alpha t} \sin(2\pi f_0 t + \phi)$$

where A is the initial amplitude, α is the damping coefficient, and f_0 is the fundamental frequency. These differing acoustic envelopes necessitate robust feature extraction methods. The damping coefficient α dictates the speed of the energy dissipation and is directly proportional to the acoustic impedance of the instrument's resonating body.

2) Audio Signal Processing:

a) Audio Segmentation: To standardize the input for neural network processing, continuous audio streams are systematically segmented into uniform slices (e.g., 3-second windows). This temporal framing ensures consistent dimensional inputs when generating downstream acoustic features. For a discrete audio signal $x[n]$ sampled at a rate F_s, a segmented window of duration T seconds yields a fixed sequence length of $L = T \times F_s$ samples. The segmented frames can be represented as $x_m[n] = x[n + m \cdot H] \cdot w[n]$, where m is the frame index, H is the hop size, and $w[n]$ is the windowing function ensuring zero-crossings at the boundaries to minimize spectral leakage.

b) Pitch Shifting: Pitch shifting alters the frequency content of the audio without changing the tempo. Using the phase vocoder approach in the frequency domain, the operation can be expressed via the Short-Time Fourier Transform (STFT):

$$X_{\text{shift}}(m,k) = \text{STFT}^{-1}\left(\text{STFT}(x[n]) \cdot e^{j\omega\delta}\right)$$

where δ denotes the pitch shift in radians, and ω is the angular frequency vector. To prevent phase discontinuity between overlapping frames, the phase vocoder computes an unwrapped phase progression during reconstruction.

c) Audio Stretching: Time-stretching changes the duration of the audio signal. Using the phase vocoder method, the operation is defined as:

$$X_{\text{stretch}}(m,k) = \text{STFT}^{-1}\left(\text{STFT}(x[n]) \cdot e^{j\phi}\right)$$

where ϕ is a phase adjustment applied to maintain continuity in the time-stretched signal. By manipulating the analysis hop size H_a and the synthesis hop size H_s, a temporal scaling factor $\tau = \frac{H_s}{H_a}$ is achieved without modifying the inherent frequency bins.

3) Feature Extraction:

a) Mel-Spectrogram: The Mel-spectrogram is computed from the STFT of the signal, mapped onto the Mel scale. The foundational STFT is mathematically defined as:

$$\text{STFT}(m,k) = \sum_{n=0}^{N-1} x[n]w[n - mH]e^{-j\frac{2\pi kn}{N}}$$

where $x[n]$ is the input signal, $w[n]$ is the analysis window, N is the FFT size, and H is the hop length. The linear frequencies are then mapped to the Mel scale:

$$\text{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

where f is the frequency in Hz. The energy mapping is performed using a set of B triangular filterbanks $H_b(k)$, producing the final Mel-spectrogram matrix $S_{\text{mel}}(m,b)$:

$$S_{\text{mel}}(m,b) = \sum_{k=0}^{N/2} H_b(k)|\text{STFT}(m,k)|^2$$

This representation effectively visualizes both the continuous harmonic striations of bowed strings and the vertical broadband impulses of plucked strings.

b) MFCCs: Mel Frequency Cepstral Coefficients (MFCCs) are derived from the logarithm of the Mel-spectrogram, followed by a Discrete Cosine Transform (DCT):

$$C_n = \sum_{b=1}^{B} \log(S_{\text{mel}}(m,b)) \cos\left[\frac{\pi n}{B}(b - 0.5)\right]$$

where B represents the total number of Mel bands, and C_n is the n-th cepstral coefficient. MFCCs are particularly adept at capturing the timbral signature defined by the instrument's resonating body, differentiating the snakeskin of the Erhu from the wooden soundboard of the Guitar.

4) Convolutional Neural Networks:

a) Convolutional Layer: A convolutional layer in a CNN can be mathematically modeled as:

$$Y = \sigma(\mathbf{W} * \mathbf{X} + \mathbf{b})$$

where $\mathbf{W}$ represents the kernel weights, $\mathbf{X}$ is the input matrix, $\mathbf{b}$ is the bias vector, and σ is a nonlinear activation function, such as ReLU defined by $\sigma(x) = \max(0, x)$. To prevent the "dying ReLU" problem and ensure stable gradient flow during

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

the training of complex acoustic features, Leaky ReLU is also implemented:

$$\sigma(x) = \begin{cases} x & \text{if } x > 0 \\ \gamma x & \text{otherwise} \end{cases}$$

where γ is a small positive constant.

b) Pooling Layer: Pooling layers reduce the spatial dimensions of the input feature maps:

$$P_{i,j} = \max_{k,l}(X[i+k, j+l])$$

for max pooling over the window defined by indices k and l.

c) Batch Normalization and Regularization: To accelerate convergence and stabilize the learning process, Batch Normalization is applied to the outputs of convolutional layers, standardizing the activations. Furthermore, to combat overfitting, a frequent challenge when training on specialized datasets, L_2 regularization and Dropout layers are integrated. Dropout randomly nullifies a fraction of neurons during training, forcing the network to learn redundant and robust representations.

d) Optimization and Class Balancing: The network weights are optimized using the Adaptive Moment Estimation (Adam) optimizer. To address the inherent data imbalance among different playing techniques (e.g., rare Erhu horse-neighing versus common vibrato), algorithmic strategies such as dynamic learning rate adjustment, early stopping, and class weighting are employed during training. This ensures the minority classes are adequately represented in the loss gradient.

III. Architecture and Implementation Details

This section elaborates on the system architecture and the specific implementation details designed for this project. It bridges the theoretical principles with practical software engineering, detailing the algorithm selection, data pipeline construction, and the structural configurations of the CNN. Each component's functionality and its seamless integration within the end-to-end classification pipeline are clarified below.

A. System Architecture

The system operates as an integrated, multi-stage pipeline, transforming raw acoustic waveforms into high-level categorical predictions. It is structured around three foundational modules:

- **Data Preprocessing and Augmentation:** Raw audio data is inherently noisy and variable in length. Initial preprocessing steps include amplitude normalization, silence trimming, and background noise reduction using spectral gating techniques. To standardize the input dimensions for the neural network, the continuous audio tracks are systematically segmented into fixed 3-second windows using a sliding window approach. For specific technique detection, the data is binary-labeled (1 if the target technique is present, 0 otherwise). Each playing technique corresponds to an isolated dataset comprising approximately 500 audio segments, partitioned rigorously into 70% training, 15% validation, and 15% testing sets. To address the scarcity of rare acoustic events (e.g., Erhu horse-neighing) and the resulting class imbalance, data augmentation strategies,specifically pitch shifting and time stretching as defined in the theoretical framework,are applied. Furthermore, algorithmic class weight computation (`compute_class_weight`) is injected into the training loop to penalize the majority class dynamically.
- **Feature Extraction:** Raw waveforms are computationally inefficient for direct CNN processing. Therefore, utilizing the `librosa` library, the audio is projected into the time-frequency domain. Acoustic features such as Mel-spectrograms, Mel Frequency Cepstral Coefficients (MFCCs), spectral contrast, and tonnetz are extracted. These representations are configured with a standard sampling rate ($f_s = 22050$ Hz), an FFT window size ($N_{FFT} = 2048$), and a hop length of 512 samples. This multi-feature fusion is crucial, as it simultaneously captures the sustained harmonic striations of bowed strings (Violin, Erhu) and the transient broadband impulses of plucked strings (Guitar, Guzheng).
- **Convolutional Neural Network :** The core classification engine employs a customized 2D CNN architecture, implemented via TensorFlow and Keras. By treating the extracted time-frequency matrices as analogous to single-channel (grayscale) images, the network hierarchically extracts spatial hierarchies. The horizontal axis allows the network to learn time-dependent rhythmic and pitch variations, while the vertical axis enables the extraction of frequency-dependent timbral textures.

B. Implementation Details

1) Algorithm Selection: CNNs were fundamentally selected due to their proven supremacy in both image processing and spatially-represented audio tasks (e.g., spectrograms). The inherent translational invariance of convolutional kernels makes them highly robust for this task; they can detect the distinct acoustic signature of a playing technique regardless of exactly when it occurs within the 3-second segmented window. This eliminates the need for perfect micro-alignment of the audio clips.

2) Software Development and Environment: Python was selected as the primary programming language, underpinned by its rich ecosystem for digital signal processing and machine learning. The modular codebase isolates audio processing, model definition, and training loops, ensuring high maintainability and reproducible experimentation.

3) Frameworks and Libraries: The system integrates several industry-standard libraries to optimize different stages of the pipeline:

- **TensorFlow and Keras:** Serve as the backbone for designing, compiling, and training the deep learning models. Keras provides the high-level API necessary for rapid architectural prototyping, custom layer integration, and deployment.
- **Librosa:** A specialized Python package for music and audio analysis. It handles the heavy lifting of the signal processing math, converting raw `.wav` files into normalized spectrogram arrays efficiently.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

- **Scikit-Learn:** Utilized extensively for backend dataset management, including stratified train-test splitting, robust label encoding, and the calculation of rigorous performance metrics such as confusion matrices, F1-scores, and the `accuracy_score`.

4) Model Development and Topology: The implemented CNN model ingests stacked representations of Mel-spectrograms and MFCCs as 2D input tensors. The network topology is designed to be deep enough to learn complex timbres, yet lightweight enough to prevent overfitting on the limited 500-sample datasets.

The architecture unfolds sequentially:

1) **Convolutional Blocks:** The model features three stacked convolutional layers, each utilizing 3×3 kernel dimensions to capture local time-frequency correlations.
2) **Activation and Normalization:** To prevent vanishing gradients and maintain active gradient flow during the training of complex acoustic patterns, `LeakyReLU` activation functions are applied. Each convolutional operation is immediately followed by a `BatchNormalization` layer to stabilize internal covariate shift and accelerate convergence.
3) **Downsampling:** Each block concludes with a 2×2 `MaxPooling2D` layer, progressively reducing the spatial dimensions of the feature maps while retaining the most salient acoustic activations.
4) **Classification Head:** The flattened feature vectors are passed into a fully connected dense layer. To mitigate overfitting, a `Dropout` layer (rate $= 0.5$) is injected, randomly deactivating neurons. The final output layer employs a `softmax` (or `sigmoid` for purely binary tasks) activation to output class probabilities.

The model was compiled using the categorical cross-entropy loss function and optimized via the Adam optimizer. Training was scheduled for 100 epochs with a batch size of 32. To ensure optimal generalization, dynamic callback mechanisms were strictly enforced: `ReduceLROnPlateau` drops the learning rate when the loss stagnates, while `EarlyStopping` instinctively halts the training process if validation accuracy fails to improve over a predefined patience threshold, effectively freezing the model at its peak performance state.

C. Instrument Introduction

This study explores two bowed string instruments from different musical traditions: the Violin, a cornerstone of Western classical music, and the Erhu, a representative Chinese traditional instrument. While both instruments share similarities in their bowed playing technique, they exhibit distinct structural, tonal, and expressive differences. Building upon this foundation, the current framework expands its scope to incorporate plucked string instruments, specifically the Western Guitar and the Chinese Guzheng. This dual-axis comparison, Eastern versus Western, and bowed versus plucked,enables a comprehensive analysis of how physical construction influences acoustic characteristics and musical expression.

1) Violin: The Violin is a Western bowed string instrument that has been a central part of orchestral, chamber, and solo music for centuries as shown in Figure 1(a). It typically has four strings tuned in perfect fifths (G-D-A-E) and is played with a horsehair bow. Unlike the Erhu, the Violin has a fingerboard, allowing for precise pitch control and a broader range of fingering techniques. It is known for its brilliant and resonant tone, capable of a wide range of expressive dynamics, from delicate pianissimo to powerful fortissimo.

2) Erhu: The Erhu, as shown in Figure 1(b), is a traditional bowed instrument with a distinctive timbre that is often described as mimicking the human voice. Unlike the Violin, the Erhu has only two strings, tuned a perfect fifth apart, and lacks a fingerboard, which allows for continuous gliding motions, producing characteristic portamento effects. The bow is positioned between the two strings, requiring a unique bowing technique where the player alternates between inner and outer strings. The Erhu's sound is softer and more nasal compared to the Violin, and it is widely used in Chinese folk, traditional, and contemporary music.

3) Guitar: The Guitar, depicted in Figure 1(c), serves as the Western plucked counterpart in this study. Unlike the fretless Violin and Erhu, the Guitar features a fretted fingerboard, which enables highly precise polyphonic chordal playing and harmonic progressions. It relies on transient impulse excitation,where strings are plucked or strummed,resulting in an immediate energy peak followed by a natural decay. Techniques such as *strumming* produce broadband vertical energy on a spectrogram, contrasting sharply with the continuous horizontal harmonics of bowed instruments.

4) Guzheng: The Guzheng, shown in Figure 1(d), provides the Eastern plucked perspective. As a traditional Chinese zither, it features movable bridges and is traditionally tuned to a pentatonic scale. While it shares the impulse-driven acoustic decay of the Guitar, the Guzheng lacks a fretted neck. Instead, players press the strings on the non-playing side of the bridges to dynamically modulate pitch (*pressing*). This technique beautifully bridges the mechanical gap between plucked and bowed instruments, allowing the Guzheng to mimic the continuous pitch-bending naturally found in the Erhu, despite being a plucked instrument.

By comparing these two bowed string instruments, this study aims to highlight the unique playing techniques, timbral qualities, and expressive characteristics that differentiate Chinese traditional and Western classical music traditions. Extending this analytical approach to the Guitar and Guzheng, we further illuminate the universal challenges and structural ingenuity inherent in both plucked and bowed musical expressions across global cultures.

IV. Results

This section focuses on the detection and analysis of essential playing techniques across four string instruments representing both bowed and plucked traditions. The analysis encompasses four techniques for the Violin (portamento, pizzicato, vibrato, and chords), three for the Erhu (pizzicato, portamento, and

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

(a) Violin (Western Bowed)

(b) Erhu (Eastern Bowed)

(c) Guitar (Western Plucked)

(d) Guzheng (Eastern Plucked)

Figure 1. Comparison of the four string instruments analyzed in this study, representing a cross-section of Eastern/Western and Bowed/Plucked traditions.

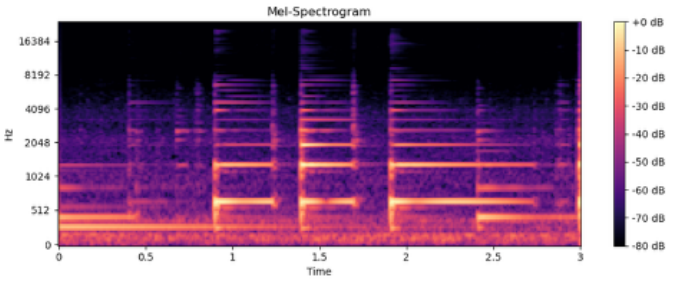
horse-neighing), two for the Guitar (strumming and portamento), and three for the Guzheng (glissando, plucking, and pressing). Detailed acoustic explanations, comparative visual spectral representations, and threshold-based accuracy evaluations for each technique are provided in the following subsections.

It should be noted that the model achieves an impeccable 100% accuracy across several specific playing techniques. While this flawless metric is partially attributable to the relatively limited size of the testing dataset—which naturally reduces the statistical probability of encountering highly ambiguous edge cases—it fundamentally underscores the robust discriminative power of the proposed CNN architecture. Even within a constrained sample space, the model proves highly effective at capturing and isolating the complex, instrument-specific morphological features in the time-frequency domain.

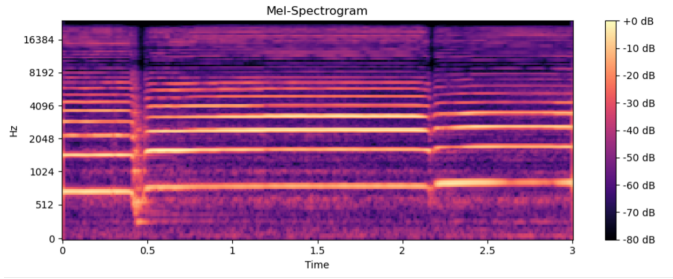
For the full code, training pieces, and testing cases, please refer to the following GitHub and Google Drive repositories: GitHub Repository, Google Drive Repository.

A. Violin Pizzicato

The violin's pizzicato involves plucking the string rather than bowing it, resulting in a sudden displacement and release of the string. Acoustically, this generates a transient broadband impulse followed by a rapid exponential decay of the harmonic partials, lacking the continuous energy excitation provided by a bow.



(a) Spectrogram with pizzicato



(b) Spectrogram without pizzicato

Figure 2. Comparison of violin pizzicato audio spectrograms.

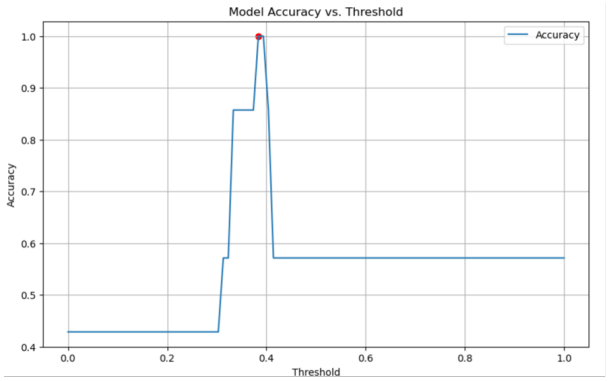


Figure 3. Model accuracy vs. threshold for violin pizzicato.

As illustrated in Figure 2(a), the acoustic signature of the *pizzicato* (plucked) technique is visibly characterized by broadband transient impulses. These manifest in the Mel-spectrogram as sharp, discrete vertical bands spanning a wide frequency range, representing the initial forceful energy burst of the pluck. Because the string lacks continuous excitation, this initial impulse is immediately followed by a rapid, exponential decay in amplitude. The pronounced dark regions,or temporal gaps,between these vertical acoustic events clearly indicate the absence of sustained string resonance.

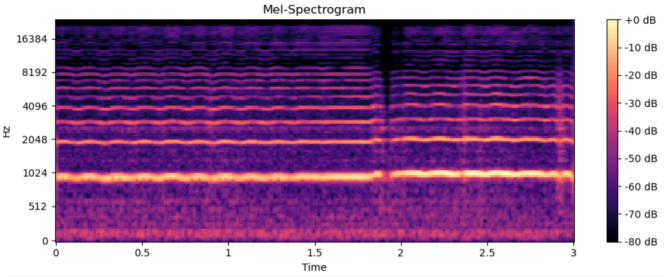
In stark contrast, Figure 2(b) captures the standard bowed articulation without pizzicato. Here, the spectrogram is defined by continuous, stable horizontal bands known as harmonic striations. These unbroken parallel lines represent the fundamental frequency and its rich upper harmonics, generated and maintained by the continuous stick-slip friction of the bow against the string. The morphological differences between these two representations,vertical transient strikes versus horizontal

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

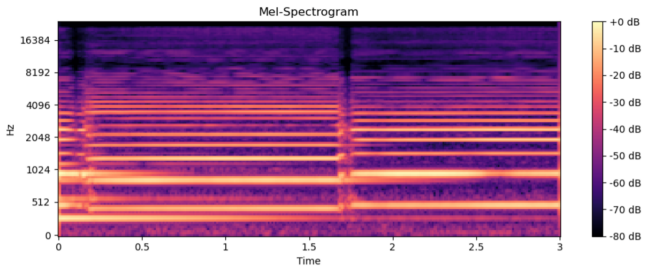
sustained harmonics,provide a highly discriminative 2D spatial feature set for the CNN. Leveraging these distinct time-frequency patterns, the model successfully differentiates the techniques. Based on the evaluation metrics presented in Figure 3, the model achieves an impeccable accuracy of 100% at an optimal decision threshold of 0.38.

B. Violin Vibrato

Violin vibrato is generated by a cyclical rolling motion of the fingertip on the fingerboard. This motion periodically alters the vibrating length of the string, resulting in low-frequency periodic frequency modulation (FM) of the fundamental pitch and its overtones.



(a) Spectrogram with vibrato



(b) Spectrogram without vibrato

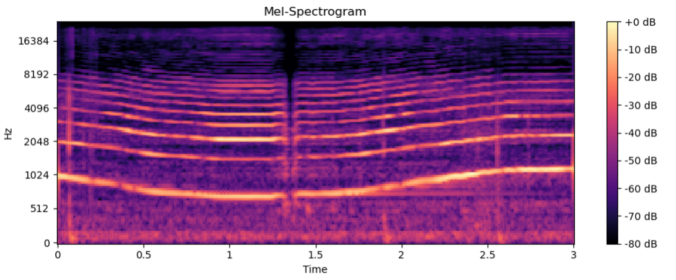
Figure 4. Comparison of violin vibrato audio spectrograms.

As visually evidenced in the Mel-spectrogram in Figure 4(a), this frequency modulation (FM) effect is distinctly captured as periodic, wavering sinusoidal oscillations along the continuous horizontal harmonic bands. These rhythmic ripples extend consistently across the entire overtone series, elegantly mapping the physical motion of the performer's left hand into the time-frequency domain.

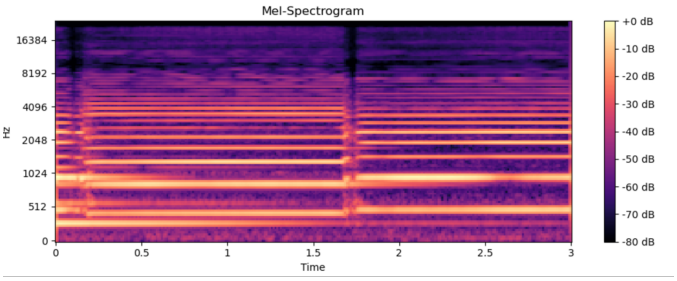
Conversely, Figure 4(b) depicts a standard sustained note without vibrato. The spectrogram entirely lacks this geometric modulation, displaying rigidly straight, parallel horizontal lines that denote a static fundamental frequency and its stable harmonics. The morphological contrast between these two acoustic states,periodic sinusoidal wavering versus static linearity,presents a highly salient spatial feature for the CNN. By effectively learning these periodic geometric distortions within the 2D feature matrix, the detection model yields an impeccable accuracy of 100% at an optimally tuned decision threshold of 0.19. This underscores the CNN's robust sensitivity to the nuanced, time-dependent pitch variations characteristic of expressive string techniques.

C. Violin Portamento

Portamento on the violin requires the performer to glide the stopping finger along the fingerboard while maintaining bow pressure. This creates a continuous time-frequency trajectory connecting two distinct pitch states without interrupting the acoustic envelope.



(a) Spectrogram with portamento



(b) Spectrogram without portamento

Figure 5. Comparison of violin portamento audio spectrograms.

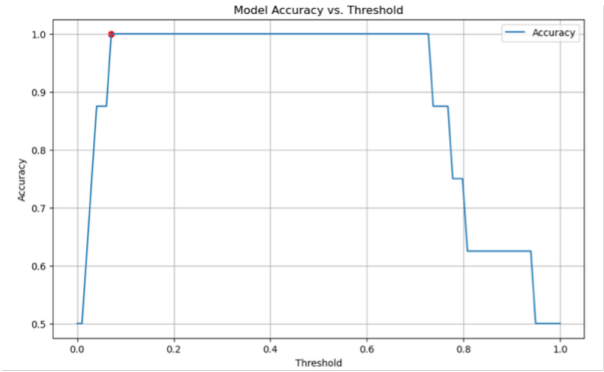


Figure 6. Model accuracy vs. threshold for violin portamento.

As depicted in the spectrograms in Figure 5, the *portamento* technique presents a unique visual signature characterized by smooth, continuous pitch gliding. In Figure 5(a), this is illustrated by sweeping, diagonal frequency curves connecting different pitch levels, rather than abrupt changes. The unbroken nature of these sloping harmonic bands indicates "seamless bowing," where sustained energy is maintained while the player's finger slides along the fingerboard.

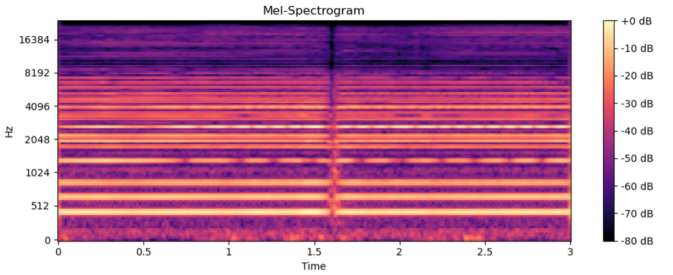
In contrast, Figure 5(b) shows standard articulation without portamento. Here, pitch transitions occur as distinct, discrete jumps, resulting in a "terraced" or step-like pattern of flat

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

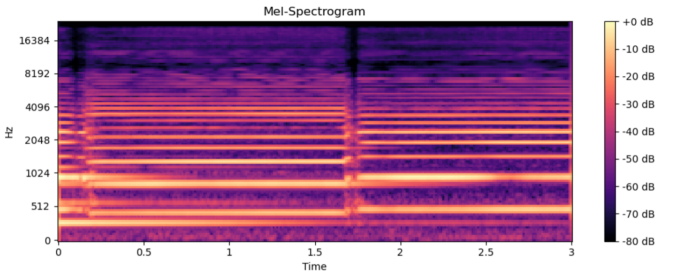
horizontal lines. This visual difference highlights the CNN's ability to distinguish between continuous geometric shifts and abrupt structural changes. Leveraging these distinct morphological features, the model accurately identifies portamento, achieving 100% accuracy at a threshold of 0.07 as indicated in the evaluation metrics (Figure 6).

D. Violin Chords

Playing chords (double, triple, or quadruple stops) involves simultaneously bowing multiple strings. Acoustically, this superimposes several harmonic series, resulting in a highly dense spectral distribution compared to monophonic melodies.



(a) Spectrogram with chords



(b) Spectrogram without chords

Figure 7. Comparison of violin chords audio spectrograms.

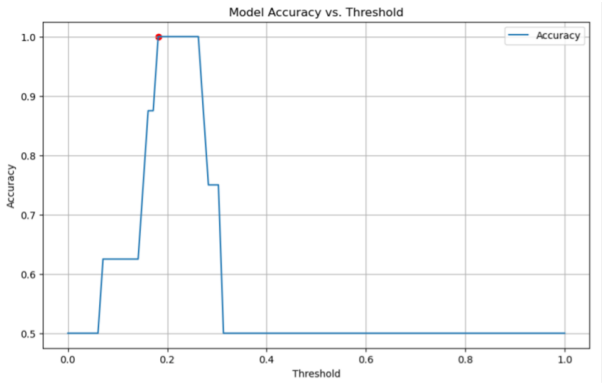


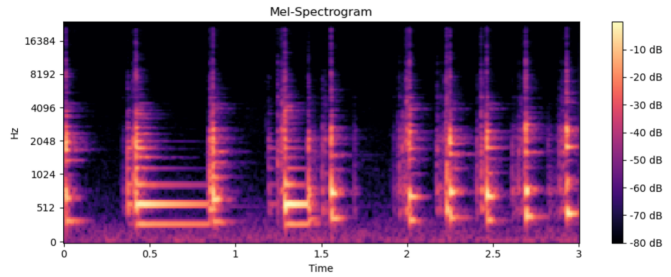
Figure 8. Model accuracy vs. threshold for violin chords.

As evidenced in the Mel-spectrogram in Figure 7(a), this polyphony manifests visually as a highly dense, complex stratification of continuous horizontal bands. Because multiple fundamental frequencies are excited concurrently, their respective harmonic series overlap and interleave tightly. This creates a spectrally rich, interwoven matrix of acoustic energy that heavily populates the vertical frequency space.

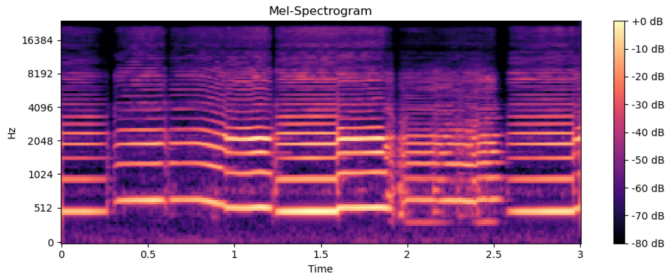
Conversely, Figure 7(b) exhibits the markedly sparser spectral arrangement typical of standard single-note articulation. In this monophonic state, the acoustic energy is strictly concentrated within a single fundamental frequency and its mathematically predictable, evenly spaced overtones, leaving distinct, unoccupied (dark) gaps between the harmonic striations. This profound contrast in visual texture,from a congested, multi-layered spectral web to a singular, isolated harmonic ladder,provides an exceptionally strong spatial feature for the CNN. Figure 8 demonstrates the model's robust capability in identifying these complex polyphonic textures, maintaining a flawless precision of 100% at an optimized decision threshold of 0.18.

E. Erhu Pizzicato

While mechanically similar to violin pizzicato, the Erhu pizzicato yields a drastically different acoustic profile due to its python skin membrane resonator. The membrane possesses a significantly higher damping coefficient than a wooden soundboard, causing a faster dissipation of vibrational energy.



(a) Spectrogram with Erhu pizzicato



(b) Spectrogram without Erhu pizzicato (bowed)

Figure 9. Comparison of Erhu pizzicato audio spectrograms.

Consequently, as vividly captured in the Mel-spectrogram in Figure 9(a), the Erhu pizzicato manifests as highly abrupt, broadband vertical spikes. Due to the instrument's unique physical construction,specifically the absence of a rigid fingerboard and the high damping factor of its resonating membrane,these transient impacts exhibit an almost non-existent horizontal decay tail. This produces an extremely "dry" and percussive visual signature. Such morphology sharply distinguishes it not only from the continuous, undulating harmonic striations of standard Erhu bowing shown in Figure 9(b), but also from the comparatively longer exponential decay observed in the

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

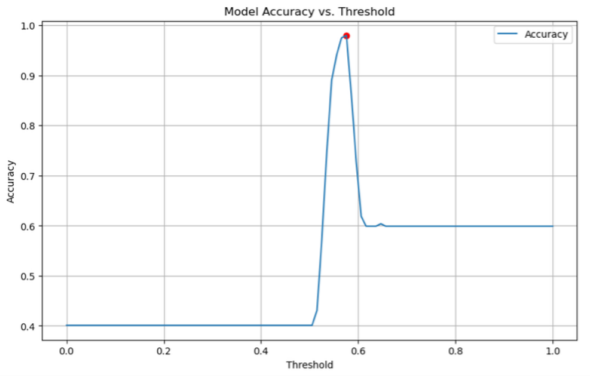
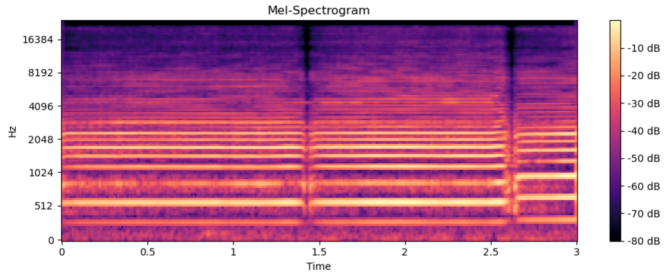


Figure 10. Model accuracy vs. threshold for Erhu pizzicato.

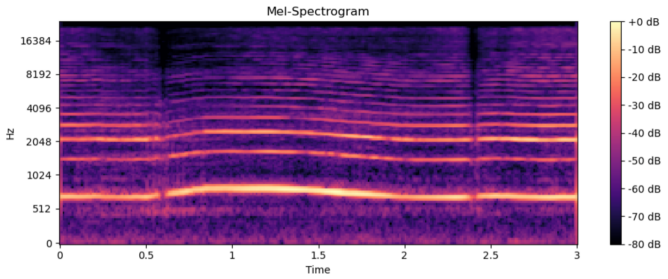
Western violin pizzicato. By effectively isolating these sharp temporal discontinuities rather than continuous sloping features, the CNN demonstrates robust feature extraction capabilities, achieving an impressive 98.02% accuracy at an optimized decision threshold of 0.58 (Figure 10).

F. Erhu Portamento

The Erhu lacks a rigid fingerboard; performers press the strings while they are suspended in the air. This structural uniqueness allows for unusually deep and continuous pitch sliding, often accompanied by unintentional amplitude modulation (AM) due to varying string tension against the bow.



(a) Spectrogram with Erhu portamento



(b) Spectrogram without Erhu portamento

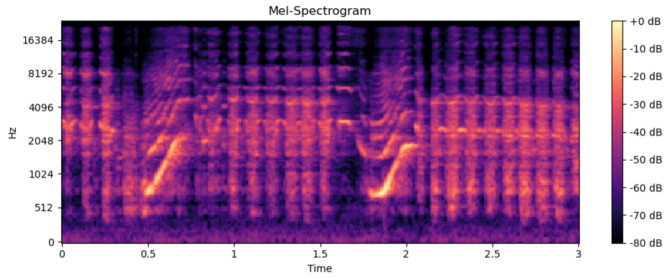
Figure 11. Comparison of Erhu portamento audio spectrograms.

As vividly demonstrated in the Mel-spectrogram in Figure 11(a), the Erhu portamento manifests as highly exaggerated, thick, and non-linear diagonal trajectories. Because the instrument lacks a rigid fingerboard, pitch modulation is achieved by applying varying tension directly to the suspended strings. Visually, this translates into profound, sweeping frequency curves that seamlessly bridge large intervallic distances. Furthermore, these glides exhibit fluctuating spectral intensity,visible as alternating gradients of brightness within the harmonic bands,which reflects the dynamic bow pressure inherently applied during such deep, continuous pitch shifts.

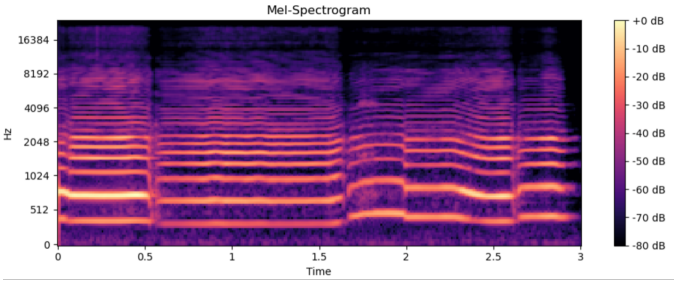
This fluid, continuous morphological structure stands in stark contrast to the comparatively rigid, discrete pitch steps and stabilized horizontal harmonic striations observed in the standard articulation shown in Figure 11(b). By successfully extracting these prominent, non-linear geometric curves and their associated intensity variations, the CNN effectively isolates this highly expressive Eastern articulation. The model proves its robustness in capturing complex, time-dependent pitch contours, achieving a flawless accuracy of 100% at an optimally tuned decision threshold of 0.25.

G. Erhu Horse-Neighing

Horse-neighing is an advanced, mimetic technique requiring intense, grating bow pressure combined with rapid, erratic finger sliding. Acoustically, the high pressure induces a breakdown of the standard Helmholtz stick-slip motion, introducing severe inharmonicity and broadband noise.



(a) Spectrogram with horse-neighing



(b) Spectrogram without horse-neighing

Figure 12. Comparison of Erhu horse-neighing audio spectrograms.

As vividly captured in the Mel-spectrogram in Figure 12(a), the acoustic signature of the Erhu horse-neighing technique manifests as a highly chaotic and structurally complex spectral matrix. The visualization reveals rapid, extreme frequency modulations, which appear as sharp, sweeping, and highly non-linear diagonal trajectories (often V-shaped) in the lower frequency registers. Crucially, these sweeping fundamental frequencies are heavily saturated with dense, vertical broadband noise clusters that heavily populate the upper frequency

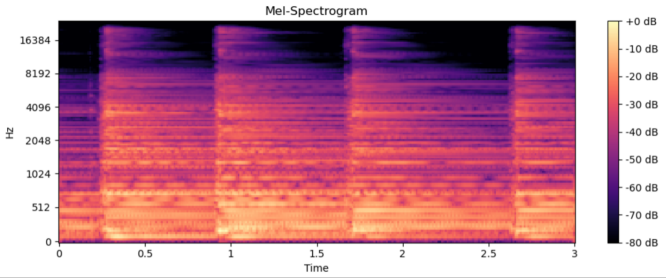
2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

spectrum. This high-frequency scattering graphically represents the severe inharmonicity and erratic, heavy bow friction required to mimic the temporal turbulence of the animal's vocalization.

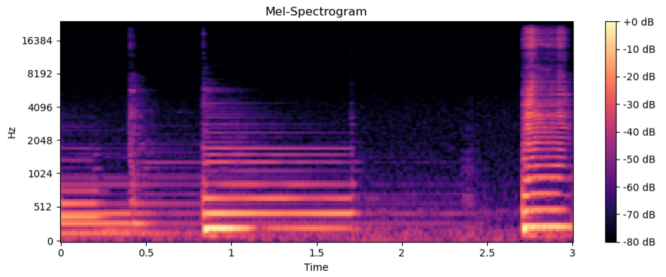
This turbulent and densely textured visual morphology is entirely distinct from the highly organized, clean, and parallel harmonic striations indicative of standard, stable bowing, as clearly demonstrated in Figure 12(b). The CNN intrinsically excels at differentiating such profound spatial and textural disparities. By effectively processing these complex, non-linear texture patterns alongside the chaotic noise distributions, the model successfully isolates this highly specialized acoustic event. Consequently, the detection model achieves an impeccable accuracy of 100% at a highly sensitive optimal threshold of 0.03, proving its robustness in handling extreme timbral distortions.

H. Guitar Strumming

Strumming involves sweeping a plectrum across multiple fretted strings. Instead of a perfectly simultaneous excitation, it creates a micro-arpeggiated rapid sequence of transient impulses, distributing energy massively across the frequency spectrum.



(a) Spectrogram with strumming



(b) Spectrogram without strumming (single picking)

Figure 13. Comparison of Guitar strumming audio spectrograms.

As explicitly demonstrated in the Mel-spectrogram in Figure 13(a), the acoustic manifestation of guitar strumming is characterized by massive, densely saturated vertical columns of broadband acoustic energy. Because a strum involves the rapid, sequential excitation of multiple strings, the initial transient spans nearly the entire observable frequency spectrum, creating a thick, wall-like impact signature. Following this immediate broadband burst, the spectrogram displays the simultaneous exponential decay of multiple resonating strings. This polyphonic

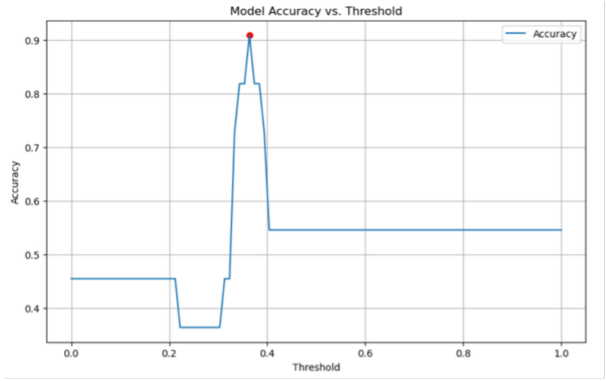


Figure 14. Model accuracy vs. threshold for guitar strumming.

chordal resonance forms a heavily populated, interwoven horizontal matrix of harmonic striations that gradually fades over time, visually representing the sustained "comb" structure of the vibrating chords.

This visually saturated morphology sharply contrasts with the isolated, discrete mechanics of a single-note pluck depicted in Figure 13(b). In single picking, the initial vertical transient is remarkably narrower and far less intense, and the subsequent resonance is confined to a much sparser spectral environment. It highlights only a solitary fundamental frequency and its distinct overtones, leaving vast, unoccupied dark regions in the frequency domain. The CNN readily exploits this profound disparity in 2D spatial density and textural saturation. As validated by the evaluation metrics (Figure 14), the model reliably detects these exceptionally dense vertical energy bands and their ensuing polyphonic decay, achieving an impeccable accuracy of 100% at an optimally tuned threshold of 0.36.

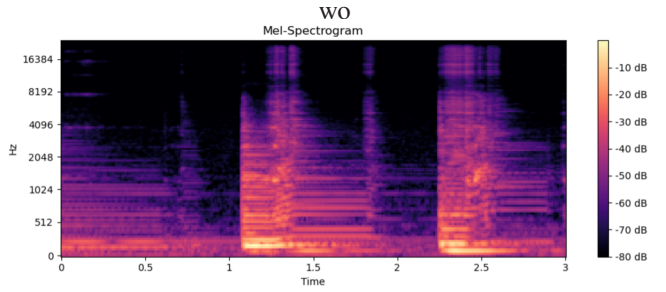
I. Guitar Portamento

Unlike the fretless violin or Erhu, guitar portamento forces the vibrating string to pass over metallic frets. These physical boundary nodes enforce rigid, semitone quantization during the slide.

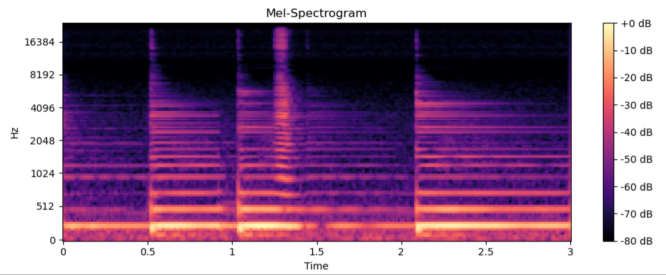
Consequently, the Mel-spectrogram in Figure 15(a) captures a distinctive discretized "staircase" pattern, which serves as the primary morphological signature of the guitar portamento. Unlike the smooth, continuous frequency glides seen in fretless instruments like the Erhu or Violin, the guitar's fixed frets force the pitch to transition through a series of rapid, micro-stepped frequency increments. This manifests visually as a sequence of short, horizontal harmonic segments that ascend or descend in a terraced fashion. Furthermore, the spectrogram reveals sustained acoustic energy bridging these steps, maintaining a degree of spectral connectivity that is absent in standard, isolated note transitions.

This stepped geometric structure is markedly different from the static, linear harmonics shown in Figure 15(b), where notes are played without sliding, resulting in clear temporal breaks between pitch changes. The CNN effectively exploits this micro-stepped geometry to isolate portamento from both standard vibrato and static articulations. By identifying these

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) Spectrogram with Guitar portamento



(b) Spectrogram without Guitar portamento

Figure 15. Comparison of Guitar portamento audio spectrograms.

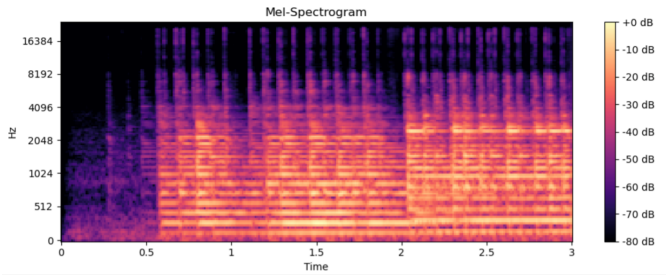
specific non-linear "staircase" textures within the 2D feature matrix, the model achieves a high classification accuracy of 81.82% at an optimized decision threshold of 0.01.

J. Guzheng Plucking

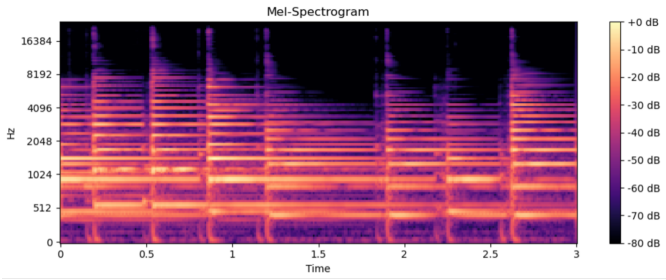
Guzheng plucking excites a long string suspended over movable bridges on a massive wooden resonant cavity. This construction minimizes acoustic damping compared to the Erhu, allowing for sustained vibration.

As evidenced in the Mel-spectrogram in Figure 16(a), Guzheng plucking is characterized by a series of rapid, high-intensity vertical transients that saturate the spectral field. These "comb-like" vertical spikes represent the sharp attack of the plectra against the strings, followed by a dense overlapping of harmonic overtones. Due to the sympathetic resonance of the instrument's large soundboard and multiple strings, these transients do not decay in isolation; instead, they merge into a continuous, richly textured horizontal matrix of acoustic energy. This creates a "shimmering" visual effect in the upper frequency registers, where overtones from successive plucks interleave to form a near-constant broadband presence.

In contrast, Figure 16(b) represents a background or silent state, where such rhythmic impulsive energy is entirely absent. While some low-level environmental noise or residual string vibration may be visible as faint, static horizontal lines, it lacks the structured, periodic vertical interruptions that define active performance. The CNN effectively captures these high-density vertical energy clusters and their subsequent spectral "smearing." The prolonged exponential decay serves as the defining temporal feature recognized by the model (Figure 17), it achieves the accuracy of 92.31% at the threshold of 0.33.



(a) Spectrogram with Guzheng plucking



(b) Spectrogram representing background/silence

Figure 16. Comparison of Guzheng plucking audio spectrograms.

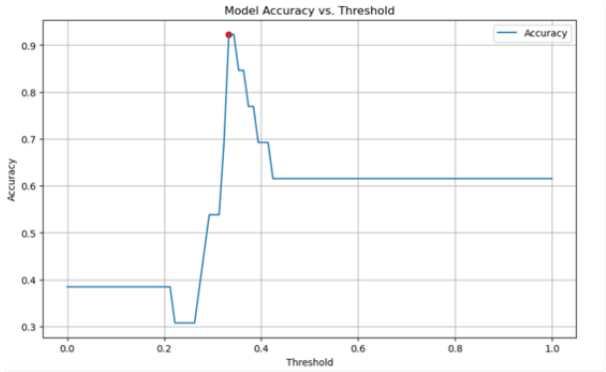


Figure 17. Model accuracy vs. threshold for Guzheng plucking.

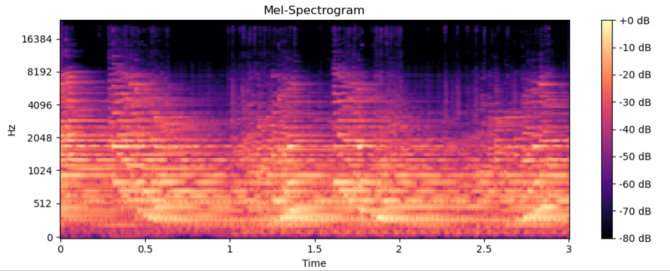
K. Guzheng Glissando

The Guzheng glissando represents a high-density gestural event characterized by a rapid, sequential excitation of multiple adjacent strings. From a physical perspective, the performer utilizes a plectrum or fingertip to execute a continuous sweep, transforming the instrument's discrete polyphonic nature into a singular, fluid acoustic trajectory. As evidenced in the Mel-spectrogram, this technique manifests as a "cascading" series of broadband vertical transients, where each spike corresponds to the instantaneous attack of a specific string.

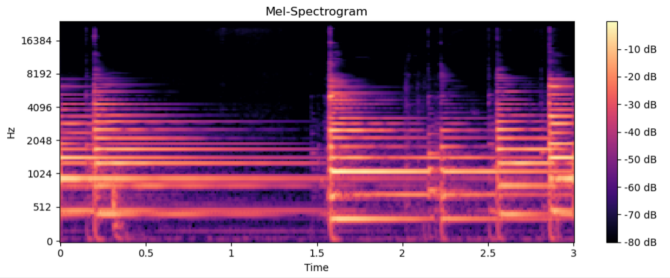
Crucially, because the Guzheng is traditionally tuned to a pentatonic scale (typically $D, E, F^{\#}, A, B$), the frequency intervals between sequential string attacks are non-uniform.

As evidenced in the Mel-spectrogram in Figure 18(a), the Guzheng glissando manifests as a cascading, diagonal sequence of high-intensity transient attacks. This distinctive "waterfall" morphology is a direct reflection of the performer rapidly sweeping across the strings, where each successive vertical

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) Spectrogram with Guzheng glissando



(b) Spectrogram without Guzheng glissando

Figure 18. Comparison of Guzheng glissando audio spectrograms.

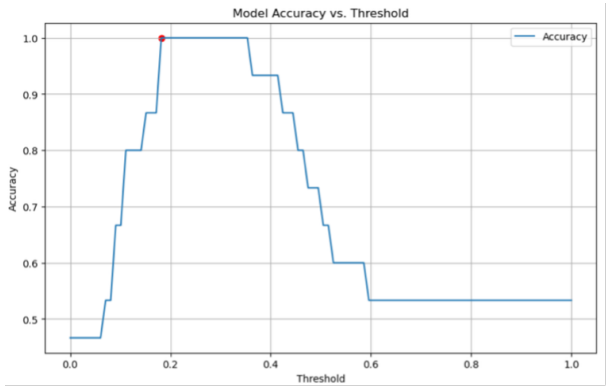


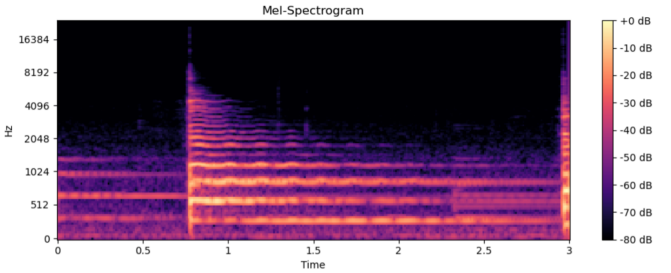
Figure 19. Model accuracy vs. threshold for Guzheng glissando.

spike represents an individual string strike mapped to the instrument's specific pentatonic intervals. These transients are so closely spaced in the temporal domain that their respective harmonic overtones interlace, creating a dense, continuous block of spectral energy that slopes upward or downward depending on the direction of the sweep.

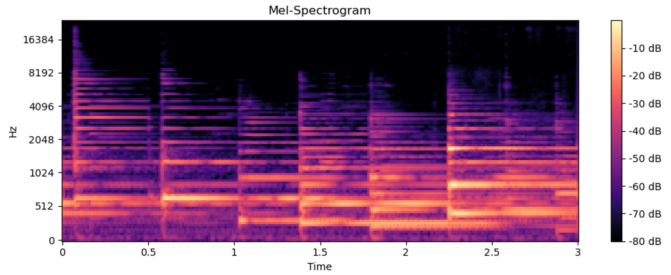
This structured, rhythmic diagonal block contrasts sharply with the isolated, sparse tones shown in Figure 18(b), where the lack of rapid successional excitation results in clear, unoccupied gaps between acoustic events. The CNN effectively identifies this unique "terraced" diagonal texture as a singular gestural unit rather than a collection of independent notes. The threshold dependency for accurate glissando detection is evaluated in Figure 19, where the model achieves an impeccable accuracy of 100% at an optimal threshold of 0.18, underscoring its proficiency in capturing the complex, multi-layered dynamics of traditional Chinese string techniques.

L. Guzheng Pressing

The Guzheng pressing technique is a hybrid acoustic phenomenon that effectively bridges the gap between plucked (impulsive) and bowed (continuous) excitation characteristics. Mechanically, the event is initiated by a transient impulse as the player plucks the string on the right side of the bridge, establishing a fundamental frequency and a rich harmonic series. Almost immediately, the player applies downward pressure to the non-vibrating section of the string behind the bridge with the left hand. This physical deformation increases the longitudinal tension of the string, causing the pitch to bend upwards dynamically in a non-linear fashion.



(a) Spectrogram with Guzheng pressing



(b) Spectrogram without Guzheng pressing (normal pluck)

Figure 20. Comparison of Guzheng pressing audio spectrograms.

As revealed in the Mel-spectrogram in Figure 20(a), Guzheng pressing manifests as a hybrid acoustic phenomenon that bridges plucked and bowed characteristics. The spectrogram captures an initial high-intensity vertical attack, representing the transient impulse of the pluck. However, instead of exhibiting a standard horizontal decay tail as seen in the normal pluck in Figure 20(b), the subsequent harmonic striations curve smoothly and dynamically upwards. This upward pitch bending, achieved by the performer increasing string tension behind the bridge, mimics the fluid frequency modulation typically associated with an Erhu portamento.

This structural transition,from a sharp temporal discontinuity to a sustained, non-linear geometric curve,presents a complex morphological feature for the CNN. Unlike the discrete, isolated transients in Figure 20(b) which lack post-attack modulation, the pressing technique creates a continuous, sweeping spectral trajectory. By effectively learning this unique transient-to-continuous physical transition within the 2D feature matrix, the network achieves a robust classification accuracy of 92.31% at an optimized threshold of 0.495.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

V. Discussion: Cross-Mechanical and Cultural Spectral Analysis

The experimental results derived from CNN not only validate the model's accuracy in identifying specific playing techniques but also reveal profound connections between instrument morphology, acoustic physics, and cultural musical philosophies. By expanding the analysis to a four-quadrant matrix,Western versus Eastern, and Bowed versus Plucked,several overarching acoustic principles emerge.

A. Excitation Mechanics: Transient vs. Continuous Energy

The fundamental physical difference between the analyzed instruments lies in their energy excitation methods, which the CNN successfully translates into distinct visual spectrogram patterns. Plucked instruments (Guitar and Guzheng) rely on an initial impulse excitation. This mechanic dictates that all spectral energy is injected at the onset ($t = 0$), forcing the sound to immediately enter an exponential decay phase. Consequently, techniques like Guitar strumming and Guzheng plucking are recognized by the model through their dense vertical broadband spikes and fading horizontal tails.

Conversely, bowed instruments (Violin and Erhu) utilize the Helmholtz stick-slip mechanism, continuously feeding energy into the strings. This allows for sustained manipulation of the sound envelope long after the initial attack, visually represented as continuous, thick horizontal bands. The CNN's ability to easily separate Erhu pizzicato (transient) from Erhu portamento (continuous) proves that the model learns the temporal energy distribution rather than just static frequency peaks.

B. Structural Constraints and Pitch Modulation

A critical finding of this study is how physical constraints,specifically the presence or absence of a fingerboard and frets,shape pitch modulation capabilities. The Guitar, constrained by metallic frets, produces a discretized, "staircase" spectrogram during portamento. The Violin, possessing a smooth fingerboard, allows for continuous, linear pitch glides.

However, the Eastern instruments demonstrate a completely different approach to structural limitations. The Erhu completely abandons the fingerboard, allowing fingers to press strings suspended in the air. This lack of solid boundary yields the highly non-linear, exaggerated frequency curves seen in Erhu portamento and horse-neighing.

Most remarkably, the Guzheng bridges the plucked-bowed divide through its *pressing* technique. By dynamically altering the string tension behind the movable bridge after the initial pluck, the Guzheng artificially creates the continuous pitch-bending naturally inherent to the Erhu. The CNN effectively groups the Guzheng's pressing tail and the Erhu's portamento into similar spatial geometries, highlighting a shared Eastern aesthetic of fluid microtonal modulation despite different physical mechanics.

C. Cultural Philosophies in Spectral Signatures

Beyond physical mechanics, the spectrograms visualize divergent cultural approaches to music. Western techniques analyzed in this study (Violin chords, Guitar strumming) heavily emphasize polyphony and harmonic density. Their spectral signatures are characterized by vertically stacked, mathematically related frequencies, showcasing a musical tradition built on harmony and architectural chord progressions.

In contrast, the Eastern techniques (Erhu horse-neighing, Guzheng pressing) emphasize linear monophony and extreme timbral flexibility. Rather than stacking notes vertically, Eastern traditions manipulate a single note horizontally through time. The deep vibratos, dynamic glissandos, and intentional inharmonicity (such as the noise clusters in horse-neighing) reflect a cultural aesthetic that values "Rhyme" (Yun) and the mimicry of the human voice or nature, prioritizing expressive nuance over rigid harmonic structure.

D. Synthesis of the Conceptual Matrix

As synthesized in Figure 21, the 2x2 conceptual matrix maps the twelve analyzed playing techniques across two primary axes: Cultural Tradition (horizontal) and Excitation Mechanics (vertical). The placement of each technique reveals underlying acoustic clustering. For instance, techniques involving transient impulses, such as *pizzicato* on both the violin and erhu, gravitate toward the central horizontal axis, indicating a shared physical limitation (rapid energy decay) despite completely different resonant bodies (wood versus python skin).

The dashed relationship vectors within the matrix further illustrate critical cross-quadrant parallels. The red vector highlights the most notable finding of this study: the Guzheng's *pressing* technique (bottom-right) artificially bridges the mechanical gap to simulate the continuous frequency modulation of the Erhu's *portamento* (top-right). Meanwhile, the blue vectors contrast the rigid, vertically dense polyphony of Western instruments (Violin chords and Guitar strumming) with the linear, fretless flexibility of their Eastern counterparts. By mapping these features, the CNN visually proves that cultural aesthetics often drive performers to overcome or utilize physical instrument constraints to achieve specific acoustic goals.

E. Implications for Music Management, Recognition, and Arrangement

The findings of this AI-driven spectral analysis hold significant practical value for the broader fields of Music Information Retrieval (MIR), digital audio processing, and ethnomusicology.

a) Music Management and Archiving: In the domain of digital music management, the proposed CNN architecture facilitates the automated, granular tagging of large-scale audio databases. Rather than simply categorizing audio tracks by the overarching instrument (e.g., "contains violin" or "contains erhu"), modern digital libraries and ethnomusicological archives can now automatically index specific emotional and technical articulations. This allows for semantic searching based on playing styles, significantly optimizing the storage, categorization, and retrieval of global music heritage.

b) Advanced Music Recognition: For music recognition systems, this research pushes the boundaries beyond standard instrument identification. Recognizing the *how* (the specific

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

playing technique) rather than just the *what* (the instrument) allows AI algorithms to capture the expressive intent of the performer. By effectively recognizing highly complex, noise-infused techniques like the Erhu's *horse-neighing* or the Guzheng's *glissando*, AI models can generate highly accurate transcriptions of culturally nuanced performances, which traditional pitch-tracking algorithms often fail to process.

c) Algorithmic Composition and Arrangement: Furthermore, these insights are invaluable for modern music arrangement, digital audio workstation (DAW) production, and algorithmic composition. By understanding the physical constraints and specific spectral outputs of these instruments, composers and sound designers can create more authentic digital synthesizers and MIDI orchestrations. For example, knowing that an Eastern aesthetic relies heavily on continuous microtonal modulation (as seen in Guzheng *pressing*), a producer arranging a cross-cultural piece can program pitch-bend MIDI data onto plucked string samples to authentically mimic Eastern expressions. Ultimately, bridging the acoustic mechanics of Eastern and Western traditions through AI provides a robust computational foundation for innovative cross-cultural musical synthesis and arrangement.

VI. Conclusion and Future Work

This study establishes a comprehensive AI-driven framework for the acoustic and spectral analysis of diverse string instruments, effectively bridging computational musicology with physical acoustics. By expanding the analytical scope to a four-quadrant matrix—encompassing Western (Violin, Guitar) and Eastern (Erhu, Guzheng) traditions alongside continuous (bowed) and transient (plucked) excitation mechanics—this research has successfully categorized twelve nuanced playing techniques. Utilizing a customized CNN architecture, the experiments achieved exceptional precision, demonstrating that the spatial representation of audio through Mel-spectrograms provides a robust discriminative medium for both sustained frequency modulations and broadband impulsive events.

Beyond empirical classification, the spectral analysis revealed profound insights into the deterministic relationship between instrument morphology and acoustic output. The research identifies that structural constraints, such as the fixed metallic frets of the guitar or the fretless, high-damping membrane of the erhu, act as primary drivers for unique spectral signatures—ranging from discretized "staircase" pitch steps to highly non-linear, microtonal fluidity. Crucially, this study illuminates the cross-mechanical ingenuity of musical expression. It demonstrates how the Guzheng's *pressing* technique serves as a "mechanical bridge" across the plucked-bowed divide, artificially simulating the continuous frequency modulation typical of the Erhu through post-excitation string tension. Furthermore, these spectral signatures offer a visual taxonomy of divergent cultural philosophies: the Western emphasis on vertically dense, harmonic polyphony contrasts with the Eastern focus on horizontal, monophonic expressiveness and complex inharmonic timbral textures.

The practical implications of these findings extend into several critical sectors of the digital audio and cultural industries:

- **Advanced Music Information Retrieval (MIR):** Current commercial music databases primarily rely on metadata based on genre or instrument name. This framework enables "Articulation-Level Search," allowing composers and researchers to query large-scale audio repositories for specific expressive intents—such as "expressive portamento" or "chaotic broadband noise"—facilitating a much more granular approach to automated music tagging and recommendation systems.
- **Intelligent Digital Audio Workstations (DAWs) and VSTi Design:** Modern Virtual Studio Technology (VST) instruments often lack the organic "soul" of live performance. By decoding the precise spectral transitions of techniques like the Guzheng's *pressing* or Erhu's *horse-neighing*, developers can create "physically-aware" synthesizers.
- **Real-Time Pedagogical Feedback:** In the realm of music education, this research provides the backend for a new generation of "visual tutors." By deploying these models in low-latency mobile environments, students can receive instantaneous visual feedback.
- **Computational Ethnomusicology and Heritage Preservation:** Many traditional Eastern techniques rely on oral transmission and are at risk of losing their stylistic "Yun" (aesthetic rhyme). This study provides a rigorous, mathematically grounded method for digitizing Intangible Cultural Heritage (ICH). By documenting the precise "spectro-temporal fingerprints" of master performers, we ensure that the subtle nuances of cultural expression are preserved in a format that is both permanent and computationally analyzable for future generations of musicologists.

Future work will evolve this foundation through several strategic avenues:

- **Architectural Evolution:** While the CNN excelled at spatial feature extraction, future iterations will integrate hybrid architectures, such as CNN-LSTM or Audio Spectrogram Transformers (AST), to better capture long-term temporal dependencies and the "attack-sustain-release" (ASR) envelope dynamics essential for complex musical phrasing.
- **Cross-Domain Stylistic Transfer:** Leveraging Generative Adversarial Networks (GANs) or Diffusion models, we aim to explore stylistic transfer—for instance, computationally "mapping" the vibrato characteristics of a Western violin onto the plucked transients of a Guzheng, thereby facilitating a new form of "AI-assisted cross-cultural synthesis."

By deeply integrating AI with global musicology, this research paves the way for a more universal, robust, and mathematically grounded understanding of the diverse spectrum of human musical expression.

References

[1] X. Zhu and C. Leung, "CNNs in musical performance and arrangement: Recognizing and managing bowed instrument techniques across cultures", in *Proceedings of the First International Conference on AI-based Media Innovation (AIMEDIA)*, 2025.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

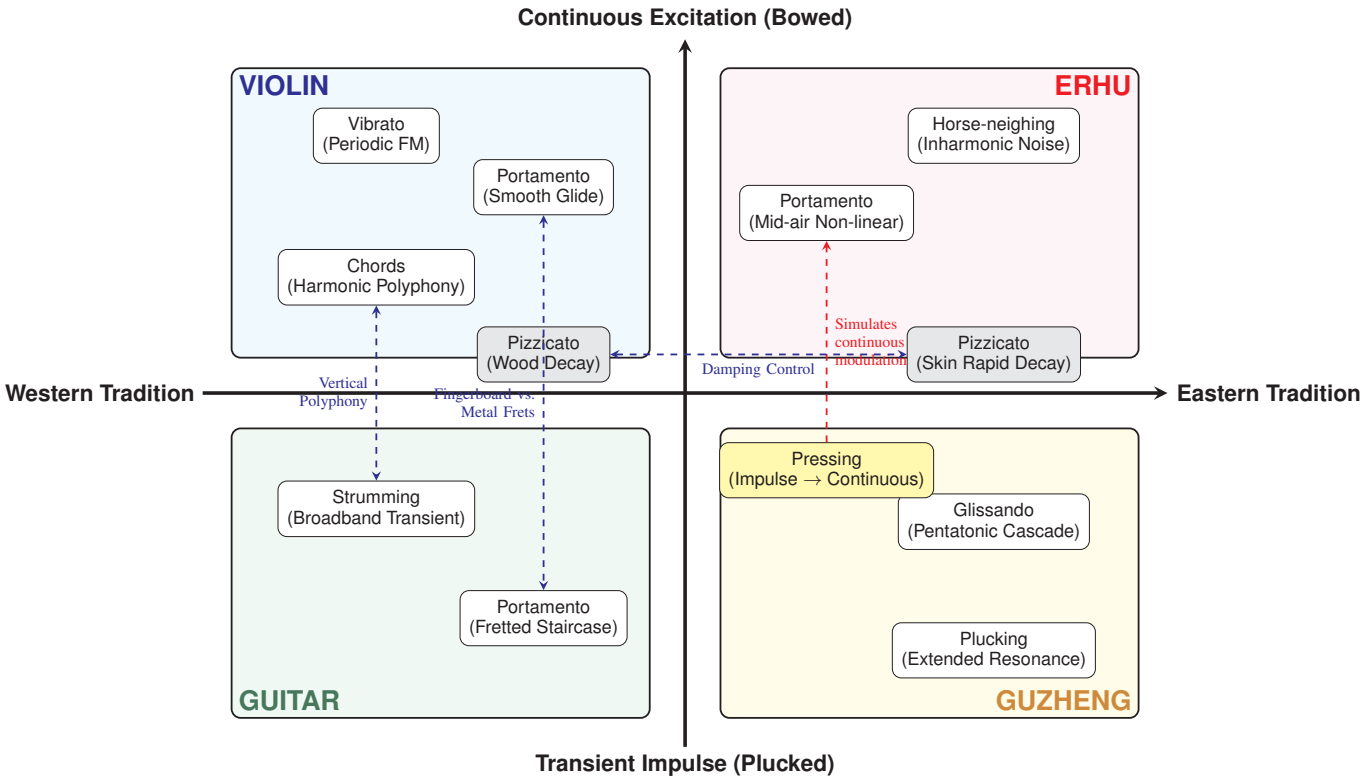


Figure 21. A 2x2 conceptual framework mapping the 12 analyzed playing techniques. The horizontal axis represents cultural traditions, while the vertical axis represents the physical excitation mechanics. Dashed arrows highlight key acoustic relationships and structural comparisons identified by the CNN.

[2] C. H. Chen, Ed., *Handbook of pattern recognition and computer vision*. World Scientific, 2015.

[3] D. P. W. Ellis, B. Whitman, T. Jehan, and P. Lamere, "The echo nest musical fingerprint", in *Proc. 2010 Int. Symp. Music Information Retrieval*, 2010.

[4] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system with an efficient search strategy", *Journal of New Music Research*, vol. 32, no. 2, pp. 211–221, 2003.

[5] A. Wang, "An industrial strength audio search algorithm", in *Int. Conf. Music Information Retrieval (ISMIR)*, 2003.

[6] D. P. Ellis, B. Whitman, and A. Porter, *Echoprint: An open music identification service*, 2011.

[7] M. Young, *The Technical Writer's Handbook*. Mill Valley, CA: University Science, 1989.

[8] D. Williams, A. Pooransingh, and J. Saitoo, "Efficient music identification using orb descriptors of the spectrogram image", *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 1–17, 2017.

[9] A. Shenoy and Y. Wang, "Key, chord, and rhythm tracking of popular music recordings", *Computer Music Journal*, vol. 29, no. 3, pp. 75–86, 2005.

[10] Z. Guo, Q. Wang, G. Liu, J. Guo, and Y. Lu, "A music retrieval system using melody and lyric", in *Proc. 2012 IEEE Int. Conf. Multimedia and Expo Workshops*, 2012, pp. 343–348.

[11] R. C. Chen, A. Ghobakhlou, and A. Narayanan, "Interpreting CNN models for musical instrument recognition using multi-spectrogram heatmap analysis: A preliminary study", *Frontiers in Artificial Intelligence*, vol. 7, p. 1 499 913, 2024.

[12] G. A. V. M. Giri and M. L. Radhitya, "Musical instrument classification using audio features and convolutional neural network", *Journal of Applied Informatics and Computing*, vol. 8, no. 1, pp. 226–234, 2024.

[13] R. Michon, J. O. Smith, M. Wright, C. Chafe, J. Granzow, and G. Wang, "Mobile music, sensors, physical modeling, and digital fabrication: Articulating the augmented mobile instrument", *Appl. Sci.*, vol. 7, no. 12, p. 1311, 2017.

[14] A. Acquilino and G. Scavone, "Current state and future directions of technologies for music instrument pedagogy", *Frontiers in Psychology*, vol. 13, 2022.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Modeling and Simulation of Autonomous Mobile Robot Fleets for Automated Baggage Handling in Airports 4.0: A Fuzzy Agent Approach

Alain-Jérôme Fougères
ECAM Louis de Broglie
IT and Telecommunications Laboratory
Rennes, France
email: alain-jerome.fougeres@ecam-ldb.fr

Ouazna Oukacha
ECAM Louis de Broglie
IT and Telecommunications Laboratory
Rennes, France
email: ouzna.oukacha@ecam-ldb.fr

Moïse Djoko-Kouam
ECAM Louis de Broglie
IETR, UMR CNRS 6164
Rennes, France
email: moise.djoko-kouam@ecam-ldb.fr

Egon Ostrosi
UTBM
ERCOS/ELLIAD EA4661
Belfort 90010, France
email: egon.ostrosi@utbm.fr

Abstract— **Airport 4.0 offers a set of methods, techniques, tools, and technologies that are naturally relevant for the development of services in the airports of the future. Integrating these different concepts into an airport is not without its challenges, given the complexity of the systems in place. Modeling and simulation phases have therefore become essential to ensure the success of these integrations. The concept of Agent-Based Modeling and Simulation (ABMS) is perfectly suited to this context. In this paper, we presented a case study involving the simulation of autonomous vehicle fleets for baggage handling in a simplified airport, in which each vehicle is simulated by an agent with fuzzy knowledge (fuzzy agent). We developed a UML (Unified Modeling Language) model of the system, providing static and dynamic views of a simplified airport circuit, vehicles, and baggage. Three different strategies were tested to determine the number of autonomous vehicles needed to most efficiently handle baggage arriving at the two entry points. Then, we presented our simulation results. These results allowed us to highlight the impact of various parameters, such as the number of simulations, the number of bags processed, the total simulation duration, and the total number of bags processed per hour. A discussion on the regulation of vehicle speeds is also included.**

Keywords- automatic baggage handling; autonomous mobile robots; ABMS; fuzzy agent; airport 4.0.

I. INTRODUCTION

This article considerably expands upon our previous conference paper [1] by providing additional information on all parts, especially by developing the fuzzy agent-based modeling section and its application for the case study. Industry 4.0 is the fourth industrial revolution after the invention of the steam engine, mechanization and mass production, computerization and robotization [2]. It brings the concepts of the Internet of Things (IoT), Cyber-Physical Systems (CPS), Machine to Machine (M2M), and intelligent robotics, including Autonomous Mobile Robots (AMR) [3]. Industry 4.0 assumes decentralized decision-making, interoperability, cyber assistance, predictive maintenance, eco-design and is user centered [4]. Industry 5.0 has complemented the previous one by amplifying the consideration of humans with Human machine connectivity and co-existence [5] and [6].

The concept of Industry 4.0 has extended to other types of structures and sectors [7], such as Supply Chain 4.0 [8] [9], Agriculture 4.0 [10] and [11], Health 4.0 [12], or Airport 4.0 [13]. However, the deployment of AMR fleets in the context of Airport 4.0 raises several challenges, all related to the actual level of autonomy of these "intelligent" vehicles: decision-making to maintain a required level of performance in carrying out tasks, traffic flow, vehicle localization, fault detection, collision avoidance, vehicle perception in changing environments, as well as acceptance by users and operators [14]. Therefore, the simulation, prior to the deployment of autonomous vehicles, makes it possible to consider the various constraints and requirements formulated by manufacturers and future airport users [15].

The use of simulation is justified by the high cost associated with developing experiments or tests on the real system, observing the behavior of static or dynamic processes and entities in the real world, or building a physical model [16]. Thus, in the context of Airport 4.0, the main advantages of simulating the operations carried out by an AMR fleet are the reduction of fleet development time and cost, the minimization of potential operational risks associated with the deployment of vehicles in a space shared with humans, and the verification of fleet performance. This makes it possible to assess the feasibility of different scenarios for the circulation of autonomous vehicles at a strategic or operational level, the possibility of a rapid understanding of the operations carried out by these vehicles, the identification of improvements in the layout configurations of vehicle circulation areas [7], and safety assessment during coexistence and possible interactions between autonomous vehicles and human operators [17].

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Numerous modeling and simulation approaches can be used in the field of Airport 4.0: Agent-Based Modelling and Simulation (ABMS), Discrete Event Simulation (DES), System Dynamics (SD), Virtual Reality (VR), Augmented Reality (AR), Artificial Intelligence (AI), Petri Nets (PN), Digital Twins (DT), and Hybrid Simulation (HS) [18]. ABMS is an increasingly used approach for modeling and simulating complex systems composed of autonomous agents interacting with each other and with their environment [19]. Thus, a strong analogy can therefore be drawn between a virtual agent and a physical (mobile) robot.

An autonomous baggage handling system in an airport is a complex and highly adaptive transportation system. So, different types of simulation frameworks and environments have been proposed for simulating complex systems involving AMRs capable of transporting luggage, mainly discrete event systems [20] or agent-based systems [21].

Numerous agent-based approaches are proposed for the modeling and simulation of both autonomous vehicles and transport systems [22]. They offer two-level simulation contexts, micro as well as macro, ranging from trajectory planning to optimal task allocation, while enabling collision and obstacle avoidance [23], addressing problems of traffic congestion, parking requirements, environmental implications or even the performance of autonomous vehicle systems [24].

One of the main problems of complex systems, such as automatic baggage handling systems in airports, results from the acquired or transmitted data, which may be uncertain, insufficient or available in a fragmented manner due to the dynamics of the environments and the variation of acquisition times. Fuzzy logic then appears as a good solution to model and simulate the uncertainty and the unknown in these complex and adaptive systems [25] and [26]. More specifically, it will allow for the adequate handling of variations in the incoming baggage flow, which requires adapting not only the number of vehicles handling the baggage, but also the distribution of these vehicles according to the different baggage entry points.

In [27] and [28], Zadeh defines fuzzy logic and the notion of fuzzy sets, introducing the notion of linguistic variables whose values are generally vague, fuzzy, or relative, such as “low”, “medium”, “high”, “most”, or “a certain number”. By defining these linguistic variables (fuzzy sets) and their potential values (fuzzy quantification), as well as rules using them (fuzzy rules), it is then possible to build Fuzzy Inference Systems (FIS).

Fuzzy set theory is therefore particularly suited to the processing of uncertain or imprecise information that should lead to decision-making by agents simulating autonomous vehicles [29]. Also, the definition of fuzzy agents to manage the levels of imprecision and uncertainty involved in modeling the behavior of simulated vehicles seems quite appropriate [30].

Fuzzy agents can track the evolution of fuzzy information from their environment and from the agents themselves. By interpreting this fuzzy information, they can act and interact within a fuzzy multi-agent system. Thus, a fuzzy agent can discriminate fuzzy interaction values or fuzzy information acquired to assess its degree of affinity with other fuzzy agents or the degree of interest of this information in a specific situation.

Most of the control tasks performed by autonomous mobile robots (in simulated or real contexts) have been the subject of performance improvement studies using fuzzy logic: motion planning, navigation, and obstacle avoidance [31]; path planning [32], localization [33], and intelligent management of energy consumption [34]. However, although the simulation of AMR behavior based on fuzzy logic has been heavily studied since the 1990s [35] and [36] to the present day [37], mainly for navigation problems (trajectories and obstacle avoidance), no study addresses the simulation of mobile robot control in dynamic and uncertain environments in an exhaustive manner using fuzzy agents. This is the objective we set for ourselves, beginning by addressing the methodological framework for modeling these fuzzy agents. Furthermore, the case study in the Airport 4.0 domain, which we develop in Section III, provides us with an ideal framework for experimenting with fuzzy agent-based simulations that are consecutive, complementary, and correlated to the modeling phase.

To enhance the readability of this case study, we provide below some details about the research context that led to this publication. The academic collaboration between the authors, focusing on the dual theme of fuzzy logic and autonomous mobile robots, played an important role in the development of this article by combining expertise from each respective field. In addition, the industrial collaboration within the framework of a multi-partner collaborative project named ALPHA also had a significant place in this same research context.

The ALPHA project brought together a four-party consortium, including two industrial partners and two research laboratories. The aim of the project was to identify and validate a mobile robotics solution for the transportation of unit baggage in airports. The ALPHA project focused particularly on two key challenges: on the one hand, optimizing an AMR fleet based on minimal energy consumption; on the other hand, minimizing the complexity of the robotic solution under the constraint of many AMRs.

In the remainder of this article, we present the context of fuzzy multi-agent modeling, which allows us to simulate complex systems such as baggage handling systems in airports. In Section III, we propose a case study on simulation of AMR fleets for baggage handling in a greatly simplified airport area. Three strategies are presented with three different data sets. In Section IV, we analyze the results obtained by each of the strategies according to the performance criteria defined in the previous Section. Then, in Section V, we discuss the results obtained by varying speeds of AMRs. Finally, we conclude on the proposed fuzzy agent-based modeling and simulation, and then we present the perspectives for improvement in the short term.

II. FUZZY AGENT-BASED MODELING

The multi-agent paradigm is well-suited for modeling and simulating complex systems such as autonomous vehicles, which interact with each other and react within

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

their dynamic environment [38]. However, for this approach to be effective, a rigorous modeling framework is necessary: the agent-based system must be characterized, quantified, and structured, and the agents' behaviors must be formally defined [19]. Furthermore, equipped with reasoning and decision-making capabilities based on fuzzy knowledge, the agents will be able to simulate more complex, and more realistic, situations [39]. The agents we design in this case study are therefore fuzzy.

A. *Fuzzy agent-based application modeling*

The modeling of an agent-based system or application, often discussed from a process or methodological point of view [40], requires adopting a local vision, to respect the fact that each agent is responsible for their knowledge and actions (agent autonomy), which are often decentralized.

Different models can be used to propose a methodology for designing an agent-based application, including: an agent model, to define the characteristics of an agent; a task model, to represent the tasks that can be performed by agents; an expertise model, to describe the knowledge of agents; a coordination model, to define the protocols and interactions between agents; an organization model, to describe the organization of the agent society; or a communication model, to describe the interactions between agents and users. Our agent modeling work mainly refers to the Agent Unified Modeling Language (AUML) [41].

In [42], we proposed a four-phase agent-based system modeling methodology (Figure 1):

1) Create use case diagrams to represent the services provided by the system.
2) For each use case diagram, create sequence diagrams specifying the interactions (message exchanges and scheduling) between the agents involved in these reference cases.
3) From the sequence diagrams, which allowed us to identify the agents, the system objects and their interactions, create the class diagram: the objects are associated with classes, the messages exchanged (service requests between objects) are translated by operations on the classes, the parameters associated with the operations are translated into class attributes (this class diagram can possibly be completed by a collaboration diagram).
4) From the class diagram, define the behavior of each agent (agent class) by means of a state or activity diagram (we also used Petri nets).

The description of the roles played by the different cooperative agents mainly focuses on collaboration and sequence diagrams. We subsequently integrated an expertise model into the previous method, particularly in the form of knowledge and fuzzy inference rules [43].

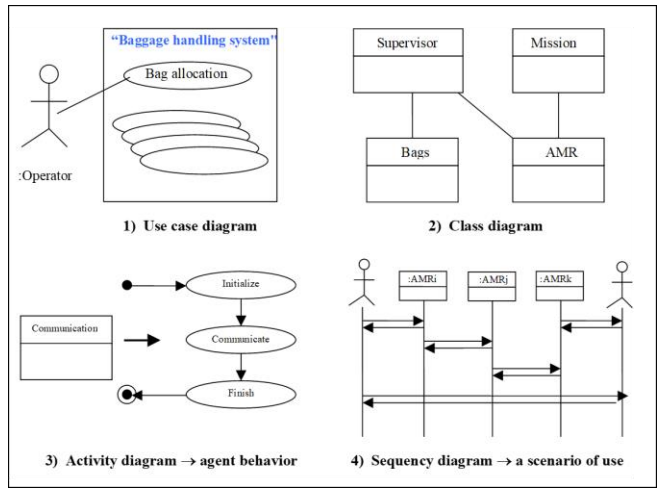


Figure 1. Agent-based system design: methodology adapted from AUML

B. *Fuzzy agent modeling*

An agent-based system is fuzzy if some of the agents that compose it sometimes have fuzzy behaviors, or if a part of the knowledge they use is fuzzy. This means that agents are modelled in such a way as to [30]:

- sometimes use fuzzy knowledge in their inferences; this knowledge then consists of fuzzy linguistic variables, fuzzy linguistic values, and fuzzy rules [29];
- sometimes adopt fuzzy behaviors, consequences of their fuzzy inferences [44];
- and potentially implement fuzzy interactions, act in fuzzy organizations, or play fuzzy roles [45].

Formally, a fuzzy agent-based system and the fuzzy agents that compose it can be defined as follows respectively (1) and (2):

$$\widetilde{M}_{\alpha} = < \widetilde{A}, \widetilde{I}, \widetilde{P}, \widetilde{O} > \quad (1)$$

Where $\widetilde{A}$ is a set of fuzzy agents; $\widetilde{I}$ is a set of fuzzy interactions between fuzzy agents; $\widetilde{P}$ is a set of fuzzy roles that fuzzy agents can perform; $\widetilde{O}$ is a set of fuzzy organizations defined for fuzzy agents (subsets of strongly linked fuzzy agents).

$$\widetilde{\alpha}_i = \left\langle \Phi_{\Pi(\widetilde{\alpha}_i)}, \Phi_{\Delta(\widetilde{\alpha}_i)}, \Phi_{\Gamma(\widetilde{\alpha}_i)}, K_{\widetilde{\alpha}_i} \right\rangle \quad (2)$$

Where, for a fuzzy agent $\widetilde{\alpha}_i$, $\Phi_{\Pi(\widetilde{\alpha}_i)}$ is its function of observation; $\Phi_{\Delta(\widetilde{\alpha}_i)}$ is its function of decision; $\Phi_{\Gamma(\widetilde{\alpha}_i)}$ is its function of action; and $K_{\widetilde{\alpha}_i}$ is its set of fuzzy knowledge.

In the case study presented in the following section, we will primarily develop the modeling of agents' fuzzy knowledge. This knowledge modeling is above all the result of methodological activity. Thus, the methodology we adopted comprises three steps:

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

1) the modeling of fuzzy data, that is to say the definition of the set *L(V, X, P)*, where *V* is the set of linguistic variables (for instance, "speed"), *X* is the set of possible values, the universe of discourse (for instance, "5 m/s"), and *P* is the set of partitions of the universe of discourse, called fuzzy subsets (for instance, "medium");
2) the quantification or fuzzification which consists of giving numerical values to the fuzzy subsets, to be able to determine that a given data has a membership value to one or more fuzzy subsets of the model (for instance, medium speed $\in$ [3, 6] in m/s);
3) the modeling of fuzzy rules so as to make fuzzy inferences from the fuzzy data supplied to the model, which consists of producing decision rules of the type: *IF A* $\in$ *V IS B* $\in$ *P THEN C* $\in$ *V IS D* $\in$ *P* (for example, "IF speed IS medium AND intersection IS near, THEN slowdown IS little").

To refine this fuzzy agent modeling methodology, we have extensively tested it in various engineering fields, such as product configuration in collaborative design process [43], integrated CAD product and adaptive manufacturing cell formation [44], collaborative design platform [45], and engineering design semantics elaboration [46].

III. CASE STUDY: SIMULATION OF AUTONOMOUS VEHICLE FLEETS FOR BAGGAGE HANDLING

In this section, we present a case study of a baggage handling system in an airport using a fleet of autonomous mobile robots. We will successively propose the agent model of vehicle behavior, the fuzzy logic modeling of their decision rules, and three baggage handling strategies to discuss the performance obtained.

A. *Presentation of the case study*

The case study presented in this article proposes the simulation of the automatic baggage handling in a greatly simplified airport area comprising two baggage entry and exit flows (via baggage conveyor belt, for example), a parking lot equipped with battery charging stations, forty AMRs, and an operator who can intervene in case of a problem such as a malfunctioning AMR or baggage on the ground. The incoming baggage flow is simulated by the arrival of aircraft carrying a variable number of bags. As for AMRs, they can carry one bag at a time (in reality, baggage is classified by size, standard or specific, but we do not consider this distinction in our simulations).

The circuit followed by AMRs to transport bags from an entry point to an exit point is a loop circuit consisting of two branches, one for the two entry points and another for the two exit points (Figure 2). This circuit presents seven specific points that require AMR agents to adapt their behavior using strategies whose performance can be measured in both centralized (supervised agents) and decentralized (autonomous agents) modes. These strategies will require the agents to implement distributed artificial intelligence capacities (DAIC), such as communication (C), knowledge (K), and, with a view to improving the current simulator, learning (L) mechanisms. The correspondence between the problems inherent in these seven specific points and the strategies developed to address them is provided in Table I.

The baggage handling circuit of the airport has two baggage entry branches and two baggage exit branches, resulting in four different routes for the AMRs, with an average length of 313 m. The AMRs operate at four different speeds along the circuit: a slow speed of 3 m/s, a slightly higher speed of 4 m/s, a medium speed of 5 m/s, and a high speed of 7.5 m/s. Furthermore, baggage pickup and drop-off on the entry and exit points take 2 seconds.

The AMRs are simulated by fuzzy agents, and the airport circuit is represented by a directed graph. This graph has 17 nodes (Figure 3): node *P0* represents the parking lot; nodes *R1* and *R2* represent the 2 baggage pick-up points; nodes *D1* and *D2* represent the 2 baggage drop-off points; and the other 12 nodes *Pi* represent characteristic points of the circuit (curve start points, curve end points, convergence points and divergence points).

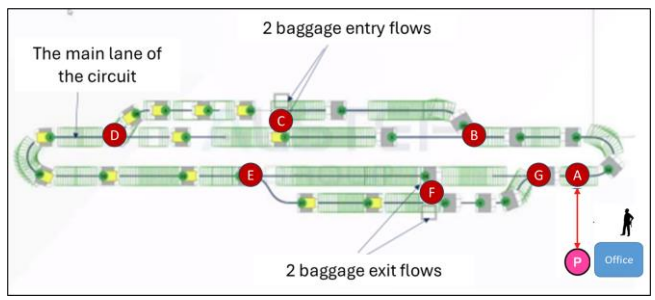


Figure 2. The circuit of the simulation Application: *A* is the input/output of the circuit; *B* and *E* are 2 points of divergence; *D* and *G* are 2 points of convergence; *C* are the 2 points of baggage collection; *F* are the 2 points of baggage drop-off; and *P* is the parking.

TABLE I. THE SEVEN PROBLEMATIC POINTS OF THE CIRCUIT

Points	Problem description	Strategies	DAIC
A	Management of AMR I/O on the circuit.	Round robin / On-demand / Prédiction	C, K, L
B	Managing the divergence between the two branches leading to R1 or R2	Random / Alternative / On-demand / Prediction	C, K
C	Baggage handling in R1 or R2	The absence of baggage is penalized.	C, K
D	Convergence management (collision avoidance) between the two branches comes from R1 or R2	Random / Alternative / Centralized or decentralized collision avoidance algorithm	C, K, L
E	Managing the divergence between the two branches leading to D1 or D2	Random / Alternative / On-demand / Prediction	C, K
F	Baggage drop-off management in D1 or D2	The absence of baggage is penalized	C, K
G	Convergence management (collision avoidance) between the two branches comes from D1 or D2	Random / Alternative / Centralized or decentralized collision avoidance algorithm	C, K, L

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

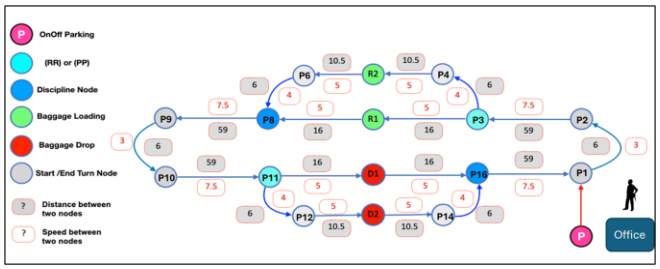


Figure 3. The circuit-directed-graph where nodes represent the characteristic points of the circuit, and edges are labeled with the segment lengths (distances between two nodes) and the permitted speeds on those segments.

The simulation application is developed in Python. Figure 4 presents the object and agent architecture of this simulator: Static objects are represented in blue, dynamic agents in red, and fuzzy agents in green. Infrastructure can be deployed in this environment. It includes a traffic plan (circuit composed of baggage entry and exit points, battery charging stations, and using a map), different types of test airports in which planes loaded with baggage arrive which will be processed by the AMRs (managing physical components, such as a Lidar, and abstract components such as a path), a supervisor which can possibly allocate baggage to AMRs according to specific strategies. Static or dynamic obstacles (e.g., operators or broken-down AMRs) can also be activated in this simulation environment.

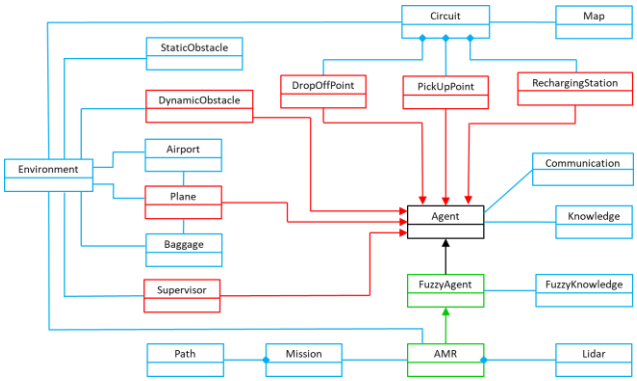


Figure 4. UML class-diagram of the fuzzy-agent based simulator.

The visualization of simulations and the evolution of performance indicators can be followed using the human-machine interface (HMI) shown in Figure 5. This HMI allows for the modeling and simulation of the dynamic behavior of the baggage handling system with its AMRs. It provides a graphical representation of the circuit elements and facilitates the analysis of flows and interactions at critical points. The main functionalities of this HMI are as follows:

- visualization of the complete circuit with the main lane, the two-baggage entry and exit branches, and the parking area,
- configuration of the parameters of AMR agents (speed, capacity, traffic logic and strategy, baggage arrival scenario),
- selection and control/tracking of a specific AMR agent,
- definition of incoming and outgoing baggage flows,
- visualization of queues/blockages at critical points in the circuit: intersections, convergence and divergence zones,
- visualization of performance indicator values distributed across three levels: simulation, bags, and AMR agent,
- monitoring battery charging in the parking lot,
- display of messages defined in the application in the console.

This interface is therefore an essential tool for testing different AMR and baggage management strategies, for analyzing the impact of flow variations on system performance, and for optimizing the circuit configuration before any physical experimentation. Thus, it allows us to measure indicators as varied as the maximum number of bags/h passing through the main track, the maximum number of bags/h passing through the two entry points, the maximum number of bags/h passing through the two exit points, or the number of AMR agents required for a maximum flow.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

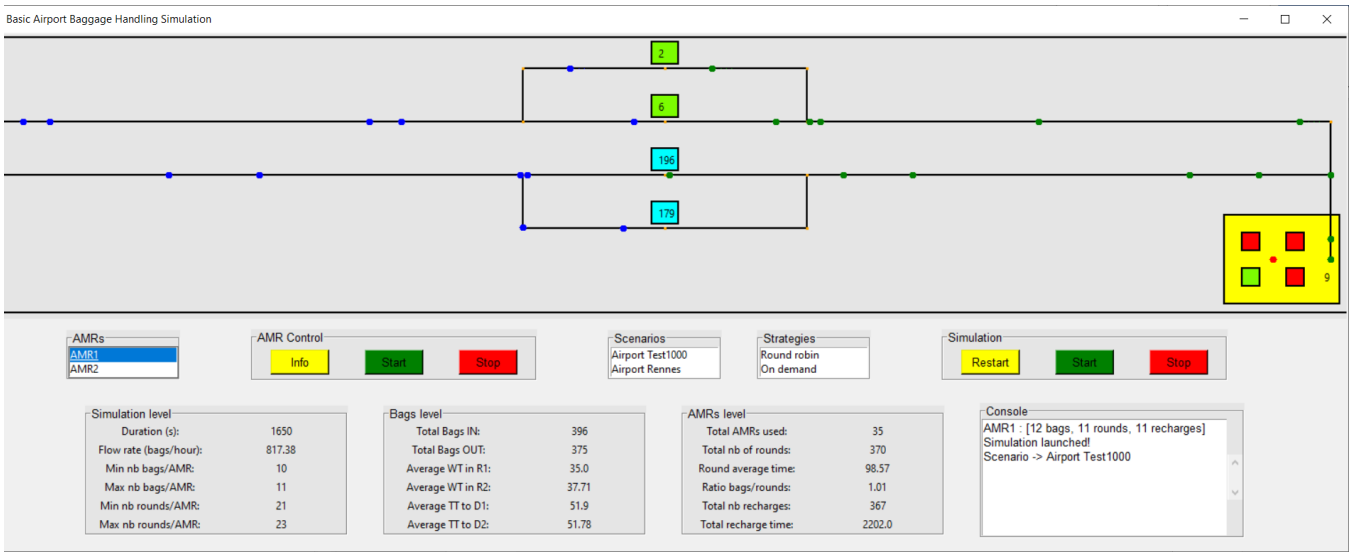


Figure 5. HMI of the simulation application.

To study the performance of the solutions considered in our simulations, we defined an optimization system where the objectives are:

1) to minimize the number of AMR, noted as x;
2) to maximize the bag throughput per hour, noted as y,
3) to minimize the recharge time of an AMR per hour, noted as z.

In ideal conditions, an AMR picks up one bag each turn, a level of performance that we will analyze in Section IV, based on three types of strategies deployed by the AMRs.

This optimization system (OS) is defined in Figure 6, where $Max(x)$ is the maximum number of AMRs; L_{avg} is the average length of the circuit; d_s is the safety distance between 2 AMRs; C_{bat} is the average capacity of a battery; t_0 is the average charging time of a battery; t_1 is the average waiting time for a battery recharge; T_0 average duration of a circuit lap; t_2 is the time to pick up a bag; t_3 is the time to drop off a bag; T_{avg} is the average number of revolutions made by an AMR during one hour; and v_{avg} is the average speed of the AMRs on the circuit.

$$OS = \begin{cases} 0 \le x \le Max(x) = {}^{L_{avg}}/_{d_s} \\ 0 \le y \le T_{avg} * Max(x) \\ 0 \le z \le 3600 \\ z = \left({}^{v_{avg}*3600}/_{C_{bat}}\right) * (t_0 + t_1) \\ T_0 = \left({}^{L_{avg}}/_{v_{avg}}\right) + (t_2 + t_3) \\ T_{avg} = {}^{(3600-z)}/_{T_0} \\ y = T_{avg} * x \end{cases}$$

Given:

$L_{avg} = 313$, $v_{avg} = 5$,
$d_s = 8$, $t_1 = t_2 = 2$
$z \approx 10\%$ of AMR's activity, if $t_1 = 0$

Then for an hour:

$z = 10\% * 3600 = 360\ s$
$Max(x) = L_{avg} / d = 313 / 8 \approx 40$ AMRs
$T_0 = (L_{avg} / v_{avg}) + (t_2 + t_3) = 313 / 5 + 4 \approx 67\ s$
$y = \left({}^{(3600-z)}/_{T_0}\right) * x = 1897$ bags

Figure 6. Optimization system defined for this study.

An initial study focused on the sizing of the problem (traffic plan, maximum number of AMRs, determination of applicable speeds, distances between AMRs, energy consumption of these AMRs, etc.) [15], followed by a Petri nets-based simulation of the AMR behavior in function of strategies of baggage handling developed (Figure 7), allowed us to set the following parameters:

- $Max(x) = 40$, $L_{avg} = 313$, $v_{avg} = 5$, $d_s = 8$, $t_2 = t_3 = 2$,
- and $z \approx 10\%$ of the AMRs operating time if t_1 is equal to zero.

Following this initial study, we conducted our first simulations to determine the optimal number of AMRs to ensure the best baggage throughput on the route. The simulations involved 1,000 bags processed by AMRs implementing the "Round Robin" strategy. We varied the number of AMRs from 23 to 45 and obtained the best result with 31 AMRs (the shortest time to process the 1,000 bags), as shown in Figure 8. Figure 9 complements this study by showing the number of rounds traveled by all these AMRs during the different simulations, thus illustrating the energy impact associated with increasing or decreasing the number of AMRs (the distance traveled being converted into electrical consumption).

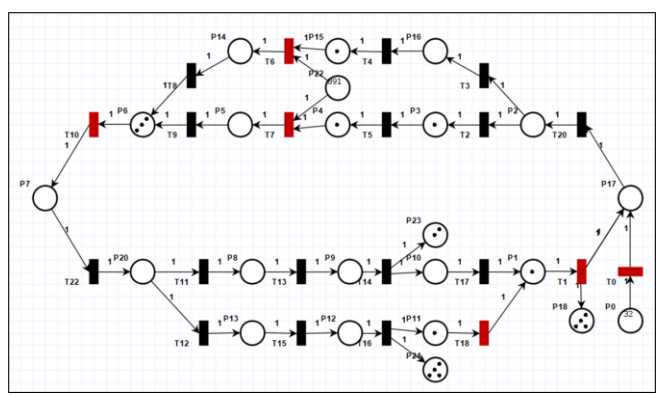


Figure 7. Petri net model of the circuit in Figure 2. It allows us to simulate a solution, while verifying that it retains the properties of liveness (non-blocking) and bounding of quantitative parameters such as the number of AMRs. The transitions in red indicate the possible evolution of the Petri net.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

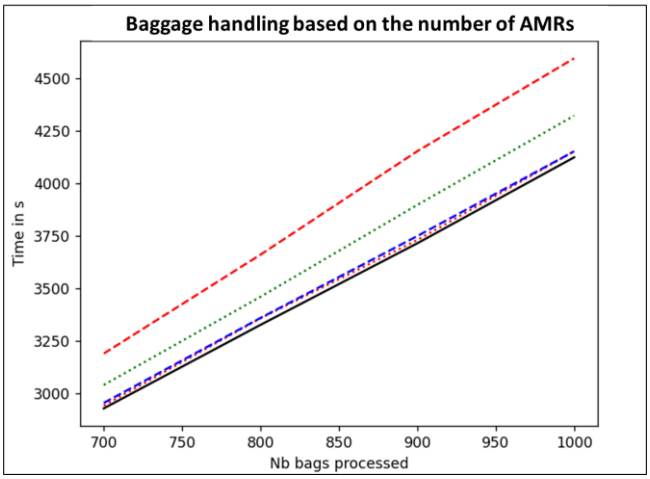


Figure 8. Partial graph representing the time taken to process 1000 bags by an AMR fleet comprising several vehicles varying from 23 to 39: 23 (red dashes), 27 (green dotted lines), 31 (solid black line), 35 (red dotted lines), and 39 (blue dashes).

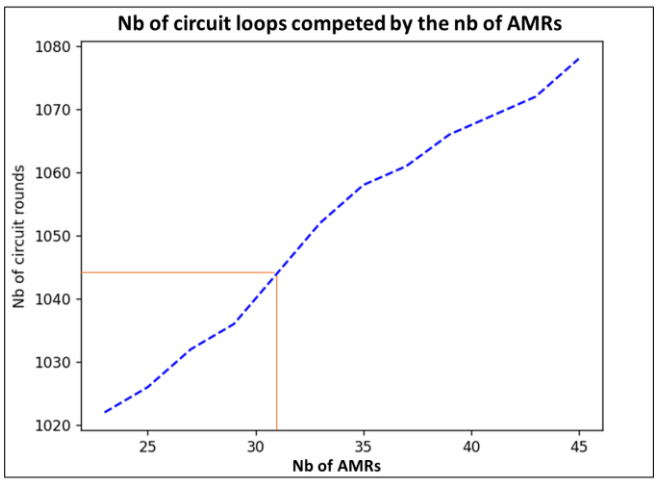


Figure 9. Graph representing the number of circuit laps completed by an AMR fleet comprising several vehicles varying from 23 to 45 (the optimum for bag processing with 31 AMRs is located on the curve).

B. *Behaviors of AMRs*

As defined in the agent-based systems design methodology, one of the steps is to model the behavior of the agents in the form of an activity or state diagram. The AMR agents in the simulation will have a behavior adapted to the strategy implemented to process the baggage.

Thus, Figures 10 to 12 present the Petri nets modeling the behavior of AMR agents, respectively: according to the "Round robin" strategy, which consists of continuously looping on the circuit as long as there is baggage to process (Figure 10); according to the "On-demand" strategy, which consists of allocating missions to AMRs when baggage arrives on the entry points (Figure 11); and according to the fuzzy logic prediction strategy (noted "FL Prediction" strategy), which consists of forecasting the need for AMRs in circulation based on baggage arrivals and the number of AMRs already present on the circuit (Figure 12).

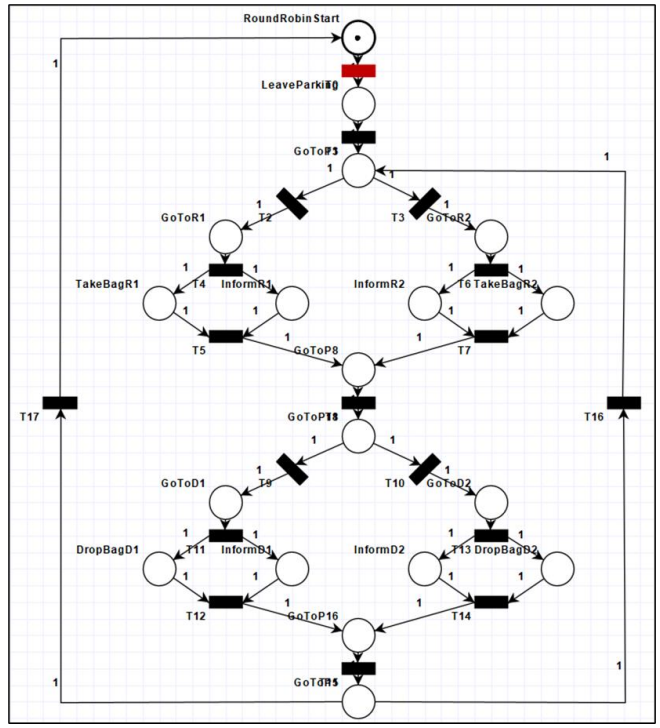


Figure 10. AMR agent behavior based on "Round robin" strategy

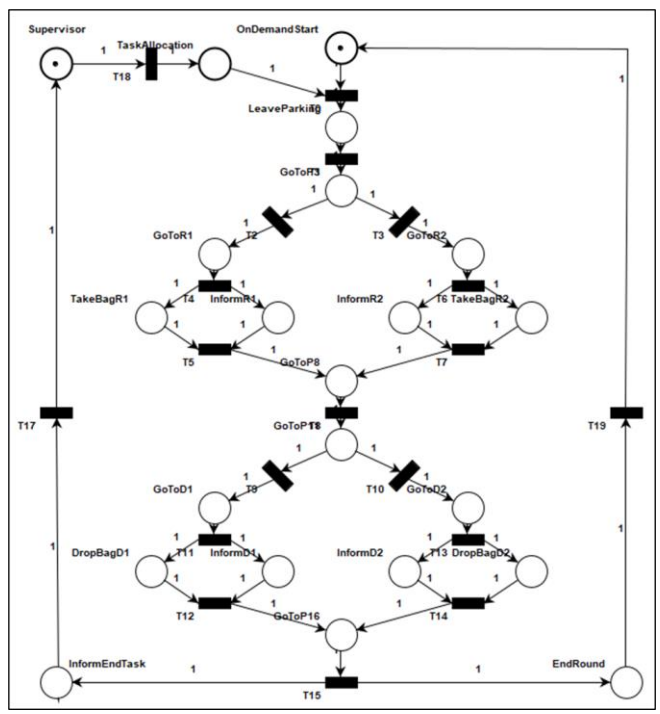


Figure 11. AMR agent behavior based on "On-Demand" strategy

C. *Fuzzy knowledge of AMR agents*

AMR agents are fuzzy agents that therefore have fuzzy knowledge. In this simulation, 2 types of fuzzy inferences are considered: the prediction of the number of AMRs that must circulate to process baggage, and the determination of

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

the branch to choose to take a baggage (passage through R1 or R2). For the fuzzy knowledge acquisition phase, and therefore the determination of the linguistic variables and their different value ranges, we did not consult a domain expert. We defined an initial ad hoc model, which we refined through our simulations.

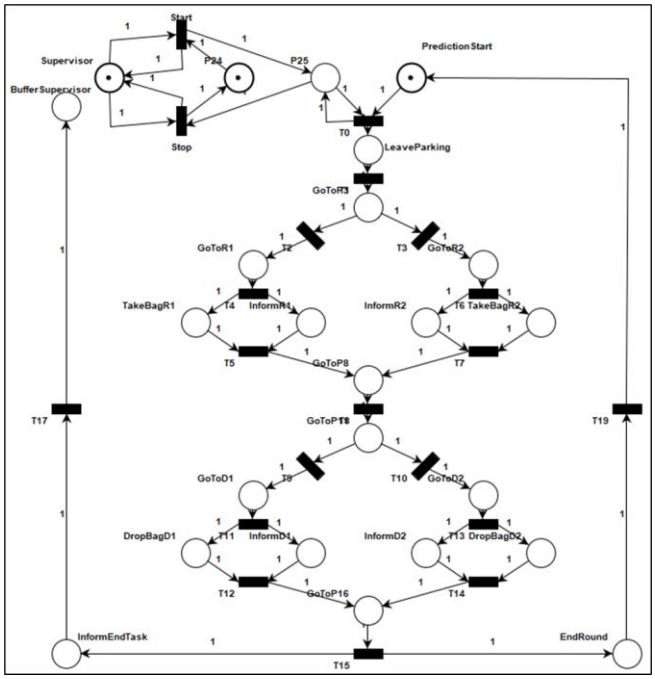


Figure 12. AMR agent behavior based on “Prediction” strategy

To predict the number of AMRs expected to circulate on the circuit at a given time, AMR fuzzy agents use 3 linguistic variables (Figure 13) and 9 rules (Table II):

- **NbBag**. This variable represents the number of bags waiting at entry points *R1* and *R2*. It is described by three linguistic values: *low*, *medium*, and *high*. The universe of discourse varies depending on the simulations and the requirement for the maximum number of bags waiting. For the present case study, we defined the interval from 0 to 20 bags. Thus, *low*, whose membership function is trapezoidal, has the value interval [0, 6]; *medium*, whose membership function is triangular, has the value interval [2, 14]; and *high*, whose membership function is trapezoidal, has the value interval [10, 20].
- **NbAMR**. This variable represents the number of AMRs circulating on the circuit at a given time. It is described by three linguistic values: *low*, *medium*, and *high*. The universe of discourse varies depending on the simulations and the circuit size (depending on the circuit, an initial study determines the maximum number of AMRs that can circulate while respecting safety distances). For the present case study, we have defined the interval from 0 to 40 AMRs. Thus, *low*, whose membership function is trapezoidal, has the interval [0, 15]; *medium*, whose membership function is triangular, has the interval [5, 25]; and *high*, whose membership function is trapezoidal, has the interval [15, 30].
- **Prediction**. This variable indicates the number of AMRs that should be brought into the circuit or removed from the circuit at a given time. It is described by five linguistic values: *highDecrease*, *decrease*, *stable*, *increase*, and *highIncrease*. The universe of discourse varies according to the frequency of the predictions (if this frequency is low, the interval will be wider than that of a higher frequency). For the present case study, we have defined the interval as -10 to 10. Thus, *highDecrease*, whose membership function is trapezoidal, has the interval [-10, -5]; *decrease*, whose membership function is triangular, has the interval [-10, 0]; *stable*, whose membership function is triangular, has the interval [-5, 5]; *increase*, whose membership function is triangular, has the interval [0, 10]; and *highIncrease*, whose membership function is trapezoidal, has the value interval [5, 10].

TABLE II. THE 9 RULES USED FOR FUZZY INFERENCE "PREDICTION"

Rule ID	Rule
R1	IF NbBag IS Low AND NbAMR IS High HEN Prediction IS HighDecrease
R2	IF NbBag IS Low AND NbAMR IS Medium THEN Prediction IS Decrease
R3	IF NbBag IS Low AND NbAMR IS Low THEN Prediction IS Stable
R4	IF NbBag IS Medium AND NbAMR IS High THEN Prediction IS Decrease
R5	IF NbBag IS Medium AND NbAMR IS Medium THEN Prediction IS Stable
R6	IF NbBag IS Medium AND NbAMR IS Low THEN Prediction IS Increase
R7	IF NbBag IS High AND NbAMR IS High THEN Prediction IS Stable
R8	IF NbBag IS High AND NbAMR IS Medium THEN Prediction IS Increase
R9	IF NbBag IS High AND NbAMR IS Low THEN Prediction IS HighIncrease

To choose which branch leading to R1 or R2 they will take at a given time, the AMR fuzzy agents use five linguistic variables: (Figure 14) and 18 rules (Table III):

- **NbBagInR1** and **NbBagInR2**. These variables represent the number of bags waiting at entry points *R1* and *R2*. They are described by three linguistic values: *low*, *medium*, and *high*. The universe of discourse varies depending on the simulations and the requirement for the maximum number of bags waiting. For the present case study, we defined the interval from 0 to 100 bags. Thus, *low*, whose membership function is trapezoidal, has the value interval [0, 20]; *medium*, whose membership function is triangular, has the value interval [0, 50]; and *high*, whose membership function is trapezoidal, has the value interval [30, 100].

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

- **NbAMRToR1** and **NbAMRToR2**. These variables represent the number of AMRs circulating on the circuit at a given time. They are described by three linguistic values: *low*, *medium*, and *high*. The universe of discourse varies depending on the simulations and the branch size. For the present case study, we have defined the interval from 0 to 10 AMRs. Thus, *low*, whose membership function is trapezoidal, has the interval [0, 3]; *medium*, whose membership function is triangular, has the interval [1, 5]; and *high*, whose membership function is trapezoidal, has the interval [3, 10].
- **GoTo**. This variable indicates the branch that the AMRs should take at a given time. It is described by two linguistic values: *R1* and *R2*. The universe of discourse is defined as the interval 0 to 10, and the two membership functions of *R1* and *R2* are trapezoidal. Thus, *R1* has the interval of value [0, 6] and *R2*, has the interval of value [4, 10].

TABLE III. THE 18 RULES USED FOR FUZZY INFERENCE "GOTO"

Rule ID	Rule
R1	`IF NbBagInR1 IS Low AND NbAMRToR1 IS High THEN GoTo IS R2`
R2	`IF NbBagInR1 IS Low AND NbAMRToR1 IS Medium THEN GoTo IS R2`
R3	`IF NbBagInR1 IS Low AND NbAMRToR1 IS Low AND NbAMRToR2 IS Medium THEN GoTo IS R1`
R4	`IF NbBagInR1 IS Medium AND NbAMRToR1 IS High THEN GoTo IS R2`
R5	`IF NbBagInR1 IS Medium AND NbAMRToR1 IS Medium THEN GoTo IS R1`
R6	`IF NbBagInR1 IS Medium AND NbAMRToR1 IS Low THEN GoTo IS R1`
R7	`IF NbBagInR1 IS High AND NbAMRToR1 IS High THEN GoTo IS R2`
R8	`IF NbBagInR1 IS High AND NbAMRToR1 IS Medium AND NbAMRToR2 IS Medium THEN GoTo IS R1`
R9	`IF NbBagInR1 IS High AND NbAMRToR1 IS Low THEN GoTo IS R1`
R10	`IF NbBagInR2 IS Low AND NbAMRToR2 IS High THEN GoTo IS R1`
R11	`IF NbBagInR2 IS Low AND NbAMRToR2 IS Medium THEN GoTo IS R1`
R12	`IF NbBagInR2 IS Low AND NbAMRToR2 IS Low AND NbAMRToR1 IS Medium THEN GoTo IS R2`
R13	`IF NbBagInR2 IS Medium AND NbAMRToR2 IS High THEN GoTo IS R1`
R14	`IF NbBagInR2 IS Medium AND NbAMRToR2 IS Medium THEN GoTo IS R2`
R15	`IF NbBagInR2 IS Medium AND NbAMRToR2 IS Low THEN GoTo IS R2`
R16	`IF NbBagInR2 IS High AND NbAMRToR2 IS High THEN GoTo IS R1`
R17	`IF NbBagInR2 IS High AND NbAMRToR2 IS Medium AND NbAMRToR1 IS Medium THEN GoTo IS R2`
R18	`IF NbBagInR2 IS High AND NbAMRToR2 IS Low THEN GoTo IS R2`

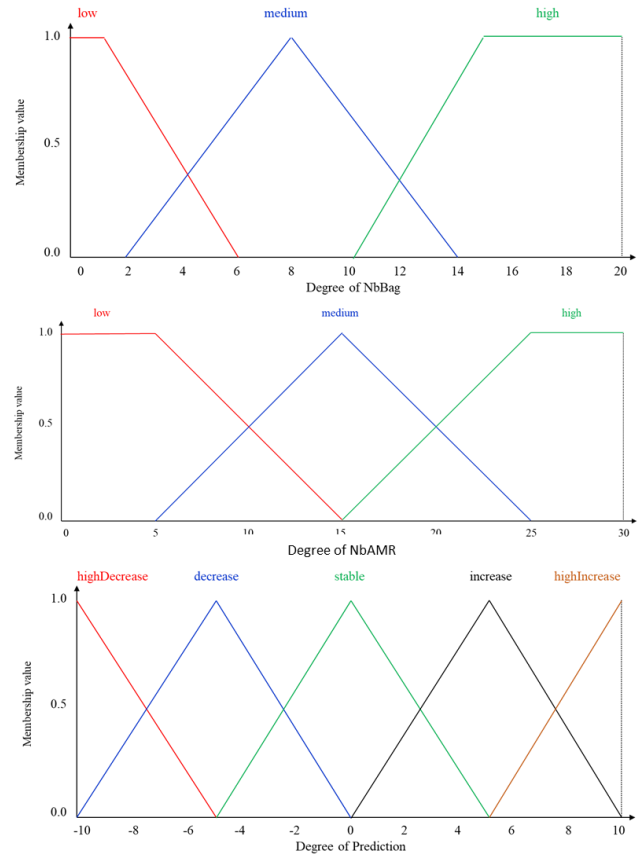


Figure 13. Linguistic variables to predict whether an AMR should circulate.

D. *The three strategies considered*

The objective is to determine an efficient strategy for handling baggage arriving at the two entry points of the circuit, based on two criteria: ensuring the best baggage throughput and limiting the energy consumption of the AMRs (i.e., minimizing the total distance traveled by the fleet's AMRs). Three distinct strategies were simulated to test them and establish their performance on the system:

- **Round robin**. The AMR agents perform a continuous rotation around the circuit and pick up baggage if any is available at the circuit's entry point according to their route (passing through R1 or R2).
- **On-demand**. Baggage arriving at R1 or R2 is assigned to available AMR agents in the parking area or along the route. An AMR agent that has dropped off baggage at D1 or D2 returns to the parking area if it has no new baggage assigned.
- **Fuzzy logic-based demand prediction**. The number of AMR fuzzy agents rotating around the circuit is calculated periodically based on predicted needs using fuzzy rules established by the operator. Other types of predictive strategies could be used, but as we have already shown the interest of fuzzy logic to efficiently solve this type of problem [47], we decided to develop this approach in this comparison of strategies.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Baggage arrival is a determining factor in the observable results for the three previous strategies. We have therefore defined different types of scenarios to cover many cases. Below we present the three main scenarios (without their variants):

- **Sc1 – Real data.** Baggage arrival is simulated based on aircraft arrivals at a sample airport, to better account for the variability of incoming baggage flow (high demand periods and low demand periods). These data cover a full day's traffic at this airport (for instance, the French airport of Rennes).

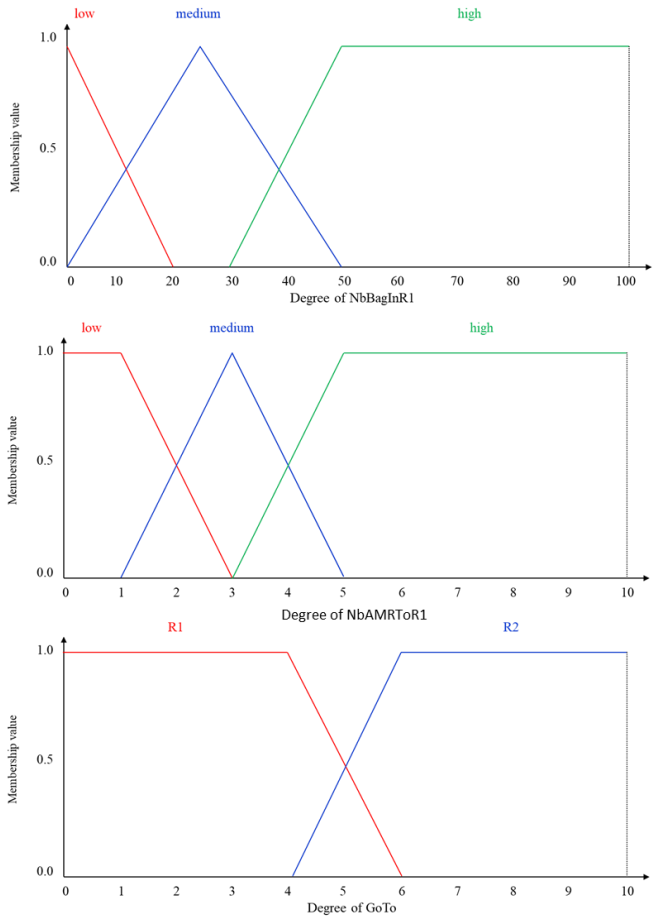


Figure 14. Linguistic variables for an AMR to choose one of the 2 branches.

- **Sc2 – Random Data**. Baggage arrival is randomly simulated over a period of one hour. This random data generation is performed upstream to create a single data set. This data set is then used to test the three strategies.
- **Sc3 – Mass data**. 1000 bags arrive continuously at the 2 entrances of the circuit. This involves performing a stress test for each strategy, regardless of the quality of each strategy's performance (e.g., a baggage waiting time threshold).

IV. RESULTS

Now, we present the results of the nine simulations carried out according to the specifications formulated in the previous Section (three scenarios were tested, each using three different strategies). The five performance criteria selected, in accordance with the optimization system presented in Figure 6, are:

1) The simulation duration, i.e., the processing time for all baggage entering the simulated scenario, as well as the battery recharge time for all AMR agents that have been in operation.
2) The hourly throughput, i.e., the number of bags processed per hour (adjusted to an hourly rate if necessary).
3) The number of circuit loops completed by the AMR agents to process all the baggage.
4) The average waiting time for baggage at entry points before it is taken over by an AMR agent.
5) The impact of AMR agent charging times on performance (this impact is estimated by default at 10% of the AMR operating time).

We will detail the results for each of the five criteria, then we will present the synthesis of these simulations.

Results regarding total simulation duration (Table IV)**.** With the exception of the 1 hour scenario in random baggage flow, the "Round robin" strategy always has the longest duration, the "On-demand" and "FL Prediction" strategies are more variable, with a shorter duration for "On-demand" when the baggage flow is continuous (test Mass data), and a lower duration for "FL Prediction" when the flow is more variable (flow depending on the arrival of aircraft at test Real data). Both "FL Prediction" and "On-demand" strategies optimize the processing time by activating AMR agents when baggage arrives by determining the right traffic branches, which is not the case for AMR agents in the "Round robin" strategy.

Results regarding the total number of bags processed (Table V)**.** Two of the scenarios have a fixed number of bags (the test airport called Real data with a fixed number of bags of 1306, and the Mass data scenario with 1000 continuous bags); for the third scenario, with a random flow over 1 hour, the "On-demand" strategy is the least efficient (1864 bags), then the "LF Prediction" strategy (1956 bags), finally the most efficient is the "Round robin" strategy (2015 bags). On this criterion, the "Round robin" strategy is overall the most satisfactory, it is also on this criterion that it has its main advantage.

Results regarding the overall throughput of the 3 strategies (Table VI)**.** The two strategies of "Round robin" and "On-Demand" have the worst overall results. In continuous and variable flows, the "Round robin" strategy performs poorly (respectively 1578 and 1713 bags/h). On the other hand, in random flow over 1 hour, the "On-demand" strategy performs the least (1864 bags/h). The average and overall flow rate is unquestionably the best with

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

the "FL Prediction" strategy (1807 bags/h on average over the 3 scenarios).

Results regarding the total number of loops made by the AMR agents (Table VII)**.** For this criterion, the "Round robin" strategy is systematically the least efficient, and this significantly. 1664 rounds on average over the 3 scenarios for the "Round robin" strategy, against 1526 rounds on average for the "FL Prediction" strategy and 1390 rounds on average for the "On-demand" strategy. The average and overall number of AMR agent loops is undoubtedly the best with the "On-demand" strategy (1390 rounds on average over the 3 scenarios); the allocation of an arriving bag to an AMR agent is indeed very efficient. For the "FL Prediction" strategy, the average results remain satisfactory (1526 rounds on average over the 3 scenarios), while those of the "Round robin" strategy are rather mediocre (1664 rounds on average over the 3 scenarios).

Results regarding the average waiting time for bags before being processed by an AMR agent (Table VIII)**.** For this criterion, the "Round robin" strategy is twice the least efficient (for the one-hour random flow and for the variable flow of the Real data test), and the "On-demand" strategy is the least efficient for the continuous flow of 1000 bags. The "On-demand" strategy gives satisfactory results on average for the 3 scenarios (22 s average wait for a bag), and the "FL Prediction" strategy also gives satisfactory performances for the three scenarios (23 s average wait for a bag).

Results regarding the performance impact of AMR agent recharge times (Table IX)**.** The AMR agents run the same algorithm to determine whether they should recharge at a charging station located in the parking lot. The principles of the algorithm are as follows: 4 stations (corresponding to 10% of the 40 AMRs), recharge every 10 laps of the circuit (every 10 x 66s = 660s, on average), recharge, if possible, when the AMR agents are in the parking lot and recharge all the AMR agents at the end of the simulation. Table IX gives the results for a simulation with scenario 2 (one-hour simulation in random mode): number of recharges, total recharge time, and total duration of the simulation (3600s + final recharge for the AMRs to be fully charged again). The results provided by the on-demand strategy are the best, although at the cost of many charges (when the AMRs return to the parking lot). The "FL Prediction" strategy offers a good compromise with interesting results that can be further improved by adjusting the values of the linguistic variables used in the fuzzy rules.

TABLE IV. SIMULATION DURATION FOR THE 3 STRATEGIES (S)

Scenarios	Real data	Random data	Mass data
Round robin	2744	3600	2280
On-demand	2686	3600	**2196**
FL prediction	**2515**	3600	2252

TABLE V. NUMBER OF BAGS PROCESSED BY THE 3 STRATEGIES

Scenarios	Real data	Random data	Mass data
Round robin	1306	**2015**	1000
On-demand	1306	1864	1000
FL prediction	1306	1956	1000

TABLE VI. FLOW RATE OF THE THREE STRATEGIES (bags/h)

Scenarios	Real data	Random data	Mass data
Round robin	1713	**2015**	1578
On-demand	1750	1864	1639
FL prediction	**1868**	1956	**1597**

TABLE VII. NUMBER OF LOOPS COMPLETED BY THE AMR AGENTS

Scenarios	Real data	Random data	Mass data
Round robin	1622	2074	1298
On-demand	**1306**	**1864**	**1000**
FL prediction	1376	2021	1182

TABLE VIII. AVERAGE WAITING TIME PER BAG BEFORE (S)

Scenarios	Real data	Random data	Mass data
Round robin	41	57	**14**
On-demand	**26**	**21**	20
FL prediction	29	22	19

TABLE IX. IMPACT OF THE RECHARGING OF AMR BATTERIES

Scenarios	Number of recharges	Total duration of recharges	Total duration of simulation
Round robin	**195**	11760	3823
On-demand	622	**9810**	**3683**
FL prediction	198	11658	3793

Finally, we can discuss the benefits of using the communication capabilities of AMR agents. Indeed, adding communication between the AMRs and the infrastructure improves the performance of the Round Robin strategy (2% higher throughput and up to 11% fewer rounds for the AMRs). This communication also improves the results of the algorithm that determines whether an AMR needs to recharge. Indeed, by passing near the parking lot where the charging stations are located, the infrastructure can communicate to the AMR if one of the 4 stations is

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

available. This information allows the AMRs to avoid waiting, especially to maintain a good level of baggage processing performance during periods of dense and continuous flow.

V. Discussion

The previous section presented the results obtained from fuzzy agent-based simulations. The objective was to evaluate the performance of the AMRs on a reference circuit (Figure 2), to validate the hypotheses formulated in the first theoretical study, to identify critical points, and to measure the impact of operational parameters on:

- Average transit/circulation times and delay distribution. The AMRs follow predefined paths, allowing the average time required to complete a mission to be measured. The influence of traffic density and priorities can therefore be quantified and then analyzed.
- Blockages and waiting times at intersections and convergence/divergence zones. Intersections and convergence points are identified as critical areas where queues form. Simulations were used to determine the average queue length and the probability of blockage for different configurations.
- Overall throughput and use of critical zones. The maximum number of AMRs that can circulate simultaneously without degrading throughput was estimated at 31, allowing the optimal circuit capacity to be determined.
- The impact of different control strategies on system operation.

This approach made it possible to compare the performance of the proposed solutions and to identify the optimal conditions for stable, smooth, and efficient operation of AMRs. Thus, the performance of a looped circuit using AMRs depends on a delicate balance between: transport capacity (entry and exit points, distances), the speed and synchronization of pickups/deposits, and flow management at intersections (convergence/divergence). Incorrect sizing or an imbalance between these factors quickly leads to saturation, while a harmonized configuration ensures the smooth, stable, and efficient flow of traffic.

Regarding the speeds of the AMRs on the circuit (speeds arbitrarily defined in Figure 3), the simulations highlighted what the preliminary analysis had already revealed:

- An average speed (5 m/s) promotes a more balanced distribution between the main and secondary tracks and allows for a higher number of AMRs without loss of flow.
- A maximum speed (7.5 m/s) locally increases the flow on the main track but requires limiting the number of AMRs on the circuit to avoid congestion.

We then sought to determine the optimal speed distribution strategy on the circuit, allowing us to reduce the duration of a mission while minimizing energy consumption. Four strategies were thus simulated:

- **StrategySpeed_1**: The speed strategy illustrated in Figure 3.
- **StrategySpeed_2**: the same speed profile as *StrategySpeed_1* for all stops, except that the average speed of 5 m/s is replaced by the maximum speed of 7.5 m/s.
- **StrategySpeed_3**: all speeds are set to the minimum value of 3 m/s.
- **StrategySpeed_4**: the same speed profile as *StrategySpeed_1*, except that on long-distance segments the maximum speed of 7.5 m/s is replaced by the average speed of 5 m/s.

The duration of an AMR mission on the circuit (Figure 3) differs for each of the four speed strategies, as it depends on a fixed travel distance and on a varying speed profile. It can therefore be shown that the maximum system throughput, defined as the number of baggage units that can be handled per unit of time, is higher when using *StrategySpeed_1*.

The maximum system throughput, denoted by λ_{max}, is given by the ratio between the maximum number of AMRs $Max(x)$ and the average duration of a circuit lap T_0. For instance, using the speed profile shown in Figure 3, if the number of AMRs present on the circuit is 31 and an AMR takes 60 s on average to complete one lap, λ_{max} will be defined by (3):

$$\lambda_{max} = Max(x) / T_0 = 31 / 60 = 0.52 \text{ bags/s} \qquad (3)$$

Assuming that the maximum number of AMRs, $Max(x)$, is fixed for all four strategies, the minimum mission durations T_0 can be ranked according to the speed strategies as follows: *StrategySpeed_2*, *StrategySpeed_1*, *StrategySpeed_4*, and *StrategySpeed_3*. This ranking can be explained by the fact that the mission duration varies only with the speed profile, since all other parameters are fixed, as shown in the mathematical model presented in Figure 6. Consequently, the maximum system throughput λ_{max} follows the same ranking.

However, *StrategySpeed_2* consumes more energy than *StrategySpeed_1*, since the cost function we established in [15] increases with speed.

VI. Conclusion and perspectives

In this article, we have presented a fuzzy agent-based modeling and simulation framework with the aim of better understanding the dynamic and uncertain situations encountered in the activity of complex systems such as baggage handling systems in airports.

We introduced the general framework for modeling fuzzy agent-based applications, highlighting the fact that in such systems, each agent is responsible for its own knowledge and actions, which makes it an autonomous entity within the system. It is also important to note that the modeling of fuzzy agents is based on the fuzzy nature of knowledge, behaviors, and interactions.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

We presented a case study involving the simulation of autonomous vehicle fleets for baggage handling in a simplified airport, in which each vehicle is simulated by a fuzzy agent. This simplification allows us to establish a baseline configuration to which we can add more complex and dynamic situations encountered in real airports, such as the introduction of baggage checkpoints requiring baggage drop-off and pick-up by the same or a different AMR.

We established a UML model of the automatic baggage handling system, providing both static and dynamic views of the circuit and the vehicles. Three different strategies were tested, with the goal of determining the number of AMRs that should circulate to handle the baggage arriving at the two entry points.

Finally, we presented our simulation results. These results allowed us to highlight the impact of various parameters, such as the number of simulations, the number of bags processed, the total simulation duration, and the total number of bags processed per hour.

We are now testing the integration of learning capabilities into fuzzy agents in the form of a basic but generic neural network or reinforcement learning processes [48] and [49]. This is to give AMR agents the ability to better adapt in unforeseen situations during their modeling, and to adjust their fuzzy knowledge when extracting it from human experts is difficult or problematic (uncertain knowledge or approximate fuzzy modeling).

Another extension of our research consists of continuing work started recently, still in a simulation approach, on the incorporation of human-robot coworking aspects to maintain system performance, simulate the management of incidents occurring on the circuit or other types of problems difficult to solve for AMRs alone, and thus improve practical credibility for Airport 4.0 stakeholders.

ACKNOWLEDGMENT

The authors thank the Brittany region for funding the ALPHA project as part of the 2022 SME call for projects entitled "Accelerating the market launch of digital technological innovations from SMEs in Western France" (project certified by the "Pôle Images & Réseaux", conv. 23003944). They also thank the three partners of this project, the companies Alstef Mobile Robotics and Secure-IC, and the IMT Atlantique engineering school, for the quality of their collaboration.

REFERENCES

[1] A.-J. Fougères et al., "Fuzzy Agent-Based Modelling and Simulation of Autonomous Vehicle Fleets for Automatic Baggage Handling in 4.0 Airports," In: The second AI-based Systems and Services (AISyS 2025). Lisbon, Portugal, Oct. 2025.

[2] M. Rüßmann et al., "Industry 4.0: The future of productivity and growth in manufacturing industries," Boston Consulting Group, vol. 9, no. 1, pp. 54-89, 2015.

[3] A. G. Frank, L. S. Dalenogare, and N. F. Ayala, "Industry 4.0 technologies: Implementation patterns in manufacturing companies," International Journal of Production Economics, vol. 210, pp. 15-26, 2019.

[4] E. Oztemel and S. Gursev, "Literature review of Industry 4.0 and related technologies," Journal of Intelligent Manufacturing, vol. 31, no. 1, pp. 127-182, 2020.

[5] X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, "Industry 4.0 and Industry 5.0—Inception, conception and perception," Journal of Manufacturing Systems, vol. 61, pp. 530-535, 2021.

[6] A. Akundi et al., "State of Industry 5.0—Analysis and identification of current research trends," Applied System Innovation, vol. 5, no. 1, pp. 27, 2022.

[7] N. Tsolakis, D. Bechtsis, and J.S. Srai, "Intelligent autonomous vehicles in digital supply chains: From conceptualisation, to simulation modelling, to real-world operations," Business Process Management Journal, vol. 25, no. 3, pp. 414-437, 2019.

[8] J. B. Perussi, F. Gressler, and R. Seleme, "Supply Chain 4.0: autonomous vehicles and equipment to meet demand," Int. J. Supply Chain Management, vol. 8, no. 4, pp. 33-41, 2019.

[9] C. Marinagi, P. Reklitis, P. Trivellas, and D. Sakas, "The impact of industry 4.0 technologies on key performance indicators for a resilient supply chain 4.0," Sustainability, vol. 15, no. 6, pp. 5185, 2023

[10] V. U. Tjhin and R. E. Riantini, "Smart farming: implementation of industry 4.0 in the agricultural sector," In Proceeding of the 6th Int. Conf. on E-Commerce, E-Business and E-Government, 2022, pp. 416-421.

[11] E. K. Gyamfi, Z. ElSayed, J. Kropczynski, M. A. Yakubu, and N. Elsayed, "Agricultural 4.0 leveraging on technological solutions: Study for smart farming sector," In 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), 2024, pp. 1-9.

[12] S. K. Sahoo and B. B. Choudhury, "A review of methodologies for path planning and optimization of mobile robots," Journal of Process Management and New Technologies, vol. 11, no 1-2, pp. 122-140, 2023.

[13] S. E. Zaharia and C. V. Pietreanu, "Challenges in airport digital transformation," Transportation Research Procedia, vol. 35, pp. 90-99, 2018.

[14] J. H. Tan and T. Masood, "Adoption of Industry 4.0 technologies in airports--A systematic literature review," arXiv preprint arXiv:2112.14333, 2021.

[15] O. Oukacha, A.-J. Fougères, M. Djoko-Kouam, and E. Ostrosi, "Computational ergo-design for a real-time baggage handling system in an airport," Sustainability, vol. 17, no. 9, pp. 3794, 2025.

[16] A.P.G. Scheidegger, T.F. Pereira, M.L.M. De Oliveira, A. Banerjee, and J.A.B. Montevechi, "An introductory guide for hybrid simulation modelers on the primary simulation methods in industrial engineering identified through a systematic review of the literature," Computers & Industrial Engineering, vol. 124, pp. 474-492, 2018.

[17] A. Hentout, M. Aouache, A. Maoudj, and I. Akli, "Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017," Advanced Robotics, vol. 33, no. 15–16, pp. 764–799, 2019.

[18] L. Serena, M. Marzolla, G. D'Angelo, and S. Ferretti, "A review of multilevel modeling and simulation for human mobility and behavior," Simulation Modelling Practice and Theory, vol. 127, 102780, 2023.

[19] C.M. Macal, "Everything you need to know about agent-based modelling and simulation," Journal of Simulation, vol. 10, no. 2, pp. 144-156, 2016.

[20] R. Skapinyecz, "Examining the collision avoidance problem of AGVs in a discrete event simulation environment," Academic Journal of Manufacturing Engineering, vol. 22, no. 4, pp. 52-65, 2024.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[21] F. Gehlhoff, N. Jobs, and V. Henkel, "Agent-Based Control of Interaction Areas in Intralogistics: Concept, Implementation and Simulation," Logistics, vol. 9, no. 2, pp. 52, 2025.

[22] J. Huang et al., "An overview of agent-based models for transport simulation and analysis," Journal of Advanced Transportation, vol. 2022, no. 1, pp. 1252534, 2022.

[23] J. Grosset, A.-J. Fougères, M. Djoko-Kouam, and J.-M. Bonnin, "Multi-agent Simulation of Autonomous Industrial Vehicle Fleets: Towards Dynamic Task Allocation in V2X Cooperation Mode," Integrated Computer-Aided Engineering, vol. 31, no. 3, pp. 249–266, 2024.

[24] P. Jing, H. Hu, F. Zhan, Y. Chen, and Y. Shi, "Agent-based simulation of autonomous vehicles: A systematic literature review," IEEE Access, vol. 8, pp. 79089-79103, 2020.

[25] A. Azadegan, L. Porobic, S. Ghazinoory, P. Samouei, and A. S. Kheirkhah, "Fuzzy logic in manufacturing: A review of literature and a specialized application," International Journal of Production Economics, vol. 132, no. 2, pp. 258-270, 2011.

[26] H. H. Tang and N. S. Ahmad, "Fuzzy logic approach for controlling uncertain and nonlinear systems: a comprehensive review of applications and advances," Systems Science & Control Engineering, vol. 12, no. 1, pp. 2394429, 2024.

[27] L. A. Zadeh, "Fuzzy sets," Information and Control, vol. 8, no. 3, pp. 338-353, 1965.

[28] L. A. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes," IEEE Transactions on Systems, Man, and Cybernetics, vol. 3, no. 1, pp. 28-44, 1973.

[29] N. Ghasem-Aghaee and T.I. Ören, "Towards Fuzzy Agents with Dynamic Personality for Human Behavior Simulation," In Proc. of SCSC 2003, Montreal, Canada, 2003, pp. 3–10.

[30] A.-J. Fougères, "A Modelling Approach Based on Fuzzy Agent," Int. J. of Computer Science Issues, vol. 9, no. 6, pp. 19-28, 2013.

[31] A. Meylani et al., "Different Types of Fuzzy Logic in Obstacles Avoidance of Mobile Robot," Int. Conf. on Electrical. Eng. and Computer Sc., 2018, pp. 93-100.

[32] B.K. Patle et al., "A review: On path planning strategies for navigation of mobile robot," Defence Technology, vol. 15, no. 4, pp. 582-606, 2019.

[33] M. Alakhras, M. Oussalah, and M. Hussein, "A survey of fuzzy logic in wireless localization," EURASIP J. on Wireless Com. and Networking, vol. 1, pp. 1-45, 2020.

[34] M.F.R. Lee and A. Nugroho, "Intelligent Energy Management System for Mobile Robot," Sustainability, vol. 14, no. 16, pp. 10056, 2022.

[35] J.Y.J. Hsu, D.C. Lo, and S.C. Hsu, "Fuzzy control for behavior-based mobile robots," In NAFIPS/IFIS/NASA'94. Proceedings of the First International Joint Conference of The North American Fuzzy Information Processing Society Biannual Conference. The Industrial Fuzzy Control and Intelligence (pp. 209-213). IEEE, 1994.

[36] T. Fukuda and N. Kubota, "An intelligent robotic system based on a fuzzy approach," Proceedings of the IEEE, vol. 87, no. 9, pp. 1448-1470, 1999.

[37] R. Singh, D.K. Nishad, S. Khalid, and A. Chaudhary, "A review of the application of fuzzy mathematical algorithm-based approach in autonomous vehicles and drones," International Journal of Intelligent Robotics and Applications, vol. 9, no. 1, pp. 344-364, 2025.

[38] W. De Paula Ferreira, F. Armellini, and L.A. De Santa-Eulalia, "Simulation in industry 4.0: A state-of-the-art review," Computers & Industrial Engineering, vol. 149, 106868, 2020.

[39] S. Hassan, L. Garmendia, and J. Pavón, "Introducing uncertainty into social simulation: using fuzzy logic for agent-based modelling," International Journal of Reasoning-Based Intelligent Systems, vol. 2, no. 2, pp. 118-124, 2010.

[40] P. K. Biswas, "Towards an agent-oriented approach to conceptualization," Applied Soft Computing, vol. 8, no. 1, pp. 127-139, 2008.

[41] B. Bauer and J. Odell, "UML 2.0 and agents: how to build agent-based systems with the new UML standard," Engineering Applications of Artificial Intelligence, vol. 18, pp. 141–157, 2005.

[42] A.-J. Fougères, "Agents to cooperate in distributed design," In the IEEE International Conference on Systems, Man and Cybernetics, (SMC'04), The Hague, Netherlands, October 10-13, 2004, vol. 3, pp. 2629-2634.

[43] E. Ostrosi, A.-J. Fougères, and M. Ferney, "Fuzzy Agents for Product Configuration in Collaborative and Distributed Design Process," Applied Soft Computing, vol. 8, no. 12, pp. 2091–2105, 2012.

[44] A.-J. Fougères and E. Ostrosi, "Holonic Fuzzy Agents for Integrated CAD Product and Adaptive Manufacturing Cell Formation," Journal of Integrated Design and Process Science, vol. 23, no. 1, pp. 77-102, 2019.

[45] A.-J. Fougères and E. Ostrosi, "Multiple Fuzzy Roles: Analysis of their Evolving in a Fuzzy Agent-Based Collaborative Design Platform," In: Lecture Notes Studies in Computational Intelligence (LNCSI), Edited by K. Madani, A. Dourado, A. Rosa, J. Filipe and J. Kacprzyk, Springer International Publishing Switzerland, 2016, p.207-226, ISBN: 978-3-319-23391-8, 978-3-319-23392-5 (e-book).

[46] A.-J. Fougères and E. Ostrosi, "Fuzzy engineering design semantics elaboration and application," Soft Computing Letters, vol. 3, pp. 100025, 2021.

[47] J. Grosset, A.-J. Fougères, M. Djoko-Kouam, and J.-M. Bonnin, "Fuzzy Agent-Based Simulations of Cooperative Strategies for Task Allocation, Collision Avoidance, and Battery Charging Management of Autonomous Industrial Vehicles," International Journal on Advances in Systems and Measurements, vol. 18, no. 1&2, pp. 8-18, 2025.

[48] Z.M. Yosif, B.S. Mahmood, and S.Z. Saeed, "Artificial techniques based on neural network and fuzzy logic combination approach for avoiding dynamic Obstacles," Journal Européen des Systèmes Automatisés, vol. 55, no. 3, pp. 339-348, 2022.

[49] H.M. Yudha, T. Dewi, and N. Hasana, "Performance comparison of fuzzy logic and neural network design for mobile robot navigation," Int. Conf. on Electrical Eng. and Comp. Sc., 2019, pp. 79-84.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Target Encoding for Object Orientation Prediction in Smart Manufacturing

Antonio Gambale, Sonya Coleman, Dermot Kerr, Philip Vance
School of Computing, Engineering & Intelligent Systems,
Ulster University, Londonderry, Northern Ireland
e-mail: {a.gambale | sa.coleman | d.kerr | p.vance}@ulster.ac.uk

Emmett Kerr
Dept. of Electronic & Mechanical Engineering,
Atlantic Technological University, Letterkenny, Ireland
e-mail: emmett.kerr@atu.ie

Cornelia Fermüller, Yiannis Aloimonos
Institute for Advanced Computer Studies,
University of Maryland, USA
e-mail: {fermulcm | yiannis}@umd.edu

Abstract*—Accurate planar object orientation prediction is critical for enabling robots to interact effectively with their environment, supporting advanced manipulation tasks beyond simple grasping. However, most existing approaches emphasize gripper pose rather than object-centric orientation, which limits both generalisability and downstream utility. Furthermore, model selection for orientation prediction is often heuristic, lacking systematic comparison or theoretical justification. This work introduces novel encoding schemes and integration strategies for robust planar orientation estimation under a unified 360° single-angle representation. Four new target encodings are proposed: base trigonometric, quadrant enhanced, polar augmented, and full composite; in conjunction with two complementary integration strategies, vector based and branched. Together, these methods address angular discontinuity challenges that hinder the performance of direct regression models.To mitigate the lack of manufacturing-oriented pose datasets, an annotation enrichment pipeline is presented, extending the MetaGraspNet dataset with geometrically derived orientation labels and rigorously annotated real-world data to assess domain transfer. Using features extracted from a pre-trained ResNet50, multiple regression architectures are benchmarked across all encoding and integration configurations. Performance is evaluated using the Mean Absolute Angular Error (MAAE) between predicted and ground truth angles. Experimental results show that the quadrant enhanced encoding combined with the vector based integration strategy achieves the highest accuracy. The XGBoost 1.7 model attains a MAAE of 8.15°, while maintaining robustness to angular discontinuities and strong computational efficiency across both synthetic and real-world testing. These findings provide practical guidance for implementing reliable and generalisable orientation predictors within industrial robotic manipulation systems.

***Keywords*-Computer Vision; Robotics; Manipulation; Machine Learning; Smart Manufacturing.**

I. Introduction

This paper is an extended version of the work presented at the IARIA ADVCOMP 2025 conference [1]. While the core findings remain consistent, this journal version provides a comprehensive methodological expansion, including detailed annotation enrichment procedures, hyperparameter optimisation protocols, and an increased in-depth analysis of encoding and integration strategy interactions. By addressing these methodological aspects, this work thereby advances the practical deployment of orientation estimation in robotic manipulation systems.

The motivation for developing robust orientation estimation methods stems from fundamental limitations in contemporary robotic grasping systems. Current approaches frequently generate arbitrary grasp poses without considering the object's pose for subsequent manipulation requirements [2], [3]. Contemporary robotic grasping research typically focuses on two-finger parallel grippers [2], [3], opting to use complex multi-dimensional grasp representations [4], [5], often referred to as grasping rectangles [6]. These representations cannot easily be implemented on vacuum or three-finger manipulators as they are inherently designed for use with parallel style grippers. Whilst these methodologies demonstrate efficacy in achieving stable grasps, they frequently neglect the requirements of subsequent manipulation tasks, such as tool manipulation or object assembly, as they focus solely on the initial grasp pose of the parallel gripper relative to the object, rather than the object pose from which a grasp can be deduced. Furthermore, many of these works require complex and expensive annotation and validation of multiple positive and negative grasping rectangles for each object [7]. These multi-parameter grasp rectangles necessitate complex approaches to predict five or more parameters, which can be computationally demanding and often result in methods unsuitable for real-time execution in resource constrained environments or embedded systems [8].

Some works have explored various approaches to simplify grasp parameters whilst maintaining effectiveness. Notably, the work in [7] has demonstrated success in representing object orientation using a single angle around the normal vector of a planar surface which is predicted through use of the ImageNet Convolutional Neural Network (CNN) for feature extraction along with a Support Vector Machine (SVM) for angle prediction. However, this approach relies on RGB-D data and complex hierarchical regression techniques to handle angle discontinuity at the 0°/360° boundary. A two-step process is employed where an SVM classifies the angle into one of four 90° intervals, followed by training separate RBF kernel support vector regressors for each interval to predict the precise angle.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

While effective, this method introduces additional complexity and computational overhead compared with direct or trigonometric encoded angle regression approaches. Moreover, the adoption of less conventional grasp representations, such as singular angles, over traditional grasping rectangles further constrains the already limited pool of comprehensive annotated datasets for robotic grasping applications [7], [8]. Consequently, for such an approach to be viable it must either involve a method to deduce single angles from existing datasets or advocate for the inclusion of more comprehensive angular information in future grasping datasets. This highlights the need for innovative data generation and annotation techniques to support advanced object orientation prediction methods.

This research builds upon these prior works with the aim of enhancing post-grasp task execution, focussing on the object pose rather than grasp pose, through the use of a methodology which represents object orientation as a single angle in a 360° coordinate system. This approach provides orientation information that not only facilitates the intended post-grasp use of the object but also enables derivation of the appropriate grasp angle. Additionally, this representation enables learning architectures to focus specifically on orientation prediction of one value (θ), circumventing the complexity of multiple parameter estimations and facilitating grasping with manipulators of all types.

To address the scarcity of annotated grasping datasets suitable for single angle object pose estimation [7], [8], this work draws direct inspiration from [9], who utilised traditional computer vision with object geometric data to generate object poses. Here, a modified version of the methodology in [9] is employed to augment the MetaGraspNet dataset: rather than relying on edge maps, ground-truth object segmentation masks are used to extract geometric features such as centroid, handle-tip displacement, and axis of symmetry, enabling the calculation of precise orientation annotations for each object instance. This automated annotation process was validated through manual inspection to ensure geometric fidelity and orientation accuracy, resulting in a high-quality synthetic dataset well-suited for model training. Recognising that synthetic datasets alone cannot capture the full complexity of real-world scenarios, a complementary real-world dataset was created and rigorously annotated. This secondary dataset, was curated to reflect a range of object configurations, occlusions, and lighting conditions encountered in practical smart manufacturing applications. The combination of these two datasets (a large, systematically annotated synthetic set and a smaller, carefully validated real-world set) enables robust evaluation of model generalisation and domain transfer, providing a more comprehensive assessment of each method's practical viability. This research proposes a suite of novel encoding schemes and integration strategies, together with a systematic evaluation of regression architectures, to advance orientation prediction performance using these enriched datasets. Cross-validation, computational benchmarking, and error analysis are then carried out to determine practical suitability, advancing the current state of planar orientation prediction for robotic manipulation by presenting a clear in depth comparison of these existing shallow learning models and clarifying the interplay between encoding, integration, and model architecture, while also providing actionable insights for deploying robust, efficient orientation predictors in real-world robotic pipelines.

II. Methodology

This section describes the methodological pipeline for planar object orientation prediction, including annotation enrichment, novel encoding strategies, integration approaches, feature extraction, model selection, and optimisation.

A. Object Orientation Prediction Pipeline

The novel pipeline presented in this work is summarised in Figure 1 and presents two solutions, hereafter referred to as integration strategies, for the task of object orientation prediction. In this orientation pipeline, an image of the object is first segmented, and learned features are extracted. These features are then input into a machine learning model as a feature vector. During training and evaluation, target variables are also provided to the models that are derived from ground truth values representing the known pose of each object instance, and represent the values the model will be predicting based on the provided learned features from the CNN.

The novelty of this work centres on the definition and structure of the target values, which specify what the model is trained to predict and constitute the ground truth used to guide predictions. For the remainder of this work these target variables are referred to as encoding schemes, and the manner in which they are presented to the shallow learning models is referred to as the integration strategy.

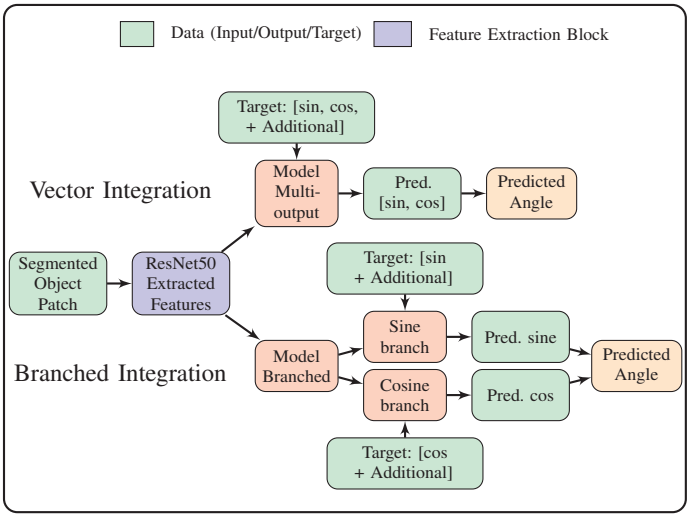


Figure 1. Object orientation prediction pipeline.

B. Target Encoding Schemes

In this study, we are encoding an object's pose angle in degrees which represents the planar orientation of the object, commonly referred to as the azimuth angle, defined over the range 0° to 360°. This angle, denoted by θ, is first encoded through trigonometric decomposition to yield sine and

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

cosine components. Additional encodings are also derived for robustness, resulting in four proposed encoding configurations:

- **Base**: The fundamental trigonometric representation:

$$\mathbf{y}_{\text{base}} = [\sin(\theta), \cos(\theta)] \tag{1}$$

  This encoding allows the model to output trigonometric components from which the orientation angle can be uniquely reconstructed, whilst naturally handling the periodic nature of angular data.
- **Quadrant**: The base encoding with quadrant confidence indicators:

$$\mathbf{y}_{\text{quad}} = [\sin(\theta), \cos(\theta), \mathbf{1}_{Q1}, \mathbf{1}_{Q2}, \mathbf{1}_{Q3}, \mathbf{1}_{Q4}] \tag{2}$$

  Here, $\mathbf{1}_{Qi}$ denotes the ground truth one-hot indicator for each quadrant, where the appropriate quadrant is assigned a value of 1 and all others are set to 0. The quadrants are defined as:
  - $Q1$: $0^\circ \leq \theta < 90^\circ$
  - $Q2$: $90^\circ \leq \theta < 180^\circ$
  - $Q3$: $180^\circ \leq \theta < 270^\circ$
  - $Q4$: $270^\circ \leq \theta < 360^\circ$

  Although the ground truth labels are strictly one-hot, the model is trained to predict a confidence value for each quadrant, where higher values indicate greater model certainty. In practice, these predicted confidences may not reach exactly 1 or 0; the final quadrant prediction is taken as the one with the highest confidence score. This approach allows for uncertainty around quadrant boundaries, providing smoother gradients and more robust learning near ambiguous angles.
- **Polar**: The base encoding is augmented with an explicit radian angle value:

$$\mathbf{y}_{\text{polar}} = [\sin(\theta), \cos(\theta), \theta_{\text{rad}}] \tag{3}$$

  The term θ_{rad} denotes the orientation angle in radians ($0 \leq \theta_{\text{rad}} < 2\pi$), providing a continuous scalar representation in addition to the trigonometric components, helping the model capture fine-grained angular differences.
- **Full**: A comprehensive encoding combining trigonometric, quadrant, and polar representations:

$$\mathbf{y}_{\text{full}} = [\sin(\theta), \cos(\theta), \mathbf{1}_{Q1}, \mathbf{1}_{Q2}, \mathbf{1}_{Q3}, \mathbf{1}_{Q4}, \theta_{\text{rad}}] \tag{4}$$

  This composite form gives the model access to all available encoded representations for each angle.

These encodings define the set of target variables that the shallow learning models are trained to predict and during inference the models only receive the feature vector derived from the input images and produce predictions for these encoded target values based on the learned features supplied by the chosen CNN. While encodings may include multiple elements (for example, both trigonometric and one-hot components in the quadrant encoding), the final predicted angle is always decoded and presented in degrees from 0° to 360°, representing the object's azimuth orientation.

C. Integration Strategies

In addition to the various target encoding schemes, two model integration strategies were proposed. These strategies determine the architectural approach taken by each model when predicting multiple target variables, structuring the regression process to either provide all outputs jointly with one model (vector-based integration) or via distinct independent models for each target variable (branched integration). The two integration strategies are depicted in Figure 1 and described below.

1) Branched Integration Strategy: In the branched integration strategy, multiple parallel models are trained independently to predict different components of the target encoding scheme. For example, one model may predict the sine component while another predicts the cosine component, with further models included as required for additional encoding targets such as quadrant confidences or polar values. Each model receives the necessary input features and auxiliary target variables relevant to its assigned output. Following prediction, the outputs, such as predicted sine and cosine values, are combined using the inverse tangent function ($\tan^{-1}$) to provide the final inferred angle. This approach allows for flexible extension to more than two models, depending on the chosen encoding configuration. The overall structure is illustrated in Figure 1.

For clarity, consider the **full** encoding:

$$\mathbf{y}_{\text{full}} = [\sin(\theta), \cos(\theta), \mathbf{1}_{Q1}, \mathbf{1}_{Q2}, \mathbf{1}_{Q3}, \mathbf{1}_{Q4}, \theta_{\text{rad}}] \tag{5}$$

In the branched integration configuration, separate models are trained to predict specific components of this target vector. For example, one model predicts the sine component along with quadrant confidences and polar angle, while a second model predicts the cosine component, also together with quadrant confidences and polar angle. Each model shares all auxiliary variables for the selected encoding, except for the trigonometric function assigned to the other model. The predicted components are then combined after inference to estimate the final orientation angle.

2) Vector Integration Strategy: The vector integration strategy employs a single model to directly predict all target variables in a unified manner. This enables the model to jointly capture dependencies and relationships among all components of the encoding. As such, all encoding targets, including both trigonometric components and any additional variables, are predicted together, sharing the same optimisation process, loss function, and model parameters [10]. Following prediction, the sine and cosine outputs are combined using the inverse tangent function ($\tan^{-1}$) to recover the angle estimate.

D. Annotation Enrichment

This section details the methodology employed for enriching annotations within the MetaGraspNet dataset framework by determining object orientations and thus providing the object orientation angles that are used as the ground truth for the target encodings discussed in Section II-B. The approach adapts principles from prior work on edge-based orientation detection [9] to segmentation-based annotation data, enabling precise angular determination of elongated objects such as screwdrivers.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

This approach addresses the lack of applicable datasets in planar object orientation estimation for manufacturing-centric applications by automating orientation determination, making it viable to efficiently add detailed annotations to existing datasets such as MetaGraspNet [11].

The annotation enrichment pipeline, as depicted in Figure 2, comprises five sequential stages: (1) COCO-format polygon parsing to extract segmentation data from annotation; (2) centroid computation via spatial moments; (3) principal axis determination using Principal Component Analysis (PCA); (4) orientation direction disambiguation; and (5) final orientation and centroid annotation.

Figure 2. Illustrative example of the orientation annotation workflow, demonstrating the key stages from input image to refined angle estimation for a segmented object.

The annotation enrichment process begins with segmentation polygons from MetaGraspNet's COCO-format annotations [11], where each object instance is defined by boundary points $\mathbf{p}_i = (x_i, y_i)$. These coordinates are first used to calculate a minimum bounding rectangle, creating a localised coordinate system with $(x_{\min}, y_{\min})$ as the origin. All polygon points are shifted to this local system: $\mathbf{p}'_i = (x_i - x_{\min}, y_i - y_{\min})$. A binary mask of the object is then generated by rasterising the shifted polygon [12] onto a blank image sized to the bounding rectangle dimensions, where pixels within the object are set to 1 and pixels outside the object representing the background are set to 0. The geometric centroid is calculated from this mask using spatial moments [13], then transformed back to the original image coordinate system providing the reference point for orientation analysis.

Principal Component Analysis (PCA) [14], [15] is then applied to identify the primary axis of elongation through eigenvector decomposition of the boundary points. The boundary points are first centred by subtracting the computed centroid creating a zero-mean coordinate system. The 2×2 covariance matrix is then computed from these centred coordinates and eigenvalue decomposition then yields two eigenvectors: the primary eigenvector corresponding to the largest eigenvalue represents the direction of maximum variance (elongation), while the secondary eigenvector represents the perpendicular direction. The initial orientation angle is calculated as $\theta = \tan^{-1}\left(\frac{v_{1y}}{v_{1x}}\right)$, where (v_{1x}, v_{1y}) are the components of the primary eigenvector. However, this PCA-derived angle presents directional ambiguity for asymmetric objects such as screwdrivers or bolts, as the eigenvector can point in either direction along the elongation axis [9]. This ambiguity arises because PCA identifies the axis of maximum variance but cannot distinguish between the tool tip and handle ends of the elongated object. To resolve this ambiguity and determine the true tip orientation, a final refinement stage is applied based on the geometric assumption that the object centroid is towards the handle due to asymmetric mass distribution in tools like screwdrivers [9].

This refinement stage defines two search corridors 10 pixels wide along directions θ (initial angle from PCA) and $\theta + 180^\circ$. For each corridor, the algorithm casts a ray from the centroid and identifies boundary points that fall within the corridor width by calculating their perpendicular distance from the ray centreline. Within each valid corridor, the furthest boundary point is identified by computing the projected distance along the ray direction using the dot product. The maximum projected distances d_+ and d_- for the two opposing directions are compared: if $d_+ > d_-$, the original PCA angle θ is retained; otherwise, θ is adjusted by 180° to point toward the tool tip (the direction with the greater extent). This refinement approach aims to resolve the directional ambiguity by exploiting the asymmetric geometry of elongated tools, where the tip extends further from the centroid than the handle region [9].

The complete pipeline processes each annotation by parsing COCO-format segmentation polygons, extracting boundary points, computing the geometric centroid through spatial moments, determining the principal elongation axis via PCA eigenvalue decomposition, resolving tip direction through ray-casting comparison, and appending the final orientation angle θ and centroid coordinates (c_x, c_y) as new annotation fields to the JSON annotation file. This automated enrichment provides orientation ground truth data that enables training of direction-aware orientation models while maintaining compatibility with existing dataset frameworks and COCO annotation standards.

E. Feature Extraction

Feature extraction serves to transform raw image data into numerical representations suitable for input into shallow learning models as part of the object orientation pipeline outline in Figure 1. A ResNet50 convolutional neural network, pre-trained on ImageNet with the classification head removed and global average pooling applied, was employed to generate consistent feature vectors for all object instances [16], [17]. The feature extraction pipeline, depicted in Figure 1, operates as the bridge between raw image data and the regression models, providing the input vectors upon which the shallow learning algorithms operate to predict the target encoding variables described Section II-B.

The feature extraction process comprises four sequential stages: (1) object patch extraction using segmentation masks, (2) background normalisation and padding, (3) standardised resizing, and (4) deep feature computation via the pre-trained CNN as described below.

Object Patch Extraction: Each object instance is isolated from its source image using the corresponding segmentation polygon from the dataset annotations, after which a bounding rectangle is computed with an additional 10% padding factor applied to ensure complete object capture. The masked object is then extracted and placed onto a uniform white background, removing potential confounding factors from the surrounding scene whilst preserving the object's visual characteristics.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Aspect-Preserving Normalisation: To maintain geometric relationships whilst ensuring consistent input dimensions, the extracted object patches undergo aspect ratio preserving resizing. Each patch is scaled such that its largest dimension fits within 224×224 pixels, with the scaled image centred and the remaining space filled with white pixels. This approach prevents distortion whilst providing the standardised input size required by the ResNet50 architecture.

Deep Feature Generation: The normalised patches are processed through the ResNet50 network pre-trained on ImageNet. The network's convolutional layers extract hierarchical visual features, which are then aggregated via global average pooling to produce a fixed-length 2048-dimensional feature vector for each object instance [18]. These resulting feature vectors constitute the input data for the shallow learning models, providing a consistent numerical representation derived from the visual content of segmented object instances. These features, combined with the target encoding variables outlined in Section II-B, enable the regression models to learn mappings from visual appearance to planar orientation estimates. For testing on real-world data, the same pipeline is applied using the custom Real-world Screwdrivers dataset, with segmentation data provided by the Segment Anything Model 2.0 (SAM) [19].

F. Model Selection and Evaluation Framework

The enriched angular annotation data derived in Section II-D enables the training and evaluation of machine learning models for planar object orientation prediction on a refined subset of MetaGraspNet. Owing to the moderate scale of the enhanced dataset, this work intentionally focuses on shallow learning models [20], which offer notable advantages in computational efficiency and interpretability over deep learning approaches. Accordingly, a diverse suite of shallow regression architectures was selected: Random Forest (RF), Support Vector Regression (SVR), Multiple-Output Support Vector Regression (M-SVR), and both the standard (single-output) and multi-output variants of XGBoost. These models were chosen for their established efficacy with moderate data volumes, rapid inference speeds, and transparency in decision-making [7], [10]. Each operates on the input features produced by the CNN feature extraction pipeline (Section II-E) and learns to predict the target variables specified by the encoding schemes and integration strategies discussed in Section II-B.

The investigative framework thus systematically evaluates:

- **Four novel target encoding schemes:** From basic trigonometric representations to rich composite encodings including quadrant and polar variables.
- **Two novel integration strategies:** Branched (predicting sine and cosine individually) and vector-based (joint multi-output regression).
- **Three regression architectures:** Decision tree-based (RF, XGBoost), kernel-based (SVR), and their multi-output extensions.

This approach supports a comprehensive exploration of how encoding design, integration strategy, and model architecture jointly influence orientation prediction accuracy and efficiency, ultimately providing actionable guidance for robust, interpretable robotic perception systems. The shallow learning models considered in this work are summarised in Table I.

TABLE I. Multi-Output Implementation of Shallow Learning Models

Model	Multi-Output Implementation
Random Forest	Native (single model, vector outputs)
SVR	Wrapper (multiple independent regressors)
M-SVR	Native (unified loss function)
XGBoost 1.7	Wrapper (multiple independent regressors)
XGBoost 2.0	Native (vector leaf nodes)

Multi-Output Prediction Implementation

The architecture of each regression model fundamentally determines how multiple target variables are handled, with critical implications for both integration strategies:

- **Native Multi-Output Models (RF, M-SVR, XGBoost 2.0):** These models inherently support simultaneous prediction of multiple targets within a single unified framework. Random Forest achieves this through ensemble averaging, where each tree can produce multi-dimensional outputs at leaf nodes. For example, when predicting the **quadrant** encoding scheme $[\sin(\theta), \cos(\theta), Q_1, Q_2, Q_3, Q_4]$, a single Random Forest model produces all six values simultaneously in one forward pass. M-SVR extends the traditional SVR with a multi-dimensional ε-insensitive loss function that jointly optimises all targets, whilst XGBoost 2.0 implements vector-valued leaf nodes where each leaf stores a complete output vector spanning all prediction targets. This unified approach enables the model to learn cross-target dependencies and implicitly enforce mathematical relationships such as $\sin^2\theta + \cos^2\theta = 1$ [10].
- **Wrapper-Based Multi-Output (SVR, XGBoost 1.7):** Models lacking native multi-output support utilise wrapper frameworks that create independent regressors for each target variable [10]. For instance, when predicting the quadrant encoding scheme, the wrapper creates six separate SVR models: one each for $\sin(\theta)$, $\cos(\theta)$, Q_1, Q_2, Q_3, and Q_4 quadrant indicators. Crucially, these regressors share identical hyperparameters (e.g., regularisation strength, kernel parameters) and receive the same input features, but each is trained independently with its respective encoded target variable output. This approach differs from a fully branched integration strategy in that hyperparameters are globally optimised across all outputs using composite scoring functions, this would mean optimisation across all target encodings for vector approaches, or separate optimisation for sine and cosine branches. However, the independent training prevents explicit enforcement of cross-target relationships, requiring any trigonometric constraints to be learned implicitly through the shared feature space [10], [21], thus the relationship $\sin^2\theta + \cos^2\theta = 1$ is not inherently enforced.
- **Practical Example - Quadrant Prediction:** To illustrate these differences: when predicting quadrant membership, native multi-output models produce confidence scores for all four quadrants (Q_1 to Q_4) simultaneously alongside sine

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

and cosine values, enabling the model to consider quadrant-trigonometric relationships during training. Wrapper-based approaches train separate regressors for each quadrant confidence score, where the final quadrant prediction is determined by selecting the regressor output with the highest confidence value. Both approaches ultimately yield the same prediction format, but the learning dynamics and cross target consistency differ substantially.

Through this systematic evaluation of model architectures and their multi-output implementations, this study provides granular insights into how structural choices impact both prediction performance and the effective exploitation of object orientation encodings within moderate-scale datasets.

G. Optimisation

Hyperparameter optimisation was carried out using grid search across the models. The parameter spaces for each model are detailed in Table II. This approach balances thorough exploration of the hyperparameter space with computational feasibility [22], [23].

TABLE II. Hyperparameter Search Spaces

Model	Hyperparameters (Values)
SVR	C: 0.01, 0.1, 1, 10, 100, 1000, 10000 Kernel: rbf, linear, sigmoid Gamma: auto, 0.001, 0.01, 0.1, 1, 10, 100, 1000 Epsilon: 10^{-5}, 10^{-4}, 10^{-3}, 0.01, 0.05, 0.1 Tolerance: 10^{-5}, 10^{-4}, 5×10^{-4}, 10^{-3}, 10^{-2}
Random Forest	Number of estimators: 100, 200, 300, 400, 500 Maximum depth: 3, 5, 10, 15, 20, 25, 30 Minimum samples split: 2, 5, 8, 11 Minimum samples leaf: 1, 2, 3, 4, 5 Maximum features: sqrt, log2 Bootstrap: True, False
M-SVR	C: 0.01, 0.1, 1, 10, 100, 1000, 10000 Kernel: rbf, linear, sigmoid Gamma: auto, 0.001, 0.01, 0.1, 1, 10, 100, 1000 Epsilon: 10^{-5}, 10^{-4}, 10^{-3}, 0.01, 0.05, 0.1 Tolerance: 10^{-5}, 10^{-4}, 5×10^{-4}, 10^{-3}, 10^{-2}
XGBoost (1.7/2.0)	Number of estimators: 100, 200, 300, 400, 500 Maximum depth: 3, 5, 10, 15, 20, 25, 30 Learning rate: 0.005, 0.01, 0.05, 0.1, 0.2 Subsample: 0.5, 0.7, 0.8, 1.0 Column sample by tree: 0.5, 0.7, 0.8, 1.0 Minimum child weight: 1, 3, 5, 7 Gamma: 0, 1, 3, 5 Regularization alpha: 0, 0.1, 0.5, 1 Regularization lambda: 0.5, 1, 1.5, 2 Tree method: hist

The optimisation process employed 5-fold cross validation with a custom angular distance scorer designed for circular data. This scorer calculates the minimum circular difference between predicted and true angles, handling the periodic nature of angular data where 359° and 0° are only 1° apart rather than 359° apart. While all models use this angular distance scorer for optimisation, they employ different loss functions: Random Forest uses Mean Squared Error (MSE), SVR utilises an epsilon-insensitive loss function, M-SVR incorporates a multi-target epsilon-insensitive loss function, and both XGBoost versions employ MSE, however only 2.0 allows for multiple targets. To assess the predictive performance of the tested models, integration strategies, and encoding approaches, the Mean Absolute Angular Error (MAAE) was employed as the primary metric. This measure was computed independently for each encoding method and dataset, ensuring a granular evaluation of angular accuracy. A comprehensive computational benchmarking analysis was also conducted to evaluate inference speed. For each model, inference was performed on identical image features, with each test comprising 1000 repetitions per run and five aggregate runs to account for variability. All experiments were executed on a Linux 6.8.0 system equipped with Python 3.8.3, an Intel Xeon w7-3445 CPU (20 cores, 40 threads), and 62.3 GB of RAM. This rigorous protocol facilitates a thorough and equitable comparison of both predictive accuracy and computational efficiency across all models and integration strategies.

III. Datasets

To address the scarcity of orientation annotations in existing datasets [24], this work employs a synthetic-to-real transfer approach using two distinct datasets. The first is a modified version of an existing synthetic dataset, used for training and validation. The second is a custom real-world dataset, created by the authors to evaluate model performance under conditions that closely resemble practical applications.

Enhanced MetaGraspNet: The initial training of the proposed methods utilised a carefully curated subset of the MetaGraspNet dataset [25] reduced to focus exclusively on Phillips and flat-head screwdrivers. The subset selected for this work corresponds to the single-class, multiple-instance configuration within the MetaGraspNet framework. The choice of screwdrivers as the focal object class is based on the fact that these objects possess geometric properties that satisfy the requirements outlined in [9] and the annotation enrichment methodology detailed in Section II-D.

The resulting curated dataset provides 7,932 annotations across 2,691 images of size 1200 x 1200 pixels [25]. To validate the calculated object orientations created through the dataset enrichment process, 10% of the datasets ground truth angle data were manually inspected with an allowable deviation threshold of 10°, following [9]. This process involved extracting objects using their segmentation data, rotating them by their calculated ground truth angle to achieve horizontal alignment, and finally a manual verification of object orientation. To aid this process, reference lines indicating the calculated angle and image horizontal were also superimposed, as illustrated in Figure 3.

This validation revealed that objects with areas below 10,000 pixels exhibited significant orientation errors (up to 180° in 184 cases). These anomalies occurred primarily in Difficulty level 5 images across all camera angles, consistent with the increased object density and occlusion at this difficulty level.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

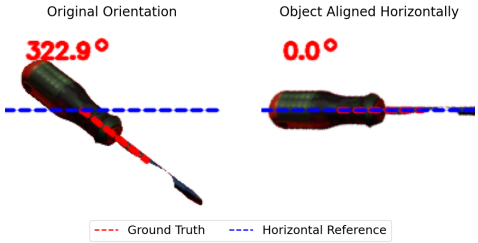


Figure 3. Illustrative example of angular validation.

Figure 4 shows an example of an object removed due to occlusion induced area reduction. To ensure that the calculated ground truth angular data used to train the models were robust, all objects with areas less than 10,000 pixels were therefore removed, resulting in the removal of 2,224 outliers and reducing dataset from 7,932 annotations to 5,709. Angular accuracy was then assessed again on a 10% sample (572 annotations).

Figure 4. Example of occluded object (highlighted in green) removed due to having an area less than 10,000 pixels.

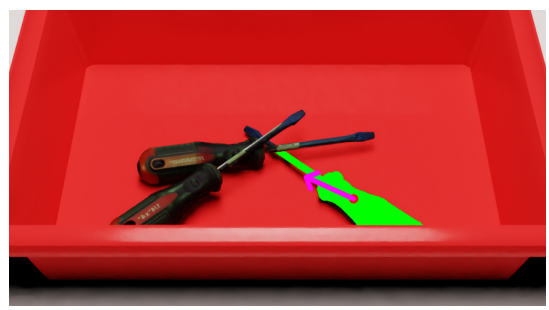
Evaluation of the reduced dataset (5,709 annotations) revealed 55 objects with incorrect detected angles, with angular errors ranging from 1° to 9°. Producing a MAAE across the 10% sample (572 annotations) of 0.313°. Table III presents the full results of the objects with errors, where Cθ refers to the camera angle, Oθ refers to the calculated ground truth angle of the objects, and Δ represents the disparity between the detected angle and the manually validated angle.

Analysis of Table III reveals several noteworthy trends:

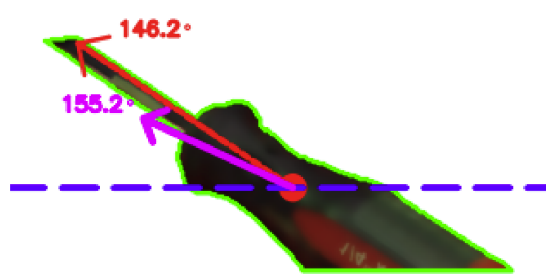
- Error distribution across camera angles is relatively even, with no strong bias towards any particular angle. However, camera angle 0 (top-down view) shows the least error, likely due to minimal occlusion from the tray.
- Object angles cover the full spectrum from 0° to 359° with no clear clustering or preference for specific orientations.
- Most angular errors are minor, typically ranging from 1-3° (Δ). The maximum angular error of 9° occurred in scene 14771, where manual validation revealed significant occlusion at both the screwdriver's tip and handle, as illustrated in Figure 5.
- Several scenes (e.g., 11131, 11519, 13228) appear multiple times with different camera angles. These cases consistently show errors related to occlusion, either from other objects or the tray sides. While these scenes maintain sufficient area (>10,000 pixels), the mass distribution is uneven and therefore affects angle calculations.

TABLE III. Scene Camera Angle Analysis

Scene	Cθ	Oθ	Δ	Scene	Cθ	Oθ	Δ
11091	6	46.57	1	12546	1	81.38	5
11124	0	234.49	1	12720	8	22.37	2
11131	1	128.99	1	12760	4	115.57	6
11131	5	210.32	1	12879	6	189.10	2
11131	8	166.57	1	12976	6	58.21	6
11165	6	348.91	2	12979	3	148.46	1
11189	7	218.57	1	13040	8	287.78	2
11353	5	150.20	6	13169	5	25.17	5
11377	7	73.41	3	13228	2	275.33	1
11377	8	29.90	2	13228	7	34.71	5
11501	4	135.76	8	13228	8	4.81	3
11519	5	359.15	5	13292	2	134.87	2
11519	6	235.90	1	13417	1	63.83	5
11519	8	150.27	5	13499	5	134.26	1
11538	2	345.07	2	13711	6	152.91	6
11538	6	3.84	4	13787	3	60.84	3
11598	1	50.02	2	13794	3	77.88	4
11623	4	109.74	5	13811	2	53.26	3
11806	3	265.60	4	13943	8	355.62	4
11933	4	341.34	2	14120	4	26.41	5
11933	6	183.35	2	14588	7	251.76	2
11933	7	147.01	4	14606	2	40.96	6
11962	1	69.80	2	14634	3	68.50	3
12129	3	198.57	3	14656	3	287.13	3
12323	3	188.74	4	14670	3	115.38	2
12441	7	136.19	4	14771	7	155.19	9
12458	4	21.96	1	14811	4	141.74	2
12488	2	48.29	3	14811	6	346.61	1



(a) Full scene visualisation with the screwdriver producing the maximum angular error highlighted in green.



(b) Isolated segmentation mask with overlays to visualise predicted (Pink) and ground truth (red) angles.

Figure 5. Occluded Screwdriver from scene 14771, camera angle 7. (a) Full scene visualisation with screwdriver mask and angle overlay. (b) Isolated object mask with overlays for predicted (pink) and ground truth (red) angles and corresponding angle values.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

It is important to note that both the maximum angular error and the MAAE are less than the 10° angle variation threshold delineated in [9]. As such, these data would not significantly compromise robotic grasping performance [26]. Therefore, the final dataset comprises 4,567 training samples and 1,142 testing samples with associated images and annotations representing an 80/20 split.

Since the MetaGraspNet dataset was not originally designed for this specific application, a final assessment of the object distribution across the 360° angular range was conducted. To accomplish this, the generated ground truth annotations were binned into 10° intervals, revealing an average of 127 images per bin for the training set and 32 objects per bin for the test set. The analysis demonstrated a consistent distribution of angles between the training and test sets, as illustrated in Figure 6. Additionally, the training set exhibited a slight imbalance between 300° and 309°, containing only 62 images compared with the average. This imbalance was less pronounced in the test set but remained present. Since the distribution of objects is not controlled in the original dataset, this imbalance will not be addressed; however, careful attention will be given to assessing model performance in these under-represented angular ranges to evaluate their impact on overall system performance.

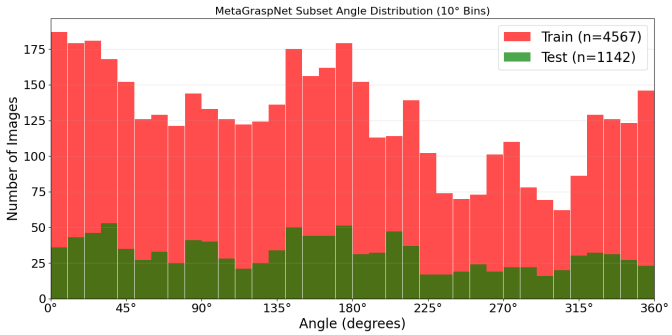


Figure 6. Distribution of ground truth angles across 10° bins for the MetaGraspNet subset.

Real-world Dataset:

To facilitate a simulation-to-reality transfer methodology, a second dataset was created for this work. This real-world dataset was constructed to mirror the structure of the curated MetaGraspNet subset used for training, closely adhering to the difficulty levels defined in [25], but utilises real images of screwdrivers rather than synthetic data.

In line with the MetaGraspNet difficulty levels, which are designed to progressively increase in complexity, levels 1 and 2 represent single objects with no occlusion, while higher levels introduce multiple objects and increasing degrees of occlusion and clutter. Consistent with the original subset selection from MetaGraspNet, which also excluded images at this level, the dataset did not include images from difficulty level 3. The annotation and data distribution by difficulty of this dataset can be seen in Table IV.

This real-world screwdriver dataset was acquired using a camera equipped with a Samsung ISOCELL GN9 sensor, capturing images of size 3072 × 4080 pixels, with a lens aperture of f/1.9, an exposure time of 1/100 second, and an ISO sensitivity of 386, with images being collected from both overhead and 45° angled perspectives. The dataset features three screwdriver variants, red/black Torx, blue/black flathead/starhead, and solid black starhead, randomly arranged on a white backdrop under uncontrolled ambient lighting conditions. To ensure natural illumination variability, 27 images were captured across daylight hours, resulting in 81 annotations; examples of such images can be seen in Figure 7.

TABLE IV. Summary of images and annotations by difficulty level

Level	No. of Images	No. of Annotations
1-2	8	8
4	5	17
5	14	56

Figure 7. Example images from the secondary dataset.

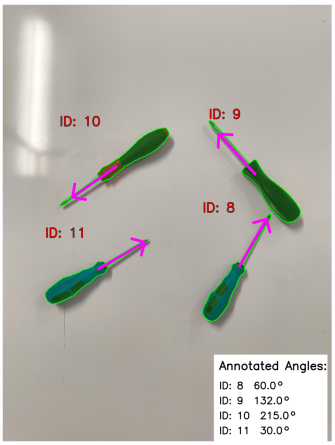


Figure 8. Sample image from the real-world dataset visualising four annotated screwdrivers.

Ground truth segmentation masks and orientation angles were annotated by three domain experts [27]. Inter-annotator agreement on segmentation masks was assessed using mean

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

IoU, yielding a high consistency of 0.95. Discrepancies were resolved via consensus review to refine the segmentation boundaries. For orientation, annotators measured shaft angles against the right horizontal axis, with a mean absolute difference of $\pm 1.8^\circ$ between predictions. Disagreements exceeding 5° were resolved by taking the circular mean for the affected objects; otherwise, the final orientation annotations were selected from a random sample of the independent measurements, except where consensus averaging was required as described. Examples of these annotations can be seen superimposed on one of the test images from the real world dataset in Figure 8 in which the segmentation masks are outlined in green, object IDs are labelled in red, and annotated orientation angles are indicated by pink arrows. The corresponding orientation values (in degrees) are listed in the lower right corner.

TABLE V. Default parameters used for the Segment Anything Model (SAM) during mask generation.

Parameter	Value
Model Type	`vit_h`
Checkpoint	`sam_vit_h_4b8939.pth`
Device	`cuda`
Points Per Side	32
Pred IoU Threshold	0.88
Stability Score Threshold	0.95
Crop N Layers	0
Crop Overlap Ratio	0.3413

To ensure relevance to downstream deployment, segmentation masks for inference on the custom real-world dataset were generated using SAM [19], rather than relying on the expert annotated masks. SAM was configured using its default parameters, which are summarised in Table V. The accuracy of these automatically generated masks was validated against the manual annotations, achieving a mean Intersection-over-Union (IoU) of 0.95. This strong agreement supported the use of SAM generated masks for subsequent stages in the pipeline.

IV. Results

The predictive performance of the encoding approaches, integration strategies and shallow learning models was evaluated using the Mean Absolute Angular Error (MAAE), as described in Section II-G. Results were computed independently for each model, encoding method, integration strategy and dataset, enabling a detailed assessment of angular accuracy.

Inference speed was benchmarked using the computational protocol detailed previously, ensuring comparability across models. The following results begin with MAAE measurements using the MetaGraspNet subset, comparing integration strategies (vector vs. branched encoding) and encoding configurations (full vs. reduced target sets). All values represent degrees of angular error, with lower values indicating better performance. The initial results are summarised in Table VI.

Following the MetaGraspNet subset results, Table VII summarises model performance using the real-world dataset and the same MAAE metric to facilitate direct comparison.

TABLE VI. MAAE results on MetaGraspNet dataset (degrees)

Model	Base	Quadrant	Polar	Full
XGBoost 1.7 (Vector)	5.15	5.15	5.15	5.15
XGBoost 1.7 (Branched)	5.15	5.15	5.15	5.15
XGBoost 2 (Vector)	**4.92**	**5.01**	5.46	5.17
M-SVR (Vector)	8.04	8.04	8.04	8.04
SVR (Vector)	5.01	5.12	**5.01**	**5.01**
SVR (Branched)	5.01	5.12	**5.01**	**5.01**
Random Forest(Vector)	5.48	**5.08**	5.87	5.67
Random Forest(Branched)	7.43	5.44	5.93	5.88

TABLE VII. MAAE results using real-world dataset (degrees)

Model	Base	Quadrant	Polar	Full
XGBoost 1.7 (Vector)	**9.61**	**8.15**	**8.96**	**9.61**
XGBoost 1.7 (Branched)	**9.61**	11.66	**9.15**	11.14
XGBoost 2 (Vector)	17.09	13.91	**7.18**	10.46
M-SVR (Vector)	28.86	28.86	28.82	28.82
SVR (Vector)	23.13	23.54	23.14	28.03
SVR (Branched)	23.15	23.54	23.14	23.28
Random Forest(Vector)	14.81	14.49	11.84	17.43
Random Forest(Branched)	16.45	**11.62**	12.90	17.50

Following the analysis of angular accuracy, Table VIII presents inference times per image in milliseconds for each model and condition. These results allow for direct comparison of computational efficiency across integration strategies and model architectures. As with previous metrics, lower values indicate superior performance, in this case, faster processing.

TABLE VIII. Inference time per image patch (milliseconds) by model and condition

Model Type	Base	Quadrant	Polar	Full
XGBoost 1.7 (Vector)	0.76	1.86	0.91	1.73
XGBoost 1.7 (Branched)	0.50	0.83	1.20	3.61
XGBoost 2 (Vector)	0.76	1.86	0.91	0.29
M-SVR (Vector)	17.76	17.78	17.80	17.76
SVR (Vector)	13.48	41.94	20.78	51.25
SVR (Branched)	13.55	70.23	27.68	75.49
Random Forest(Vector)	59.27	56.50	57.90	45.42
Random Forest(Branched)	119.15	117.08	117.83	91.04

The bolded values in Tables VI and VII highlight the configurations yielding the lowest Mean Absolute Angular Error (MAAE) for each dataset and encoding condition, representing the top-performing approaches in terms of angular prediction accuracy. Across both synthetic and real-world evaluations, the results show a clear benefit for methods employing vector-based integration, with XGBoost 2 and SVR providing the best performances on the MetaGraspNet test set, and XGBoost 1.7 (vector-based, quadrant encoding) and XGBoost 2 (vector-based, polar encoding) achieving the most accurate results on real-world data. The initial takeaway is that encoding scheme and integration strategy significantly influence predictive accuracy, with performance differences becoming more pronounced when models are tested on real-world as opposed to synthetic data.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

V. Discussion | Evaluation

While the MAAE metrics presented in Section IV form the core of this discussion, it is essential to first address the primary challenge that this work seeks to overcome: the issue of boundary discontinuity. To thoroughly assess each method's performance in this regard a more granular analysis of the inference data on the custom dataset is therefore required, as visualised in Figure 9. Figure 9 displays the prediction error as a function of the ground truth angle for each model and encoding configuration. Specifically, the x-axis represents the ground truth angle of the object, while the y-axis shows the signed angular error, calculated as the shortest difference between the predicted and actual angles, wrapped to the interval $[-180°, 180°]$. Each point corresponds to a single prediction, positioned horizontally by its ground truth angle and vertically by the deviation from the true value.

A detailed inspection of these plots reveals substantial differences in error distributions between models, which are not always reflected in the aggregate MAAE values. The M-SVR model exhibits pronounced and frequent error spikes at the 0° and 359° boundaries, indicating a persistent struggle with boundary discontinuity and a lack of robustness in these critical regions. Beyond isolated boundary failures, M-SVR demonstrates erratic behaviour across much of the angular range, with large, abrupt deviations that suggest poor generalisation and reliability. In contrast, XGBoost 1.7 (both vector and branched variants) stands out for its consistent and stable error profile. Across nearly the entire angular range, prediction errors remain tightly clustered around zero, with only occasional moderate spikes, mostly at angular boundaries. This stability is indicative of a model that not only achieves a low mean error, as seen in Table VII, but also avoids catastrophic failures, making it more suitable for real-world deployment where reliability is paramount.

XGBoost 2, while achieving competitive MAAE values in some configurations, displays a more volatile error pattern. Notably, it exhibits significant errors not only at the 0° and 359° boundaries but also around 180°, suggesting that its generalisation may be compromised at multiple critical angles. This behaviour underscores the importance of evaluating models beyond mean metrics, as a low MAAE can mask underlying instability. The SVR models, both vector and branched, provide mixed results. While their errors are generally moderate, both models are prone to sporadic, large prediction failures at various angles, particularly near the boundaries and occasionally in the mid-range. These large spikes indicate that, although SVR may perform adequately on average, it is susceptible to unpredictable outliers that could undermine its practical utility. Random Forest models show moderate stability, with the vector variant generally outperforming the branched version in terms of error consistency. While occasional spikes are present, these are less frequent and less severe than those observed in M-SVR or SVR, positioning Random Forest as a reasonable compromise between stability and accuracy reflected by its low MAAE when paired with quadrant encoding and a branched integration strategy as seen in Table VII.

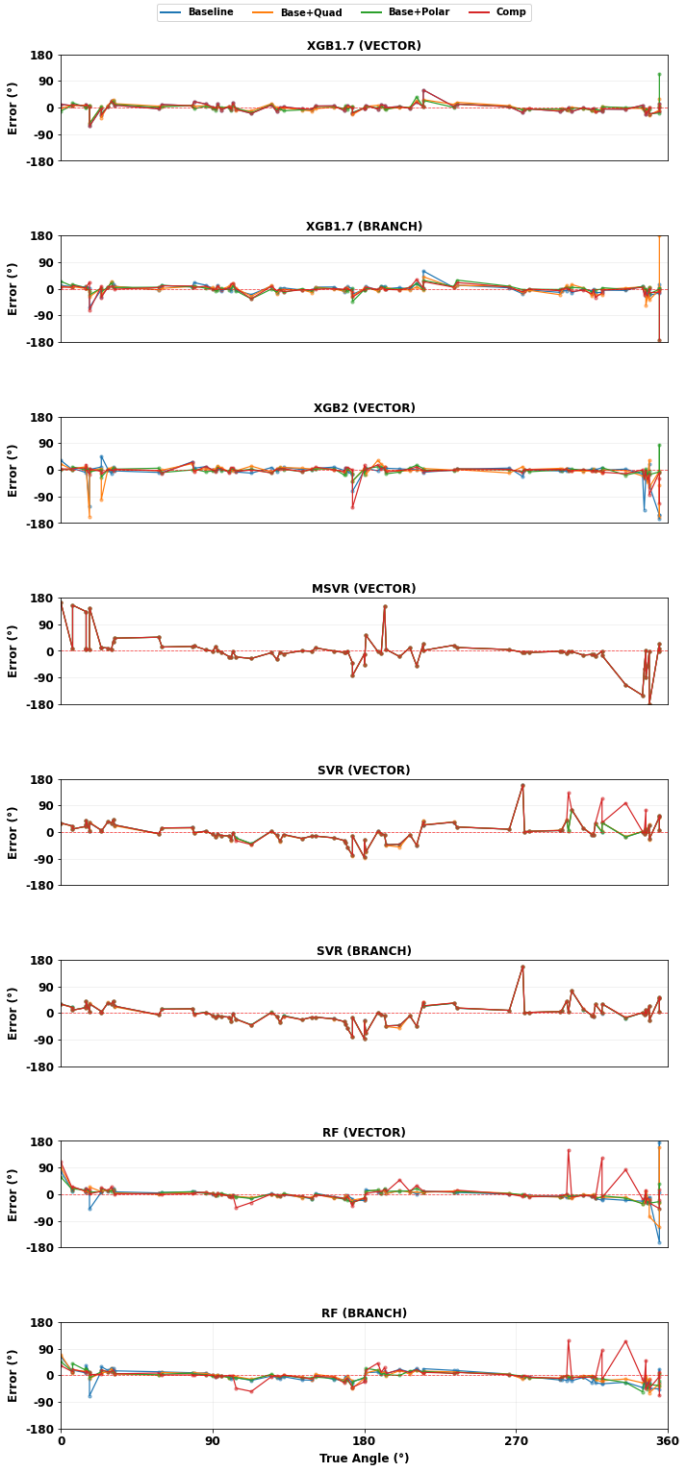


Figure 9. Model prediction errors against ground truth angle on the real-world dataset.

It is important to note that all models were trained exclusively on a synthetic dataset which is inherently simpler and more controlled than the real-world test set. As a result, the MAAE values achieved using the synthetic MetaGraspNet subset (Table VI) are substantially lower across all models, reflecting

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

the training-domain familiarity. For example, XGBoost 1.7 achieves an MAAE of 5.15° using the synthetic data, compared with 8.15–9.61° using the real-world dataset, while M-SVR and SVR also show marked increases in error when transitioning to real-world evaluation. This domain gap highlights the challenge of generalising from synthetic to real data, and underscores the value of robust error analysis using the real-world test set.

In addition to predictive accuracy, computational efficiency was also evaluated. Table VIII quantifies inference times per image patch, demonstrating XGBoost 1.7's superior performance with configurations predominantly achieving sub-2ms inference times (one outlier excepted). In contrast, the M-SVR/SVR models exhibit 10-100× slower performance (13.48-75.49ms), while Random Forest demonstrates the poorest efficiency, consistently exceeding 45ms and frequently surpassing 100ms. This efficiency advantage positions XGBoost 1.7 as optimal for applications requiring both rapid inference and angular reliability.

To understand the architectural origins of these performance differences, Figure 10 isolates each model's baseline capability using exclusively the fundamental (sin, cos) encoding, before any encoding augmentation is applied. This comparison reveals why certain models benefit substantially from advanced encodings while others show minimal improvement.

XGBoost 1.7's baseline performance (MAAE: 9.6° in both integration strategies) is remarkably strong, exhibiting stable error profiles with no large spikes at the 0°/360° boundary, demonstrating that this architecture inherently manages circular discontinuities effectively even with basic trigonometric features alone. This explains why XGBoost 1.7 maintains consistent performance across encodings in Figure 9: the architecture is already operating near its performance ceiling with baseline encoding. Conversely, M-SVR's baseline (MAAE: 28.9°) exhibits extreme error spikes exceeding ±150° concentrated at boundaries and scattered throughout the range, revealing fundamental architectural incapacity for circular prediction. Even with augmented encodings in Table VII, M-SVR cannot overcome these baseline limitations, remaining above 24° MAAE across all configurations.

Similarly, SVR's baseline (MAAE: 23.1°) shows pronounced spikes including severe failures around 180°, explaining its persistent high error (19-23°) regardless of encoding sophistication. XGBoost 2 (baseline MAAE: 17.1°) and Random Forest (baseline MAAE: 14.8-16.5°) occupy intermediate positions, with XGBoost 2's weaker baseline explaining its greater encoding sensitivity; requiring polar coordinates to achieve its best performance (7.18°) as seen in Table VII. This baseline comparison establishes that models unable to manage circular boundaries with simple (sin, cos) encoding cannot be rescued through encoding sophistication alone, directly informing model selection for angular prediction tasks.

A key focus of this work is the interplay between encoding strategies, model architectures, and integration approaches. The results reveal that model responsiveness to encoding varies considerably:

- **XGBoost 1.7** demonstrates strong robustness across all encodings, but achieves its best real-world performance with quadrant encoding in the vector integration configuration, yielding the lowest MAAE (8.15°). Notably, the addition of quadrant or polar information generally improves performance over the base (sin, cos) encoding, suggesting that XGBoost 1.7 can leverage richer target representations to better manage boundary effects and reduce systematic errors. The full encoding does not consistently outperform quadrant or polar, indicating that it may introduce redundancy or noise for this architecture.
- **XGBoost 2.0** is more sensitive to the encoding choice, achieving its best real-world MAAE (7.18°) with the polar encoding. However, its error distribution is less stable, with notable spikes at both boundary and mid-range angles, suggesting that while certain encodings can lower mean error, they may not guarantee robust predictions.
- **SVR and M-SVR** models show limited benefit from more complex encodings. Both models exhibit high MAAE and frequent large errors regardless of encoding, indicating that their architectures are less capable of exploiting additional target information to improve generalisation, especially in the presence of boundary discontinuities.
- **Random Forest** benefits moderately from quadrant encoding, particularly in the branched configuration. It does not reach the accuracy or stability of XGBoost models but presents performance more consistent than SVR/M-SVR.

The choice between branched and vector integration strategies also influences model performance:

- For **XGBoost 1.7**, the vector strategy paired with quadrant encoding is optimal, while the branched approach is less effective, especially when combined with more complex encodings.
- **Random Forest** shows a preference for the branched strategy when used with quadrant encoding, achieving its lowest MAAE (11.62°), but remains slower and less accurate than XGBoost.
- **SVR** and **M-SVR** do not benefit significantly from either integration strategy, with both approaches yielding high error and erratic predictions.

These results demonstrate that XGBoost 1.7, with vector integration and quadrant encoding, offers the best trade-off between accuracy, robustness to boundary discontinuities, and computational efficiency using real-world data. The baseline analysis reveals that this superiority stems from inherent architectural capacity to handle circular predictions, rather than merely benefiting from sophisticated encoding schemes. While more complex encodings can marginally improve mean error for models with weaker baselines (e.g., XGBoost 2), the quadrant encoding strikes an effective balance between informativeness and generalisation. Models such as M-SVR and SVR are fundamentally limited by their inability to manage angular discontinuities at the architectural level, and consequently do not benefit meaningfully from richer encodings or alternative integration strategies. Random Forest provides moderate stability but cannot match the overall reliability or

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

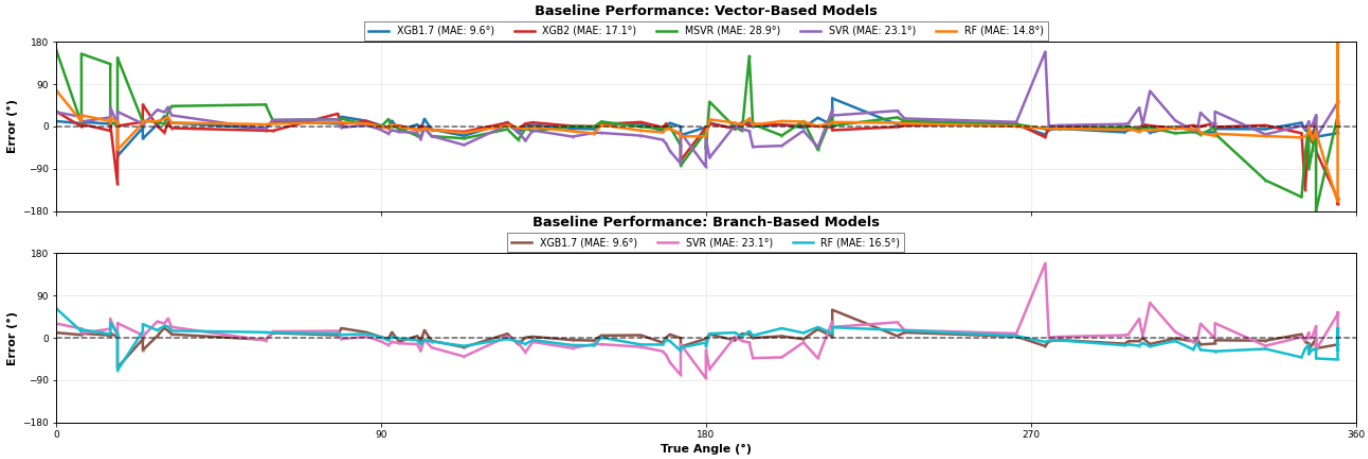


Figure 10. Baseline performance comparison across all models using only (sin, cos) encoding. Top: Vector-based integration strategies. Bottom: Branch-based integration strategies. MAAE values quantify performance without encoding augmentation, revealing inherent architectural capacity for circular prediction.

speed of XGBoost 1.7. These findings underscore the need to consider both aggregate metrics and detailed error distributions when selecting models for applications requiring consistent, timely performance across all angular ranges.

VI. Conclusion and Future Work

This study systematically evaluated novel target encoding schemes, integration strategies, and a variety of shallow regression models for the challenging task of planar object orientation prediction, using both synthetic and real-world datasets. The results demonstrate that among the configurations tested, XGBoost 1.7 combined with the vector based integration strategy and quadrant encoding consistently achieves the best balance of predictive accuracy, robustness to angular discontinuities, and computational efficiency when applied to real-world data. While more complex encodings such as full composite or polar augmented representations can marginally improve mean error for certain models, the quadrant encoding offers superior generalisability by effectively balancing informativeness with simplicity and avoiding overfitting or feature redundancy. Building on these insights, future work will focus on deploying the optimal model configuration within an industrial robotic grasping context, thereby advancing automation reliability and precision in manufacturing systems. These findings provide actionable guidance for designing efficient and robust orientation prediction modules that can integrate seamlessly into intelligent robotic manipulation pipelines in smart factories.

References

[1] A. Gambale *et al.*, "Orientation prediction in robotics: A study of trigonometric decomposition methods across synthetic and real-world datasets", in *Proceedings of the 19th International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP)*, 2025, pp. 1–7.

[2] X. Gao and Z. Wen, "Task-oriented robotic grasping for intelligent manufacturing", in *2023 3rd International Symposium on Artificial Intelligence and Intelligent Manufacturing (AIIM)*, IEEE, 2023, pp. 101–104.

[3] H. Sekkat, O. Moutik, L. Ourabah, B. ElKari, Y. Chaibi, and T. A. Tchakoucht, "Review of reinforcement learning for robotic grasping: Analysis and recommendations", *Statistics, Optimization & Information Computing*, vol. 12, pp. 571–601, 2024.

[4] L. Chen, P. Huang, Y. Li, and Z. Meng, "Detecting graspable rectangles of objects in robotic grasping", *International Journal of Control, Automation and Systems*, vol. 18, pp. 1343–1352, 2020.

[5] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks", in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2015, pp. 1316–1322.

[6] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks", in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2017, pp. 769–776.

[7] M. Schwarz, H. Schulz, and S. Behnke, "Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features", in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1329–1335. DOI: 10.1109/ICRA.2015.7139363.

[8] A. Wong, Y. Wu, S. Abbasi, S. Nair, Y. Chen, and M. Shafiee, "Fast graspnext: A fast self-attention neural network architecture for multi-task learning in computer vision tasks for robotic grasping on the edge", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2293–2297.

[9] A. Gambale *et al.*, "A comparative study of hough transform and pca for bolt orientation detection", in *2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)*, IEEE, 2024, pp. 1–6.

[10] J. Brownlee, *Ensemble Learning Algorithms With Python: Make Better Predictions with Bagging, Boosting, and Stacking*. Machine Learning Mastery, 2021.

[11] Y. Chen, E. Z. Zeng, M. Gilles, and A. Wong, "Metagraspnet_v0: A large-scale benchmark dataset for vision-driven robotic grasping via physics-based metaverse synthesis", *arXiv preprint arXiv:2112.14663*, 2021.

[12] G. Bradski, "The opencv library", *Dr. Dobb's Journal of Software Tools*, 2000.

[13] O. Team, *Image moments (ana huamán tutorial)*, https://docs.opencv.org/4.x/d0/d49/tutorial_moments.html, Accessed: 2025-06-26, 2023.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[14] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] K. Pearson, "On lines and planes of closest fit to systems of points in space", *Philosophical Magazine*, vol. 2, pp. 559–572, 1901. DOI: 10.1080/14786440109462720.

[16] J. Zhang, B. Yin, Y. Zhong, Q. Wei, J. Zhao, and H. Bilal, "A two-stage grasp detection method for sequential robotic grasping in stacking scenarios", *Mathematical Biosciences and Engineering*, vol. 21, pp. 3448–3472, 2024.

[17] I. M. Baltruschat, A. Saalbach, M. P. Heinrich, H. Nickisch, and S. Jockel, "Orientation regression in hand radiographs: A transfer learning approach", in *Medical Imaging 2018: Image Processing*, SPIE, vol. 10574, 2018, pp. 473–480.

[18] A. Depierre, E. Dellandréa, and L. Chen, "Scoring graspability based on grasp regression for better grasp prediction", in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 4370–4376.

[19] A. Kirillov *et al.*, "Segment anything", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[20] S. Robles Herrera, M. Ceberio, and V. Kreinovich, "When is deep learning better and when is shallow learning better: Qualitative analysis", *International Journal of Parallel, Emergent and Distributed Systems*, vol. 37, pp. 589–595, 2022.

[21] H. Borchani, G. Varando, C. Bielza, and P. Larranaga, "A survey on multi-output regression", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, pp. 216–233, 2015.

[22] A. Morales-Hernández, I. Van Nieuwenhuyse, and S. Rojas Gonzalez, "A survey on multi-objective hyperparameter optimization algorithms for machine learning", *Artificial Intelligence Review*, vol. 56, pp. 8043–8093, 2023.

[23] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization", *Journal of Machine Learning Research*, vol. 13, 2012.

[24] Z. Wang, Z. Zhang, T. Pang, C. Du, H. Zhao, and Z. Zhao, "Orient anything: Learning robust object orientation estimation from rendering 3d models", *arXiv preprint arXiv:2412.18605*, 2024.

[25] M. Gilles, Y. Chen, T. R. Winter, E. Z. Zeng, and A. Wong, "Metagraspnet: A large-scale benchmark dataset for scene-aware ambidextrous bin picking via physics-based metaverse synthesis", in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2022, pp. 220–227.

[26] A. Gambale, E. Kerr, D. Kerr, and S. Coleman, "Computing the orientation of hardware components from images using traditional computer vision methods", *Engineering Proceedings*, vol. 65, p. 8, 2024.

[27] A. Gambale, *Screwdrivers dataset*, https://app.roboflow.com/mvtec-uihve/screwdrivers-kmij6/2, Accessed: 2025-06-13.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Representing Human Capabilities in Occupational Databases: An Embedding-Based Audit of O*NET, ESCO and Vocational Standards

Ilona Erika Negro
Doctoral Program of Applied Artificial Intelligence
Alma Mater Europaea, Maribor, Slovenia
e-mail: ilona.negro@almamater.si

***Abstract*—Labour-market analyses of automation and artificial intelligence frequently rely on occupational databases such as O*NET and ESCO to characterise the task content of work and model technological substitution or exposure. Yet less attention has been paid to how these data infrastructures represent different dimensions of human expertise. This study uses embedding-based semantic similarity analysis to conduct a representational audit of three widely used occupational taxonomies. Drawing on the Human-Centred Resilience Model (HCRM), the study examines four capability domains: cognitive–technical judgement, embodied–sensorimotor calibration, trust-calibrated interpersonal engagement, and situated adaptive decision-making. Behavioural anchor statements operationalising the four pillars are compared with descriptors from three occupational representation systems, namely O*NET Work Activities, ESCO skills, and national vocational training standards, across three occupations: electrician, dog groomer, and massage therapist. Consistently, O*NET descriptors show the strongest alignment with the capability pillars, followed by vocational standards and ESCO descriptors. A sensitivity analysis confirms that this ordering is robust to the choice of aggregation rule and embedding model. The adaptability pillar shows weaker alignment than cognitive and interpersonal capabilities. Qualitative inspection indicates a more abstract representation of adaptability rather than explicit descriptions of real-time behavioural adjustment. These findings reveal a systematic representational asymmetry in occupational taxonomies: situated adaptive decision-making is consistently encoded in a more compressed and abstract form than cognitive and interpersonal capabilities, suggesting that automation exposure estimates derived from such taxonomies may not fully capture the tacit and context-dependent dimensions of human expertise that are most difficult to replicate computationally.**

Keywords-occupational taxonomies; semantic embeddings; automation research; artificial intelligence; tacit knowledge.

I. Introduction

Artificial intelligence and automation research frequently relies on structured occupational taxonomies as an empirical foundation for estimating labour-market exposure and task automatability. Systems such as O*NET in the United States and ESCO in Europe provide standardised, machine-readable descriptors of occupational skills, abilities, and work contexts.

In previous work, the Human-Centred Resilience Model (HCRM) was developed to conceptualise dimensions of capabilities hypothesised to resist automation, particularly in embodied and relational service roles [1]. The model originally comprised three pillars:

1) Embodied Dexterity and Skilled Movement,
2) Agency and Adaptability, and
3) Emotional Labour.

These dimensions capture forms of expertise associated with tacit knowledge, embodied coordination, and interpersonal interaction that are often difficult to formalise computationally.

When the HCRM was empirically examined in subsequent work [2], the Adaptability pillar produced weaker differentiation than anticipated for occupations expected to rely heavily on situational flexibility and real-time adjustment.

This discrepancy prompted a methodological question. If occupational taxonomies serve as empirical inputs for automation modelling, it becomes important to understand whether they adequately represent the dimensions of work on which such models rely. The present study, therefore, asks: to what extent do occupational representation systems capture tacit, situational, and embodied forms of expertise?

Despite the widespread use of occupational taxonomies in automation modelling, relatively little research has examined how the internal representational structure of these datasets may shape the capability profiles derived from them. The present study addresses this question through a representational audit of three major sources of occupational description:

1) O*NET (United States) [3]
2) ESCO (European Skills, Competences, Qualifications and Occupations) [4]
3) National vocational standards from six European countries, used as comparatively detailed descriptions of occupational practice.

Recent advances in sentence embedding models allow the quantitative evaluation of semantic similarity between conceptual constructs and textual descriptors across large corpora. Using embedding-based semantic similarity analysis, the study evaluates the extent to which these representation systems align with theoretically defined human-centred capability pillars. Rather than redefining the core premises of the HCRM, this research refines its operationalisation for measurement purposes. Specifically, the original adaptability construct is decomposed into analytically distinguishable components that capture different aspects of situated action and decision-making.

The aim of the study is not to critique existing taxonomies normatively, but to quantify how different dimensions of human-centred work are represented within widely used occupational description systems. By examining where representational alignment is strong and where it is weak, the analysis contributes to a more systematic understanding of the data foundations on which empirical automation assessments are built. To the best of the author's knowledge, this study represents one of the

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

first systematic attempts to evaluate occupational taxonomies through embedding-based representational auditing, assessing how conceptual capability constructs align with the descriptor structures of widely used labour-market datasets.

The remainder of this paper is organised as follows. Section II presents the state of the art in automation research and occupational taxonomy use. Section III introduces the theoretical foundations of the HCRM and the empirical anomaly that motivated the present study. Section IV situates the representational audit within the broader context of occupational taxonomy use in automation modelling and describes the three representation systems examined. Section V presents the four-pillar capability framework used as the basis for the analysis. Section VI describes the data sources and corpus preparation procedures. Section VII details the embedding-based methodology. Section VIII reports the results of the representational audit across occupations and capability pillars. Section IX discusses the findings in relation to tacit knowledge, automation modelling, and human-centred work. Section X presents the conclusions and directions for future research.

II. Related Work

Empirical studies estimating automation exposure frequently rely on occupational databases to approximate the task structure of work. For example, influential early estimates of computerisation probabilities trained machine-learning classifiers on O*NET task characteristics and "bottleneck variables" describing perceptual, creative, and social capabilities [5]. Subsequent studies have continued to rely on similar occupational representations when estimating the exposure of occupations to digital technologies or artificial intelligence [6] [7]. More recently, assessments of Large Language Model (LLM) exposure have used O*NET Detailed Work Activities to approximate the task structure of occupations [8]. Beyond these landmark studies, more recent work has examined the labour-market effects of artificial intelligence through task-displacement and task-reinstatement frameworks [9], and through direct assessment of the skills and activities most exposed to AI capabilities [10]. Across this body of work, O*NET remains the dominant empirical input for approximating occupational task structures in the United States, while ESCO fulfils a comparable role in European labour-market research [4].

However, occupational taxonomies are not neutral mirrors of real work; they are representational infrastructures designed to render labour markets legible to statistical and computational analysis by translating complex work practices into standardised descriptors. In order to standardise occupational descriptions across sectors and regions, these systems represent certain aspects of work with high granularity while compressing or simplifying others. Consistent with theories of routine-biased technological change [11] and longstanding observations about the difficulty of formalising tacit knowledge [12], occupational databases may tend to privilege formally articulated, context-stable cognitive activities. By contrast, work that depends on embodied coordination, adaptive judgement under uncertainty, or the relational dynamics of emotional labour may be only partially captured. These structural tendencies have been documented empirically: Handel [13] identified redundancy between importance and level scales, construct overlap, and substantial abstraction in item wording as specific measurement limitations of the O*NET content model — properties that are particularly consequential when fine distinctions between structured cognitive and tacit capabilities are required.

Because automation models frequently rely on these structured representations to approximate occupational capability profiles, any systematic representational imbalance may influence how automation potential is assessed. In particular, forms of expertise that are difficult to codify, such as tacit sensorimotor skill, situated decision-making, and relational coordination, may be underrepresented in the data structures used to model technological substitution. Whether and how these representational properties systematically vary across occupational databases, and what this means for automation modelling, has received limited direct empirical attention. The present study addresses this gap through an embedding-based representational audit of three widely used occupational description systems.

III. Background: The Human-Centred Resilience Model

The theoretical point of departure for the present representational audit is the HCRM, introduced in previous work examining the resilience of embodied service and craft occupations in the context of automation [1] [2]. The HCRM was developed to explain why some occupations, particularly those in the leisure, care, and skilled service economy, appear to retain value despite rapid advances in automation. Examples include wellness practitioners, pet care professionals, and skilled tradespeople, which share characteristics with the occupations analysed in the present study. These roles often combine physical skill, interpersonal interaction, and situational variability, characteristics that are difficult to formalise within conventional task-based automation frameworks.

Conceptually, the model draws on Self-Determination Theory [14] and Bourdieu's concepts of symbolic capital [15] and cultural distinction [16]. From the perspective of Self-Determination Theory, occupations that support autonomy, competence, and relatedness can generate intrinsically motivated forms of work for practitioners while also shaping demand, as clients often value services that provide meaningful interaction, personalised attention, and opportunities for human connection. At the same time, Bourdieu's concepts of symbolic capital and distinction help explain how scarcity, authenticity, and embodied expertise may increase the perceived value of certain forms of human work as automated services become more widespread.

In its original formulation, the HCRM proposed that the resilience of these occupations arises from the combined presence of three capability domains:

1) Embodied Dexterity and Skilled Movement
2) Agency and Adaptability
3) Emotional Labour

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

The key theoretical insight of the model was that resilience does not emerge from any single attribute in isolation, but rather from the interaction of these dimensions during real-world task execution, particularly in occupations that require close interaction with people, animals, or physical environments.

A. Empirical Anomaly and Measurement Challenge

Initial empirical testing of the HCRM using occupational data revealed a counterintuitive pattern. While the embodied skill and emotional labour dimensions broadly aligned with theoretical expectations, the Agency and Adaptability pillar did not clearly distinguish HCRM occupations from a broader comparison group [2]. At the aggregate level, the comparison group even produced slightly higher overall alignment scores on the adaptability dimension than the occupations expected to rely most strongly on situational flexibility.

This result was surprising because many leisure, care, and personal service roles rely on continuous, real-time judgement exercised in situ, including responding to non-verbal cues, adapting techniques to client preferences or animal behaviour, and operating in variable, often unpredictable environments. Such contexts require ongoing judgement and situational adjustment rather than reliance on predetermined task scripts.

This empirical pattern raised an important methodological question: whether the concept of adaptability was inadequately represented in the occupational taxonomies used for measurement. In particular, descriptors in databases such as O*NET and ESCO appeared to represent adaptability primarily through descriptors of cognitive planning, evaluation, and problem-solving, rather than through descriptions of real-time behavioural adjustment in dynamic environments.

B. Increasing Construct Resolution

To enable a more precise representational analysis of occupational datasets, this study refines the operationalisation of the HCRM. Rather than altering the theoretical foundations of the model, the original Agency and Adaptability pillar is analytically decomposed into two distinct components grounded in the literature on expertise, decision-making, and adaptive performance in complex environments [17] [18] [19].

This decomposition yields a four-pillar capability framework:

1) *Cognitive–Technical Judgement* — The application of structured knowledge, formal rules, and analytical reasoning in diagnosing problems and determining solutions.
2) *Embodied–Sensorimotor Calibration* — Skilled physical performance based on continuous integration of tactile, spatial, and tool-mediated feedback.
3) *Trust-Calibrated Interpersonal Engagement* — The capacity to establish and maintain cooperative, trust-based interactions with clients or stakeholders during service delivery.
4) *Situated Adaptive Decision-Making* — The ability to adjust actions dynamically in uncertain or evolving environments, selecting among multiple possible courses of action without relying on predetermined scripts.

These pillars define the capability constructs used to examine how occupational taxonomies represent different forms of expertise. Fig. 1 illustrates the conceptual structure of the HCRM and the four capability pillars derived from its three foundational domains: embodied dexterity and skilled movement, agency and adaptability, and emotional labour.

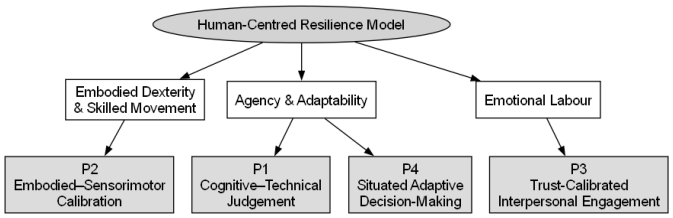


Figure 1. Human-Centred Resilience Model (HCRM) showing the decomposition of the original three-pillar structure into four capability domains.

IV. From HCRM to Representational Audit of Occupational Taxonomies

Automation modelling and algorithmic labour-market analysis rely heavily on occupational description systems that translate work into structured task, skill, and capability descriptors. Such approaches use formalised descriptions of occupational requirements as inputs to empirical automation models to estimate technological substitution or exposure [5] [6] [7] [8] [10]. The validity of these assessments, therefore, does not only depend on model design but also on the representational properties of the occupational data on which they rely.

A. Standard Taxonomies: O*NET and ESCO

The dominant occupational representation systems in this domain are O*NET in the United States and ESCO in the European Union. O*NET provides a large survey-based architecture covering knowledge, skills, abilities, work activities, and work contexts [3]. However, the database also exhibits several measurement challenges, including redundancy between importance and level scales, overlap across constructs, and substantial abstraction in item wording [13]. These properties are particularly relevant in the present context because the analysis focuses on fine distinctions between structured cognitive descriptors and more tacit, situated capabilities.

ESCO, by contrast, provides a multilingual ontology linking occupations to standardised skill and competence descriptors [4]. Its primary function is more to support interoperability among labour-market, education, and policy systems than survey-based measurement. Although its architecture differs from O*NET's, ESCO similarly relies on standardised, context-independent descriptors designed to make occupational characteristics machine-readable and comparable across labour-market systems.

When labour-market and automation studies use such databases as empirical inputs or feature spaces — as seen in influential models of computerisation probability [5] and recent assessments of Large Language Model exposure [8] — they also inherit the representational assumptions embedded in these systems. If an occupational taxonomy primarily represents work through abstract, standardised, or decontextualised descriptors,

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

then downstream models may reflect those abstractions rather than the full situational complexity of occupational practice.

B. Comparative Reference Corpus: National Vocational Standards

To evaluate the representational coverage of standard taxonomies, this study introduces a third source of occupational description: national vocational standards from six countries: Austria, Belgium, France, Germany, Hungary and the Netherlands. Unlike statistical or policy taxonomies, vocational standards are training- and qualification-oriented documents developed to specify the competencies, procedures, safety requirements, and learning outcomes associated with professional practice. In competence-based vocational education systems, training regulations typically define occupational competences and performance requirements in terms of observable learning outcomes linked to workplace activities [20] [21] [22].

Vocational standards should not be treated as direct records of practice. However, because these documents are designed to guide training, assessment, and certification of practitioners, they often include detailed descriptions of procedures, materials, tools, and performance criteria that reflect how work is expected to be carried out in practice.

C. The Representational Audit

The present study, therefore, conducts a representational audit comparing O*NET, ESCO, and national vocational standards across three service and craft occupations.

The audit covers three occupations selected because they rely heavily on embodied skill, interpersonal coordination, and situational variability: electrician, dog groomer, and massage therapist. These occupations are described in detail in Section VI.

V. Pillar Framework for Representational Analysis

To operationalise the capabilities associated with human-centred service and craft work, the HCRM is translated here into four analytically distinct but interrelated capability domains. These pillars capture different forms of expertise through which skilled practitioners engage with technical systems, physical environments, and other people in the course of work. While occupational taxonomies tend to privilege structured cognitive tasks and discrete work activities, research on tacit knowledge, expertise, embodied cognition, and situated action suggests that many forms of skilled performance emerge through the dynamic integration of perception, action, and social interaction in context [12] [18] [19] [23] [24]. This perspective is also reflected in Rasmussen's Skill–Rule–Knowledge framework, which distinguishes between embodied skill-based behaviour, rule-guided action, and knowledge-based reasoning in complex work environments [17]. These traditions collectively emphasise that skilled performance cannot be fully understood as the execution of isolated tasks; it emerges from the coordinated interaction of cognitive judgement, embodied action, and social engagement within dynamic contexts.

The four pillars, therefore, represent complementary dimensions of human capability: cognitive–technical judgement, embodied–sensorimotor calibration, trust-calibrated interpersonal engagement, and situated adaptive decision-making. Table I summarises the four capability pillars used in the representational analysis and the primary forms of expertise associated with each domain.

TABLE I. Capability Pillars Used in the Representational Analysis

Pillar	*Capability Domain*	*Core Capability Mechanisms*
P1	Cognitive–Technical Judgement	Diagnosis of technical problems, interpretation of standards, rule-based and knowledge-based reasoning
P2	Embodied–Sensorimotor Calibration	Skilled bodily movement, tactile feedback, coordination of tools, materials, and physical environments
P3	Trust-Calibrated Interpersonal Engagement	Interpretation of verbal and non-verbal cues, emotional regulation, and maintenance of cooperative interaction
P4	Situated Adaptive Decision-Making	Real-time adjustment to evolving conditions, selection among action paths, integration of situational cues

A. Cognitive–Technical Judgement

Cognitive–Technical Judgement refers to the capacity to understand technical systems, diagnose faults, interpret standards, and apply structured domain knowledge under operational constraints. This pillar captures forms of rule-based and knowledge-based reasoning commonly required in technical occupations, where practitioners must interpret signals, diagnose problems, and apply established procedures or technical standards [17].

In labour-market research, such structured forms of problem-solving align closely with task-based approaches to work, which conceptualise occupations as bundles of tasks that require different combinations of cognitive and manual capabilities [25]. Within this framework, technological change has been shown to substitute for tasks that can be expressed through explicit procedural rules while complementing more complex problem-solving activities [11].

This pillar, therefore, corresponds closely to the kinds of structured cognitive activities typically captured in occupational taxonomies such as the O*NET content model, including analysing information, interpreting technical documentation, and determining solutions to technical problems [3].

B. Embodied–Sensorimotor Calibration

Embodied–Sensorimotor Calibration refers to the capacity to perform skilled physical actions through continuous sensorimotor adjustment, integrating tactile, spatial, and tool-mediated feedback in real time. Skilled manual activity requires the dynamic coordination of perception and movement, allowing practitioners to adapt bodily actions to the constraints imposed by tools, materials, and physical environments. Research in ecological psychology emphasises that such behaviour emerges from direct interaction with environmental affordances rather than from explicit cognitive representations [26].

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Similarly, theories of embodied cognition highlight that many forms of practical intelligence arise from the coupling of perception and action within situated bodily experience [27]. In skilled work, this coupling enables practitioners to calibrate force, position, timing, and movement patterns continuously while responding to subtle environmental cues. Classical research in motor control further demonstrates that complex skilled movements are not executed through rigid programmes but are dynamically organised through the coordination of multiple degrees of freedom within the body–environment system [28].

Taken together, these perspectives describe embodied skill as adaptive sensorimotor regulation in which perception, movement, and environmental feedback are tightly integrated during task execution. Such capabilities are central to many craft and service occupations, although they are often described only indirectly in occupational taxonomies, which tend to emphasise tasks and outcomes rather than the underlying sensorimotor processes through which skilled performance is achieved.

C. Trust-Calibrated Interpersonal Engagement

Trust-Calibrated Interpersonal Engagement (P3) refers to the capacity to establish and sustain cooperative, trust-based interactions with clients, colleagues, or other stakeholders in physically co-present environments. This dimension includes the ability to interpret verbal and non-verbal cues, regulate emotional expression, and adjust communicative behaviour in response to the expectations and affective states of others.

Research on emotional labour highlights that many service occupations require workers to actively manage and coordinate emotional displays to produce appropriate interpersonal experiences for clients [29]. In such contexts, effective interaction depends not only on technical competence but also on the ability to sustain socially meaningful encounters through subtle forms of relational coordination and affective regulation.

Sociological studies of face-to-face interaction further emphasise that social encounters rely on continuous mutual monitoring and the maintenance of interactional trust between participants [30]. Workers in client-facing occupations must therefore continuously adjust their communicative behaviour, tone, and attentional focus in order to sustain cooperative engagement.

From a labour-market perspective, these interpersonal capabilities form part of a broader category of social skills that have become increasingly important in modern economies and are associated with occupations involving collaboration, service provision, and client interaction [31]. Such skills rely heavily on situational awareness and tacit interpersonal knowledge, making them difficult to fully capture in standardised occupational descriptors that typically emphasise discrete tasks or communication activities.

D. Situated Adaptive Decision-Making

Situated Adaptive Decision-Making refers to the capacity to operate effectively in complex and uncertain environments without predetermined scripts. It involves selecting among multiple possible courses of action, revising strategies dynamically, and integrating technical knowledge with situational judgement in real time. In many forms of skilled practice, effective performance does not depend solely on the application of explicit rules but also on the ability to interpret evolving environmental cues and adjust behaviour accordingly.

Research in cognitive systems engineering emphasises that practitioners working in complex environments must continuously coordinate multiple interacting constraints while adapting their actions in response to changing conditions [17] [19]. Studies of situated action further demonstrate that real-world activity frequently unfolds through ongoing interaction with the environment rather than through the execution of prespecified plans [24]. Similarly, research on naturalistic decision making and recognition-primed decision-making shows that experienced practitioners often rely on recognition-based judgement, rapidly identifying patterns in unfolding situations and selecting feasible courses of action without explicitly comparing alternatives [18]. Analyses of expertise further emphasise that such adaptive behaviour relies heavily on tacit, experience-based knowledge that cannot be fully articulated through formal rules or procedures [12] [23].

To operationalise this pillar for empirical analysis, adaptability is represented through behavioural anchor statements describing real-time situational adjustment in dynamic environments, derived from the pillar's conceptual definitions. Examples of the behavioural anchors are shown in Table II.

VI. Data Sources

This section describes the three occupational representation systems used in the analysis and the procedures applied to prepare each corpus for embedding-based comparison.

A. O*NET

Occupational descriptors were drawn from the O*NET Database 30.2 [3] Work Activities taxonomy for the selected occupations, with the analysis focusing on the Work Activities component. Each Work Activities descriptor consists of a short activity statement describing a generic work behaviour (e. g., "Monitoring and reviewing information from materials, events, or the environment, to detect or assess problems."). O*NET provides structured occupational descriptions derived from surveys of job incumbents and occupational experts.

In this study, the analysis focused on the 41 descriptors contained in the O*NET Work Activities taxonomy. The taxonomy consists of standardised activity categories applied uniformly across occupations. Because these descriptors are identical across occupations, the textual corpus used in the embedding analysis remains constant for all three cases.

The importance and level scores associated with the Work Activities taxonomy were not incorporated into the similarity analysis, as the objective was to evaluate the semantic representational coverage of capability constructs, not the relative weighting of activities within occupations.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE II. EXAMPLE BEHAVIOURAL ANCHOR STATEMENTS USED TO OPERATIONALISE THE HCRM CAPABILITY PILLARS

Pillar	*Behaviour Anchors*
Cognitive–Technical Judgement (P1)	Diagnoses technical problems using structured reasoning and technical knowledge. ● Interprets technical documentation, diagrams, or operational instructions. ● Applies established procedures and standards to solve technical problems. ● Evaluates system behaviour to determine appropriate corrective actions. ● Analyses technical information to identify faults or system malfunctions. ● Uses domain knowledge to determine appropriate technical solutions. ● Interprets signals, measurements, or system indicators to guide decision-making. ● Applies formal rules or procedures to complete technical tasks.
Embodied–Sensorimotor Calibration (P2)	Adjusts hand movements in response to tactile or spatial feedback during manual tasks. ● Coordinates bodily movement and tool use to perform precise physical actions. ● Continuously adjusts force, position, and timing during manual work. ● Uses tactile and spatial feedback to guide tool manipulation. ● Calibrates bodily actions while interacting with tools or materials. ● Performs skilled manual actions requiring fine motor control. ● Adjusts physical technique based on subtle environmental cues. ● Maintains precise bodily coordination while performing physical tasks.
Trust-Calibrated Interpersonal Engagement (P3)	Establishes trust through clear and respectful communication with clients. ● Interprets verbal and non-verbal cues during interpersonal interaction. ● Adjusts communication style in response to the emotional state of others. ● Maintains cooperative relationships with clients or colleagues. ● Provides reassurance or explanation to reduce uncertainty for clients. ● Listens carefully and responds appropriately to client concerns. ● Maintains calm and professional interaction in emotionally charged situations. ● Coordinates work activities through interpersonal communication.
Situated Adaptive Decision-Making (P4)	Adjusts actions in real time in response to changing environmental conditions. ● Selects among multiple possible courses of action based on emerging cues. ● Revises task strategies when routine procedures prove insufficient. ● Adapts behaviour continuously while interacting with people, animals, or physical environments. ● Responds flexibly to unexpected events during task execution. ● Integrates technical knowledge with situational judgement in real time. ● Modifies task execution as conditions evolve. ● Navigates uncertain environments without relying on predetermined scripts.

B. ESCO

The ESCO skill and competence descriptors used for each occupation were extracted from ESCO v1.2.1 (accessed 8 Feb. 2026) [4]. Individual occupation concept URIs and the full extracted descriptor sets are provided in the supplementary repository: https://github.com/Ilona-woody/Hcrm-representational-audit/tree/main/esco. ESCO provides a multilingual classification of occupations, skills, competences, and qualifications designed to support labour-market interoperability across European policy and employment systems.

For each selected occupation, the associated skill and competence descriptors listed in the ESCO ontology were extracted and treated as individual textual descriptors for embedding comparison.

C. National Vocational Standards

National vocational training and qualification documents for the selected occupations were collected from official national or regional sources in Austria, Belgium, France, Germany, Hungary and the Netherlands. These documents included qualification frameworks, curricula, competency standards, and training specifications describing the competencies, performance criteria, and learning outcomes associated with professional practice. Documents were chosen that provided formal competency specifications for the selected occupations within national qualification frameworks or accredited training programmes. The electrician vocational standards corpus contains substantially more descriptors than the other corpora, reflecting the higher level of procedural specification in electrical installation standards.

Descriptor counts, therefore, differ across representation systems because documentation practices differ, not because the occupations differ in underlying capabilities.

To ensure comparability with O*NET and ESCO, only content-bearing competency and task descriptions were retained. Administrative, legal, and procedural sections that did not describe occupational capability requirements were excluded. For longer documents, only sections containing learning outcomes, competence statements, performance criteria, or task-relevant descriptions were extracted.

Each extracted segment was recorded with provenance metadata, including issuing authority, year, regulatory level, source URL, and language. Each extracted segment was treated as a potential descriptor unit for subsequent semantic analysis. Non-English source documents were translated into English using a consistent machine-translation workflow (DeepL) [32] [33] [34] [35]. The English translations were then used for embedding-based comparison to enable cross-language analysis using an English sentence-transformer model. Full provenance metadata for all vocational standard documents used in the analysis, including issuing authority, year, regulatory level, source URL, and language, is provided in the supplementary repository: https://github.com/Ilona-woody/Hcrm-representational-audit/tree/main/data.

Table III summarises the occupational representation systems examined in the analysis, the type of data each provides, and the textual unit used as the basis for embedding comparison.

TABLE III. OCCUPATIONAL REPRESENTATION SYSTEMS AND UNITS OF ANALYSIS

Representation System	*Data Type*	*Textual Unit Used in Analysis*
O*NET Work Activities	structured task descriptors	activity statement
ESCO Skills	skills taxonomy	skill labels
Vocational Standards	vocational training standards	competency statements

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

D. Corpus Preparation

Each corpus was segmented into descriptor-level textual units to enable embedding-based comparison. Where documents contained longer narrative passages, text was first divided into sentence-level segments using punctuation and line-break boundaries so that similarity scoring could later be performed at the segment level while retaining the descriptor as the reporting unit.

VII. Method

The empirical analysis implements an embedding-based representational audit of occupational databases. Descriptor texts from three representation systems, O*NET Work Activities, ESCO skills, and vocational standards (Regulatory and Professional Documentation Corpus, RPDC), are embedded using a transformer-based sentence embedding model. Behavioural anchor statements operationalising the four human-centred capability pillars are embedded in the same semantic space. Representational alignment is then quantified using cosine similarity between anchors and descriptor segments, with descriptor-level scores defined by the highest observed semantic correspondence. This procedure enables systematic comparison of how strongly different occupational representation systems encode the capability constructs defined by the HCRM. In this sense, the analysis evaluates the extent to which occupational data infrastructures make human-centred capabilities computationally visible.

A. Representational Approach

Embedding-based semantic similarity analysis provides a method for examining how conceptual constructs are represented within textual corpora by mapping linguistic content into high-dimensional vector spaces. In vector-space models of language, words and sentences are represented as numerical embeddings, with their relative positions capturing semantic relationships between expressions [36]. Transformer-based language models, such as BERT, generate contextual embeddings that encode linguistic meaning based on the surrounding context [37]. Sentence-level models, such as Sentence-BERT (SBERT), extend this approach by producing embeddings optimised for semantic similarity comparisons between short texts [38]. In such embedding spaces, semantically related expressions occupy nearby regions. Embedding-based techniques have increasingly been applied in computational social science to analyse conceptual relationships within textual corpora [39].

This study applies embedding-based semantic similarity analysis to evaluate how occupational representation systems encode the four theoretically defined pillars of human-centred capability introduced in Section V. Sentence embeddings were generated using the pretrained Sentence-Transformers model all-MiniLM-L6-v2, a compact transformer architecture designed for efficient semantic similarity and sentence-embedding tasks [40]. The model produces 384-dimensional embeddings optimised for semantic comparison.

Each representation corpus, comprising O*NET Work Activities, ESCO skills, and vocational standards drawn from national training documents (hereafter the Regulatory and Professional Documentation Corpus, RPDC), was converted into descriptor-level textual units. Each pillar was operationalised using a set of behavioural anchor sentences derived from the conceptual definitions of the four capability pillars, as listed in Table II. The complete anchor sets are provided in the supplementary repository: https://github.com/Ilona-woody/Hcrm-representational-audit.

Semantic alignment between pillars and representation systems was estimated using cosine similarity between pillar anchor embeddings and descriptor embeddings. Each pillar is represented by multiple anchor statements. Descriptor-level pillar alignment is operationalised as the maximum cosine similarity between any descriptor segment and any anchor statement within that pillar, capturing the strongest semantic correspondence within the descriptor. Where descriptors contained longer narrative passages, they were segmented into sentence-like units using punctuation and line-break boundaries. Sentence embeddings were computed for these segments, and the maximum similarity value across segments was retained for the original descriptor. This procedure reduces semantic dilution in longer descriptions that contain multiple behavioural components.

Retaining the maximum similarity constitutes a presence-sensitive detection strategy. The objective of the audit is to determine whether construct-relevant capability content appears anywhere within a descriptor, even when embedded within broader narrative descriptions. Maximum similarity, therefore, provides a conservative test of representational coverage: if a representation system fails to produce strong similarity even under this most permissive detection rule, this indicates that the construct is likely compressed, abstracted, or omitted within the representation.

For each pillar $\times$ representation system $\times$ occupation combination, the resulting distribution of descriptor-level similarity scores was summarised using mean, median, 10th percentile (p10), and 90th percentile (p90) statistics. Cosine similarity ranges from -1 to 1. Negative values indicate semantic dissimilarity between descriptor and anchor embeddings and are expected for descriptors unrelated to a given capability domain. Standard deviation and range statistics were also computed as robustness checks and are reported in Table V. Mean similarity is used as the primary summary indicator of overall representational density, while median and percentile statistics are reported to assess distributional robustness and the presence of high-alignment descriptors within each system.

ESCO descriptors frequently consist of short skill labels: across the three occupational corpora examined, the median descriptor length was 3–4 words for ESCO, compared with 14 words for O*NET Work Activities and 5–11 words for RPDC vocational standards. Such short phrases contain substantially less contextual information than longer task descriptions and may therefore produce lower semantic similarity scores independently of construct alignment. This compression of contextual information is itself a representational feature of ontology-based taxonomies and may systematically limit

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

the extent to which tacit, situational, or embodied capabilities can be expressed within such systems. The script used to compute descriptor length statistics is provided in the supplementary repository: https://github.com/Ilona-woody/Hcrm-representational-audit/tree/main/scripts.

Because the O*NET descriptor set is identical across occupations, O*NET similarity distributions remain constant across cases; observed cross-occupation variation therefore arises primarily from the RPDC and ESCO corpora.

B. Preprocessing and Unit of Analysis

Occupational representation systems differ substantially in textual granularity. ESCO descriptors typically consist of short skill labels, O*NET Work Activities are structured task descriptions, and vocational standards often contain longer narrative passages describing competencies, training requirements, or performance criteria. To ensure comparability across sources, all corpora were transformed into descriptor-level analysis units prior to embedding.

Text was cleaned through whitespace normalisation and the removal of empty or near-empty strings. O*NET Work Activities and ESCO skills were retained as individual descriptors reflecting their structured format. Vocational standard descriptors were segmented as described in Section VI. Extremely short fragments lacking meaningful lexical content were excluded. For reporting purposes, similarity scores are recorded at the descriptor level; however, for segmented descriptors, the reported value reflects the highest-scoring sentence-level segment contained within that descriptor.

Vocational standards were collected from publicly available national and regional training and qualification documents across six countries and translated into English. Translation was necessary to enable cross-corpus comparison within a single embedding space, though minor linguistic nuances in the original documents may be attenuated in translation. Only sections describing occupational tasks, competencies, or behavioural expectations were retained, while administrative and regulatory sections were excluded.

The resulting descriptor sets served as the unit of analysis for similarity computation. Descriptor counts, therefore, vary across representation systems, reflecting differences in document structure and descriptive granularity.

C. Anchor Construction and Use

Each pillar was represented by an anchor set comprising short behavioural statements designed to capture the conceptual content of the pillar (for example, real-time situated adaptive decision-making in dynamic environments for Pillar 4). Multiple anchor statements were used for each pillar to capture different behavioural expressions of the construct and reduce dependence on any single phrasing. The same anchor sets were applied across occupations and representation systems to ensure comparability.

Anchors were finalised prior to running the cross-occupation comparisons and were not tuned to individual corpora, ensuring that similarity scores reflect representational alignment rather than optimisation artefacts. Table II provides examples of the anchor statements used for each pillar. The complete anchor sets for all four pillars are provided in the supplementary repository: https://github.com/Ilona-woody/Hcrm-representational-audit.

D. Analytical Objective and Interpretation

The objective of the analysis is to assess representational coverage: the extent to which a representation system semantically encodes the capability constructs defined by the four pillars.

Higher similarity scores indicate stronger semantic correspondence between pillar anchors and the descriptors contained in a representation system. Lower similarity scores may indicate abstraction, compression, or omission of construct-relevant content.

Because embedding similarity reflects semantic proximity rather than causal importance, the results are interpreted as indicators of representational emphasis and descriptive granularity. The analysis is descriptive–comparative, not inferential, focusing on systematic differences in representational coverage through occupational databases. The quantitative similarity analysis is complemented by qualitative inspection of high-similarity descriptor matches retrieved from each representation system.

E. Robustness and Sensitivity Analysis

To assess the robustness of the main findings, two sensitivity analyses were conducted. The first varied the aggregation rule: in addition to the maximum cosine similarity used in the primary analysis, mean similarity across all segment-anchor pairs was computed for each descriptor. The second varied the embedding model: a parallel analysis was conducted using the larger all-mpnet-base-v2 model under maximum similarity aggregation, alongside the primary all-MiniLM-L6-v2 model.

The core finding, that O*NET descriptors show the strongest semantic alignment across all pillar–occupation combinations, proved fully stable across all conditions tested: O*NET remained the highest-scoring system in all 36 pillar–occupation–condition combinations examined. The secondary ordering of RPDC above ESCO showed minor variation only in the dog groomer corpus under mean aggregation, consistent with the two exceptions already noted in the main analysis and attributable to the small margin of difference in that corpus. These results indicate that the reported representational ordering is not an artefact of the specific aggregation rule or embedding model selected.

Table IV summarises the results across all three conditions.

VIII. Results

The results are presented in five steps. First, Section VIII.A compares how the three occupational representation systems (RPDC, O*NET, and ESCO) align with the four capability pillars across the selected occupations. Second, Section VIII.B examines pillar-level representation patterns, analysing how each capability dimension is captured by the systems. Third, Section VIII.C focuses on Situated Adaptive Decision-Making (Pillar 4), which exhibits a distinct cross-system representation

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE IV. SENSITIVITY ANALYSIS: STABILITY OF REPRESENTATIONAL ORDERING ACROSS AGGREGATION RULES AND EMBEDDING MODELS

Condition	*Model*	*Aggregation*	*O*NET highest (n/12)*	*Full ordering holds (n/12)*
Primary analysis	all-MiniLM-L6-v2	Maximum	12/12	10/12
Sensitivity 1	all-MiniLM-L6-v2	Mean	12/12	9/12
Sensitivity 2	all-mpnet-base-v2	Maximum	12/12	11/12

Note. O*NET highest = O*NET produces the highest mean similarity score across all 12 pillar–occupation combinations. Full ordering holds = O*NET > RPDC > ESCO across all 12 combinations. Deviations occur exclusively in the dog groomer corpus, consistent with exceptions noted in the main analysis. $n = 12$ (4 pillars $\times$ 3 occupations).

profile and therefore requires more detailed examination. Fourth, Section VIII.D presents a qualitative inspection of high-similarity descriptors for Pillar 4, providing interpretive validation of the quantitative patterns. Fifth, Section VIII.E summarises the representational patterns observed across occupations and representation systems. Finally, Section VIII.F reports a robustness and sensitivity analysis examining the stability of the main findings across alternative aggregation rules and embedding models.

O*NET similarity distributions are identical among occupations because the same set of 41 O*NET Work Activity taxonomy descriptors was used for all analyses. Consequently, cross-occupation variation is driven by differences in the occupation-specific corpora (vocational standards and ESCO skills), while the O*NET alignment profile remains constant. The electrician vocational standards corpus contains substantially more descriptors than the other corpora, reflecting the higher level of procedural specification in electrical installation standards.

Fig. 2 shows the mean cosine similarity scores across pillars and representation systems.

Table V summarises the similarity statistics for each pillar–representation system combination for the three occupations. The table reports mean similarity values, the number of descriptors analysed in each corpus, and distributional statistics, including median, standard deviation, and percentile values. Across representation systems, similarity distributions showed moderate dispersion (standard deviation typically between approximately 0.07 and 0.12), indicating that the observed patterns are not driven by a small number of extreme descriptors.

Across occupations and representation systems, several consistent patterns emerge. First, descriptors associated with Trust-Calibrated Interpersonal Engagement (Pillar 3) show strong semantic alignment, particularly within the O*NET activity taxonomy. Second, Cognitive–Technical Judgement (P1) also exhibits relatively high alignment across systems, reflecting the traditional emphasis of occupational classifications on procedural and problem-solving tasks. Third, Embodied–Sensorimotor Calibration (P2) is moderately represented by the systems and is relatively more visible in RPDC, where procedural and bodily techniques are described in operational detail. Finally, Situated Adaptive Decision-Making (P4) shows a more uneven

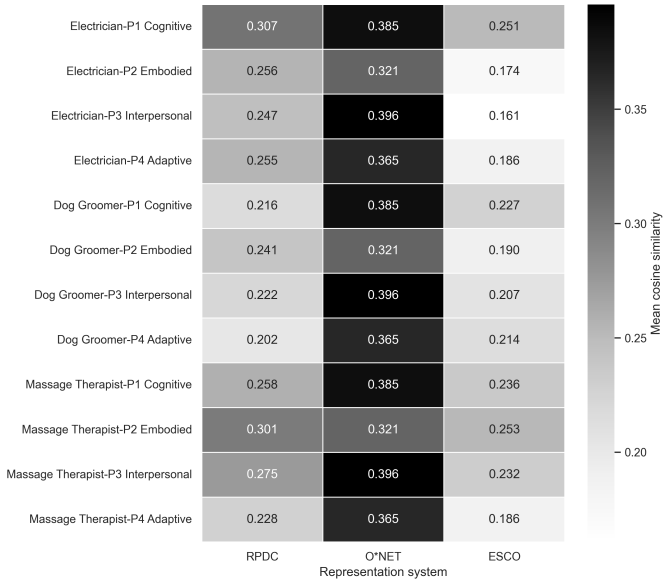


Figure 2. Heatmap of mean cosine similarity scores across pillars and representation systems.

representational profile across systems. While moderately represented in O*NET activities, it exhibits comparatively lower alignment in RPDC and ESCO descriptors. Overall, these patterns suggest that existing occupational taxonomies capture codified tasks, interpersonal interaction, and some forms of procedural embodiment more readily than context-sensitive adaptive behaviour.

The following subsections examine these patterns in greater detail, beginning with cross-system comparisons and then turning to pillar-specific differences.

A. Cross-System Representational Alignment

To evaluate how different occupational representation systems represent the four conceptual pillars, cosine similarity was computed between pillar anchor embeddings and descriptor embeddings extracted from three sources: vocational standards (RPDC), O*NET Work Activities, and ESCO skills. The analysis was conducted for three occupations representing human-centred service and craft work: electrician, dog groomer, and massage therapist.

Across the analysed occupation–pillar combinations, a consistent ordering of representational alignment is observed: O*NET > RPDC > ESCO. In most cases, O*NET descriptors exhibit the highest semantic similarity to the pillar anchors, vocational standards (RPDC) show intermediate alignment, and ESCO descriptors produce lower similarity scores (see Fig. 3a, 3b and 3c). This ordering holds for 10 of the 12 pillar–occupation combinations analysed. Two minor deviations occur in the dog groomer corpus: for Cognitive–Technical Judgement (P1), ESCO (0.227) slightly exceeds RPDC (0.216), and for Situated Adaptive Decision-Making (P4), ESCO (0.214) is marginally higher than RPDC (0.202). These differences are small in magnitude and do not alter the overall cross-system pattern.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE V. SIMILARITY STATISTICS FOR PILLAR–SYSTEM COMPARISONS ACROSS OCCUPATIONS

Occupation	*Pillar*	*System*	*n*	*Mean*	*Median*	*Std*	*Min*	*Max*	*p10 / p90*
Dog Groomer	P1	RPDC	161	0.216	0.209	0.106	0.004	0.502	0.072 / 0.372
	P1	O*NET	41	0.385	0.374	0.090	0.219	0.615	0.278 / 0.491
	P1	ESCO	46	0.227	0.218	0.095	0.051	0.457	0.114 / 0.358
	P2	RPDC	161	0.241	0.233	0.097	0.034	0.519	0.118 / 0.372
	P2	O*NET	41	0.321	0.313	0.106	0.153	0.650	0.192 / 0.429
	P2	ESCO	46	0.190	0.184	0.096	0.067	0.518	0.079 / 0.287
	P3	RPDC	161	0.222	0.202	0.104	0.027	0.518	0.106 / 0.370
	P3	O*NET	41	0.396	0.385	0.078	0.247	0.624	0.309 / 0.494
	P3	ESCO	46	0.207	0.173	0.096	0.070	0.506	0.122 / 0.359
	P4	RPDC	161	0.202	0.193	0.090	0.038	0.450	0.089 / 0.341
	P4	O*NET	41	0.365	0.357	0.084	0.235	0.527	0.255 / 0.482
	P4	ESCO	46	0.214	0.202	0.085	0.076	0.459	0.109 / 0.322
Electrician	P1	RPDC	2341	0.307	0.299	0.122	0.001	0.743	0.150 / 0.467
	P1	O*NET	41	0.385	0.374	0.090	0.219	0.615	0.278 / 0.491
	P1	ESCO	41	0.251	0.242	0.106	0.093	0.554	0.125 / 0.390
	P2	RPDC	2341	0.256	0.249	0.098	0.002	0.734	0.135 / 0.380
	P2	O*NET	41	0.321	0.313	0.106	0.153	0.650	0.192 / 0.429
	P2	ESCO	41	0.174	0.166	0.111	0.000	0.540	0.057 / 0.259
	P3	RPDC	2341	0.247	0.245	0.109	−0.003	0.661	0.104 / 0.391
	P3	O*NET	41	0.396	0.385	0.078	0.247	0.624	0.309 / 0.494
	P3	ESCO	41	0.161	0.148	0.082	0.011	0.330	0.063 / 0.264
	P4	RPDC	2341	0.255	0.254	0.100	−0.013	0.634	0.127 / 0.387
	P4	O*NET	41	0.365	0.357	0.084	0.235	0.527	0.255 / 0.482
	P4	ESCO	41	0.186	0.168	0.107	−0.016	0.443	0.076 / 0.329
Massage Therapist	P1	RPDC	180	0.258	0.240	0.120	0.058	0.683	0.123 / 0.412
	P1	O*NET	41	0.385	0.374	0.090	0.219	0.615	0.278 / 0.491
	P1	ESCO	53	0.236	0.245	0.098	0.072	0.480	0.113 / 0.367
	P2	RPDC	180	0.301	0.293	0.090	0.111	0.557	0.192 / 0.425
	P2	O*NET	41	0.321	0.313	0.106	0.153	0.650	0.192 / 0.429
	P2	ESCO	53	0.253	0.229	0.092	0.081	0.493	0.157 / 0.390
	P3	RPDC	180	0.275	0.251	0.123	0.008	0.637	0.135 / 0.448
	P3	O*NET	41	0.396	0.385	0.078	0.247	0.624	0.309 / 0.494
	P3	ESCO	53	0.232	0.201	0.106	0.048	0.506	0.104 / 0.372
	P4	RPDC	180	0.228	0.216	0.097	0.057	0.528	0.110 / 0.370
	P4	O*NET	41	0.365	0.357	0.084	0.235	0.527	0.255 / 0.482
	P4	ESCO	53	0.186	0.187	0.073	0.060	0.359	0.092 / 0.290

For example, in the electrician corpus, mean similarity values for Pillar 1 (Cognitive–Technical Judgement) were 0.384 for O*NET, 0.307 for RPDC, and 0.251 for ESCO. The same ordering is observed for the remaining pillars and across the other two occupations.

At the same time, the similarity distributions vary systematically by pillar. Trust-Calibrated Interpersonal Engagement (P3) generally achieves the highest alignment, especially within O*NET, while Situated Adaptive Decision-Making (P4) tends to be among the weaker represented dimensions in RPDC and ESCO. In O*NET, however, P4 is not the weakest pillar; it scores above Embodied–Sensorimotor Calibration (P2). This indicates that existing occupational taxonomies do not underrepresent all non-cognitive capabilities equally; rather, they vary in how they represent the interpersonal, embodied, and adaptive aspects of work (see Table V). Spanning systems, the pillars therefore indicate distinct representational profiles: cognitive–technical judgement and interpersonal engagement exhibit relatively strong alignment, embodied calibration shows moderate alignment, and situated adaptive decision-making displays the most variable representation. Fig. 3a, 3b and 3c visualise these differences by showing the mean similarity values for each pillar for the three representation systems for the selected occupations.

B. Pillar-Level Representation Patterns

The following subsections examine each capability pillar individually in order to identify systematic differences in how the dimensions are represented in occupational description systems.

1) Cognitive–Technical Judgement (P1): Cognitive–Technical Judgement was relatively well represented across all systems. O*NET produced the highest similarity values for this pillar in all three occupations, with a mean similarity of approximately 0.385 in each case. RPDC corpora also captured this dimension to a substantial degree, particularly for electrician (0.307) and massage therapist (0.258), indicating the presence of analytical and technical reasoning elements in vocational descriptions. ESCO descriptors showed weaker but still measurable alignment, ranging from 0.216 for dog groomer to 0.251 for electrician.

Overall, the results indicate that structured cognitive and technical judgement is comparatively well represented in the language used by existing occupational description systems.

2) Embodied–Sensorimotor Calibration (P2): Embodied–Sensorimotor Calibration showed moderate representation by all sources, with greater variation across occupations than observed for Pillar 1. O*NET produced the highest mean similarity values overall for this pillar (approximately 0.321 across occupations), although within the O*NET profile, this

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

was the lowest-scoring pillar relative to the other three. This indicates that embodied skill is represented within O*NET, but less strongly than cognitive, interpersonal, and adaptability-related descriptors.

Vocational standards came close in occupations involving more explicit bodily technique, particularly for massage therapist, where RPDC achieved 0.301, compared with 0.321 for O*NET. For the electrician and dog groomer, RPDC values were lower (0.256 and 0.241, respectively), but remained above or near ESCO. ESCO showed the weakest representation of embodied skill overall, although the strength of this pillar varied by occupation, from 0.174 for electrician to 0.253 for massage therapist. These results suggest that embodied competence is represented across all three systems, but more strongly where corpora contain explicit references to manual technique, dexterity, bodily coordination, or physical handling.

3) Trust-Calibrated Interpersonal Engagement (P3): Trust-Calibrated Interpersonal Engagement showed the strongest representation in O*NET in all occupations. Mean similarity values were approximately 0.396 for electrician, dog groomer, and massage therapist (reflecting the identical O*NET descriptor set across occupations), making Pillar 3 the highest-scoring O*NET pillar in all three cases. RPDC also captured this dimension, particularly for massage therapist (0.275), but less strongly for electrician (0.247) and dog groomer (0.222). ESCO descriptors showed weaker overall alignment, ranging from 0.162 for electrician to 0.232 for massage therapist. These results indicate that interpersonal coordination, communication, and related social interaction descriptors are represented relatively strongly in O*NET and, to a lesser extent, in vocational standards.

C. Representation of Situated Adaptive Decision-Making (P4)

Because adaptability showed a distinct cross-system pattern of representation, Pillar 4 is analysed separately. Across all three occupations, this pillar produced lower similarity values than the cognitive and interpersonal pillars. Its relation to embodied skill, however, differed across representation systems. In O*NET, Pillar 4 (mean $\approx$ 0.365 across occupations) scored higher than Pillar 2 (mean $\approx$ 0.321), whereas in RPDC it was the lowest-scoring pillar for dog groomer and massage therapist and slightly below Pillar 2 in the electrician corpus. This indicates that adaptability is represented differently by the systems: in O*NET, it aligns more closely with evaluative, monitoring, and planning language, whereas in vocational standards, it appears less prominently specified than embodied competence. Table VI reports mean cosine similarity values for Pillar 4 across occupations and representation systems.

TABLE VI. MEAN COSINE SIMILARITY VALUES FOR P4

Occupation	*RPDC*	*O*NET*	*ESCO*
Electrician	0.255	0.365	0.186
Dog groomer	0.202	0.365	0.214
Massage therapist	0.228	0.365	0.186

O*NET still produced the highest values among the three systems, but the representational profile of this pillar remained below that observed for Pillars 1 and 3. Within RPDC, Pillar 4 was the lowest-scoring pillar for both dog groomer and massage therapist, and slightly below embodied capabilities (Pillar 2) for electrician. In ESCO, Pillar 4 was also among the weakest represented dimensions, particularly for electrician and massage therapist.

Fig. 4 presents the representation of Situated Adaptive Decision-Making across occupations and representation systems. It shows mean cosine similarity for Pillar 4 only, highlighting the comparatively weaker and more variable representation of this dimension outside O*NET.

These results indicate that adaptability, when defined as situated, non-scripted, real-time behavioural adjustment under uncertainty, is less directly represented in existing occupational taxonomies than cognitive reasoning or interpersonal coordination, and that its representation depends strongly on the descriptive logic of the system considered.

D. Qualitative Inspection of High-Similarity Descriptors for Pillar 4

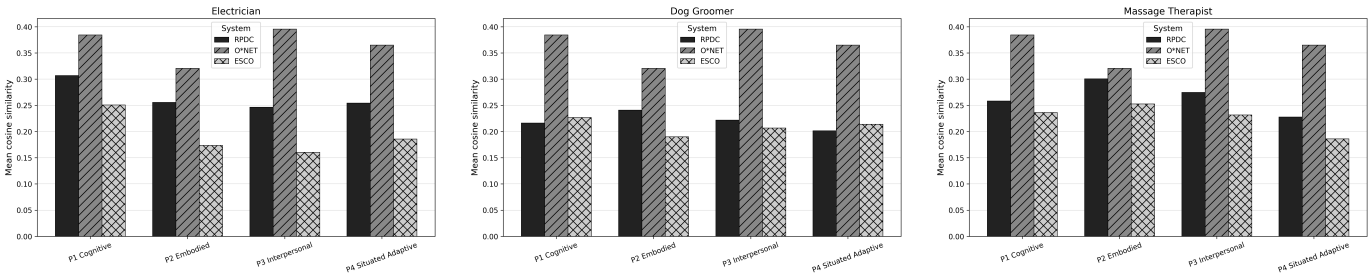
To complement the quantitative similarity analysis shown in Fig. 2–4, the top 15 descriptors per occupation and representation system with the highest cosine similarity to Pillar 4 were examined qualitatively. This step provides interpretive validation by allowing the semantic structure captured by the embeddings to be examined directly in the descriptor language. The descriptors reported here correspond to the highest-ranked matches produced by the similarity computation. Qualitative inspection of nearest neighbours is a common interpretive step in embedding-based analyses, where retrieved matches are examined to interpret the semantic structure captured by vector similarity measures [40] [41]. Similar approaches have also been used in computational analyses of occupational task structures and automation exposure [42].

For all three occupations, O*NET retrieved highly similar descriptors centred on analytical evaluation, monitoring, and planning, including examples such as: “Using relevant information and individual judgment to determine whether events or processes comply with laws, regulations, or standards”, “Monitoring and reviewing information from materials, events, or the environment, to detect or assess problems”, “Analysing information and evaluating results to choose the best solution and solve problems”, “Identifying information by categorising, estimating, recognising differences or similarities, and detecting changes in circumstances or events”, “Developing specific goals and plans to prioritise, organise, and accomplish your work”.

These examples suggest that O*NET primarily encodes adaptability through descriptors of evaluation, information processing, judgment, and planning.

Vocational standards showed a different pattern. In the electrician corpus, high-similarity RPDC matches included more explicit references to situational adjustment and troubleshooting, such as “Takes environmental conditions into account (temperature, humidity, etc.)”, “Easily adapting behaviour or working methods to changing situations and different people”, “Technical understanding (e. g., analysing and troubleshooting faults in electrical and electronic systems)”. In the dog groomer

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) Electrician (b) Dog Groomer (c) Massage Therapist

Figure 3. Mean cosine similarity scores for the four capability pillars across occupational representation systems for (a) electrician, (b) dog groomer, and (c) massage therapist. Bars compare RPDC, O*NET, and ESCO descriptors.

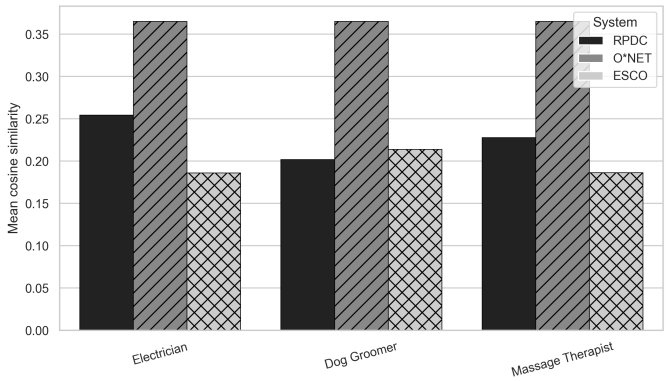


Figure 4. Representation of Situated Adaptive Decision-Making (Pillar 4) across occupations and representation systems. Bars represent mean cosine similarity between pillar anchor sets and occupational descriptors extracted from RPDC, O*NET Work Activities, and ESCO skills.

and massage therapist corpora, high-similarity RPDC matches included descriptors such as “Reaction skills”, “Dealing with change”, “Ability to work independently”, “Goal-oriented working”.

Other RPDC matches combined broader professional characteristics with technical or organisational language, for example, references to analytical skills, individual treatment plans, appointment coordination, and flexibility. All in all, these descriptors acknowledge adaptability, but often do so through a combination of broader professional attributes, technical planning, and occupation-specific competence formulations rather than by consistently explicit descriptions of real-time behavioural adjustment.

ESCO descriptors were more compressed and typically appear as short skill labels rather than process-oriented descriptions. High-similarity examples included “handle customer complaints”, “control animal movement”, “develop personalised massage plan” and “manage a small-to-medium business”. In some cases, ESCO also retrieved descriptors that were relevant but highly condensed, such as “cope with challenging circumstances in the veterinary sector” or “make decisions regarding the animal’s welfare.” Largely, however, the ESCO matches function primarily as categorical labels rather than explicit descriptions of dynamic adaptive behaviour.

Overall, the qualitative inspection shows that all three representation systems capture aspects of adaptability but represent it differently: O*NET through evaluative and planning language, vocational standards through a mixture of occupation-specific planning, adjustment, and broader professional formulations, and ESCO through short categorical skill labels.

E. Summary of Representational Patterns

Across the three occupations, three consistent patterns emerge. First, O*NET provides the strongest overall semantic alignment with the four capability pillars, followed by RPDC and then ESCO, a pattern stable across occupations and distributional statistics. Second, Cognitive–Technical Judgement and Trust-Calibrated Interpersonal Engagement show comparatively strong alignment, while Embodied–Sensorimotor Calibration shows moderate alignment with relatively stronger representation in RPDC for occupations involving explicit bodily technique. Third, Situated Adaptive Decision-Making exhibits the most uneven representational profile, scoring below cognitive and interpersonal pillars quantitatively and differing in semantic form across systems, as confirmed by qualitative inspection.

F. Robustness and Sensitivity Analysis

To assess whether the observed representational ordering is robust to methodological design choices, two sensitivity analyses were conducted as described in Section VII.E. Table IV summarises the results across all three conditions. The finding that O*NET descriptors show the strongest semantic alignment proved fully stable: O*NET remained the highest-scoring system across all 36 pillar–occupation combinations examined, regardless of aggregation rule or embedding model. The secondary ordering of RPDC above ESCO held in the large majority of combinations and showed minor variation only in the dog groomer corpus, consistent with the two exceptions already documented in Section VIII.A and attributable to the small margin of difference in that corpus.

These results indicate that the reported representational ordering is not an artefact of the specific aggregation rule or embedding model selected, and that the core findings are robust to the methodological choices examined.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

IX. Discussion

The following subsections interpret the empirical findings in light of the theoretical literature on tacit knowledge, automation modelling, and human-centred work, and consider implications for occupational taxonomy design and AI labour-market research.

A. Adaptability, Tacit Knowledge, and the Limits of Occupational Representations

This analysis was motivated by an unexpected finding in the HCRM. Based on the theoretical and empirical literature, occupations characterised by intensive interpersonal interaction, embodied skill, and situational variability were expected to exhibit high scores on adaptability. Professions such as electricians, dog groomers, and massage therapists routinely operate in environments that require real-time adjustment to changing physical, social, and contextual conditions. Practitioners must simultaneously interpret signals from clients, animals, or physical environments while executing skilled manual tasks, often in private, uncontrolled settings. Such contexts require continuous decision-making without predetermined scripts and involve the coordination of cognitive, embodied, and interpersonal processes.

From the perspective of automation and expertise research, such open-ended adaptive behaviour has long been recognised as difficult to formalise computationally. Early discussions of this challenge appear in the literature on artificial intelligence and skill acquisition, particularly in the work of Dreyfus and Dreyfus [23], who argued that expert human performance often depends on tacit, context-sensitive judgment rather than explicit rule-following. Similarly, Polanyi's well-known observation that "we know more than we can tell" highlights the central role of tacit knowledge in skilled activity [12]. Within labour economics, this insight has been reflected in discussions of Polanyi's paradox, which emphasises the difficulty of automating tasks that rely on implicit knowledge and situational judgement [43].

Similar observations appear in cognitive systems research on human expertise. In the Skill–Rule–Knowledge framework, Rasmussen [17] describes skilled performance as operating through multiple levels of control, ranging from automatic sensorimotor skill execution to rule-guided procedures and analytical problem-solving. Importantly, experts flexibly shift between these modes depending on the demands of the situation. Such dynamic coordination between embodied action, procedural knowledge, and analytical judgement is characteristic of many craft and service occupations, yet it is difficult to capture within static occupational taxonomies that describe work primarily through discrete tasks or procedural categories, as illustrated by the present analysis.

Against this background, the HCRM framework led to the expectation that occupations characterised by direct human interaction and embodied service provision would display particularly high scores on the adaptability dimension when measured using O*NET adaptability-related indicators. Contrary to this expectation, the empirical analysis revealed slightly lower adaptability scores for the selected occupations. This discrepancy prompted a closer examination of the representational systems used to describe work, specifically the occupational taxonomies and databases that frequently underpin empirical automation analyses.

The results of the present study suggest that the unexpected HCRM finding may reflect a limitation of occupational representation systems, rather than an absence of adaptive capacity in the occupations themselves. Across the three occupations and representation systems examined (O*NET, ESCO, and vocational standards), the pillar representing situated adaptive decision-making generally showed lower semantic alignment than the cognitive and interpersonal pillars and, in several cases, was among the weakest represented dimensions.

Qualitative inspection of high-similarity descriptors further illustrates how adaptability is represented in these systems. The descriptors most closely aligned with the adaptability anchors often emphasised evaluating information, monitoring conditions, planning, or broad professional flexibility, rather than explicit descriptions of real-time behavioural adjustment within dynamic environments. While such descriptors capture aspects of decision-making, they predominantly frame adaptability as deliberative evaluation and planning rather than as real-time behavioural adjustment within dynamic environments.

This distinction is theoretically significant. As Suchman [24] demonstrated in her critique of classical AI planning models, human expertise rarely relies on the execution of formal, prespecified cognitive scripts. Instead, action emerges through ongoing interaction with the immediate conditions of the environment, a phenomenon she describes as situated action. Similarly, the ecological psychology tradition emphasises that skilled behaviour emerges from the coupling between agents and environmental affordances, enabling fluid adjustment to changing constraints [26]. From this perspective, expertise often manifests not as explicit problem solving but as situated responsiveness embedded within continuous perception–action loops.

The present findings, therefore, indicate that such forms of situational adaptability may be compressed or indirectly represented within occupational taxonomies. Because taxonomies must translate complex activities into discrete, standardised categories, they tend to privilege formally articulated procedures and explicit decision points. In the terminology of resilience engineering, this reflects the well-known distinction between work-as-imagined (formal procedures and plans) and work-as-done (the situated, adaptive practices through which work is actually carried out) [44].

B. Occupational Taxonomies and Automation Modelling

These representational dynamics are particularly relevant in the context of labour-market automation research. Empirical research on automation has increasingly emphasised that the economic effects of new technologies depend not only on their technical feasibility but also on how work is organised into tasks within occupations [9]. This perspective builds on the task-based framework in labour economics, which

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

conceptualises occupations as bundles of heterogeneous tasks rather than fixed job categories [25]. Within this framework, occupational databases such as O*NET are frequently used to approximate the task structure of work, providing the empirical inputs for models estimating technological substitution, task exposure, or augmentation potential. Many influential studies estimating the susceptibility of occupations to automation rely on occupational databases such as O*NET to characterise task content and skill requirements. For example, Frey and Osborne [5] used descriptors derived from the O*NET database to train machine-learning classifiers estimating the probability of computerisation across occupations. Related approaches have been adopted in subsequent studies analysing the effects of artificial intelligence, robotics, and digital technologies on employment structures [6] [7] [10]. More recently, analyses of generative AI exposure have relied on O*NET work activities to approximate the task structure of occupations [8].

In such frameworks, occupational characteristics derived from taxonomies function as empirical inputs to models estimating automation risk. If certain dimensions of work are under-represented or abstracted in these representations, automation estimates may partly reflect the representational structure of the taxonomy used to describe work, rather than the full complexity of real-world work activities.

An important contrast emerged between embodied skill and situated adaptability. As shown in the qualitative inspection (Section VIII.D), adaptability in O*NET is primarily represented through abstract descriptors related to evaluation, monitoring, and planning.

In the original HCRM analysis, the embodied dexterity pillar produced the strongest differentiation for the selected occupations. This suggests that although O*NET represents manual activity through general descriptors, such as manual dexterity and finger dexterity, these indicators still capture core aspects of skilled manual work and therefore contribute to higher scores for occupations involving hands-on practice. By contrast, adaptability in O*NET is largely represented by cognitive descriptors related to problem-solving, planning, and decision-making. This formulation may favour professional occupations and may partly explain why the expected adaptability advantage did not appear for the selected service and craft occupations.

In vocational standards, by contrast, embodied skill was often more strongly represented than adaptability, while situated real-time adjustment remained weakly specified. This divergence indicates that different representation systems privilege different aspects of work: survey-based taxonomies more readily represent abstract decision processes, whereas training-oriented documents more readily encode physical technique than unscripted behavioural adaptation.

Taken together, these findings suggest that situated adaptability is one of the capability domains most vulnerable to under-specification in occupational taxonomies. Many service and craft occupations rely on adjustment to evolving biological, interpersonal, and physical conditions that cannot be fully specified in advance. Such adaptive processes have historically posed major challenges for automation. Robotics research has long recognised the difficulty of replicating human adaptability in unstructured environments, an asymmetry famously captured by Moravec's paradox, which observes that high-level reasoning is computationally easier than sensorimotor interaction with complex physical environments [45] [46]. Despite major advances in machine learning and robotics, the real-time coordination of perception, action, and social interaction in dynamic environments remains a central technical challenge.

From this perspective, the representational structure of occupational taxonomies becomes particularly important. If adaptive work practices are primarily represented through abstract cognitive descriptors, occupations that rely heavily on embodied or situational adaptation may appear more amenable to automation than they are under real deployment conditions. This does not invalidate taxonomy-based automation research; rather, it suggests that such models should be interpreted with awareness of the representational assumptions embedded in their input data. The structure of occupational representation systems may influence how automation risk is estimated, highlighting the importance of examining how human capabilities are encoded in the datasets used to model technological change.

C. Implications for Human-Centred Work and AI Integration

More broadly, these findings contribute to ongoing debates about the resilience of human-centred service occupations in the context of artificial intelligence and technological change. Occupations involving close physical proximity, interpersonal trust, and real-time interaction with unpredictable environments occupy a distinctive position within the contemporary labour market. Such roles often require workers to coordinate technical skill, social communication, and embodied responsiveness simultaneously.

In many service contexts, practitioners must continuously interpret subtle signals, such as a client's discomfort, an animal's behavioural cues, or constraints imposed by the physical environment, while adapting their actions accordingly. These adaptive processes are frequently acquired through experience rather than formal instruction and may therefore remain only partially codified in occupational standards, consistent with longstanding insights on the role of tacit knowledge in professional practice [47] [48]. The present analysis suggests that such forms of situated adaptability may therefore be only indirectly represented within existing occupational taxonomies.

These observations do not imply that such occupations are immune to technological change. Rather, they suggest that the interaction between artificial intelligence and human-centred work may be better understood through augmentation frameworks than simple substitution models. Recent labour-economic analyses increasingly emphasise that digital technologies may restructure tasks within occupations rather than eliminate occupations entirely [49]. In this perspective, AI systems may automate certain cognitive components of work while increasing the relative importance of human capabilities that remain difficult to formalise computationally.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Accordingly, in many HCRM-type occupations, AI technologies are likely to function primarily as complementary tools supporting planning, documentation, scheduling, or information retrieval, leaving situational judgement and embodied interaction relatively more dependent on human practitioners. Understanding how occupational representation systems capture, or fail to capture, these capabilities is therefore important for designing AI systems that complement rather than attempt to replace human expertise. In this sense, the HCRM provides a framework for identifying capability domains that may remain under-specified in occupational datasets. The present analysis constitutes an initial step toward a broader empirical programme examining how human-centred capabilities are encoded and measured within the data infrastructures used to model technological change in labour markets.

D. Limitations

This study provides an exploratory analysis of how occupational taxonomies represent different dimensions of human work and encode occupational capabilities in datasets commonly used to model automation risks. The empirical analysis was limited to three occupations and three representation systems. Expanding the range of occupations would help determine whether the observed representational patterns generalise across broader segments of the labour market.

A further limitation concerns the institutional and geographical scope of the occupational representation systems examined. The analysis combines O*NET, which reflects the structure of the U.S. labour market, with ESCO and national vocational standards drawn primarily from European training frameworks. While these datasets represent some of the most widely used occupational taxonomies in labour-market research, they may not fully capture how skills and competencies are defined in other institutional or cultural contexts. In particular, vocational training systems outside Europe and North America may structure occupational knowledge differently.

The embedding-based methodology employed in this study estimates semantic similarity between textual descriptors and conceptual anchor sets, but does not analyse the full structural organisation of occupational taxonomies. A sensitivity analysis reported in Section VIII.F confirms that the core findings are robust to the choice of aggregation rule and embedding model, though the scope of this robustness check remains limited to the conditions examined.

X. Conclusion and Future Work

This study examined how occupational representation systems encode different dimensions of human work, with particular attention to situational adaptability. The analysis was motivated by an unexpected finding in the HCRM: occupations characterised by highly variable interpersonal and physical environments did not exhibit the anticipated high scores on the adaptability pillar. Because occupations such as electricians, dog groomers, and massage therapists require continuous real-time adjustment to changing circumstances, this discrepancy prompted a closer examination of the occupational representational systems used to describe work activities.

Using embedding-based semantic similarity analysis, the study compared three widely used occupational representation systems: O*NET Work Activities, ESCO skills, and national vocational training standards. Across all three occupations, a consistent pattern emerged. O*NET descriptors exhibited the strongest semantic alignment with the conceptual capability pillars, vocational standards showed intermediate alignment, and ESCO descriptors generally produced lower similarity values, with minor exceptions in the dog groomer corpus. At the same time, the pillar representing non-scripted adaptive decision-making showed a weaker alignment than the cognitive and interpersonal pillars and was among the weaker represented dimensions across the examined representation systems.

Qualitative inspection of high-similarity descriptors further clarified this pattern. Adaptability in occupational taxonomies is typically represented through abstract cognitive processes, such as analysing information, evaluating outcomes, or planning actions. While these descriptors capture important aspects of decision-making, they tend to frame adaptability primarily as deliberative reasoning rather than as situated behavioural adjustment within dynamic environments. In many service and craft occupations, however, expertise manifests through continuous perception–action coordination in response to evolving physical, biological, and interpersonal conditions.

Taken together, the findings point to a representational asymmetry in occupational description systems. Dimensions of work that can be explicitly articulated, such as analytical reasoning, planning, and communication, are comparatively well represented, whereas tacit and situational forms of adaptive expertise appear more weakly specified. Because many labour-market automation studies rely on occupational taxonomies to operationalise task content, these representational asymmetries may influence automation exposure estimates for any occupation containing tacit, situational, or context-dependent components of expertise, not only those characterised by highly interpersonal or embodied work, though the present analysis focuses on three illustrative cases from this domain.

Importantly, the results should not be interpreted as evidence that existing taxonomies are inadequate. Systems such as O*NET provide rich and systematically structured descriptions of work activities and remain indispensable resources for labour-market research and policy analysis. Rather, the findings highlight the inherent difficulty of representing certain forms of human expertise, particularly those grounded in tacit knowledge, embodied interaction, and context-sensitive judgement, within standardised classification frameworks.

More broadly, the study contributes to emerging discussions on the data infrastructures underlying automation research. A sensitivity analysis confirmed that the core representational ordering is robust to the choice of aggregation rule and embedding model, supporting the reliability of the findings within the scope of the present study. As artificial intelligence systems perform an expanding range of cognitive tasks, understanding how occupational capabilities are represented in the datasets

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

used to model technological change becomes increasingly important. Embedding-based representational audits, such as the one presented here, offer a systematic method for examining how different dimensions of human work are encoded in occupational databases and for identifying capability domains that may remain underspecified in existing representations of work.

Future research should extend this representational audit to a broader range of occupations and occupational databases, including taxonomies from labour-market regimes outside Europe and North America. Complementary methodological approaches, such as network analysis, ontology comparison, and multi-modal observational datasets including task recordings and sensor-based studies of manual work, could provide additional insight into how embodied and situated capabilities are formally represented and organised within occupational systems. Extending embedding-based audits across multiple labour-market datasets may contribute to a more systematic understanding of how occupational knowledge is encoded in the data infrastructures that increasingly inform economic forecasting and AI capability assessment.

Acknowledgement

The author thanks Prof. Matjaž Gams for his insightful guidance and valuable feedback on the original HCRM. The author is also grateful to Professors Dr Matej Mertik and Dr Maciej Wielgosz for their continued intellectual support throughout the development of this research. Their stimulating discussions, thoughtful questions, and encouragement helped refine many of the ideas presented in this work.

This research was conducted as part of the Applied Artificial Intelligence doctoral programme at Alma Mater Europaea, and the supportive academic environment is gratefully acknowledged. Any opinions, findings, or conclusions expressed in this paper are those of the author and do not necessarily reflect the views of the affiliated institutions.

References

[1] I. E. Negro, 'The resilience of the leisure and care economy: Human-centred niches in an AI-driven labour market', in *AISyS 2025, The Second International Conference on AI-based Systems and Services*, Lisbon, Portugal: IARIA, 2025, pp. 1–3, ISBN: 978-1-68558-303-3.

[2] I. E. Negro, 'The slow food of labour: A task-structural analysis of Human-centred resilience in AI-driven labour markets', in *Proceedings of the Future Technologies Conference (FTC) 2026*, in press, Berlin, Germany, 2026.

[3] N. G. Peterson et al., 'Understanding work using the occupational information network (O*NET): Implications for practice and research', *Personnel Psychology*, vol. 54, no. 2, pp. 451–492, Jun. 2001. DOI: 10.1111/j.1744-6570.2001.tb00100.x

[4] European Commission, 'European skills, competences, qualifications and occupations (ESCO)', 2025, Accessed: 2026-02-08. [Online]. Available: https://esco.ec.europa.eu/en

[5] C. B. Frey and M. A. Osborne, 'The future of employment: How susceptible are jobs to computerisation?', *Technological Forecasting and Social Change*, vol. 114, pp. 254–280, Jan. 2017. DOI: 10.1016/j.techfore.2016.08.019

[6] M. Arntz, U. Zierahn and T. Gregory, 'The risk of automation for jobs in OECD countries: A comparative analysis', OECD, Social, Employment and Migration Working Papers 189, May 2016. DOI: 10.1787/5jlz9h56dvq7-en

[7] E. Brynjolfsson, T. Mitchell and D. Rock, 'What can machines learn, and what does it mean for occupations and the economy?', *AEA Papers and Proceedings*, vol. 108, pp. 43–47, May 2018. DOI: 10.1257/pandp.20181019

[8] T. Eloundou, S. Manning, P. Mishkin and D. Rock, 'GPTs are GPTs: An early look at the labor market impact potential of large language models', 2023, Accessed: 2026-03-07. [Online]. Available: https://arxiv.org/abs/2303.10130

[9] D. Acemoglu and P. Restrepo, 'Artificial intelligence, automation and work', National Bureau of Economic Research, Cambridge, MA, Tech. Rep. w24196, Jan. 2018. DOI: 10.3386/w24196

[10] M. Webb, 'The impact of artificial intelligence on the labor market', Nov. 2019, Accessed: 2026-03-07. [Online]. Available: https://ssrn.com/abstract=3482150

[11] D. H. Autor, F. Levy and R. J. Murnane, 'The skill content of recent technological change: An empirical exploration', *Quarterly Journal of Economics*, vol. 118, no. 4, pp. 1279–1333, Nov. 2003. DOI: 10.1162/003355303322552801

[12] M. Polanyi, *The Tacit Dimension*. Chicago: University of Chicago Press, 1966.

[13] M. J. Handel, 'The O*NET content model: Strengths and limitations', *Journal for Labour Market Research*, vol. 49, no. 2, pp. 157–176, Oct. 2016. DOI: 10.1007/s12651-016-0199-8

[14] R. M. Ryan and E. L. Deci, 'Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being', *American Psychologist*, vol. 55, no. 1, pp. 68–78, 2000. DOI: 10.1037/0003-066X.55.1.68

[15] P. Bourdieu, *Distinction: A Social Critique of the Judgement of Taste*. Cambridge, MA: Harvard University Press, 1984.

[16] P. Bourdieu, *The Logic of Practice*. Stanford, CA: Stanford University Press, 1990.

[17] J. Rasmussen, 'Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models', *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no. 3, pp. 257–266, 1983. DOI: 10.1109/TSMC.1983.6313160

[18] G. Klein, *Sources of Power: How People Make Decisions*. Cambridge, MA: MIT Press, 1998.

[19] D. D. Woods, 'The theory of graceful extensibility: Basic rules that govern adaptive systems', *Environment Systems and Decisions*, vol. 38, no. 4, pp. 433–457, Dec. 2018. DOI: 10.1007/s10669-018-9708-3

[20] M. Mulder, T. Weigel and K. Collins, 'The concept of competence in the development of vocational education and training in selected EU member states: A critical analysis', *Journal of Vocational Education & Training*, vol. 59, no. 1, pp. 67–88, 2007. DOI: 10.1080/13636820601145630

[21] C. Bohne, F. Eicker and G. Haseloff, 'Competence-based vocational education and training (VET): An approach of shaping and networking', *European Journal of Training and Development*, vol. 41, no. 1, pp. 28–38, 2017. DOI: 10.1108/EJTD-07-2016-0052

[22] International Labour Organization, 'Competency-based training (CBT): An introductory manual for practitioners', Oct. 2020, Accessed: 2026-03-11. [Online]. Available: https://www.ilo.org/publications/competency-based-training-cbt-introductory-manual-practitioners

[23] H. L. Dreyfus and S. E. Dreyfus, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press, 1986.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[24] L. A. Suchman, *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, UK: Cambridge University Press, 1987.

[25] D. Autor, 'The "Task approach" to labor markets: An overview', National Bureau of Economic Research, Cambridge, MA, Tech. Rep. w18711, Jan. 2013. DOI: 10.3386/w18711

[26] J. J. Gibson, *The Ecological Approach to Visual Perception: Classic Edition*, 1st ed. New York: Psychology Press, 1979.

[27] F. J. Varela, E. Rosch and E. Thompson, *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press, 1991. DOI: 10.7551/mitpress/6730.001.0001

[28] N. A. Bernstein, *The Co-ordination and Regulation of Movements*. Oxford: Pergamon Press, 1967.

[29] A. R. Hochschild, *The Managed Heart: Commercialization of Human Feeling*. Berkeley: University of California Press, 1983.

[30] E. Goffman, *Interaction Ritual: Essays in Face to Face Behavior*. New York: Doubleday, 1967.

[31] D. J. Deming, 'The growing importance of social skills in the labor market', *Quarterly Journal of Economics*, vol. 132, no. 4, pp. 1593–1640, Nov. 2017. DOI: 10.1093/qje/qjx022

[32] M. Deng and X. Fan, 'Post-editing efficiency and quality assessment: A comparative analysis of Google translate and DeepL', *IOSR Journal of Humanities and Social Science*, vol. 30, no. 1, pp. 10–24, Jan. 2025. DOI: 10.9790/0837-3001021024

[33] M. I. Kamaluddin, M. W. K. Rasyid, F. H. Abqoriyyah and A. Saehu, 'Accuracy analysis of DeepL: Breakthroughs in machine translation technology', *Journal of Education, English, and Foreign Language*, vol. 4, no. 2, pp. 122–126, Jun. 2024. DOI: 10.29303/jeef.v4i2.681

[34] L. Li, 'Artificial intelligence translator DeepL translation quality control', *Procedia Computer Science*, vol. 247, pp. 710–717, 2024. DOI: 10.1016/j.procs.2024.10.086

[35] S. Sabrina, M. Fadhillah, F. O. Chawdri, C. Safara and L. Juniarti, 'Contrastive analysis of DeepL translation vs. Google translate's performance in rendering academic texts', *Jurnal Sinestesia*, vol. 26, no. 1, pp. 73–82, Apr. 2025. DOI: 10.32672/jsi.v26i1.2502

[36] T. Mikolov, K. Chen, G. Corrado and J. Dean, 'Efficient estimation of word representations in vector space', 2013, Accessed: 2026-03-07. [Online]. Available: https://arxiv.org/abs/1301.3781

[37] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding', 2018, Accessed: 2026-03-07. [Online]. Available: https://arxiv.org/abs/1810.04805

[38] N. Reimers and I. Gurevych, 'Sentence-BERT: Sentence embeddings using siamese BERT-networks', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3980–3990. DOI: 10.18653/v1/D19-1410

[39] A. C. Kozlowski, M. Taddy and J. A. Evans, 'The geometry of culture: Analyzing meaning through word embeddings', 2018, Accessed: 2026-03-07. [Online]. Available: https://arxiv.org/abs/1803.09288

[40] W. Wang et al., 'MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers', 2020, Accessed: 2026-03-07. [Online]. Available: https://arxiv.org/abs/2002.10957

[41] W. L. Hamilton, J. Leskovec and D. Jurafsky, 'Diachronic word embeddings reveal statistical laws of semantic change', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1489–1501. DOI: 10.18653/v1/P16-1141

[42] P. Brandes and R. Wattenhofer, 'Opening the Frey/Osborne black box: Which tasks of a job are susceptible to computerization?', 2016, Accessed: 2026-03-07. [Online]. Available: https://arxiv.org/abs/1604.08823

[43] D. Autor, 'Polanyi's paradox and the shape of employment growth', National Bureau of Economic Research, Cambridge, MA, Tech. Rep. w20485, Sep. 2014. DOI: 10.3386/w20485

[44] E. Hollnagel, 'Resilience engineering: A new understanding of safety', *Journal of the Ergonomics Society of Korea*, vol. 35, no. 3, pp. 185–191, Jun. 2016. DOI: 10.5143/JESK.2016.35.3.185

[45] H. Moravec, *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press, 1988.

[46] R. A. Brooks, 'Intelligence without representation', *Artificial Intelligence*, vol. 47, no. 1–3, pp. 139–159, Jan. 1991.

[47] I. Nonaka and H. Takeuchi, *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. New York: Oxford University Press, 1995.

[48] M. Eraut, 'Non-formal learning and tacit knowledge in professional work', *British Journal of Educational Psychology*, vol. 70, no. 1, pp. 113–136, Mar. 2000. DOI: 10.1348/000709900158001

[49] D. Autor, 'Applying AI to rebuild middle class jobs', National Bureau of Economic Research, Cambridge, MA, Tech. Rep. w32140, Feb. 2024. DOI: 10.3386/w32140

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A Modular and Reproducible Framework Based on Plugin Composition for Simulating Cloud Systems

Rubén Luque
Department of Informatics
University of Oviedo
Gijón, Spain
e-mail: luqueruben@uniovi.es

José Luis Díaz
Department of Informatics
University of Oviedo
Gijón, Spain
e-mail: jldiaz@uniovi.es

Joaquín Entrialgo
Department of Informatic Systems
Polytechnical University of Madrid
Madrid, Spain
e-mail: j.entrialgo@upm.es

Rubén Usamentiaga
Department of Informatics
University of Oviedo
Gijón, Spain
e-mail: rusamentiaga@uniovi.es

***Abstract*—Simulating cloud systems requires reproducible execution models that capture latency dynamics and network contention. This paper presents a discrete-event simulation framework designed to enforce experimental reproducibility through a modular, plugin-based architecture. A key innovation is its execution engine, which replaces traditional callback-based simulation processes with asynchronous coroutines (async/await), enabling structured concurrency and simplified exception handling for complex control flows. We incorporate a communication layer using full-duplex channels with configurable directional latencies, allowing the explicit modeling of end-to-end round-trip times. Simulation scenarios are defined declaratively, treating experimental conditions as versioned data. We validate the framework through a case study on container lifecycle management, comparing timing-sensitive 'drain mode' termination policies under stochastic workloads. Results demonstrate that accounting for network delays and bounded queuing reveals significant discrepancies in compliance with service-level objectives compared to network-agnostic simulations, even when operational cost estimations remain similar. We also show that naive load balancing in cost-optimized heterogeneous deployments induces queue imbalance and increased tail latency, resulting in persistent latency objective violations despite sufficient aggregate capacity, thus revealing a blind spot in cost-driven optimizers.**

***Keywords*-Cloud simulation; Discrete-event simulation; Reproducible research; Plugin-based architecture; Network-aware modeling**

I. Introduction

The growing heterogeneity of cloud infrastructures, including container orchestration platforms, serverless computing, and hybrid cloud-edge deployments, introduces new challenges for modeling and evaluating such systems in a systematic and repeatable manner. In this context, discrete-event simulation remains a fundamental tool for studying resource allocation policies, autoscaling strategies, load balancing mechanisms, and cost evaluation models.

To address these limitations, we propose a modular simulation framework designed to enforce reproducibility by design. Unlike monolithic predecessors, our framework, internally named *Nuberu*, decouples the simulation kernel from the policy logic using a strict plugin system. Scenarios are defined declaratively, treating experimental conditions as versioned data rather than code. This approach allows researchers to systematically explore the impact of diverse architectural inputs, from resource allocators to network topologies, without modifying the underlying engine.

This article is an extended version of the work presented in [1], where the foundational principles of the declarative plugin architecture were introduced. In this extended manuscript, we expand the framework's capabilities and evaluation. Specifically, we introduce: (1) a simulation kernel based on Python's native asynchronous coroutines, replacing the traditional generator-based approach to improve the modeling of complex concurrent flows; (2) a communication layer based on full-duplex channels that accounts for network latency and protocol overheads; (3) an extended experimental evaluation that analyzes the impact of stochastic network delays and load balancing strategies on optimized resource allocations; and (4) the framework's full source code, now made publicly available to support reproducible research [2].

The remainder of this paper is organized as follows: Section II reviews related simulation frameworks. Section III presents the architectural foundations, focusing on the native async/await simulation kernel, the component model, and the duplex channel communication primitives; Section IV describes the plugin system, specifically the discovery mechanisms, the interface contracts via protocols and the development workflow. Section V validates the design by comparing termination policies and load balancing strategies under stochastic workloads. Section VI concludes the paper, summarizing key findings and outlining directions for future work.

II. State of the Art

Simulation has long been a fundamental tool for evaluating cloud infrastructures, as real-world experimentation is often prohibitively expensive, time-consuming, and difficult to reproduce. Numerous simulation frameworks have been developed to support the study of cloud systems, each focusing

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

on specific aspects, such as resource provisioning, scheduling policies, or cost modeling.

CloudSim [3] is one of the most established simulators, providing a general-purpose Java framework for modeling datacenters, Virtual Machines (VMs), and application workloads. Despite its configurability, CloudSim lacks a plugin architecture, is tightly coupled to Java workflows, and requires code modification to explore alternative policies, which limits its adaptability and reproducibility.

SimGrid [4] is another mature toolkit for modeling large-scale distributed systems, supporting diverse paradigms, such as High Performance Computing (HPC) and Grid computing. While it enables precise modeling of network and computing resources and has been widely adopted in the systems research community, its focus is broader than cloud infrastructures, and its extensibility relies on low-level Application Programming Interfaces (APIs) rather than composable modules.

Beyond these foundational tools, several recent surveys [5]–[9] systematically review cloud simulation frameworks, identifying common limitations and areas for future research. Mansouri et al. [5] evaluated 33 simulators and concluded that no single tool covers all required dimensions, calling for improvements in Mobile Cloud Computing (MCC) [5][10], federated environments [11], and emerging paradigms, such as edge, fog, and Internet Of Things (IoT) [12][13]. Other studies [6][7] stress the lack of integrated support for security, dynamic behavior, or complex task prioritization, and emphasize the need for reproducibility, flexibility, and modularity in future frameworks.

More recently, several simulators written in Python have gained attention for their accessibility and extensibility. *Yet Another Fog Simulator (YAFS)* [12] simulates microservice deployments over user-defined network topologies, using the SimPy engine, and supports modular control over service placement and routing policies. Although it exhibits high flexibility for customizing placement, routing, and scheduling strategies, allowing dynamic scenario definition via class extension and functions integration, it lacks an explicit plugin system for external and decoupled integration of new core components. As a result, adding new functionality in YAFS often requires more intrusive modifications to the core codebase. *Cloudy* [14], by contrast, introduces a hybrid discrete-time and event-driven simulator with native Graphics Processing Unit (GPU) support and integration plans for optimization and machine learning (ML) libraries. However, its extensibility depends on manual template duplication, and it lacks a unified declarative configuration system. Similarly, *ECLYPSE* [15] focuses on composable cloud architectures but bypasses a formal plugin ecosystem entirely. Instead, it ties extensibility directly to Python's internal mechanisms, specifically inheritance and decorators. While this utilizes language features, it requires users to navigate the object hierarchy rather than interacting with decoupled extension interface.

Moreover, a significant limitation persists regarding experimental reproducibility. Existing frameworks typically adopt a *code-as-configuration* approach, where scenario definitions are embedded within imperative Python scripts or rely on unversioned external libraries. While YAFS supports JSON for topology definition, dynamic policies remain coupled to the codebase [12]. Similarly, ECLYPSE scenarios are constructed by programmatically instantiating configuration objects. This practice complicates compliance with FAIR (Findable, Accessible, Interoperable, and Reusable) principles [16], as the experimental setup is not an immutable artifact. In contrast, a declarative configuration model decouples the simulation kernel from the experiment definition, treating scenarios as static, auditable data.

These Python-based initiatives highlight the community's growing interest in modern, flexible, and scriptable simulation platforms. However, to the best of our knowledge, none of them adopts a modular, plugin-based architecture as the one we propose for simulating cloud environments.

TABLE I. SIMULATORS COMPARISON.

Simulator	Modularity	Declarative Config	Plugin Support	Reproducibility	Limitations	Use Cases
CloudSim	○	○	○	◐	(a)	VM scheduling, provisioning
SimGrid	◐	◐	◐	◐	(b)	HPC, distributed simulation
YAFS	◐	◐	○	●	(c)	Fog, IoT strategy evaluation
Cloudy	◐	○	○	◐	(d)	Cloud + ML, GPU workloads
ECLYPSE	●	○	◐	●	(e)	Edge-cloud prototyping
Nuberu (proto)	●	●	●	●	(f)	Reproducible workflows

Legend of symbols:
● Fully supported
◐ Partially supported
○ Not supported

Limitation notes:
(a) Rigid architecture, low modularity.
(b) Low extensibility, low-level abstractions.
(c) No plugin interface, core modification required.
(d) No declarative config, manual extension required.
(e) Tightly coupled modules, no plugin API.
(f) See Section VI.

Table I summarizes and contrasts key features of representative simulators in the domain, highlighting their support for modularity, configuration mechanisms, extensibility, and reproducibility, along with their limitations and typical application areas. As shown, none of the existing solutions fully meets all desired characteristics, especially in terms of reproducibility and plugin support.

III. ARCHITECTURAL FOUNDATIONS

The proposed framework targets cloud simulation and is based on a discrete-event simulation engine, a global event bus, and a set of pluggable components. This design minimizes coupling between simulation logic and system policies, allowing researchers to prototype, compare, and reproduce complex deployment strategies with minimal implementation effort.

Figure 1 presents the concrete architecture of the framework. Driven by a declarative YAML configuration, the system relies on a Simulation Kernel that integrates the `aSimPy` environment (a custom fork of the SimPy discrete-event engine adopting an `async/await`-based syntax. See Section III-A)

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

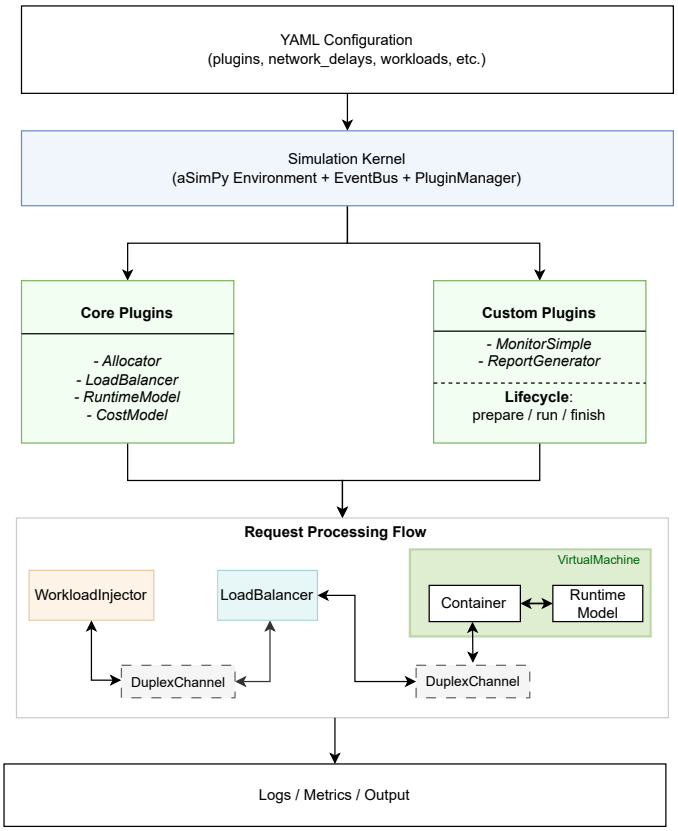


Figure 1. Detailed architecture of the Nuberu framework.

for asynchronous execution, the EventBus for decoupled control signaling, and the PluginManager for dynamic extension loading.

The diagram distinguishes between Core Plugins (which provide essential logic like allocation or load balancing) and Custom Plugins (focused on observability and lifecycle management). Furthermore, it illustrates the Request Processing Flow, visualizing how the Workload Injector, Load Balancer, and Virtual Machines interact through the DuplexChannel primitives to route traffic and generate simulation outputs.

A. Core Simulation Engine: Native Async/Await

The simulation kernel follows a discrete-event model built upon a custom fork of the `SimPy` library [17], referred to here as `aSimPy`. The fork replaces `SimPy`'s generator-based process definitions with Python's native async/await syntax, while preserving the original event scheduling semantics and execution model.

The primary motivation for this refactoring is to improve the expressiveness and clarity of simulation code. In scenarios involving multiple interacting components—such as virtual machines, containers, and request dispatchers—the async/await syntax enables a more linear and compositional representation of process logic than explicit yield-based suspension.

From a semantic perspective, `aSimPy` preserves `SimPy`'s execution model. Although simulation processes are expressed as coroutines, their execution remains governed by the underlying discrete-event scheduler. Suspension and resumption points correspond to simulated events rather than to concurrency primitives managed by Python's native asynchronous runtime.

B. Component Model and Plugin Architecture

This framework supports two user roles: developers, who create alternative implementations of pluggable components by writing plugins that conform to predefined interfaces, and analysts, who design simulation scenarios by selecting among available plugins without modifying the core system. In practice, a single user may assume both roles, developing custom components and designing simulation scenarios.

The simulation framework distinguishes between core components, which define the structure and control flow of the system, and pluggable components, which encapsulate specific, customizable behaviors. Core components include the discrete-event simulation engine, the communication primitives (e.g., event bus and channels) and essential modules, such as the workload injector and the allocator. These components are not pluggable themselves, but delegate critical functionality—such as workload characteristics, allocation strategy, or cost modeling—to user-defined plugins.

The mechanisms for dynamic discovery, registration, and static validation of these plugins are detailed in Section IV.

C. Communication Primitives

To achieve realistic modeling of distributed systems, the architecture moves beyond idealized method calls between objects. Instead, it employs two distinct communication patterns: a global Event Bus for global signals and Full-Duplex Channels for data plane interactions.

1) Event Bus and Inter-component Communication: Global system events, such as VM_STARTED or SIMULATION_FINISHED, are propagated via an asynchronous EventBus. This provides strict decoupling: auxiliary components, such as metric collectors, trace loggers or user custom plugins, can subscribe to REQUEST_COMPLETED events without the Workload Injector or Load Balancer being aware of their existence. This mechanism models an out-of-band communication mechanism, preventing observability instrumentation from introducing artificial blocking in the operational flow of requests.

2) Duplex Channels with Configurable Latency: Operational request-response interactions use explicit point-to-point *Duplex Channels*. The proposed framework models network propagation delays as configurable parameters. A duplex channel D connecting two components A and B is composed of two unidirectional buffered queues:

$$D_{A \leftrightarrow B} = \langle C_{\text{fwd}}(\delta_{\text{fwd}}), C_{\text{bwd}}(\delta_{\text{bwd}}) \rangle \tag{1}$$

where each channel component C represents a directional path, forward ($A \rightarrow B$) or backward ($B \rightarrow A$), and is configured with an independent transmission delay δ. This abstraction specifically supports the modeling of asymmetric network conditions (where $\delta_{\text{fwd}} \neq \delta_{\text{bwd}}$) and distinct

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Listing 1: send() method in Channel

```
async def send(self, message) -> None:
  await self.env.timeout(self.delay) # Enforces configured
                                     # propagation delay
  await self.store.put(message)
```

infrastructure overheads per direction. The implementation enforces these delays asynchronously using the simulation environment's timeout primitives, as shown in Listing 1.

a) Formal Response Time Model: The total response time T_{response} for a request traversing the system is formally defined as the aggregation of network transit, queuing, and service time:

$$T_{\text{response}} = T_{\text{network}} + T_{\text{queue}} + T_{\text{service}} \tag{2}$$

where:

- $T_{\text{network}} = \delta_{wi \to lb} + \delta_{lb \to vm} + \delta_{vm \to lb} + \delta_{lb \to wi}$ represents the accumulated network round-trip time (RTT). Note that forward and backward delays are configured separately in the scenario.
- T_{queue} is the waiting time in the runtime model's buffer.
- T_{service} is the nominal service time provided by the performance model or benchmark.

T_{queue} is an *emergent variable* determined by the system's runtime state (congestion) and the queuing policy, whereas T_{network} is a configuration parameter specified declaratively. This distinction allows researchers to isolate the impact of network topology from resource contention.

Figure 2 illustrates the component interaction throughout the request lifecycle. Network delays (δ) translate into asynchronous timeouts during transmission phases, while service (T_{service}) and queuing (T_{queue}) times depend on the internal execution of the Runtime Model. The diagram confirms the symmetry of the communication path and the isolation of latency sources.

b) Hierarchical Latency Propagation: The operational lifecycle of a request is modeled through a two-tier hierarchy of duplex channels, representing distinct boundaries:

1) *Client–Load Balancer layer:* Bidirectional communication governed by the round-trip delay wi_to_lb.
2) *Load Balancer–Container layer:* Intra-datacenter communication governed by lb_to_vm_default.

To illustrate how these delays aggregate, Table II presents a breakdown of a single request's lifecycle. Each transition appends a timestamped entry to the request's timeline attribute (see Listing 2), creating an auditable trace that adheres to FAIR principles [16].

c) Declarative Configuration and Overrides: Delays are specified in the scenario YAML, enabling version-controlled experimentation. The framework supports granular overrides to simulate heterogeneous environments, such as multi-region deployments (Listing 3).

TABLE II. REQUEST LATENCY BREAKDOWN ACROSS CHANNEL HIERARCHY.

Event	Comp.	Time (ms)	δt	Delay Source
Req. generation	WI	0.0	—	—
LB rx request	LB	10.0	+10.0	`wi_to_lb` (fwd)
C rx request	C	15.0	+5.0	`lb_to_vm` (fwd)
RM complete	RM	25.0	+10.0	Service
LB rx response	LB	35.0	+5.0	`lb_to_vm` (bwd)
Client rx response	WI	40.0	+10.0	`wi_to_lb` (bwd)

Listing 2: RequestTimestamp structure for FAIR traceability

```
@dataclass
class RequestTimestamp:
    time: float            # Simulated time
    component_id: str      # e.g., "Container_c1"
    event: str             # e.g., "receive_from_lb"
    metadata: dict         # queue_length, vm_id, etc.
```

d) Advanced Channel Semantics: The duplex channel architecture incorporates two features essential for cloud orchestration:

Implicit Response Routing: By encapsulating the return path within the channel state, the architecture avoids complex correlation ID management. When the Load Balancer dispatches a request, the response inherently propagates backward through the same topological path. This simplifies component logic, as containers remain agnostic of the upstream topology.

Half-Close Semantics: The channels implement half-close semantics to support graceful shutdown. A component can close its output direction (signaling no further requests will be sent) while keeping the input direction open to process pending messages. This is the foundational mechanism for the 'drain' termination policy detailed in Section V.

D. Simulation Configuration

The simulation runtime is configured declaratively via a YAML file specifying parameters such as the simulation duration, the names of the plugins to be loaded for each functional component, and the input data, such as workloads, infrastructure specifications, performance data or allocation strategies. It can also define external data sources, such as workload traces in custom formats, to be parsed and injected

Listing 3: Network configuration with heterogeneous overrides

```
network_delays:
  wi_to_lb: 10.0          # Default Client-LB latency
  lb_to_vm_default: 5.0   # Default intra-datacenter latency
  vm_container_overhead: 2.0  # Container overhead

# Overrides for specific topology scenarios
vm_network_configs:
  - name: "vm_remote_az"
    network_delay: 50.0  # Higher latency for cross-AZ VMs
```

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

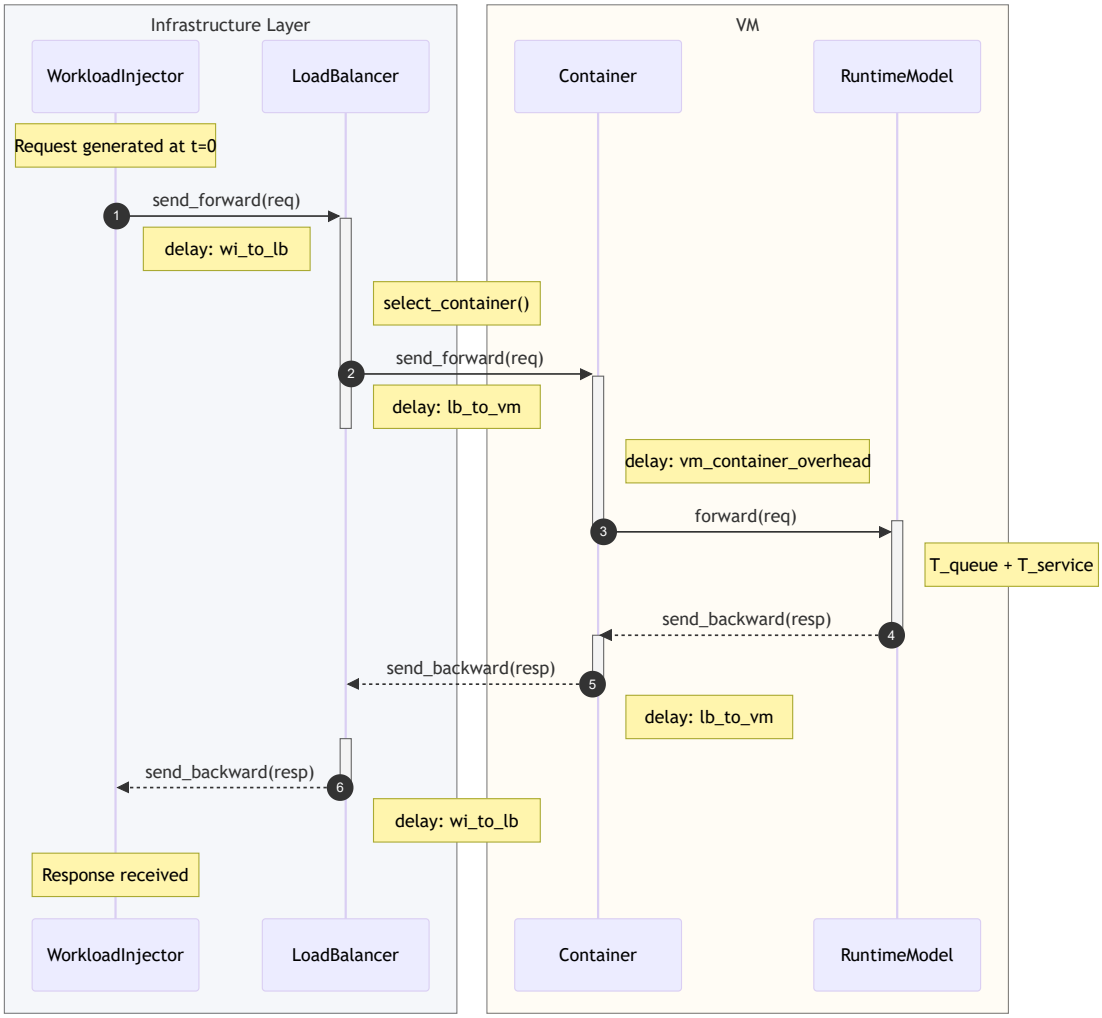


Figure 2. Sequence diagram of the request lifecycle showing asymmetric network delays, container overhead, and internal runtime states ($T_{\text{queue}}, T_{\text{service}}$).

at runtime by compatible plugins. An full example of such a configuration file is provided in Appendix A. This enables integration with external tools, such as cost optimizers, whose solutions can be imported through the appropriate plugin.

The framework follows a *microkernel-inspired* design, in which the simulation engine acts as a lightweight orchestrator. Pluggable components are dynamically instantiated, operate in isolated asynchronous processes, and communicate exclusively through the primitives defined in Section III-C. This design allows for flexible composability and simplifies the development and integration of experiment-specific logic without entangling it with the simulation kernel.

1) Reproducibility through Declarative Specification: Experimental reproducibility is achieved by capturing the complete simulation specification as a version-controlled YAML file. Formally, each simulation run is parameterized by a configuration tuple:

$$\Theta = \langle \theta_{\text{kernel}}, \theta_{\text{plugins}}, \theta_{\text{inputs}}, \theta_{\text{net}} \rangle \tag{3}$$

where:

- θ_{kernel} specifies simulation duration and stop conditions.
- θ_{plugins} declares the mapping of functional roles to specific plugin implementations.
- θ_{inputs} specifies workload traces and infrastructure specifications.
- θ_{net} specifies the network latency parameters defined in Section III-C2.

By versioning Θ as a single artifact, experimental variations can be expressed as controlled perturbations $\Delta\Theta$, such as sweeping a latency parameter or switching a load-balancing policy, rather than as implicit code modifications. This design makes experiment-defining assumptions explicit and auditable, adhering to the FAIR principles [16].

IV. Plugin System and Extensibility

A central design principle of Nuberu is decoupled extensibility: users should be able to alter simulation behavior without modifying the core source code. The architecture leverages the `pluggy` library [18] for dynamic discovery and Python Protocols [19] for static interface enforcement.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A. Discovery and Validation Mechanism

The framework uses the `pluggy` library [18] to support dynamic plugin discovery using Python's entry_points mechanism. Each plugin is an installable python package which registers itself in the `pyproject.toml` file under a specific namespace (e.g., `application_model.llm`, `cost_model.default`), which enables the simulation framework to identify the type and logical name of each component. Once installed in the Python environment, plugins are automatically discovered at runtime without requiring any additional code modification.

Multiple plugins of the same type can be installed and selected declaratively through the YAML configuration file.

If a plugin declared in the YAML configuration cannot be found or does not conform to the expected interface, the simulation engine is designed to abort execution and issue a descriptive error. This validation occurs at startup time, before any event is generated, ensuring that core simulation behavior remains consistent and reproducible, even when user-defined extensions are used in the configuration. Non-essential third-party plugins, such as auxiliary observers or loggers, may fail gracefully with a warning, allowing the simulation to proceed when their absence does not compromise correctness.

B. Interface Contracts via Protocols

Each pluggable component in the framework adheres to two complementary interface mechanisms. First, a *hook specification* (`hookspec`) is defined using `pluggy`, which declares the methods that a plugin must implement to be properly registered and invoked at runtime. Second a Python `Protocol` interface is used for each plugin type, enabling static type checking and improved developer experience. These protocols specify the required methods (e.g., get_workloads(), apply_allocation(), compute_cost()) and allow for static verification using tools, such as `mypy`.

This dual-layer interface ensures runtime compatibility via `pluggy`, while also providing static guarantees, editor support, and better documentation through `Protocol`. Together, these mechanisms improve reliability, reduce integration errors, and facilitate the rapid development of new components.

Table III summarizes the primary protocols defined in the framework. Each protocol enforces a specific set of methods that ensure the component complies with the simulation kernel interfaces.

C. Plugin Taxonomy: Core vs. Custom

The framework unifies the extension mechanism using a strict *Factory Pattern* for all plugin types. Regardless of their category, all plugins expose a factory hook that the kernel invokes to instantiate the component. This design ensures that runtime dependencies (such as the `Environment` or `EventBus`) are injected only when the simulation is fully initialized, while allowing configuration parameters to be captured earlier from the YAML definition.

Listing 4: Example of a Core Plugin implementation using the Factory Pattern.

```
from nuberu.plugins import PluginBase, hookimpl

class RuntimeModelSimple:
    # Implementation details omitted for brevity
    ...

class RuntimeModelPlugin(PluginBase):
  plugin_name = "runtime_model_simple"

  @hookimpl
  def get_runtime_model_factory(self, conf: dict):
    # Capture config from YAML
    q_size = conf.get("queue_size", 1000)

    # Factory receives runtime dependencies
    def factory(env, container, channel):
      return RuntimeModelSimple(env, container, channel,
                                q_size, self.logger)
    return factory
```

While the instantiation pattern is shared, plugins are classified into two categories based on the nature of the object they return.

1) Core Plugins: These plugins provide essential implementations for system modules such as *Load Balancers*, *Allocators*, *Runtime Models* or *Workload Injectors*. The factory returns a functional component responsible for specific simulation decisions (e.g., routing a request).

Listing 4 presents a simplified implementation of a Runtime Model plugin. The factory captures the `queue_size` from the configuration closure and returns a callable that the kernel will later invoke with the necessary dependencies.

2) Custom Plugins: Custom plugins run concurrently with the simulation to perform auxiliary tasks, such as real-time monitoring or live visualization. Their factory returns an object that adheres to the `CustomPluginProtocol`, enforcing a specific lifecycle:

- `prepare()`: Invoked once before the simulation starts (e.g., to subscribe to events).
- `run()`: Invoked at $t = 0$ to spawn asynchronous background processes.
- `finish()`: Invoked after the simulation ends for cleanup.

D. Implementation Example

Listing 5 demonstrates a minimal Custom Plugin that monitors request completion events. Note that the `RequestMonitorPlugin` defines the factory hook, which instantiates the `RequestMonitor` runtime object. This separation allows the runtime object to safely interact with the `env` and `event_bus` injected by the factory.

E. Project Structure and Development Workflow

To support the extensibility model described above, the framework is organized as a *monorepo workspace*. This structure enforces a physical separation between the simulation

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE III. KEY PROTOCOL INTERFACES DEFINING THE EXTENSION POINTS OF THE FRAMEWORK.

Component Type	Protocol	Factory Method	Functional Role
Infrastructure	`InfrastructureSpecs`	get_infrastructure_factory()	Defines Apps, InstanceClasses, and ContainerClasses for the simulated environment.
Allocation	`AllocationSpecs`	get_allocations_factory()	Maps container replicas to VMs based on optimization results or heuristics.
Performance	`PerformanceSpecs`	get_rps_for_app_factory()	Computes RPS capacity for a given app on an instance class.
Load Balancer	`LoadBalancerSpecs`	get_load_balancer_factory()	Implements routing logic (Round-Robin, SWRR) to distribute requests.
Workload	`WorkloadSpecs`	get_distribution_params_for_app_factory()	Extracts traffic parameters (num_reqs, time_slot) from traces.
Distribution	`ArrivalDistributionSpecs`	get_interarrival_times_factory()	Generates arrival times based on statistical distributions (Poisson, etc.).
Runtime Model	`RuntimeModelSpecs`	get_runtime_model_factory()	Models service time, queuing logic, and execution behavior.
Cost Model	`CostModelSpecs`	get_cost_model_factory()	Calculates infrastructure costs based on VM usage.
Custom Plugin	`CustomPluginSpecs`	get_factory()	Provides auxiliary processes for monitoring, logging, or visualization.

```
nuberu-monorepo/
├─ pyproject.toml              Workspace config
├─ core/
│  ├─ pyproject.toml           Deps: asimpy, pluggy
│  └─ src/nuberu/
│     ├─ core/                 EventBus, NetworkDelays
│     ├─ interfaces/           Protocols definition
│     ├─ channels/             DuplexChannel
│     └─ plugins/             PluginManager
├─ plugins/
│  ├─ nuberu-plugin-smoothwrr-lb/
│  │  └─ pyproject.toml        Entry: load_balancer swrr
│  └─ nuberu-plugin-conlloovia-dp/
│     └─ pyproject.toml        Deps: conlloovia
└─ examples/
   └─ example_v3.1/configs/
      └─ example_v3.1.yaml     Declarative config
```

Figure 3. Directory structure of the monorepo workspace.

kernel and the extension ecosystem, mirroring the logical decoupling provided by the architecture.

The repository is structured into two primary directories, as illustrated in Figure 3:

- `core/`: Contains the simulation core built on top of the aSimPy discrete-event library, the definitions of Protocol interfaces, and the Event Bus implementation. It maintains minimal external dependencies to ensure a lightweight and stable core.
- `plugins/`: Hosts the extensions ecosystem. Each subdirectory correspond to an independent Python package with its own `pyproject.toml` configuration file.

This organization enables strict *dependency isolation*. For instance, a data-provider plugin that depends on computationally intensive libraries (e.g., `pandas` or `scikit-learn` for trace parsing and preprocessing) can declare these requirements locally, without propagating them to the core engine or to other lightweight plugins.

From a workflow perspective, the build system (based on `uv`) treats the repository as a unified workspace. This approach allows developers to install the core and a specific subset of plugins in editable mode, thereby supporting rapid prototyping and iterative development while preserving the stability of the simulation kernel.

V. EXPERIMENTAL VALIDATION: COMPARING OPTIMIZED ALLOCATIONS WITH SIMULATED BEHAVIOR

Optimizers based on mathematical models, such as linear programming, often rely on idealized assumptions about workload, resource performance, and system behavior. This section investigates to what extent such optimized allocations remain effective when deployed in a more realistic simulated environment. By simulating the deployment plan produced by the optimizer under multiple runtime conditions, we aim to identify discrepancies, stress points, and potential modeling oversights. This not only validates the practical viability of the computed solution but also highlights the role of simulation as a complementary tool for refining optimization strategies.

To demonstrate how the proposed architecture supports rigorous, scenario-driven evaluation, we present a case study executed with *Nuberu*. The goal of this validation is twofold: first, to demonstrate the structural fidelity of the simulator by verifying that it produces physically consistent results under stress (e.g., network contention and queuing); and second, to show how these realistic simulations expose hidden assumptions in static optimization models, guiding their refinement.

A. Description of the scenario to simulate

The scenario is derived from the output of *Conlloovia* [20], an optimizer that allocates container replicas on VMs to minimize cost while ensuring the throughput expressed in rps (requests per seconds) of each application reaches its 95th percentile (p95). Inputs include VM instance classes, container classes, and performance matrices (see Table IV). We analyze a one-hour segment involving two applications: *app0* (stable load, p95=44 rps) and *app1* (variable load, p95=117 rps), as shown in Figure 4.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

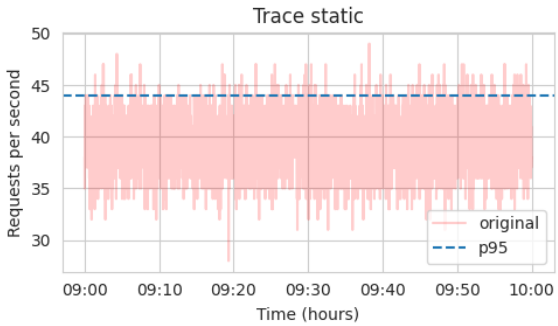
Listing 5: Implementation of a Custom Monitoring Plugin

```
from nuberu.plugins import PluginBase, hookimpl
from nuberu.core.events import EventTopic

class RequestMonitor:
    """Implementing␣CustomPluginProtocol."""
    def __init__(self, env, event_bus, logger):
        self.env = env
        self.event_bus = event_bus
        self.logger = logger
        self.queue = None

    def prepare(self) -> None:
        """Phase␣1:␣Subscribe␣to␣relevant␣topics."""
        self.queue = asimpy.Store(self.env)
        self.event_bus.subscribe(
            EventTopic.REQUEST_COMPLETED,
            self.queue)

    def run(self) -> None:
        """Phase␣2:␣Spawn␣concurrent␣collection␣loop."""
        self.env.process(self._monitor_loop())

    async def _monitor_loop(self) -> None:
        while True:
            event = await self.queue.get()
            self.logger.info(
            f"Req␣{event.payload['id']}␣done")

    def finish(self) -> None:
        """Phase␣3:␣Finalize␣artifacts."""
        self.logger.info("Sim␣finished.")

class RequestMonitorPlugin(PluginBase):
    """Plugin␣Entry␣Point."""
    plugin_name = "request_monitor"

    @hookimpl
    def get_factory(self, config: dict):
        # Factory injects dependencies
        def factory(env, event_bus):
            return RequestMonitor(
            env,
            event_bus,
            self.logger)
        return factory
```

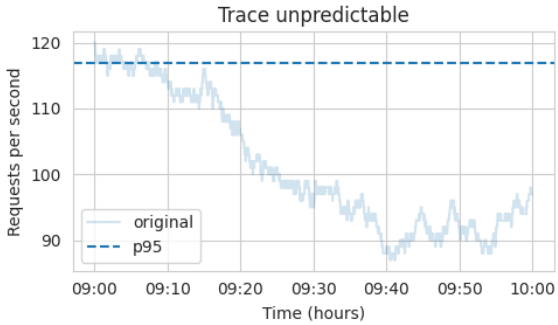
TABLE IV. PERFORMANCE IN REQUEST PER SECOND (RPS) OF EACH CONTAINER CLASS ON EVERY VM INSTANCE CLASS

C. Class / VM I. Class	cc0app0	cc0app1	cc1app0	cc1app1	cc2app0	cc2app1
c5.2xlarge	2.10	0.46	4.30	0.96	6.35	1.63
c5.large	2.10	0.46	4.30	0.96	6.35	1.63
c5.xlarge	2.10	0.46	4.30	0.96	6.35	1.63
c6i.2xlarge	2.29	0.50	4.71	1.02	6.82	1.76
c6i.large	2.29	0.50	4.71	1.02	6.82	1.76
c6i.xlarge	2.29	0.50	4.71	1.02	6.82	1.76

It is important to clarify the relationship between the optimizer's constraints and the simulator's metrics. *Conlloovia*'s objective is to minimize cost subject to a throughput constraint: the allocated capacity must meet or exceed the 95th percentile of the forecasted load. While the optimizer does not explicitly model response times, latency Service Level Objectives (SLOs) are implicit in the input Performance Matrix. The maximum rps defined for each container class represent a



(a) Workload for app0



(b) Workload for app1

Figure 4. Plot of the workloads for each application

benchmarking-derived limit, below which the container is known to satisfy its latency requirements.

In contrast, *Nuberu* simulates the full request lifecycle, including network hops and queuing delays, allowing us to measure the 99th percentile of Response Time (p99 RT) as an emergent output. This distinction is crucial: the experiments validate whether the optimizer's static assumption—that satisfying throughput constraints guarantees acceptable latency—holds under dynamic runtime conditions such as network contention or variability.

The deployment plan computed by *Conlloovia* is visualized in Figure 5. The solution involves 38 VMs hosting a total of 116 containers.

A critical detail in this deployment is the performance heterogeneity introduced by the cost-minimization objective. While the majority of App1 containers are of the high-performance class (*cc1app1*, dark blue, $\approx$ 1.02 rps), the optimizer allocated two instances of a lower-performance class (*cc0app1*, light blue, $\approx$ 0.50 rps). This decision is perfectly rational from a throughput perspective: these cheaper, slower containers provided exactly the marginal capacity needed to reach the target p95 load without paying for a full-sized instance. However, this creates a potential problem: while the aggregate throughput is sufficient, the presence of these slow nodes introduces a risk of head-of-line blocking if the load balancer is unaware of their limited capacity.

Using a custom plugin, *Nuberu* bootstraps the exact infrastructure proposed by Conllovia to inject traffic and collect run-

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

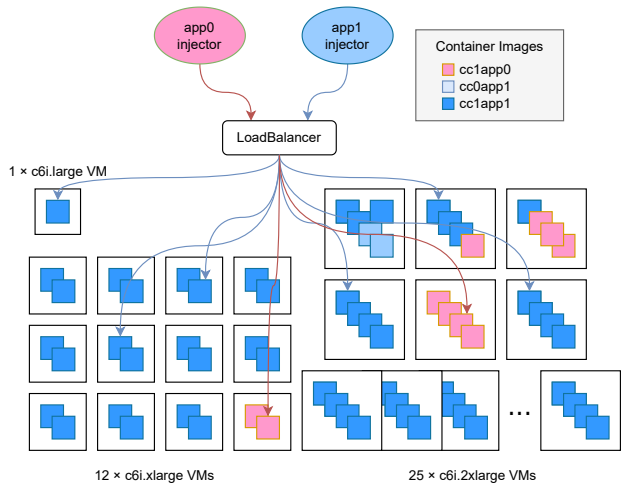


Figure 5. Optimized Deployment Plan. The solution mixes different container classes to minimize cost.

time metrics, allowing us to compare the theoretical "optimal" performance against simulated reality.

B. Experimental design

To assess the analytical power of the simulator, we expand the scenario into 32 configurations by combining five binary dimensions. This design allows us to isolate specific architectural factors:

1) **Load injection (Load)**: either the original *trace* (realistic variability) or a synthetic *Poisson* process matching the p95 throughput which is the target of the optimizer.
2) **Load Balancing (LB)**: either a naive Round-Robin (*RR*) or a Smooth Weighted Round-Robin (*SWRR*) [21], which accounts for container performance heterogeneity.
3) **Queuing model (Q)**: either *zero buffer* ($Q = 0$, requests dropped if busy) or *large queues* ($Q = 1000$).
4) **Termination policy (Term)**: Specifies the shutdown strategy, defined as either *hard* (immediate termination) or *drain* (graceful termination). Under the *drain* policy, containers continue processing until their request queues are empty, ensuring the underlying VMs remain active until all their containers are shut down.
5) **Network Latency (Lat)**: either *ideal* (0 ms) or *realistic* (20 ms per channel). In Nuberu's topology, a request traverses four network hops, as seen in Figure 2, namely: Client $\rightarrow$ LB $\rightarrow$ Container $\rightarrow$ LB $\rightarrow$ Client. Therefore, a 20 ms channel latency implies a theoretical round-trip time (RTT) floor of $4 \times 20 = 80$ ms.

All experiment definitions and results are available in the public repository [22].

C. Discussion

Table V summarizes three selected cases that illustrate the divergence between theoretical optimization models and runtime behavior. We analyze these results also to validate the structural fidelity of the simulator. The complete results for all 32 configurations are detailed in Table VI.

1) Baseline Validation: Intrinsic Time and Buffering: The first block of results (Case A) serves as a baseline validation, confirming that the simulator outcomes align with the optimizer's predictions under idealized conditions. In the absence of queues ($Q = 0$) and network delays ($Lat = 0$), the simulator isolates the raw computational performance of the allocated resources. The observed p99 response time of 0.98 s represents the *intrinsic processing time* (T_{service}) of the container mix chosen by *Conlloovia*.

However, this case with $Q = 0$ also reveals a critical limitation of static capacity planning. Despite sufficient aggregate capacity to handle the load volume, the completion rate drops to 94.8%. This discrepancy occurs because the optimizer assumes a uniform arrival rate, whereas the simulator models the actual stochastic nature of traffic (Poisson arrivals). In a zero-buffer system ($Q = 0$), this stochasticity produces momentary micro-bursts, where inter-arrival times momentarily drop below service times, causing immediate request drops due to the lack of a backlog to absorb the variance.

When adequate buffering is introduced ($Q = 1000$), the completion rate rises to 100.0%. This validates that the allocated capacity was indeed sufficient, provided that the infrastructure includes queues to smooth out arrival jitter. The cost of this stability is a slight increase in tail latency (from 0.98 s to 1.12 s) due to queuing time.

2) Network Overhead: Case B introduces the network topology constraints. Comparing the ideal baseline ($L = 0$) with the network-aware simulation ($L = 20$ ms), the p99 response time degrades from 1.45 s to 1.53 s.

The introduction of the network produces two distinct effects on the Cumulative Distribution Function (CDF):

1) **Latency Floor:** The curve shifts to the right. The minimum response time increases by exactly 80 ms, as shown Figure 6, validating the correct modeling of the 4-hop topology (4×20 ms RTT).
2) **Tail Expansion:** In some configurations (e.g., the top right four configurations for $Q = 1000$ in Table VI) the gap widens significantly at the 99th percentile. This confirms that network latency is not merely additive; it increases the "time-in-system" for each request, indirectly increasing queue occupancy and further inflating wait times.

3) The heterogeneity problem: Policy Efficiency: The most dramatic finding appears in Case C, where the p99 response time explodes to near 990 seconds under the Simple Round-Robin (RR) policy. This failure is a direct consequence of the cost-optimization performed by *Conlloovia*. As noted in Section V, the optimizer minimized costs by mixing high-performance containers with a few low-performance instances (0.5 rps) to meet the marginal demand. While mathematically optimal, this heterogeneity creates a trap for naive load balancers. The RR policy blindly directs traffic to these slow containers, creating massive Head-of-Line blocking. Figure 7 illustrates the magnitude of this collapse using a logarithmic scale. The *Time-To-Drain* (TTD) metric reveals that after the workload ends, the infrastructure must remain active for

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE V. **Selected Simulation Cases.** Comparison of theoretical baseline vs. realistic constraints. *(App1 results shown)*

		Configuration			App1 Outcomes	
Case Type	**Workload**	**Buffer (Q)**	**Lat (L)**	**LB**	**% Completed**	**p99 RT (s)**
A. Ideal Baseline	Poisson	0	0 ms	SWRR	94.8%	0.98
(Validation)	Poisson	1000	0 ms	SWRR	**100.0%**	1.12
B. Network Impact	Trace	1000	0 ms	SWRR	100.0%	1.45
	Trace	1000	20 ms	SWRR	100.0%	1.53
C. Policy Saturation	Trace	1000	20 ms	RR	99.8%	**989.77**

TABLE VI. Complete Experimental Results for all 32 Simulation Scenarios (Split by Queue Capacity).

Scenarios with Zero Buffer / No Queuing ($Q = 0$)

				Completed		RT (p99)		Cost
Term	LB	Load	Lat	app0	app1	app0	app1	
drain	RR	poisson	0	82.6%	95.0%	0.212	0.980	$10.63
			20	82.4%	94.9%	0.293	1.061	$10.63
		trace	0	100.0%	98.5%	0.212	0.980	$10.63
			20	99.7%	98.5%	0.293	1.061	$10.63
	SWRR	poisson	0	82.6%	94.8%	0.212	0.980	$10.63
			20	82.4%	94.7%	0.293	1.061	$10.63
		trace	0	100.0%	94.4%	0.212	0.980	$10.63
			20	99.7%	94.3%	0.293	1.061	$10.63
hard	RR	poisson	0	82.6%	94.9%	0.212	0.980	$10.62
			20	82.4%	94.9%	0.293	1.061	$10.62
		trace	0	100.0%	98.5%	0.212	0.980	$10.62
			20	99.7%	98.5%	0.293	1.061	$10.62
	SWRR	poisson	0	82.6%	94.8%	0.212	0.980	$10.62
			20	82.4%	94.7%	0.293	1.061	$10.62
		trace	0	100.0%	94.3%	0.212	0.980	$10.62
			20	99.7%	94.3%	0.293	1.061	$10.62

Scenarios with Large Queues ($Q = 1000$)

				Completed		RT (p99)		Cost
Term	LB	Load	Lat	app0	app1	app0	app1	
drain	RR	poisson	0	100.0%	99.8%	0.355	842.870	$10.82
			20	100.0%	99.8%	0.436	852.609	$10.82
		trace	0	100.0%	99.8%	0.212	977.790	$10.82
			20	100.0%	99.8%	0.293	989.771	$10.82
	SWRR	poisson	0	100.0%	100.0%	0.355	1.118	$10.63
			20	100.0%	100.0%	0.436	1.200	$10.63
		trace	0	100.0%	100.0%	0.212	1.448	$10.63
			20	100.0%	100.0%	0.293	1.529	$10.63
hard	RR	poisson	0	100.0%	99.2%	0.355	5.702	$10.62
			20	100.0%	99.2%	0.436	5.783	$10.62
		trace	0	100.0%	99.2%	0.212	1.062	$10.62
			20	100.0%	99.2%	0.293	1.143	$10.62
	SWRR	poisson	0	100.0%	100.0%	0.355	1.118	$10.62
			20	100.0%	100.0%	0.436	1.199	$10.62
		trace	0	100.0%	100.0%	0.212	1.448	$10.62
			20	100.0%	100.0%	0.293	1.529	$10.62

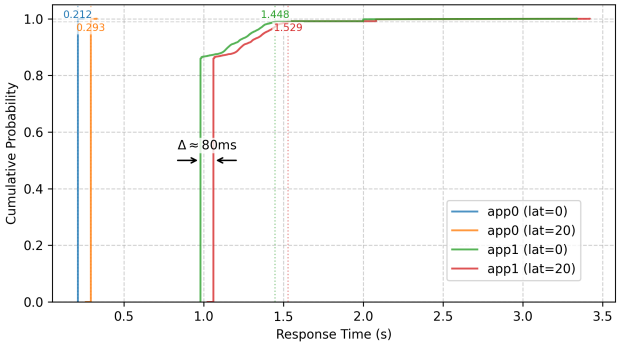


Figure 6. Impact of Network Latency on Response Time CDF ($Q = 0$). The shift highlights the 80 ms RTT floor introduced by the 4-hop topology.

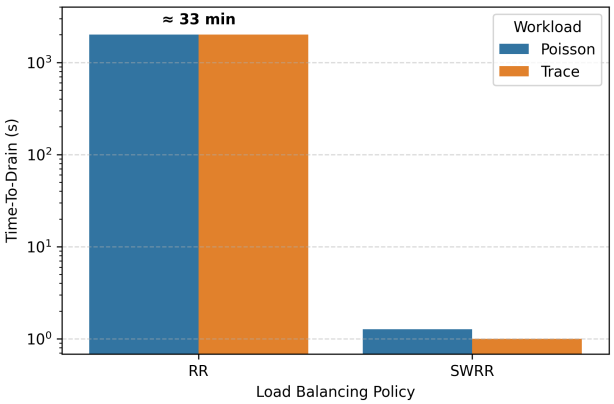


Figure 7. Impact of Load Balancing Strategy on Time-To-Drain (TTD). Naive RR saturates slow containers, forcing extended VM uptime.

$\approx$ 2000 seconds (33 minutes) to process the backlog created by RR, compared to just $\approx$ 1 second for SWRR. Crucially, despite this, the operational cost remains nearly identical to the optimal case ($< 2\%$ difference). This is explained because the extra time required by draining is consumed by a single VM instance, the one containing the slow containers, while the other 37 VMs were already shut down after 1 s of draining.

This highlights a dangerous blind spot: a system can be cost-

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

efficient according to billing metrics while being functionally broken for the end-user.

Finally, an analysis of the saturated RR runs (see Table VI) reveals extreme sensitivity to initial conditions. For instance, in the configurations with $Q = 1000$, drain policy and RR load balancer, with trace based workload, the mere addition of 20 ms of network latency causes the p99 response time to jump from ≈ 978 s to ≈ 990 s, as can be seen in the first two rows on the right-hand side of Table VI.

This 12-second discrepancy, triggered by a 0.08 s (80 ms RTT) structural perturbation, represents a non-linear amplification factor of nearly $300\times$. This confirms that the instability is not random noise, but a structural property of saturated queues: in the heavy tail of the distribution, the CDF curve becomes nearly horizontal (slope ≈ 0). In this region, microscopic shifts in probability, whether from network jitter or workload shape, translate into massive discrepancies on the time axis. This validates *Nuberu*'s ability to reproduce the non-linear sensitivity of real-world congestion.

4) Temporal Dynamics and Workload Fidelity: Finally, we analyze the impact of workload modeling on system stability over time. Static optimization relies on aggregated metrics (e.g., global p95 throughput), implicitly assuming that load fluctuations are smoothed out. Figure 8 challenges this assumption by contrasting a synthetic Poisson process against a real-world trace.

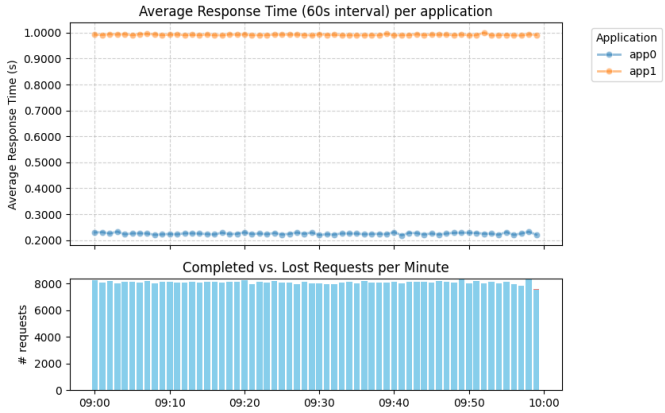
In the configurations using synthetic workload (Figure 8a), the response time remains artificially flat. Since the Poisson process is stationary with a fixed λ, it fails to capture the temporal correlations and burstiness of real traffic. In contrast, the trace-driven simulation (Figure 8b) reveals the true dynamic behavior. Here, we observe a transient congestion event during the first 10 minutes for *app1* (orange line), where the response time spikes significantly above the baseline. This corresponds to a period where the specific input rate momentarily exceeded the p95 threshold used by the solver.

This figure allows us to visualize when and how long the system violates its SLOs under these conditions. While *app0* (blue line) remains stable throughout due to adequate headroom, *app1*'s degradation highlights the temporal risks that static provisioning metrics cannot predict. This confirms the simulator's value not just for computing averages, but for stress-testing the deployment against the specific shape of the workload over time.

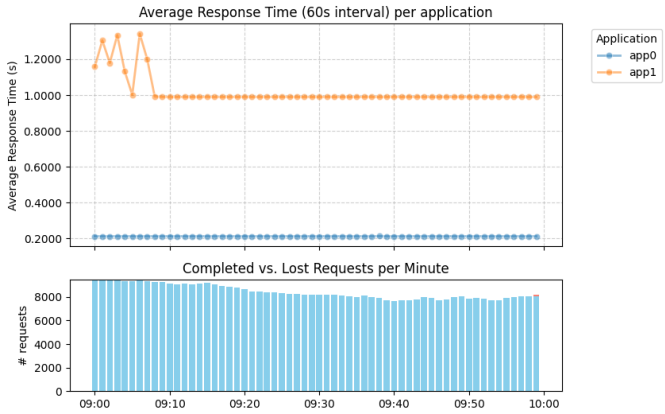
D. Performance and Scalability

The simulation framework's efficiency was evaluated using the 32 configurations presented above. On a server equipped with 32 cores, a single 1-hour simulation configuration completes in approximately 1 minute and 30 seconds. Thanks to the discrete-event independence and the modular design, the 32 configurations were executed in parallel, completing the entire experimental suite in 3 minutes and 21 seconds.

The current bottleneck identified in the prototype is the monitoring layer. The default observer plugin dumps every request lifecycle event into a JSONL file for offline debugging



(a) Synthetic workload with Poisson arrivals



(b) Trace based workload

Figure 8. Response time and number of requests completed for the configurations with SWRR balancing, large queues and 'hard' termination (last two rows of Table VI)

and metric generation (as used in Figure 8). These files can reach sizes of up to 1.3 GB per configuration, introducing significant I/O overhead. However, *Nuberu*'s architecture allows this monitor to be easily replaced by more efficient plugins (e.g., streaming to an external time-series database or performing on-the-fly aggregation), which would likely further accelerate the simulation. While the current implementation also exhibits high memory consumption due to non-optimized internal loops (e.g., $O(N \times M)$ component iterations), these are identified as targets for future engineering optimizations and do not compromise the logical validity of the results.

VI. Conclusion and Future Work

This paper has introduced a modular, extensible framework for cloud simulation designed to support reproducible and composable experimentation. Based on decoupled components, dynamic plugin discovery, and declarative configuration, the design enables researchers to prototype and compare alternative models for workload generation, resource allocation, and cost evaluation without modifying the simulation core.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Experimental results confirmed that ignoring network structural latency underestimates tail response times, validating the importance of the proposed bidirectional channel modeling. Furthermore, the evaluation highlighted that naive load balancing in cost-optimized heterogeneous clusters leads to extreme performance degradation during termination. Notably, this degradation in system behavior has minimal impact on total billing costs, suggesting that cost-driven optimization approaches may overlook service quality issues arising from queue imbalances.

From a software engineering perspective, the framework combines runtime plugin discovery, static interface contracts via Python `Protocols`, and version-controlled configuration. Together, these elements align the framework with the FAIR principles, enabling both local reproducibility and broader community validation of alternative orchestration strategies.

Building on the preliminary performance analysis presented, future work will focus on engineering optimizations to address the identified I/O and memory bottlenecks, specifically targeting the monitoring layer efficiency. Simultaneously, we plan to expand the plugin ecosystem to support emerging paradigms such as Large Language Model (LLM) serving, spot-instance strategies, and hybrid cloud-edge deployments. These extensions will further validate the framework's ability to model complex, heterogeneous environments.

Appendix A
Example Simulator Configuration

Listing 6 shows the configuration file for an example scenario, with Poisson arrivals, Round-Robin load balancing, and custom network delays. This configuration use specific latencies: 20 ms for both client-to-balancer and balance-to-VM channels. The runtime model is configured with a bounded queue capacity of 1000 requests to model resource contention, as discussed in the experimental evaluation. Furthermore, although this baseline sets global defaults, Nuberu allows for fine-grained network modeling: the user can specify per-VM or per-application overrides (via vm_network_overrides) to simulate heterogeneous topologies.

Acknowledgment

This research was funded by the project PID2021-124383OB-I00 of the Spanish National Plan for Research, Development and Innovation from the Spanish Ministerio de Ciencia e Innovación.

References

[1] R. Luque, J. L. Díaz, J. Entrialgo, and R. Usamentiaga, "Modular and reproducible simulator architecture for composable cloud systems," in *Proceedings of SIMUL 2025: The Seventeenth International Conference on Advances in System Modeling and Simulation*, Lisbon, Portugal: IARIA / ThinkMind Digital Library, 2025, pp. 82–87, ISBN: 978-1-68558-300-2. Available from: https://www.thinkmind.org/library/SIMUL/SIMUL_2025/simul_2025_1_180_50098.html.

[2] R. Luque and J. L. Díaz, "Nuberu: A modular cloud simulation framework," version 0.4.0, 2026, Available from: https://github.com/asi-uniovi/nuberu-releases [retrieved: February, 2026].

[3] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. D. Rose, and R. Buyya, "Cloudsim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and Experience*, vol. 41, pp. 23–50, 2011. DOI: https://doi.org/10.1002/spe.995.

[4] H. Casanova, A. Legrand, and M. Quinson, "SimGrid: A generic framework for large-scale distributed experiments," *Tenth International Conference on Computer Modeling and Simulation (uksim 2008)*, pp. 126–131, 2008. DOI: 10.1109/UKSIM.2008.28.

[5] N. Mansouri, R. Ghafari, and B. M. H. Zade, "Cloud Computing simulators: A comprehensive review," *Simulation Modelling Practice and Theory*, vol. 104, p. 102 144, Nov. 2020, ISSN: 1569-190X. DOI: 10.1016/j.simpat.2020.102144.

[6] I. Bambrik, "A survey on Cloud Computing simulation and modeling," *SN Computer Science*, vol. 1, no. 5, p. 249, Aug. 2020, ISSN: 2661-8907. DOI: 10.1007/s42979-020-00273-1.

[7] S. Lata and D. Singh, "Cloud simulation tools: A survey," *AIP Conference Proceedings*, vol. 2555, no. 1, p. 030 003, Oct. 2022, ISSN: 0094-243X. DOI: 10.1063/5.0109181.

[8] F. Fakhfakh, H. H. Kacem, and A. H. Kacem, "Simulation tools for cloud computing: A survey and comparative study," *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pp. 221–226, 2017. DOI: https://doi.org/10.1109/ICIS.2017.7959997.

[9] S. V. Margariti, V. V. Dimakopoulos, and G. Tsoumanis, "Modeling and simulation tools for fog computing - a comprehensive survey from a cost perspective," *Future Internet*, vol. 12, p. 89, 2020. DOI: https://doi.org/10.3390/fi12050089.

[10] M. Shiraz, A. Gani, R. H. Khokhar, and E. Ahmed, "An extendable simulation framework for modeling application processing potentials of smart mobile devices for mobile cloud computing," in *2012 10th International Conference on Frontiers of Information Technology*, Dec. 2012, pp. 331–336. DOI: 10.1109/FIT.2012.66.

[11] A. Núñez et al., "iCanCloud: A flexible and scalable cloud infrastructure simulator," *Journal of Grid Computing*, vol. 10, no. 1, pp. 185–209, Mar. 2012, ISSN: 1572-9184. DOI: 10.1007/s10723-012-9208-5.

[12] I. Lera, C. Guerrero, and C. Juiz, "YAFS: A simulator for IoT scenarios in fog computing," *IEEE Access*, vol. 7, pp. 91 745–91 758, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2927895.

[13] X. Zeng et al., "IOTSim: A simulator for analysing IoT applications," *Journal of Systems Architecture*, Design Automation for Embedded Ubiquitous Computing Systems, vol. 72, pp. 93–107, Jan. 2017, ISSN: 1383-7621. DOI: 10.1016/j.sysarc.2016.06.008.

[14] A. Siavashi and M. Momtazpour, "Cloudy: A Pythonic cloud simulator," *2024 32nd International Conference on Electrical Engineering (ICEE)*, pp. 1–5, 2024. DOI: https://doi.org/10.1109/ICEE63041.2024.10667881.

[15] J. Massa et al., "ECLYPSE: A Python Framework for Simulation and Emulation of the Cloud-Edge Continuum," *Journal of Software: Evolution and Process*, vol. 38, no. 1, e70081, Jan. 2026. DOI: 10.1002/smr.70081.

[16] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160 018, 2016. DOI: 10.1038/sdata.2016.18.

[17] SimPy Development Team, "Simpy: Discrete event simulation for python," Version 4.1.2, 2025, Available from: https://simpy.readthedocs.io/ [retrieved: July, 2025].

[18] pytest-dev, "Pluggy: A minimalist production ready plugin system," Version 1.5.0, 2025, Available from: https://pluggy.readthedocs.io/ [retrieved: July, 2025].

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Listing 6. Example YAML simulation configuration

```yaml
config_version: "3.2"
simulation:
  seed: 12345
  stop_time: 3600
  eager_start: true
  drain_pending_requests: true

network_delays:
  wi_to_lb: 20
  lb_to_vm_default: 20

logging:
  log_to_file: true
  log_to_console: false
  level: critical
  component_levels:
    plugins.report_generator: info
  output_path: logs/
  log_file_name: "example.log"

custom_plugins:
  - plugin_name: monitor_simple
    plugin_config:
      dump_file: "example.jsonl"
  - plugin_name: report_generator
    plugin_config:
      output_format: jsonl
      output_path: "example.report"

infrastructure:
  plugin_name: conlloovia
  plugin_config:
    pickle_path: "conlloovia_data.p"

runtime_model:
  plugin_name: simple
  plugin_config:
    queue_size: 1000

performance_data:
  plugin_name: conlloovia
  plugin_config:
    pickle_path: "conlloovia_data.p"

allocation:
  plugin_name: "conlloovia"
  plugin_config:
    pickle_path: "conlloovia_data.p"

workloads:
  - app: app0
    distribution:
      plugin_name: poisson
      plugin_config:
        num_reqs: 124427
        time_slot_size: 3600
  - app: app1
    distribution:
      plugin_name: poisson
      plugin_config:
        num_reqs: 362587
        time_slot_size: 3600

load_balancer:
  plugin_name: "roundrobin"
  plugin_config: {}
```


[19] I. Levkivskyi, J. Lehtosalo, and L. Langa, "Pep 544: Protocols: Structural subtyping (static duck typing)," Python Enhancement Proposal, Python-Version: 3.8, Final, 2017, Available from: https://peps.python.org/pep-0544/.

[20] J. Entrialgo, M. García, J. García, J. M. López, and J. L. Díaz, "Joint autoscaling of containers and virtual machines for cost optimization in container clusters," *Journal of Grid Computing*, vol. 22, pp. 1–24, 2024. DOI: https://doi.org/10.1007/s10723-023-09732-4.

[21] M. Dounin, "Upstream: Smooth weighted round-robin balancing," Commit, nginx source tree, May 2012, Available from: https://github.com/phusion/nginx/commit/27e94984486058d73157038f7950a0a36ecc6e35 [retrieved: July, 2025].

[22] R. Luque, J. L. Díaz, J. Entrialgo, and R. Usamentiaga, "Nuberu simulation results – experiments repository," 2026, Available from: https://github.com/asi-uniovi/nuberu-experiments-results/tree/simul2026 [retrieved: February, 2026].

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Memory-Driven Dynamics and Repeated Patterns in Online Toxicity: A Fractal-Fractional SEIQR Modeling Approach

Nitin Agarwal*†
*COSMOS Research Center, University of Arkansas at Little Rock, Arkansas, USA
†International Computer Science Institute, University of California, Berkeley, California, USA
e-mail: nxagarwal@ualr.edu

***Abstract*—This study employs an epidemiological framework called SEIQR, which stands for Susceptible, Exposed, Infected, Quarantine, and Recovered, to analyze the spread of toxicity in online environments, integrating toxicity intensity stratification and fractal-fractional dynamics to capture the complexity of toxicity propagation. Using datasets from Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), commonly known as coronavirus disease (COVID-19), and social movement discussions from the Latin American countries of Brazil and Peru, we conduct sensitivity analysis to evaluate key parameters influencing toxicity diffusion. The results reveal that splitting toxicity into moderate and high levels significantly reduces model error rates, enhancing predictive accuracy across all datasets. Additionally, our findings indicate that the basic reproduction number (R_0) is highly sensitive to exposure and quarantine rates, emphasizing the critical role of enhanced moderation and adaptive quarantining in suppressing toxicity. By incorporating fractal-fractional operators, the model further accounts for memory-driven toxicity spread, long-range dependencies, and network complexity. The study demonstrates that reducing the fractal dimension (τ) and fractional order (α) slows toxicity propagation, highlighting the influence of user behavior history and structural dynamics in toxicity persistence. Moreover, quarantine interventions and content demotion strategies are shown to significantly curb toxicity intensity while maintaining engagement dynamics. These insights provide a foundation for policy-driven interventions, enabling social media platforms to implement optimized content moderation, algorithmic intervention, and network-level strategies to mitigate online toxicity and promote healthier digital ecosystems.**

***Keywords*—mathematical model; toxicity; social media; epidemiological modeling; hate speech; COVID-19; Brazil; Peru.**

I. Introduction

The rapid evolution of social media platforms has transformed the way users engage in communication, share information, and participate in public discourse. These platforms serve as catalysts for the swift dissemination of news, educational resources, and discussions on critical societal matters. By enabling real-time interaction across diverse audiences, social media fosters global connectivity and awareness. However, alongside these beneficial aspects, social media has also unintentionally become a breeding ground for the amplification of toxic behaviors [1]. This toxicity manifests in various forms, including hate speech, misinformation, harassment, cyberbullying, and online extremism. The unrestricted and algorithm-driven nature of social platforms often facilitates the viral spread of harmful content, influencing not only individual users but also wider societal dynamics [2], [3], [4].

The proliferation of toxicity on digital platforms has far-reaching consequences that extend beyond online environments, affecting psychological well-being, social cohesion, and even real-world actions. Misinformation-driven narratives have contributed to public unrest, political polarization, and radicalization, demonstrating the tangible effects of unchecked digital toxicity [5]. In extreme cases, the rapid circulation of false information has led to public health crises, economic disruptions, and coordinated disinformation campaigns aimed at manipulating public opinion [6]. As a result, addressing the spread and impact of online toxicity has become an urgent research priority that demands multidisciplinary approaches and advanced analytical frameworks [3], [4], [7].

Given the increasing reliance on digital communication and algorithmic content curation, it is imperative to understand the underlying mechanisms that drive toxicity propagation across interconnected online communities [8]. The primary challenge lies in developing intervention strategies that effectively mitigate toxic interactions while upholding fundamental rights to freedom of speech and avoiding excessive restrictions on public discourse [1]. Striking a balance between content regulation and open dialogue is crucial to maintaining healthy digital ecosystems that encourage constructive engagement.

The complexity and scale of modern social networks necessitate the adoption of computational modeling and data-driven methodologies to analyze the spread, persistence, and impact of online toxicity. By leveraging mathematical modeling, network analysis, and artificial intelligence, researchers can identify key transmission patterns, predict emerging toxicity trends, and develop targeted interventions that curb digital toxicity while preserving online freedoms [9]. These insights will help policymakers, social media platforms, and researchers formulate evidence-based strategies to foster safer and more inclusive online environments.

Mathematical models have been widely employed in epidemiology to study the spread of infectious diseases. The same principles can be applied to model the dissemination of online toxicity, given their shared characteristics: contact-based transmission and behavioral contagion. Various epidemic models, including the Susceptible-Infected-

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Recovered (SIR), Susceptible-Infected-Susceptible (SIS), and Susceptible-Exposed-Infected-Recovered (SEIR) models, have been adapted to analyze social media toxicity.

In this study, we extend the Susceptible, Exposed, Infected, Quarantined, Recovered (SEIQR) model by incorporating toxicity intensity levels, classifying individuals into moderately and highly toxic categories. This refinement allows for more precise intervention strategies, as different toxicity levels necessitate distinct mitigation approaches. We validate our model using real-world datasets from online discussions related to Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), commonly known as coronavirus disease (COVID-19), and social justice movements in the Latin American countries of Brazil and Peru, enabling the identification of critical patterns in toxicity spread.

Traditional epidemiological models typically assume homogeneous transmission dynamics. However, online toxicity propagation exhibits several unique features that necessitate a more advanced modeling framework. Specifically:

i. **Fractal network topology:** Social media networks exhibit self-similarity and scale-free structures, where a small number of *superpropagators* significantly influence the spread of toxicity. Fractal-fractional models better represent these non-uniform dynamics.
ii. **Memory effects in user interactions:** Unlike disease transmission, online toxicity is affected by past exposures and historical user behaviors. Fractional calculus captures these long-range dependencies and persistent memory effects.
iii. **Superpropagation and feedback loops:** Toxicity dissemination is often amplified through echo chambers and algorithmic recommendations. The fractal-fractional approach models these reinforcement mechanisms more effectively than integer-order differential equations.

By integrating fractal-fractional operators into the SEIQR model, we capture the irregularity, persistence, and non-locality of online toxicity spread. This approach allows for improved predictive capabilities and better policy recommendations for online moderation.

This study addresses the following key research questions:

ii **RQ1:** What are the key parameters in the SEIQR model that have the most significant impact on the *basic reproduction number* R_0, and how can these parameters be controlled to mitigate toxicity spread?
i **RQ2:** How do variations in toxicity intensity (moderate vs. high) influence its dissemination, and how does intensity impact the SEIQR model's performance?
iii **RQ3:** How does the integration of fractal-fractional operator influence the accuracy and predictive capabilities of toxicity spread models compared to traditional integer-order models?
iv **RQ4:** What are the effects of superpropagation and long-term memory on toxicity dissemination, and how can quarantine strategies be optimized in this context?

The remainder of the paper is organized as follows: Section II provides a review of related work on toxicity propagation in social media and epidemiological modeling techniques. Section III introduces the methodology, detailing the SEIQR model formulation and the fractal-fractional extensions. Section IV presents the experimental results and discusses the impact of different toxicity intensities. Section V presents the results and discussion. Finally, Section VI concludes the study with recommendations for future research and policy interventions.

II. Related Work

The rise of online toxicity in social media environments has prompted extensive research into its dynamics, effects, and mitigation strategies. With the increasing spread of harmful or toxic content, researchers have employed various computational, epidemiological, and network-based models to understand and control the propagation of toxicity.

A. Machine Learning Approaches to Toxicity Detection

Several studies have leveraged machine learning techniques to detect, classify, and analyze toxicity in online discussions. For example, researchers in [10], [11] analyzed COVID-19 misinformation and toxicity across social media platforms, revealing notable differences in the extent and nature of toxic behaviors on Twitter, Reddit, and Facebook. Their findings emphasized the role of key user groups (super-spreaders) in amplifying misinformation and toxic content. Similarly, natural language processing models such as BERT, GPT, and Detoxify have been used to detect toxicity in real time [12], [13]. These models have been instrumental in the identification of various forms of toxic language, including hate speech, harassment, cyberbullying, and misinformation. Furthermore, studies in [14] investigated the dynamics of discussion threads on Reddit, where toxicity spreads hierarchically within nested conversations. However, these machine learning-based approaches largely focus on static classification of toxicity, often overlooking long-term propagation effects and feedback loops that sustain toxic environments.

B. Epidemiological Models for Online Toxicity Spread

Epidemiological modeling has proven to be an effective framework for understanding and predicting the spread of online toxicity by drawing analogies between toxic content dissemination and infectious disease transmission. Traditional models such as SIR, SIS, and SEIR have been adapted to analyze how toxicity spreads within digital communities [4]. The Susceptible-Toxic-Recovered-Susceptible (STRS) model has further refined this analogy by incorporating recovery mechanisms where users disengage from toxic interactions [15], [16], [17].

Moreover, quarantine-based control strategies have been widely adopted in epidemiological models to curb infectious spread, an approach that has direct applications in content moderation and toxicity mitigation. Authors in [18] studied the impact of content quarantining on social media toxicity,

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

demonstrating that isolating toxic users reduces the overall infection rate in a network-based toxicity model. Similarly, authors in [15] applied epidemiological modeling to study COVID-19 misinformation spread, highlighting the role of user quarantine, moderation interventions, and algorithmic content suppression.

C. Fractal-Fractional Modeling for Toxicity Dynamics

A growing body of research has focused on fractal-fractional approaches to model complex social dynamics. Unlike classical integer-order epidemiological models, fractal-fractional models capture memory effects, non-locality, and irregular diffusion patterns [19], [20]. These properties are particularly relevant to online toxicity modeling, where the following are true:

1) Fractal network topology leads to self-similarity and scale-free structures, where toxicity is disproportionately amplified by a small subset of highly influential users (superpropagators).
2) Memory effects in user interactions suggest that toxicity is influenced by past exposures and behavioral reinforcement, rather than instantaneous interactions.
3) Superpropagation and feedback loops occur due to algorithmic content amplification and echo chamber effects, making toxicity diffusion highly persistent.

Fractal-fractional epidemiological models have focused primarily on disease transmission and financial systems, with limited application to online toxicity dynamics. For disease models, see [21], [22] and references therein. Despite extensive research in toxicity detection, epidemiological modeling, and fractal-fractional approaches, there are several gaps. To address these gaps, this study introduces a fractal-fractional extension of the SEIQR epidemiological model, uniquely designed to capture memory effects, long-range dependencies, and heterogeneous toxicity diffusion patterns inherent to online platforms. By unifying fractal-fractional operators, which encode non-local interactions and irregular propagation dynamics—with a stratified SEIQR framework that distinguishes between moderate- and high-intensity toxicity, this work advances a novel computational framework for analyzing and mitigating persistent toxicity. To our knowledge, this represents the first integration of fractal-fractional calculus with toxicity-specific epidemiological modeling, enabling a granular understanding of how algorithmic amplification and user memory shape long-term propagation.

Building on this foundation, the study also addresses critical limitations in prior SEIQR analyses by conducting a rigorous sensitivity analysis. We identify the parameters most influential to the basic reproduction number $\mathcal{R}_0$, thereby pinpointing actionable levers—such as platform moderation efficacy or user recovery rates—for disrupting toxicity spread. This dual focus on mechanistic modeling (via fractal-fractional dynamics) and practical control strategies (via sensitivity-driven interventions) bridges a key gap between theoretical epidemiology and real-world social media governance.

III. Methodology

This section outlines the data collection framework, analytical methods, and computational procedures employed to validate the proposed fractal-fractional SEIQR model.

A. Data Collection and Analysis

To validate the model, we analyzed two distinct datasets: (1) discourse related to the COVID-19 pandemic (spanning February 2020–June 2021) and (2) social movements in Brazil (spanning November 1, 2022–February 25, 2023) and Peru (spanning December 7, 2022–January 31, 2023). Public posts were collected via X's (formerly Twitter) Academic API, focusing on keywords and hashtags associated with polarized discourse.

- **COVID-19 Dataset**: Included topics such as pandemic policies, face mask mandates, lockdowns, and 5G conspiracy theories. Key hashtags: #fckthecovid/s, #fckyourmask/s, #f*cklockdown/s, #5GCoronavirus.
- **Brazilian Protests**: Centered on post-election unrest following the October 30, 2022, presidential election, with hashtags like #semanistia and #SOSbrasil reflecting demands for military intervention and counter-protests.
- **Peruvian Protests**: Captured anti-government demonstrations after President Pedro Castillo's removal on December 7, 2022, using hashtags such as #peruprotest/s.

These datasets have less sparsity in terms of toxic discussions, which is important for this study, where the objective is to examine spread of toxicity in high vs moderate toxic environments. Contexts where toxicity is negligible or highly sparse are outside the scope of this study.

1) Toxicity Classification: Toxicity scores for each post were computed using Detoxify [23], a pre-trained deep learning model that evaluates text for harmful content. Detoxify employs convolutional neural networks and semantic embeddings to assign a toxicity probability between 0 (non-toxic) and 1 (highly toxic). Posts with scores of 0.5 or more were classified as toxic; those below 0.5 were deemed non-toxic.

To analyze toxicity intensity, the toxic subset was further stratified:

- **Moderate Toxicity**: Posts with scores below the dataset-specific average toxicity score.
- **High Toxicity**: Posts exceeding the average score.

Table I summarizes the distribution of high/moderate toxic posts and average scores across datasets.

2) User Activity and Quarantining: To identify superpropagators, we isolated the top 10% of users by activity level, defined as the number of retweets per post [18]. These high-engagement users were algorithmically transferred to a quarantine compartment in the SEIQR model, simulating platform-level interventions to curb toxicity spread.

B. Model Formulation

Online toxicity is a growing epidemic on digital platforms, characterized by high transmission rates, latent behavior, and

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I: Toxicity Dataset Statistics

Dataset	Num. of Posts	Avg. Toxicity	Num. of High Toxic	Num. of Moderate Toxic
F*Covid	28,131	0.91	4,684	2,082
F*Mask	2,423	0.91	538	217
F*Lockdown	1,995	0.82	598	493
5G	33,403	0.84	1,096	703
Brazil Anti	405,160	0.70	1,221	2,309
Brazil Pro	44,415	0.75	105	131
Peru	195,290	0.71	511	546

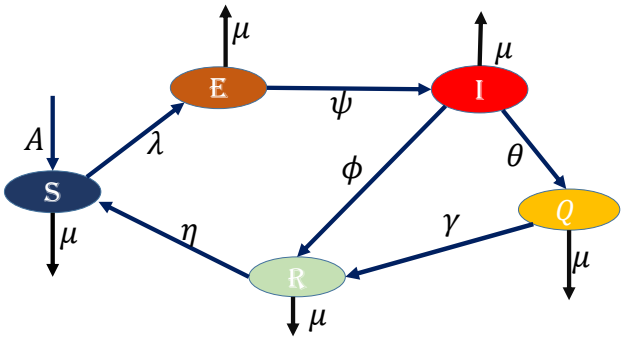


Fig. 1: Transfer diagram for the toxicity spread on the social network platform.

infectiousness. To analyze this, [18] used the SEIQR model as the following system of ordinary differential equations.

$$\begin{cases} \frac{dS(t)}{dt} = \mathcal{A} + \eta R - \frac{\beta IS}{N(t)} - \mu S, \\ \frac{dE(t)}{dt} = \frac{\beta IS}{N(t)} - (\mu + \psi)E, \\ \frac{dI(t)}{dt} = \psi E - (\mu + \phi + \theta)I, \\ \frac{dQ(t)}{dt} = \theta I - (\gamma + \mu)Q, \\ \frac{dR(t)}{dt} = \gamma Q + \phi I - (\mu + \eta)R, \end{cases} \tag{1}$$

where $\lambda = \frac{\beta IS}{N(t)}$, the entire population, and we define the quantity $N(t)$ by

$$N(t) = S(t) + E(t) + I(t) + Q(t) + R(t).$$

and initial conditions

$$S(0) = S_0 \geq 0, \quad E(0) = E_0 \geq 0, \quad I(0) = I_0 \geq 0,$$
$$Q(0) = Q_0 \geq 0, \quad R(0) = R_0 \geq 0.$$

C. Fractal-Fractional Model

Fractal-fractional models combine the advantages of fractal and fractional calculus to account for the irregular, dynamic, and memory-driven nature of online interactions. The fractal dimension captures the self-similar structure of user networks, content propagation pathways, and spatial-temporal patterns of toxic content dissemination. Meanwhile, the fractional calculus component incorporates long-range temporal dependencies and non-local effects, reflecting the persistence of toxicity and its amplification through feedback loops inherent in social media dynamics. While the classical SEIQR model in [18] effectively captured the dynamics of online toxicity spread, it failed to account for the complex, irregular, and memory-dependent nature of user interactions in social media environments. To address this limitation, we extend the model using fractal-fractional derivatives in the Caputo sense as follows:

$$\begin{cases} {}_0^{FFP}D_t^{\alpha,\tau} S(t) = \mathcal{A} + \eta R - \frac{\beta I}{N} S - \mu S, \\ {}_0^{FFP}D_t^{\alpha,\tau} E(t) = \frac{\beta I}{N} S - (\mu + \psi)E, \\ {}_0^{FFP}D_t^{\alpha,\tau} I(t) = \psi E - (\mu + \phi + \theta)I, \\ {}_0^{FFP}D_t^{\alpha,\tau} Q(t) = \theta I - (\gamma + \mu)Q, \\ {}_0^{FFP}D_t^{\alpha,\tau} R(t) = \gamma Q + \phi I - (\mu + \eta)R, \end{cases} \tag{2}$$

where ${}_0^{FFP}D_t^{\alpha,\tau}(.)$ is the fractal-fractional derivative with the fractional order $0 < \alpha \leq 1$, and fractal dimension $0 < \tau \leq 1$ in the Caputo sense with power law type kernel and the variables are assumed to be non-negative with appropriate initial conditions.

IV. Model Parameterization and Data Fitting Analysis

Ensuring accurate model validation and precise parameter estimation is fundamental in mathematical modeling when working with real-world data. The main challenge lies in determining the most suitable parameter values from empirical data, making parameter fitting a crucial step in model formulation. One widely used approach for parameter estimation in nonlinear models is the Non-Linear Least Squares Method (NLSM), which minimizes the discrepancy between observed and predicted values.

Consider a nonlinear mathematical model

$$y_i = f(x_i, \Theta) + \epsilon_i, \quad i = 1, 2, ..., n$$

where

- y_i represents the observed data points,
- $f(x_i, \Theta)$ is a nonlinear function that depends on the parameter set Θ,
- x_i are the independent variables, and
- ϵ_i is an error term assumed to be normally distributed.

The goal of NLSM is to minimize the sum of squared residuals (RSS), defined as

$$Z(\Theta) = \sum_{i=1}^{n} [y_i - f(x_i, \Theta)]^2$$

where

- $Z(\Theta)$ is the objective function to be minimized,
- y_i are the actual observed values,
- $f(x_i, \Theta)$ represents the model's predicted values, and

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE II: Explanation of the Model Parameters

Parameter	Value	Source	Explanation
$\mathcal{A}$	[100]	fitted	recruitment rate of human
β	0.0006	fitted	effective contact rate
ψ	0.047	fitted	the rate at which exposed become infected
θ	0.020	fitted	the rate at which I transfer to quarantine class
η	0.04	fitted	the rate at which recovery becomes susceptible
γ	0.002	fitted	the rate at which Q transfer to recovery
ϕ	0.1	fitted	the rate at which I transfer to recovery
μ	0.1	fitted	the rate at which people exit autonomously

- Θ is the set of unknown model parameters.

To evaluate model fitting, we use the relative error, given by

$$E_{\text{rel}} = \frac{\|I_{\text{est}}(t_i) - I_{\text{data}}(t_i)\|_2}{\|I_{\text{data}}(t_i)\|_2}. \tag{3}$$

This relative error, E_{rel}, quantifies the discrepancy between the estimated number of infected users $I_{\text{est}}(t_i)$ and the actual recorded infected users $I_{\text{data}}(t_i)$ at various time points t_i. The norm $\|\cdot\|_2$ represents the Euclidean distance, allowing for an objective measure of the overall deviation relative to actual data. A lower error value signifies a better model fit, affirming its accuracy in depicting the dynamics of toxicity spread.

The findings indicate that we can better capture the propagation of toxicity when the dataset is segmented (Table II). This pattern is consistent across different contexts, such as COVID-19 and social movements, suggesting that the findings are applicable in various scenarios. This approach provides policymakers with insights on prioritizing areas requiring focused interventions and allocating resources to effectively reduce the spread of toxicity.

A. Preliminaries

In this section, we will review some fractal-fractional calculus concepts that were covered earlier.

Definition 4.1 ([24], [25]) Suppose that $\mathcal{M}(t)$ is continuous and fractal differentiable on (a, b) with order τ. Then, the fractal–fractional derivative of $y(t)$ with order α in the Riemann–Liouville sense, having a power-law type kernel, is defined as follows:

$$^{FFP}D^{\alpha,\tau}_{a,t}(\mathcal{M}(t)) = \frac{1}{\Gamma(y-\alpha)}\frac{d}{dt^{\tau}}\int_a^t (t-s)^{y-\alpha-1}\mathcal{M}(s)\,ds,$$

where $\mathcal{M}(t)$ is a suitable integer ensuring the definition holds, $\mathcal{M}\ -1 < \alpha,\ \tau \leq y \in \mathbb{N}$, and

$$\frac{d\mathcal{M}(s)}{ds^{\tau}} = \lim_{t\to s}\frac{\mathcal{M}(t)-\mathcal{M}(s)}{t^{\tau}-s^{\tau}}.$$

Definition 4.2 ([25], [26]) Suppose that $\mathcal{M}(t)$ is continuous on an open interval (a, b), then the fractal-fractional integral of $\mathcal{M}(t)$ with order α, having a power-law type kernel, is defined as follows:

$$^{FFP}D^{\alpha}_{0,t}(\mathcal{M}(t)) = \frac{\tau}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}s^{\tau-1}\mathcal{M}(s)\,ds,$$

where $\Gamma(\alpha)$ is the Gamma function.

These operators are referred to as *fractal-fractional integral operators*.

Lemma 4.3 ([20], [26]) If $\mathcal{M}$ is continuous on (a, b), then the following fractal-fractional derivative

$$^{FFP}D^{\alpha,\tau}_{0,t}(\mathcal{M}(t)) = \mathcal{M}(t)$$

has a unique solution

$$\mathcal{M}(t) = \mathcal{M}(0) + \frac{\tau}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}s^{\tau-1}\mathcal{M}(s)\,ds.$$

We represent the Banach space by $\Omega = P\times P\times P\times P\times P$ and $P = C(\mathcal{T}, \mathbb{R}^5)$, $0 \leq t \leq T < \infty$, under the norm given by

$$\|W\| = \sup_{t\in\mathcal{T}}|\mathcal{M}(t)|, \quad \text{for } \mathcal{M}\in Z,$$

$$|W(t)| = |S(t) + E(t) + I(t) + Q(t) + R(t)|,$$

where $S, E, I, Q, R \in C(\mathcal{T}, \mathbb{R})$.

B. Model Analysis: Existence-Uniqueness

When formulating epidemiological models, a fundamental question arises: Does the model have a well-defined and unique solution? Before performing any analysis, it is essential to ensure that the system of equations governing the model is mathematically valid. This is typically achieved through fixed-point theorems, such as Banach's fixed point theorem, which guarantees the existence and uniqueness of solutions under specific conditions.

In this study, we employ Banach's fixed point theorem to demonstrate that the fractal-fractional model in equation (2) has a unique solution. We define the function space as: $\Omega = P\times P\times P\times P\times P$ and $P = C(\mathcal{T}, \mathbb{R}^5)$, $0 \leq t \leq T < \infty$, under the norm given by

$$\|W\| = \sup_{t\in\mathcal{T}}|\mathcal{M}(t)|, \quad \text{for } \mathcal{M}\in Z,$$

$$|W(t)| = |S(t) + E(t) + I(t) + Q(t) + R(t)|,$$

where $S, E, I, V, R \in C(\mathcal{T}, \mathbb{R})$.

From equation (2), the right-hand side of the fractal-fractional $SEIQR$ model can be rewritten for simplicity as

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE III: Error Rates for SEIQR Model with and without Dataset Splitting

Dataset	Moderate toxicity	High toxicity	Without split
F*Covid	0.0011	0.0021	0.081
F*Mask	0.021	0.049	0.073
Lockdown	0.032	0.045	0.061
5G	0.0029	0.0011	0.003
Brazil Anti	0.062	0.055	0.094
Brazil Pro	0.060	0.061	0.088
Peru	0.095	0.124	0.41

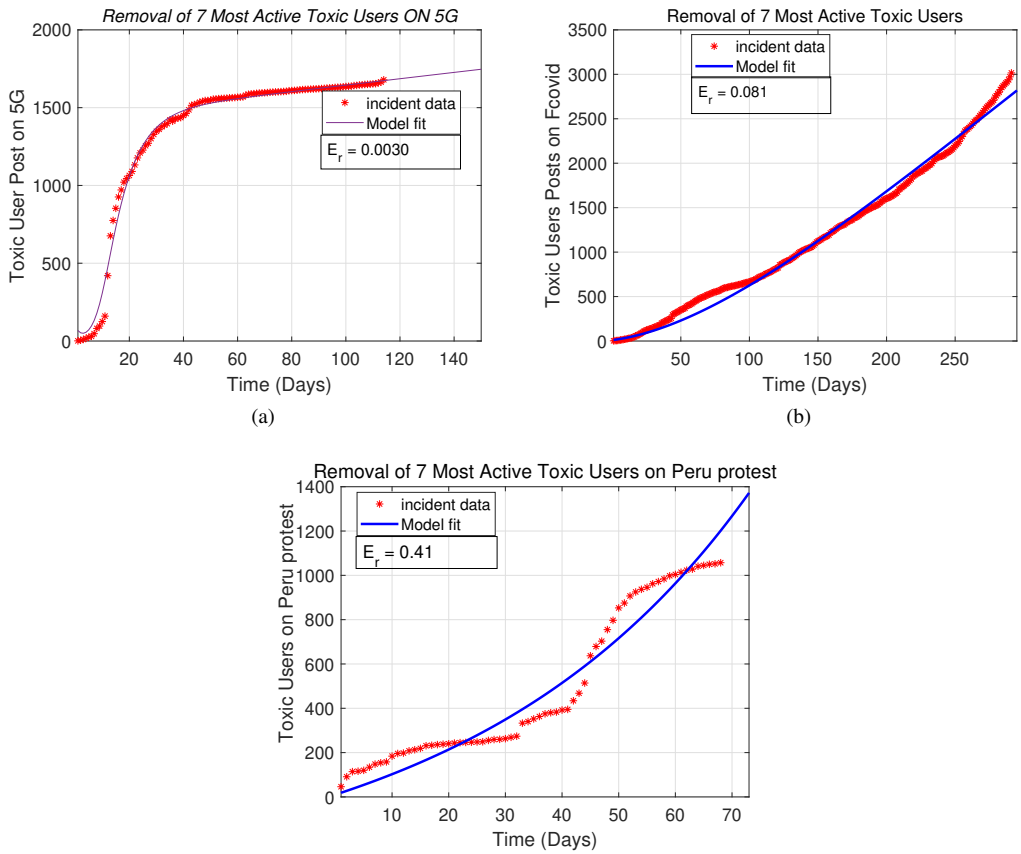


Fig. 2: SEIQR model fitted to infected without splitting toxicity in the datasets.

$$\begin{cases} \mathcal{M}_1(t,S(t)) = \mathcal{A} + \eta R - \frac{\beta I}{N} S - \mu S, \\ \mathcal{M}_2(t,E(t)) = \frac{\beta I}{N} S - (\mu+\psi)E, \\ \mathcal{M}_3(t,I(t)) = \psi E - (\mu+\phi+\theta)I, \\ \mathcal{M}_4(t,Q(t)) = \theta I - (\gamma+\mu)Q, \\ \mathcal{M}_5(t,R(t)) = \gamma Q + \phi I - (\mu+\eta)R. \end{cases} \tag{4}$$

The fractional integral possesses differentiability, and if the given equation (2) can be transformed into the Volterra form, then the fractal-fractional derivative, expressed in the Riemann–Liouville framework, can be represented as follows

$$\begin{cases} {}_0^{RL}D_t^{\alpha,\tau} S(t) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dt} \int_a^t (t-s)^\alpha S(s)\, ds \frac{1}{st^{s-1)}}, \\ {}_0^{RL}D_t^{\alpha,\tau} E(t) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dt} \int_a^t (t-s)^\alpha E(s)\, ds \frac{1}{st^{s-1)}}, \\ {}_0^{RL}D_t^{\alpha,\tau} I(t) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dt} \int_a^t (t-s)^\alpha I(s)\, ds \frac{1}{st^{s-1)}}, \\ {}_0^{RL}D_t^{\alpha,\tau} Q(t) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dt} \int_a^t (t-s)^\alpha Q(s)\, ds \frac{1}{st^{s-1)}}, \\ {}_0^{RL}D_t^{\alpha,\tau} R(t) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dt} \int_a^t (t-s)^\alpha R(s)\, ds \frac{1}{st^{s-1)}}, \end{cases} \tag{5}$$

Due to the fact that the integrals in equation (5) are differentiable, the system can be expressed as:

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

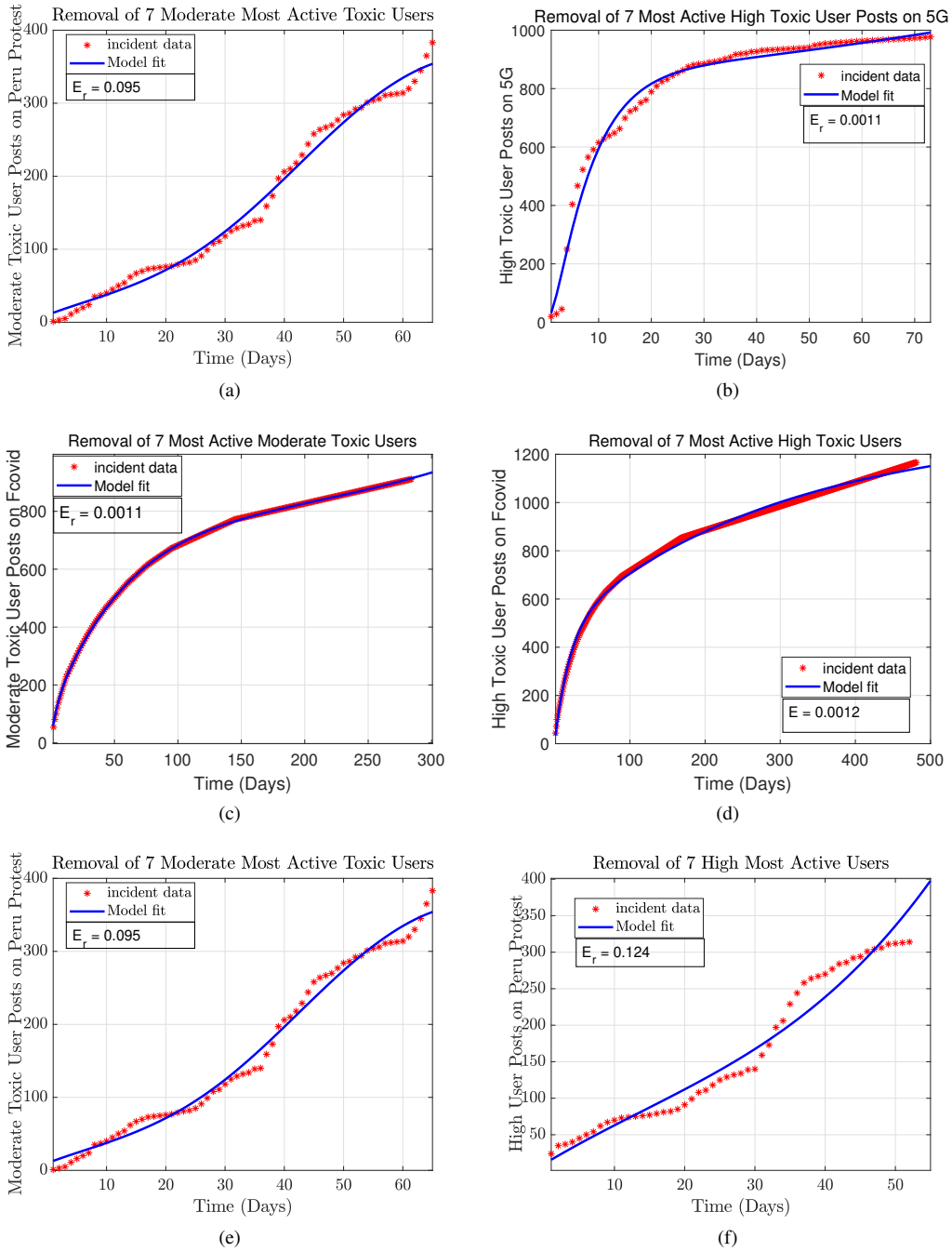


Fig. 3: SEIQR model fitted to moderate and high toxic posts.

$$\begin{cases} {}_0^{RL}D_t^{\alpha,\tau}S(t) = st^{s-1}[\mathcal{A} + \eta R - \frac{\beta I}{N}S - \mu S], \\ {}_0^{RL}D_t^{\alpha,\tau}E(t) = st^{s-1}[\frac{\beta I}{N}S - (\mu+\psi)E], \\ {}_0^{RL}D_t^{\alpha,\tau}I(t) = st^{s-1}[\psi E - (\mu+\phi+\theta)I], \\ {}_0^{RL}D_t^{\alpha,\tau}Q(t) = st^{s-1}[\theta I - (\gamma+\mu)Q], \\ {}_0^{RL}D_t^{\alpha,\tau}R(t) = st^{s-1}[\gamma Q + \phi I - (\mu+\eta)R], \end{cases} \tag{6}$$

To ensure that the initial conditions are defined in terms of integer-order derivatives, we replace the Riemann-Liouville derivative with the Caputo derivative. This modification facilitates the formulation of well-posed initial conditions. Subsequently, we apply the Riemann–Liouville fractional integral on both sides of the equation, leading to the following integral

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

system:

$$\begin{cases} S(t) = S(0) + \frac{\tau}{\Gamma(\alpha)} \int_a^t s^{\tau-1}(t-s)^\alpha \mathcal{M}_1(s, S(s))\, ds, \\ E(t) = E(0) + \frac{\tau}{\Gamma(\alpha)} \int_a^t s^{\tau-1}(t-s)^\alpha \mathcal{M}_2(s, E(s))\, ds, \\ I(t) = I(0) + \frac{\tau}{\Gamma(\alpha)} \int_a^t s^{\tau-1}(t-s)^\alpha \mathcal{M}_3(s, I(s))\, ds, \\ Q(t) = Q(0) + \frac{\tau}{\Gamma(\alpha)} \int_a^t s^{\tau-1}(t-s)^\alpha \mathcal{M}_4(s, Q(s))\, ds, \\ R(t) = R(0) + \frac{\tau}{\Gamma(\alpha)} \int_a^t s^{\tau-1}(t-s)^\alpha \mathcal{M}_5(s, R(s))\, ds, \end{cases} \tag{7}$$

C. Positivity and Biological Invariant Region

From an epidemiological perspective, ensuring that all model variables remain positive is essential for the realism and validity of the model. A fundamental approach to achieving this is by demonstrating that each variable maintains a non-negative value throughout the analysis. This guarantees that the solution to the proposed model in equation (2) is both positive and bounded, reinforcing its applicability to real-world scenarios.

Lemma 3.4.1 The boundedness and biologically acceptance region for the presented model (1) are expressed in the region $\mathfrak{B} \subset \mathbb{R}^5_+$, and are given by

$$\mathcal{H} = \left\{ (S(t), E(t), I(t), Q(t), R(t)) \in \mathbb{R}^5_+ \mid 0 \leq N(t) \leq \frac{\mathcal{A}}{\mu} \right\}.$$

Proof The suggested model (1) is evaluated in order to study the first equation,

$$\frac{dS}{dt} = \mathcal{A} + \eta R - \frac{\beta I}{N} S - \mu S,$$

and from that it follows that

$$\frac{dS}{dt} \geq -\frac{\beta I}{N} S - \mu S,$$

in accordance with the exponential growth criterion and the provision of integration

$$\frac{dS}{dt} \geq S(0) e^{-\int_0^t (\frac{\beta I}{N} + \mu) dt},$$

then, we have:

$$S(t) \geq 0.$$

Similarly, we have $E(t) \geq 0$, $I(t) \geq 0$, $Q(t) \geq 0$, and $R(t) \geq 0$ for all $t \geq 0$, so that the solution to the model does not deviate from the planes $S = E = I = Q = R = 0$. Due to the fact that the whole population is represented by the function $N(t)$, we are able to write the system (1) as follows:

$$\frac{dN(t)}{dt} = \mathcal{A} - \mu N(t),$$

It is possible that users become negligible, which means that none autonomously exit, then

$$\frac{dN(t)}{dt} \leq \mathcal{A} - \mu N(t).$$

The solution of the above equation turns to

$$N(t) \leq \frac{\mathcal{A}}{\mu} - \left(\frac{\mathcal{A}}{\mu} - N_0 \right) e^{-\mu t}.$$

Taking advantage of [27], we can say that if $N_0 < \frac{\mathcal{A}}{\mu}$ then as $t \to \infty$, asymptotically, $N(t) \to \frac{\mathcal{H}}{\mu}$ implies that $0 \leq N(t) \leq \frac{\mathcal{A}}{\mu}$. Hence, all of the feasible solutions for the model will eventually converge in the region $\mathcal{H}$. □

D. Equilibria, Basic Reproduction number, and Their Stability

In this section, we present the computations and explanation of the equilibria and basic reproduction number for the suggested model in equation (2), as well as explore how locally and globally asymptotically stable.

E. Disease-free Equilibrium Point

In order to maintain a disease-free equilibrium, the variables $E = I = Q = R = 0$ are established. In this specific case, the system (1) demonstrates that the disease-free equilibrium point is given as

$$\text{DFE} = \left(\frac{\mathcal{A}}{\mu}, 0, 0, 0, 0 \right).$$

F. Basic Reproduction Number, $\mathfrak{R}_0$

We apply the next-generation matrix technique to a dynamic system, where we consider the emergence rate of new infections, denoted as $\mathcal{F}$, and the transmission rate per individual, $\mathcal{V}$, under conditions where the disease is not present in the population. The following represents this calculation

$$\mathcal{F} = \begin{pmatrix} 0 & \beta S & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\mathcal{V} = \begin{pmatrix} -(\mu + \psi) & 0 & 0 \\ \psi & -(\mu + \phi + \theta) & 0 \\ 0 & \theta & -(\gamma + \mu) \end{pmatrix},$$

and

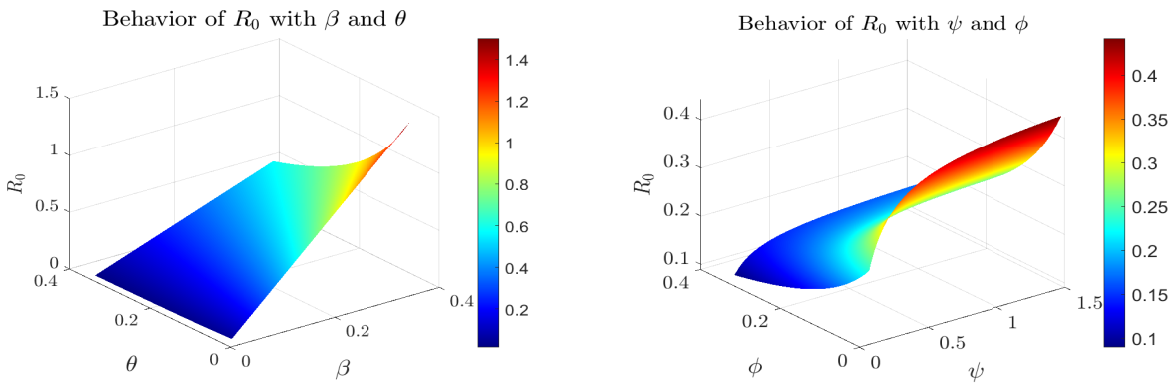
$$\mathcal{V}^{-1} = \begin{pmatrix} \frac{-1}{(\mu+\psi)} & 0 & 0 \\ \frac{\psi}{(\mu+\psi)(\mu+\phi+\theta)} & \frac{-1}{(\mu+\phi+\theta)} & 0 \\ 0 & \frac{\theta}{(\mu+\phi+\theta)(\gamma+\mu)} & \frac{-1}{(\gamma+\mu)} \end{pmatrix}$$

Thus, the basic reproduction number $\mathfrak{R}_0$ is the spectral radius of $\mathcal{F}\mathcal{V}^{-1}$ or

$$\mathfrak{R}_0 = \frac{\psi \beta \mathcal{A}}{\mu(\mu + \psi)(\mu + \phi + \theta)}.$$

Figures 4 (a) and (b) present a 3D visualization of the basic reproduction number $\mathfrak{R}_0$, generated using MATLAB, to illustrate its dependence on various model parameters. In Figure 4 (a), an increase in β leads to a corresponding rise in $\mathfrak{R}_0$, indicating that platforms where toxicity spreads more rapidly experience greater persistence of toxicity over time. Conversely, as θ increases, $\mathfrak{R}_0$ declines, demonstrating that an enhanced quarantine transfer rate plays a crucial role

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) (b)

Fig. 4: Effect of $\mathfrak{R}_0$ on (a) θ and β, and (b) ϕ and ψ

in limiting the spread of toxicity. Similarly, Figure 4 (b) depicts the relationship between $\mathfrak{R}_0$, the content exposure rate ϕ, and the transition rate ψ. The analysis suggests that by carefully regulating these parameters—such as implementing strategies to reduce exposure to toxic content—it is possible to significantly mitigate or even reverse the proliferation of harmful interactions in online spaces.

Theorem 3.4.2. The disease-free equilibrium point is locally asymptotically stable if $\mathfrak{R}_0 < 1$, and otherwise unstable.

Proof. Consider the following Jacobian matrix in regard to the system in order to demonstrate the stability criterion for the proposed system (1) at the DFE point:

$$J = \begin{bmatrix} -(\mu+\psi) & \frac{\beta A}{\mu} & 0 \\ \psi & -(\mu+\phi+\theta) & 0 \\ 0 & \theta & -(\gamma+\mu) \end{bmatrix}.$$

The characteristic equation for the matrices J is then derived in the following manner:

$$\lambda^2 + (2\mu+\psi+\phi+\theta)\lambda + (\mu+\psi)(\mu+\phi+\theta) - \frac{\beta A}{\mu}\psi = 0.$$

The stability of the DFE depends on the roots of this characteristic equation. According to the Routh-Hurwitz stability criterion, all roots (i.e., eigenvalues) have negative real parts if and only if all coefficients of the characteristic equation are positive.

- The linear term $(2\mu+\psi+\phi+\theta)$ is always positive since all parameters are positive.
- The constant term $(\mu+\psi)(\mu+\phi+\theta) - \frac{\beta A}{\mu}\psi$ determines the stability.

By defining the **basic reproduction number** as

$$\mathfrak{R}_0 = \frac{\psi\beta A}{\mu(\mu+\psi)(\mu+\phi+\theta)}, \tag{8}$$

we rewrite the constant term as

$$(\mu+\psi)(\mu+\phi+\theta)(1-\mathfrak{R}_0). \tag{9}$$

- If $\mathfrak{R}_0 < 1$, this term is positive, ensuring that all roots of the characteristic equation have negative real parts. This confirms that the disease-free equilibrium is locally asymptotically stable.
- If $\mathfrak{R}_0 > 1$, this term becomes negative, implying that at least one eigenvalue has a positive real part, leading to instability.

The disease-free equilibrium (DFE) is locally asymptotically stable if and only if $\mathfrak{R}_0 < 1$. If $\mathfrak{R}_0 > 1$, the system becomes unstable, allowing the disease to persist and spread in the population. This result confirms the epidemiological interpretation that controlling disease spread requires keeping $\mathfrak{R}_0$ below the critical threshold of 1.

G. Existence of Endemic Equilibrium Point

In this section, we will look at whether an endemic equilibrium point exists, which will be represented by the notation $(S^*, E^*, I^*, Q^*, R^*)$. The second equilibrium point is the endemic equilibrium point, as presented below:

$$\begin{cases} 0 = \mathcal{A} + \eta R - \frac{\beta I}{N} S - \mu S, \\ 0 = \frac{\beta I}{N} S - (\mu+\psi)E, \\ 0 = \psi E - (\mu+\phi+\theta)I, \\ 0 = \theta I - (\gamma+\mu)Q, \\ 0 = \gamma Q + \phi I - (\mu+\eta)R, \end{cases} \tag{10}$$

Thereby, this brings us to a further conclusion, which is given below:

Theorem 6.2. The system (1) has a unique endemic equilibrium point given by:

$$S^* = \frac{N(\mu+\psi)(\mu+\phi+\theta)}{\psi\beta},$$

$$E^* = \frac{\beta S^* I^*}{N(\mu+\psi)}, I^* = \frac{\psi}{\mu+\phi+\theta} \cdot \frac{\beta S^* I^*}{N(\mu+\psi)}$$

$$I^* = \frac{\psi}{\mu+\phi+\theta} \cdot \frac{\beta S^* I^*}{N(\mu+\psi)}$$

$$Q^* = \frac{\theta I^*}{\gamma+\mu}$$

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

$$R^* = \frac{\gamma\theta I^*}{(\gamma+\mu)(\mu+\eta)} + \frac{\phi I^*}{\mu+\eta}$$

H. Global Stability Analysis at Endemic Equilibrium Point

To achieve global stability and rule out periodic orbits, we apply Dulac's criterion.

Theorem 6.3. There are no periodic orbits in the system (1).

Proof. We introduce a Dulac function $B(S,E,I,Q,R)$ to transform the system into a divergence form where the total divergence is of a definite sign. A common choice of a Dulac function is

$$B(S,E,I,Q,R) = \frac{1}{SI}$$

Multiplying the system by $B(S,E,I,Q,R)$, we compute the divergence:

$$\frac{\partial}{\partial S}(BS') + \frac{\partial}{\partial E}(BE') + \frac{\partial}{\partial I}(BI') + \frac{\partial}{\partial Q}(BQ') + \frac{\partial}{\partial R}(BR').$$

Taking the partial derivatives, the divergence simplifies to

$$\nabla\cdot(B\mathcal{H}) = \frac{\partial}{\partial S}\left(\frac{A+\eta R - \frac{\beta I}{N}S - \mu S}{SI}\right) + \frac{\partial}{\partial I}\left(\frac{\psi E - (\mu+\phi+\theta)I}{SI}\right).$$

After simplifications, this leads to

$$\nabla\cdot(B\mathcal{H}) = -\frac{\mu}{SI} - \frac{(\mu+\phi+\theta)}{SI}.$$

Since all parameters are **positive**, the divergence function is **negative definite**, i.e.,

$$\nabla\cdot(B\mathcal{H}) < 0.$$

Since the divergence of the vector field is negative everywhere in the domain of interest, Dulac's criterion guarantees that there are no periodic orbits in the system. This implies that global stability holds for the endemic equilibrium. Thus, the endemic equilibrium is globally asymptotically stable, confirming that all solutions will ultimately converge to it in the long run.

I. Sensitivity Analysis of Important Parameters

To answer **RQ1**, we begin by identifying the parameter values that are most influential in determining social media toxicity spread. It is vital to discover numerous aspects that contribute to the toxicity spread and prevalence to decide the best technique for minimizing the number of affected users. To determine the dependence of each parameter on the SEIQR model, a sensitivity analysis of each parameter was performed using the Latin Hypercube Sampling-Partial Rank Correlation Coefficient (LHS-PRCC) method. The PRCC corresponds directly to the degree of statistical influence. A positive value indicates that an increase in this parameter leads to a positive influence on the SEIQR model. In contrast, a negative value indicates that an increase in this parameter leads to a negative influence on the SEIQR model. As shown in Figure 5 and Table IV, among the ϕ, β, ψ, μ, $\mathcal{A}$, and θ parameters, ϕ, β, ψ, $\mathcal{A}$, and θ have positive effects on $\mathcal{R}_0$. In contrast, parameter μ has a negative effect on $\mathcal{R}_0$. The red line represents the p-value, and in the context of social media toxicity, $p < 0.05$ indicates that the results of the sensitivity analysis are statistically significant. Specifically, it means that there is less than a 5% probability that the observed effects (relationships between the parameters and the effective reproduction number $\mathcal{R}_0$) occurred by chance. These behaviors are significant in executing emergency management measures. Let

TABLE IV: Sensitivity Index of Each Parameter on $\mathcal{R}_0$

Parameters	Sensitivity Indices	Relationship
ψ	0.863213	+ve
β	0.902867	+ve
$\mathcal{A}$	0.8617	+ve
μ	-0.9167	-ve
ϕ	0.89574	-ve
θ	0.1913	+ve

$$C_p^{\mathcal{R}_0} = \frac{\partial R_0}{\partial p} \times \frac{p}{\mathcal{R}_0}$$

where p is the parameter being studied for its sensitivity. Utilizing this index, one can deduce the sensitivity index corresponding to every parameter integrated into the expression for $\mathcal{R}_0$. We compute their derivatives as follows:

$$C_\beta^{\mathcal{R}_0} = \frac{\partial R_0}{\partial \beta} \times \frac{\beta}{\mathcal{R}_0},$$

such that

$$\frac{dR_0}{d\psi} = \frac{A\beta(\mu+\phi+\theta-\psi)}{\mu(\mu+\psi)^2(\mu+\phi+\theta)}$$

$$\frac{dR_0}{d\beta} = \frac{A\psi}{\mu(\mu+\psi)(\mu+\phi+\theta)}$$

$$\frac{dR_0}{dA} = \frac{\beta\psi}{\mu(\mu+\psi)(\mu+\phi+\theta)}$$

$$\frac{dR_0}{d\mu} = -\frac{A\beta\psi(\mu+\psi+\phi+\theta)}{\mu^2(\mu+\psi)(\mu+\phi+\theta)}$$

$$\frac{dR_0}{d\phi} = \frac{A\beta\psi}{\mu(\mu+\psi)(\mu+\phi+\theta)^2}$$

$$\frac{dR_0}{d\theta} = \frac{A\beta\psi}{\mu(\mu+\psi)(\mu+\phi+\theta)}$$

In Figure 5, the positive sensitivity indices, associated with ψ, ζ, β, ϑ, σ, and ϵ, predominantly influence the frequency of online toxicity manifestations, with β having the highest positive impact index, indicating its crucial role in online

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

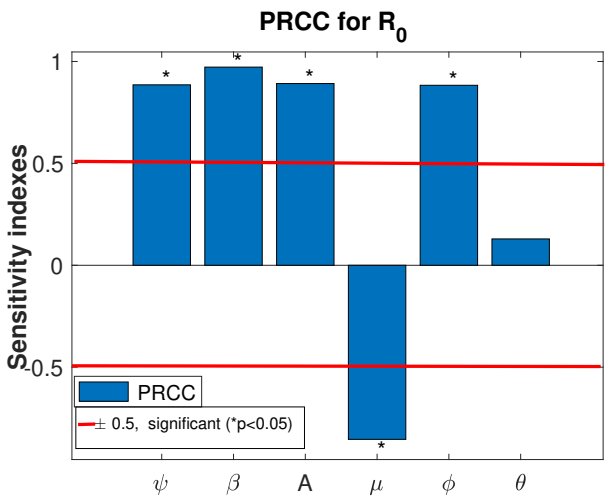


Fig. 5: Sensitivity of $\mathcal{R}_0$ of the online toxicity contagion

toxicity spread. In contrast, negative sensitivity indices linked to parameters μ and ϕ attenuate the online toxicity spread. Remarkably, parameters such as the infection rate of exposed users (ϑ), recovery rates of users who hadn't been quarantined (σ), the rate at which infected users transfer to rehabilitation (ϵ), and the rate at which infected users are banned from platform (ϕ) all curb the dissemination of online toxicity. This highlights the importance of improving infected users banned and rehabilitation measures to control the spread of the online toxicity effectively.

J. Existence and Uniqueness

We employ the Picard–Lindelöf method along with **fixed point theory** to analyze the qualitative properties of the solution for model (2). The proposed model equation ((2)) can be reformulated as:

$$\begin{cases} {}^{FFP}D^{\alpha,\tau}\mathcal{M}(t) = \Upsilon(t,\mathcal{M}(t), & 0<\alpha,\ \tau\leq 1 \\ \mathcal{M}(0)=\mathcal{M}_0\geq 0, & t\in[0,T] \end{cases} \tag{11}$$

where,

$$\mathcal{M}(t) = \begin{cases} S(t), \\ E(t), \\ I(t), \\ Q(t), \\ R(t), \end{cases} \quad \mathcal{M}_0 = \begin{cases} S(0), \\ E(0), \\ I(0), \\ Q(0), \\ R(0), \end{cases}$$

$$\Upsilon(t,\mathcal{M}(t)) = \begin{cases} \mathcal{M}_1(t,S,E,I,Q,R), \\ \mathcal{M}_2(t,S,E,I,Q,R), \\ \mathcal{M}_3(t,S,E,I,Q,R), \\ \mathcal{M}_4(t,S,E,I,Q,R), \\ \mathcal{M}_5(t,S,E,I,Q,R). \end{cases}$$

Taking advantage from Lemma 4.3, system (11) yields

$$\mathcal{M}(t) = \mathcal{M}_0(t) + \frac{1}{\Gamma(\alpha)}\int_0^t s^{\tau-1}\Upsilon(s,\mathcal{M}(s))(t-s)^{\alpha-1}ds. \tag{12}$$

such that Υ satisfies

$$|\Upsilon(t,\mathcal{M}(t)) - \Upsilon(t,\tilde{\mathcal{M}}(t))| \leq \mathbf{C}_\upsilon[|\mathcal{M}-\tilde{\mathcal{M}}|],\quad \mathbf{C}_\upsilon>0. \tag{13}$$

Theorem 7.1 *Supposed that equation (13) holds. Then the equation (11) has a unique solution if*

$$\mathcal{D} := \frac{\tau B(\alpha,\tau)}{T^{\tau+\alpha-1}\Gamma(\alpha)}C_\upsilon < 1, \quad \text{where } B(\cdot,\cdot) \text{ is the beta function.}$$

Proof. Considering Picard operator, denoted by

$$\Omega : \mathcal{Z}\to\mathcal{Z},$$

and defined as

$$\Omega\mathcal{M}(t) = \mathcal{M}_0 + \tfrac{\tau}{\Gamma(\alpha)}\int_0^t s^{\tau-1}(t-s)^{\alpha-1}\Upsilon(s,\mathcal{M}(s))ds \tag{14}$$

and set

$$\sup_{t\in\mathcal{T}}\mathcal{M}(t,0) = \mathcal{M}_0.$$

It is important to point out that the solution to the system (11) is bounded, i.e.,

$$\|\Omega\mathcal{M}(t)-\mathcal{M}(0)\| = \sup_{t\in T}|\Omega\mathcal{M}(t)-\mathcal{M}(0)|.$$

$$\|\Omega\mathcal{M}(t)-\mathcal{M}(0)\| = \sup_{t\in\mathcal{B}}\left|\frac{\tau}{\Gamma(\alpha)}\int_0^t s^{\tau-1}(t-s)^{\alpha-1}\Upsilon(s,\mathcal{M}(s))ds\right|,$$

$$\leq \sup_{t\in\mathcal{T}}\frac{\tau}{\Gamma(\alpha)}\int_0^t s^{\tau-1}(t-s)^{\alpha-1}|\Upsilon(s,\mathcal{M}(s))|ds.$$

$$\leq \sup_{t\in\mathcal{T}}\frac{\tau}{\Gamma(\alpha)}\int_0^t \xi^{\tau-1}(t-\xi)^{\alpha-1}|\Upsilon(s,\mathcal{M}(s)) - \Upsilon(t,0)|ds,$$

$$\leq \frac{\tau B(\alpha,\tau)}{\Gamma(\alpha)}T^{\tau+\alpha-1}(C_\upsilon\|\mathcal{M}\|_\infty + \Upsilon_0) := \mathcal{R} < \infty.$$

By utilizing the definition of the Picard operator (14) and considering any functions $\mathcal{M},\tilde{\mathcal{M}}\in\mathcal{Z}$, we derive the following bound:

$$\|\Omega\mathcal{M}-\Omega\tilde{\mathcal{M}}\| = \sup_{t\in\mathcal{T}}\|\Omega\mathcal{M}(t)-\Omega\tilde{\mathcal{M}}(t)\|.$$

$$\leq \sup_{t\in\mathcal{T}}\frac{\tau}{\Gamma(\alpha)}\int_0^t s^{\tau-1}(t-s)^{\alpha-1}|\Upsilon(\mathcal{M}(s)) - \Upsilon(\tilde{\mathcal{M}}(s))|ds,$$

$$\leq \sup_{t\in\mathcal{T}}\frac{\tau}{\Gamma(\alpha)}\int_0^t s^{\tau-1}(t-s)^{\alpha-1}C_\upsilon|\mathcal{M}(s)) - \tilde{\mathcal{M}}(s)|ds,$$

$$\leq \frac{\tau B(\alpha,\tau)}{\Gamma(\alpha)}T^{\tau+\alpha-1}C_\upsilon|\mathcal{M}-\tilde{\mathcal{M}}|,$$

which leads to the conclusion that $\|\mathcal{M}-\tilde{\mathcal{M}}\|\leq\mathcal{D}\|\mathcal{M}-\tilde{\mathcal{M}}\|$. Consequently, since $\mathcal{D}$ is a contraction mapping, it follows from Banach's contraction principle that system 11 has a unique solution.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

K. Numerical Analysis and Simulation

In this section, we present the numerical approximation of the proposed fractal-fractional $SEIQR$ model using a fractal-fractional derivative in the Caputo sense with a power-law kernel and simulate the long-term behavior of toxicity propagation in online social networks, analyzing the effects of quarantine interventions, mitigation policies, and fractal-fractional fluctuations in toxicity intensity. The step-by-step proof of this scheme can found in [24], [25], and [26].

Hence, the fractal-fractional $SEIQR$ model scheme is derived as follows. Given the initial conditions, the forward time-step approximation is formulated as:

$$S(t_{n+1}) = S(0) + \frac{\tau(\Delta t)^{\alpha}}{\Gamma(\alpha+2)} \sum_{m=0}^{n} \left[t_m^{\alpha-1} \Upsilon_1(t_m, S(t_m)) \right] \left[(n+1-m)^{\alpha}(n-m+2+\alpha) - (n-m)^{\alpha} (n-m+2+2\alpha) \right] - t_{m-1}^{\alpha-1} \Upsilon_1(t_{m-1}, S(t_{m-1})) \left[(n+1-m)^{\alpha+1} - (n-m)^{\alpha}(n-m+1+\alpha) \right] \tag{15}$$

$$E(t_{n+1}) = E(0) + \frac{\tau(\Delta t)^{\alpha}}{\Gamma(\alpha+2)} \sum_{m=0}^{n} \left[t_m^{\alpha-1} \Upsilon_2(t_m, E(t_m)) \right] \left[(n+1-m)^{\alpha}(n-m+2+\alpha) - (n-m)^{\alpha} (n-m+2+2\alpha) \right] - t_{m-1}^{\alpha-1} \Upsilon_2(t_{m-1}, E(t_{m-1})) \left[(n+1-m)^{\alpha+1} - (n-m)^{\alpha}(n-m+1+\alpha) \right] \tag{16}$$

$$I(t_{n+1}) = I(0) + \frac{\tau(\Delta t)^{\alpha}}{\Gamma(\alpha+2)} \sum_{m=0}^{n} \left[t_m^{\alpha-1} \Upsilon_3(t_m, I(t_m)) \right] \left[(n+1-m)^{\alpha}(n-m+2+\alpha) - (n-m)^{\alpha} (n-m+2+2\alpha) \right] - t_{m-1}^{\alpha-1} \Upsilon_3(t_{m-1}, I(t_{m-1})) \left[(n+1-m)^{\alpha+1} - (n-m)^{\alpha}(n-m+1+\alpha) \right] \tag{17}$$

$$Q(t_{n+1}) = Q(0) + \frac{\tau(\Delta t)^{\alpha}}{\Gamma(\alpha+2)} \sum_{m=0}^{n} \left[t_m^{\alpha-1} \Upsilon_4(t_m, Q(t_m)) \right] \left[(n+1-m)^{\alpha}(n-m+2+\alpha) - (n-m)^{\alpha} (n-m+2+2\alpha) \right] - t_{m-1}^{\alpha-1} \Upsilon_4(t_{m-1}, Q(t_{m-1})) \left[(n+1-m)^{\alpha+1} - (n-m)^{\alpha}(n-m+1+\alpha) \right] \tag{18}$$

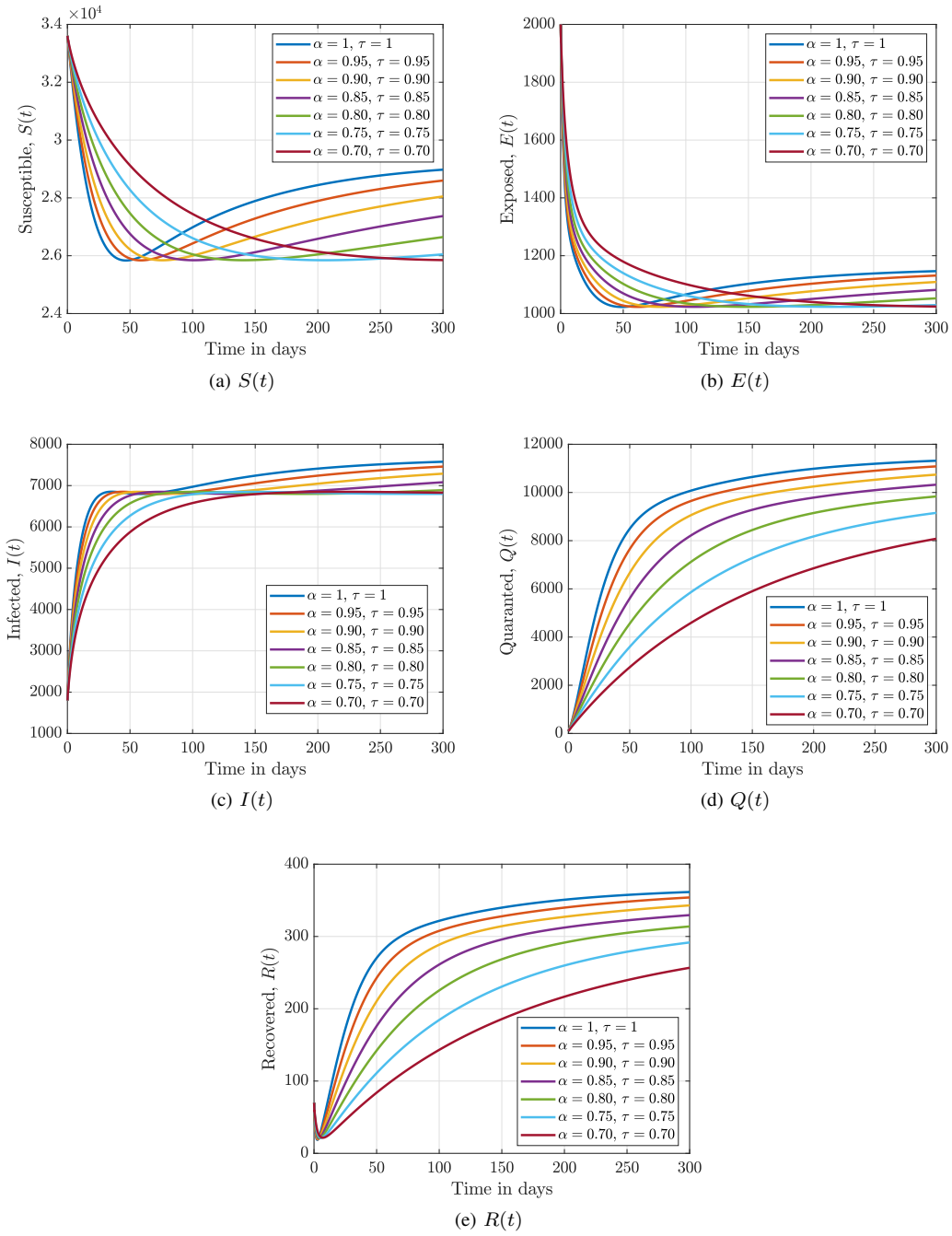
$$R(t_{n+1}) = R(0) + \frac{\tau(\Delta t)^{\alpha}}{\Gamma(\alpha+2)} \sum_{m=0}^{n} \left[t_m^{\alpha-1} \Upsilon_5(t_m, R(t_m)) \right] \left[(n+1-m)^{\alpha}(n-m+2+\alpha) - (n-m)^{\alpha} (n-m+2+2\alpha) \right] - t_{m-1}^{\alpha-1} \Upsilon_5(t_{m-1}, R(t_{m-1})) \left[(n+1-m)^{\alpha+1} - (n-m)^{\alpha}(n-m+1+\alpha) \right] \tag{19}$$

In the next section, we apply this numerical approach to simulate the long-term behavior of toxicity propagation in online social networks, analyzing the effects of quarantine interventions, mitigation policies, and fractal-fractional fluctuations in toxicity intensity.

V. Results and Discussion

In order to answer **RQ2**, by employing the outlined methodology, we applied the SEIQR model to seven distinct datasets to assess their performance and alignment with observed

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) $S(t)$ (b) $E(t)$

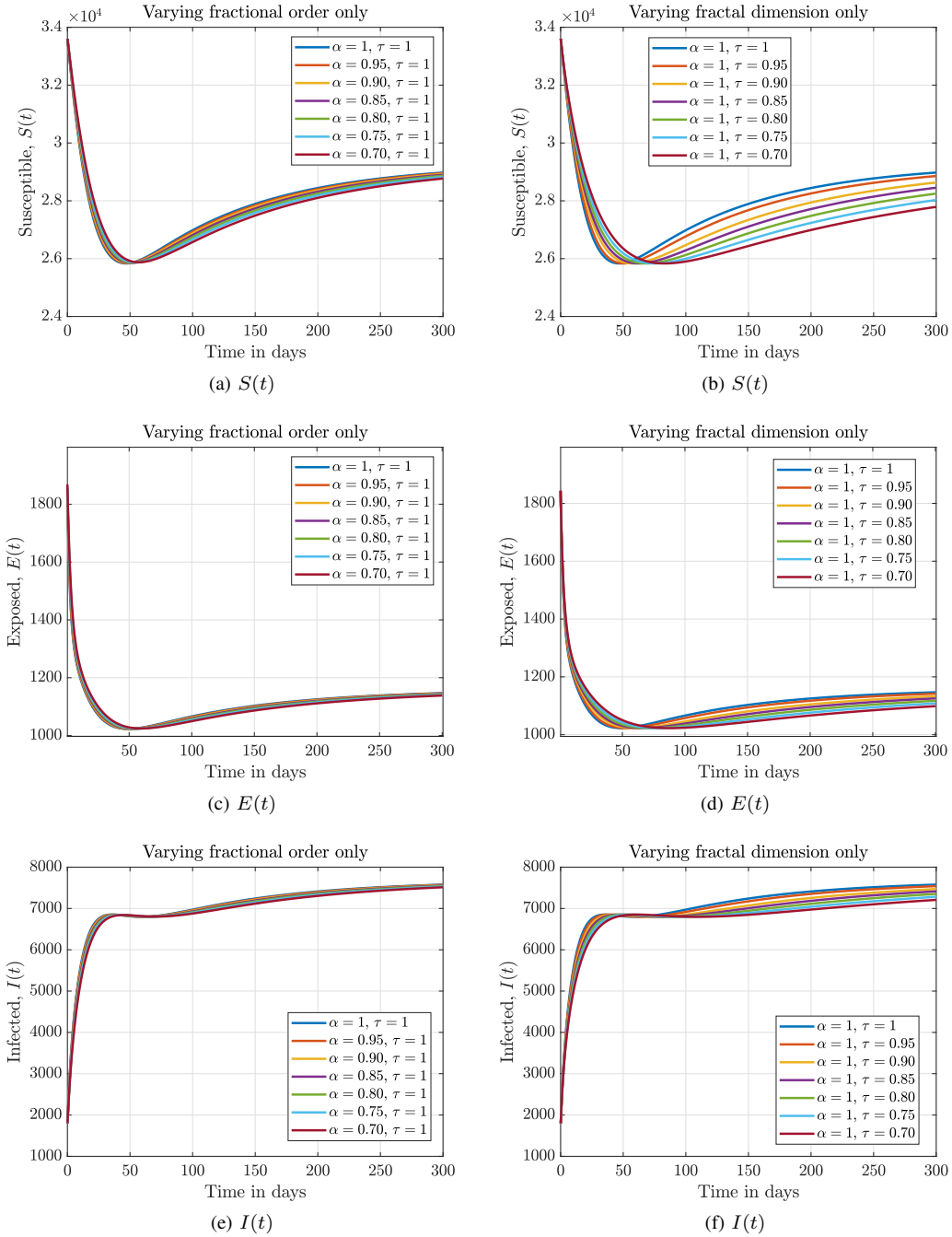
(c) $I(t)$ (d) $Q(t)$

(e) $R(t)$

Fig. 6: Numerical trajectories of $SEIQR$ model under the Caputo fractal-fractional operator.

data. Table III presents the error rates obtained from these evaluations using equation (2); it shows the error rate for moderate and high toxicity levels, as well as the error rate without splitting the data. Figure 2 shows the error rates and fitted models' representations for the three datasets—5G, Peru, and F*covid—that use the SEIQR model without splitting the data are depicted in. In the figures, the red stars represent the observed data, while the blue line depicts our fitted model. The plot illustrates how well our model aligns with the observed data. The plots in Figure 3 indicate the fitted plots for the same datasets—5G, Peru, and F*covid—using the SEIQR model with a split in the data based on the toxicity intensity. For simplicity, we excluded the fitted model for the rest of the datasets and summarized all in Table III. We compared the error rates of the SEIQR model with and without dataset splitting. The consistently low error rates across datasets when splitting the data into moderate and high toxicity levels, as compared to using the model without this division, underscore

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) $S(t)$ (b) $S(t)$

(c) $E(t)$ (d) $E(t)$

(e) $I(t)$ (f) $I(t)$

Fig. 7: Numerical comparison of fractional dynamics only and fractal dynamics only for Susceptible (S(t)), Exposed (E(t)), and Infected (I(t)) compartment under the Caputo fractal-fractional operator.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

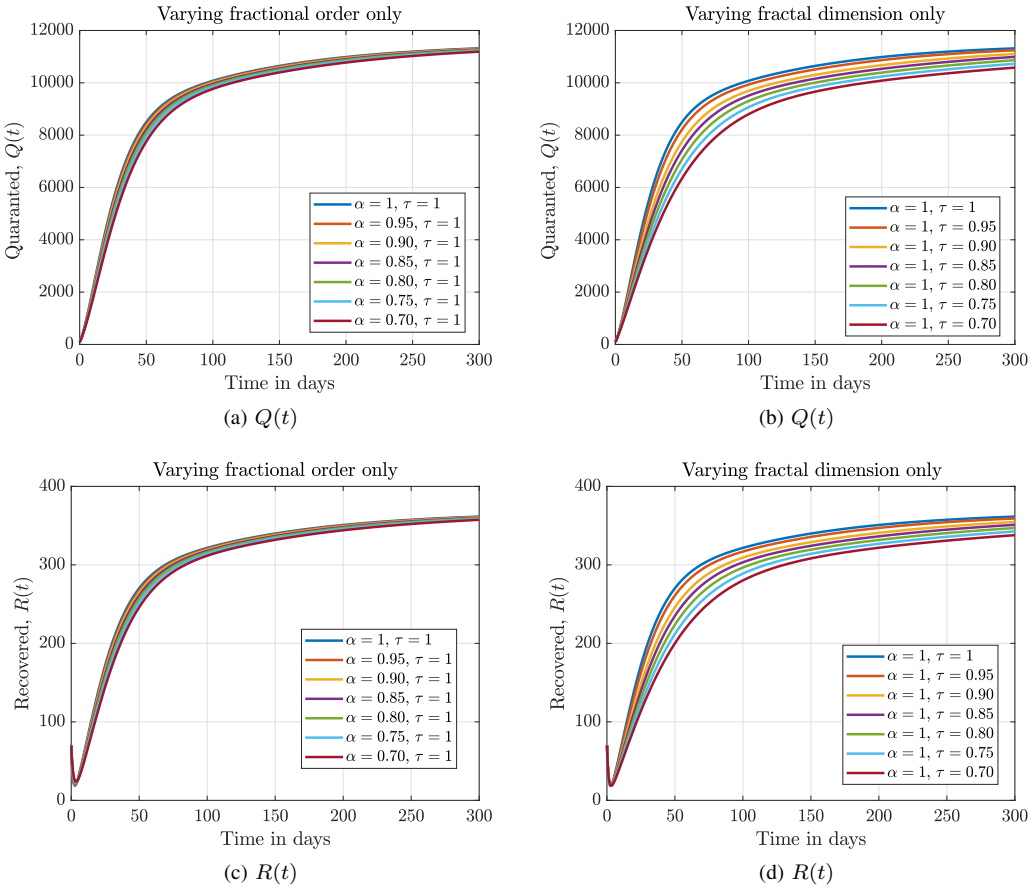


Fig. 8: Numerical comparison of fractional dynamics only and fractal dynamics only for Quarantine (Q(t)) and Recovery (R(t)) compartment under the Caputo fractal-fractional operator.

the effectiveness of the SEIQR model in capturing the dynamics of toxicity diffusion. The findings suggest that employing the SEIQR model and categorizing toxicity into moderate and high levels is reliable for understanding and predicting the spread of toxicity in various contexts. Hence, these analyses offer an answer to **RQ2**.

To analyze the evolution of toxicity spread on social media over time, numerical simulations were performed using parameter values in Table II together with the initial values ($S = 33403; I = 1799;\ E = 2000; Q = 1000; R = 700$). The behavior of toxicity propagation, influenced by the model parameters, aligned with the predictions of the SEIQR model. The results shown in Figure 6 represent the fractal-fractional spread model. The simulation results show that reducing the fractional order (α) and the fractal dimension (τ) from 1, leads to a decrease in the dynamics of the SEIQR model. This result means that the system has longer memory and more complex network structure, which slow down the spread of toxicity as past interactions continue to influence current behavior. Interventions aimed at increasing the memory effect and enhancing network complexity could be effective in reducing the spread of toxicity.

In Figures 7 and 8, we show the numerical comparison of fractional dynamics only and fractal dynamics only on the toxicity spread SEIQR model. From Figures 7(a) and 7(b), we can see that the fractional-only dynamics exhibit more asymptotic behavior compared to the fractal-only dynamics. This means that the system approaches a steady state more gradually when considering memory effects. This asymptotic behavior can be seen from Figures 7(c) to 7(f) and Figures 8(a) to 8(d). The numerical comparison of fractional and fractal dynamics in the SEIQR model highlights the importance of considering both memory effects and network structure in understanding and mitigating the spread of toxicity on social media. These analyses offer an answer to **RQ3**.

Responding to **RQ4**, in Figures 9(a) to 9(e), we present the numerical sensitivity analysis of the toxicity spread SEIQR model under the Caputo fractal-fractional operator with order and dimension ($\alpha = \tau$). This analysis was conducted by varying the rate at which toxic users were transferred to quarantine. The results revealed a significant increase in the number of users in each compartment, except for the infected users, as shown in Figures 9(a), 9(b), 9(c), 9(d), and 9(e), respectively. These findings suggest that reducing the number of infected users, by leveraging memory effects (fractional dynamics) and addressing repeated behavioral patterns (fractal

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

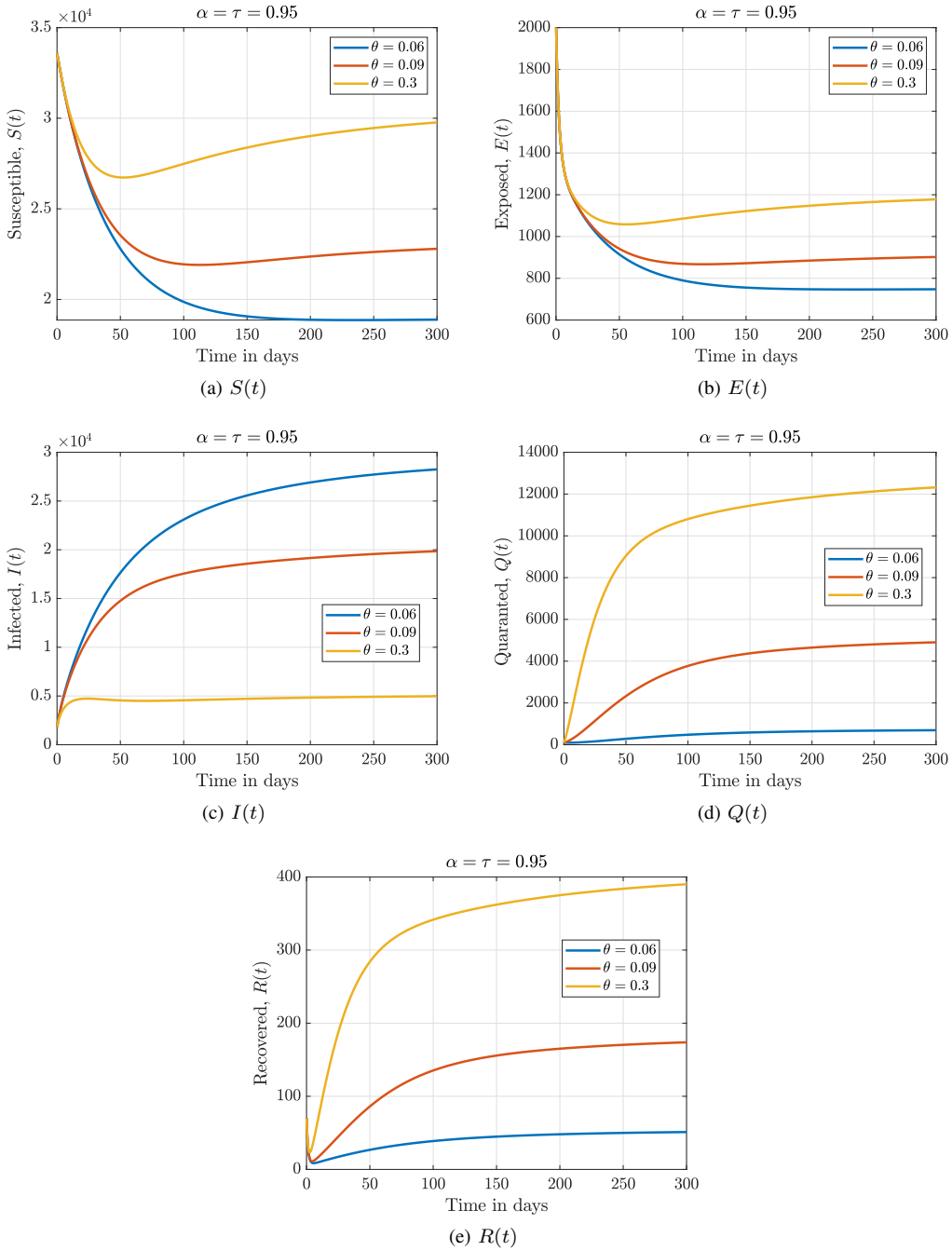


Fig. 9: Numerical trajectories of $SEIQR$ model under the Caputo fractal-fractional operator with order and dimension $(\alpha = \tau = 0.95)$ when the rate at which users transfer to quarantine.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

dynamics), will effectively mitigate the spread of toxicity on social media.

Building on the insights from the SEIQR model's numerical simulations and sensitivity analysis, several strategic interventions are recommended to control the spread of toxicity on social media platforms. These measures leverage fractional and fractal dynamics to enhance content moderation and behavioral management. **First**, we suggest adaptive moderation through behavioral memory and recurrence patterns. To effectively manage toxicity, platforms should implement temporary restrictions on accounts engaged in harmful behavior. These restrictions may include suspensions, limited posting capabilities, or increased scrutiny. This approach acknowledges the role of memory effects in user behavior, ensuring that past interactions influence future activities. By allowing monitored users to reform while preventing unchecked toxicity from gaining momentum, this method balances community safety with user rehabilitation. **Secondly**, we suggest algorithmic content demotion and network complexity reduction. Reducing the visibility of toxic content using algorithmic content demotion can limit its reach and engagement. By prioritizing non-toxic interactions and reducing amplification through engagement restrictions (e.g., limiting likes, shares, and comments on flagged content), the network's fractal structure becomes less conducive to toxicity proliferation. This intervention effectively slows down the contagion effect, aligning with the SEIQR model's findings that lowering the fractal dimension reduces toxicity spread. **Thirdly**, we suggest cross-platform collaboration for unified toxicity control. Toxic content often spreads across multiple social media networks, requiring cross-platform collaboration for consistent moderation. A unified approach—incorporating shared databases, technological innovation, and standardized policies—can enhance intervention efficiency. Platforms can integrate collective quarantine mechanisms, ensuring that users flagged for toxicity on one platform face restrictions across multiple services, thereby mitigating repeated behavioral patterns. **Lastly**, we suggest structured quarantine and reintegration mechanisms. The SEIQR model suggests that quarantining toxic users is an effective strategy to limit further spread. Platforms should establish dedicated discussion spaces where quarantined users can engage in monitored interactions rather than simply migrating to another platform. This could include controlled forums, restricted access groups, or supervised comment sections, ensuring that users receive guidance toward positive behavioral reform.

By implementing these memory-aware and network-structured interventions, social media platforms can effectively mitigate the spread of toxicity, reduce harmful interactions, and foster a healthier digital environment.

VI. Conclusion and Future Works

This study applied the SEIQR epidemiological model to analyze the propagation of toxicity on social media, integrating fractal-fractional dynamics to capture the complexity of online interactions. By categorizing toxicity into moderate and high intensities, the model significantly improves prediction accuracy, as evidenced by lower error rates across diverse datasets (as seen in Table III). This granular classification aligns with real-world observations that toxicity manifests in varying degrees of severity, each requiring distinct intervention strategies. The sensitivity analysis further underscores the critical role of parameters such as the contact rate (β) and quarantine rate (θ) in modulating the basic reproduction number ($\mathcal{R}_0$), offering actionable insights for platform regulators.

The fractal-fractional approach addresses two key limitations of traditional models:

1) The **memory effect** of user interactions, where past behaviors influence future toxicity spread, and
2) The **heterogeneous network structure** of social media, characterized by super-spreaders and echo chambers.

Simulations (seen in Figures 6–9) demonstrated that reducing the fractal dimension (τ) and fractional order (α) slowed toxicity propagation, highlighting the importance of algorithmic adjustments to demote toxic content and disrupt feedback loops. These findings resonate with prior epidemiological studies on quarantine efficacy but extend their applicability to digital behaviors through fractal-fractional operators. The summary of our key contributions include the following:

1) **Improved Predictive Accuracy:** Splitting toxicity into moderate and high levels reduces model error rates by up to 90% (e.g., Peru dataset error drops from 0.41 to 0.095), enabling precise resource allocation.
2) **Parameter Sensitivity:** The positive correlation of β, ψ, and $\mathcal{A}$ with $\mathcal{R}_0$ underscores the need to limit exposure to toxic content, while the negative impact of μ and θ validates quarantine policies.
3) **Fractal-Fractional Dynamics:** By embedding memory effects and network complexity, the model reveals that interventions targeting behavioral persistence (e.g., temporary suspensions) and algorithmic amplification (e.g., content demotion) are critical for mitigation.

However, the study has limitations. The reliance on threshold-based toxicity classification (e.g., Detoxify scores > 0.5) may oversimplify nuanced human communication. Additionally, the model assumes homogeneous mixing within compartments, which may not fully capture the fragmented nature of online communities. Future iterations could incorporate network-specific topology data to refine compartmental transitions.

Acknowledgments

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078, W911NF-25-1-0147), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency, the Australian Department of Defense Strategic Policy

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Grants Program, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment, and the Donaghey Foundation at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

REFERENCES

[1] N. Agarwal, "Modeling Quarantine Intervention for Varied Toxic Intensities," in *Proc. 15th Int. Conf. Soc. Media Technol., Commun., and Informatics (SOTICS)*, Lisbon, Portugal, Sep. 2025.

[2] N. Agarwal and T. Falade, "Beyond Algorithmic Detection: How Community Expertise Moderates Toxic Algospeak via Dual Cognitive Mechanisms," in *Proc. 32nd Americas Conf. Inf. Syst. (AMCIS)*, Reno, NV, USA, Aug. 2026.

[3] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the ACM SIGCOMM International Measurement Conference*, 2009.

[4] S. Shajari, M. Alassad, and N. Agarwal, "Characterizing suspicious commenter behaviors," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. Kusadasi, Turkiye: ACM, Nov. 2023, pp. 631–635.

[5] L. Dai, X. Liu, and Y. Chen, "Global dynamics of a fractional-order sis epidemic model with media coverage," *Nonlinear Dynamics*, vol. 111, pp. 19 513–19 526, 2023.

[6] T. C. Falade and N. Agarwal, "Weighted Focal Structure Analysis for Coordinated Toxicity Propagation in Social Networks," in *Proc. 14th Int. Conf. Complex Netw. Their Appl. (Complex Networks)*, New York, NY, USA, Dec. 2025.

[7] K. DiCicco, N. B. Noor, N. M. Yousefi, B. Spann, and N. Agarwal, "Toxicity and networks of covid-19 discourse communities: A tale of two social media platforms," in *Proceedings of the 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval*, 2023, pp. 30–42.

[8] T. C. Falade and N. Agarwal, "Probabilistic Soft Logic for Toxic Intent Prediction in Conversation: A Moral Foundations Theory-Driven Neural-Symbolic Framework," in *Proc. 59th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Maui, HI, USA, Jan. 2026.

[9] N. Agarwal and T. C. Falade, "Context-Dependent Toxicity Propagation in Online Discourse: A Generational Analysis Framework for Characterizing Conversation Dynamics in Reddit," *IEEE J. Soc. Comput.*, 2026.

[10] D. DeMarsico, N. Bounoua, R. Miglin, and N. Sadeh, "Aggression in the digital era: Assessing the validity of the cyber motivations for aggression and deviance scale," *Assessment*, vol. 29, no. 4, pp. 764–781, 2022.

[11] N. B. Noor, N. Yousefi, B. Spann, and N. Agarwal, "Comparing toxicity across social media platforms for covid-19 discourse," in *Proceedings of the Ninth International Conference on Human and Social Analytics*, 2023, pp. 21–26.

[12] N. Yousefi, N. B. Noor, B. Spann, and N. Agarwal, "Towards developing a measure to assess contagiousness of toxic tweets," in *Proceedings of the International Workshop on Combating Health Misinformation for Social Wellbeing*, 2023.

[13] N. Yousefi, N. Noor, B. Spann, and N. Agarwal, "Examining toxicity's impact on reddit conversations," in *Proceedings of the Complex Networks & Their Applications XII*, H. Cherifi, L. M. Rocha, C. Cherifi, and M. Donduran, Eds. Cham: Springer Nature Switzerland, 2023, pp. 401–411.

[14] T. Li, S. Wang, and B. Li, "Research on suppression strategy of social network information based on effective isolation," *Procedia Computer Science*, pp. 131–138, 2018.

[15] R. Das and W. Ahmed, "Rethinking fake news: Disinformation and ideology during the time of covid-19 global pandemic," *IIM Kozhikode Society & Management Review*, vol. 11, no. 1, 2021.

[16] G. A. Ngwa and M. I. Teboh-Ewungkem, "A mathematical model with quarantine states for the dynamics of ebola virus disease in human populations," *Computational and Mathematical Methods in Medicine*, 2016.

[17] N. Yousefi and N. Agarwal, "Studying the influence of toxicity intensity on its propagation using epidemiological models," in *Proceedings of the 30th Americas Conference on Information Systems*, 2024.

[18] E. Addai, N. Yousefi, and N. Agarwal, "Seiqr: An epidemiological model to contain the spread of toxicity using memory-index," in *Proceedings of the 18th International AAAI Conference on Web and Social Media (ICWSM): CySoc 2024 – 5th International Workshop on Cyber Social Threats*, 2024, pp. 1–12.

[19] K. Diethelm and N. J. Ford, "Analysis of fractional differential equations," *Journal of Mathematical Analysis and Applications*, vol. 265, no. 2, pp. 229–248, 2002.

[20] A. Atangana and D. Baleanu, "New fractional derivatives with nonlocal and non-singular kernel: Theory and application to heat transfer model," *Thermal Science*, vol. 21, no. 2, pp. 1035–1045, 2017.

[21] A. Atangana, "Modeling the spread of covid-19 with new fractal-fractional operators: Can the lockdown save mankind before vaccination?" *Chaos, Solitons & Fractals*, vol. 136, p. 109860, 2020.

[22] A. Atangana and M. A. Khan, "Modeling and analysis of competition model of bank data with fractal-fractional caputo-fabrizio operator," *Alexandria Engineering Journal*, vol. 59, pp. 1985–1998, 2020.

[23] L. Hanu and Unitary Team, "Detoxify," https://github.com/unitaryai/detoxify, 2020, last accessed: May 2026.

[24] K. Diethelm, N. J. Ford, and A. D. Freed, "A predictor-corrector approach for the numerical solution of fractional differential equations," *Nonlinear Dynamics*, vol. 29, pp. 3–22, 2004.

[25] A. Atangana and D. Baleanu, "New fractional derivatives with non-local and non-singular kernel: Theory and application to heat transfer model," *Thermal Science*, vol. 20, pp. 763–769, 2016.

[26] Z. Odibat and N. T. Shawagfeh, "Generalized taylor's formula and its application on a fractional differential equation," *Applied Mathematics and Computation*, vol. 197, pp. 467–477, 2008.

[27] M. Blanc and O. Bournez, "A characterization of polynomial time computable functions from the integers to the reals using discrete ordinary differential equations," in *International Conference on Machines, Computations, and Universality*. Springer, 2022, pp. 58–74.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Developing a Framework for Circular Economy Products and Processes within the Construction Industry

Alan Martin Redmond PhD, CSEP
Auto-Entrepreneur
Montigny-le-Bretonneux, France
e-mail: a.redmond@numeriquetech.eu

***Abstract*— This article presents a framework that will bring a new perspective to circular economy products and processes within the construction industry. The framework will utilize Federated Enterprise Architecture approach that is traditionally used in the aerospace, automotive, and oil and gas industries, for built environment needs. The framework will address 5 main objectives : i) to analyze and develop a decentralized Federated Framework for construction and renovation processes; ii) to optimize the re-usability and recycling of building materials and components; iii) to investigate the validation process for framework solutions across large-scale pilots in diverse contexts ; iv) to promote and implement innovative tools for stakeholder collaboration and green finance integration; and v) to provide guidance for policy, standardization, and stakeholder adoption. This article acts as prerequisite to prepare the construction industry for transition to Federated Enterprise Architecture practices.**

Keywords – Federated Framework; interoperability; adaptability; re-use and deconstruction of buildings, life cycle assessment; construction and renovation, circular economy; pilot studies

I. INTRODUCTION

This article is a technical advancement of a published conference paper: 'MODERN SYSTEMS 2025' article "A Framework for Adaptability, re-use and Deconstruction of Buildings, Aligned with the Principles of Circular Economy"[1]. In this article the reader will understand how previous model representation projects supported the development of MBSE tools and techniques. Furthermore, the attributes of Federal Enterprise Architecture Framework (FEAF) is identified to demonstrate future architecture interface requirements, while also defining matrix criteria that aligns with the architecture types. Moreover, in this article the demonstrations are further developed and cross referenced to the case studies. In addition, this article presents the 'Tools Evaluation and Holistic Impact Assessment' platform that summarizes the what-if scenario analysis to meet the clients requirements.

It is estimated that half of the total CO2 emissions of a building arise during the construction process and the production of the material components. Furthermore to reduce the construction industry's contribution to global warming, its immediate decarbonisation is necessary and cannot only depend on incremental CO2 savings over the life cycle of a building but must include the planning and construction process. There is growing acknowledgement for circular construction as a means to develop, use, and re-use buildings, sites and infrastructure without unnecessarily exploiting natural ressources, polluting the environment and damaging ecosystems [2]. Moreover, The European Environment Agency recognizes that "almost 75% of the building stock is currently energy inefficient and more than 85% of today's buildings are likely to still be in use in 2050. Energy renovation of buildings is ongoing but at a very slow rate"[3]. In addition, 'Ecochain' has identified that in many industries, the supply chain accounts for more than 80% of the environmental impact and in addition the supply chain managers that source from different suppliers contribute to a massive impact on their product footprint [4]. In 2018, 'The European Commission' led an intiative to focus on supporting regions and EU countries to develop national bioeconomy strategies that will enhance knowledge on biodiversity and ecosystems, monitor progress towards a sustainable bioeconomy, promote good practices to operate the bioeconomy and enhance the benefits of biodiversity. In order to unlock investments and markets, the commission mobilised stakeholders in developing sustainable biobased solutions while launching a €100 million circular bioeconomy thematic investment platform [5]. According to the World Economic Forum 'Bioeconomy is emerging as a transformative force for sustainable development, leveraging biological resources and innovative technologies to address global environmental challenges [6]. Building upon these challenges and initiatives there is a growing need for solutions that extend the service life of buildings, support material reuse and recycling, and improve stakeholder collaboration through shared data and digital tools. This article outlines the development of a Federated Enterprise Architechure framework for developing solutions, methods and processes to meet the outlined objectives and align with the European Comission's Built4People Partnership [7] contributing to : (1) Increased adaptability of buildings, (2) Reduction of Waste, (3) Support for local and regional economic development, and (4) Policy evolution. The article has three main sections covering the design of the framework (objectives etc.) in Section II, the methodology in Section III and Section IV highlights the potential demonstrations. The acknowledgement and conclusions close the article.

II. DESIGNING A FRAMEWORK

The main goal of the framework is to deliver a comprehensive and sustainable framework for circular

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

construction and renovation, ensuring adaptability, reuse, and deconstruction of building components while minimizing environmental impact and maximizing stakeholder value. The outcome is an innovative approach that integrates digital technologies, advanced processes, and novel materials into an interoperable and decentralized Federated Framework for (de)construction and renovation. This framework shifts away from monolithic, proprietary systems and creates an interconnected ecosystem where tools, platforms, and processes maintain autonomy while contributing to a unified, efficient workflow. The Federated Framework is designed to support decision-making across all stages of a building's life cycle from design and construction to reuse and deconstruction. Moreover, at the centre of the framework is Model-Based Systems Engineering (MBSE).

According to [8] MBSE is a paradigm that uses formalized representations of systems, known as models, to support and facilitate the performance of Systems Engineering [SE] tasks throughout a system's life cycle. In addition, [9] acknowledged that "Models are representations that are used to capture, analyze, and/or communicate information about a system or concept. They can vary in scope, purpose, and type, and can be utilized both individually as stand-alone entities as well as in concert with each other as part of an integrated set." [10] The CLARITY project is recognised for developing and structuring an international ecosystem to support open innovation in the domain of MBSE. Furthermore, the Capella open-source tool that was provided for the project was a Thales initiative that had been developed in house, and deployed in various Business Units. It was implemented as a key solution for complex systems design providing "technological breakthroughs for MBSE, methodological support and guidance for systems engineers through the underlying integrated methodology Arcadia." Since the CLARITY project in 2015, the Capella open-source tool and Arcadia methodology have advanced to be a global leader in defining, analysing, designing & valiadting System, Software, and Hardware Architecture.

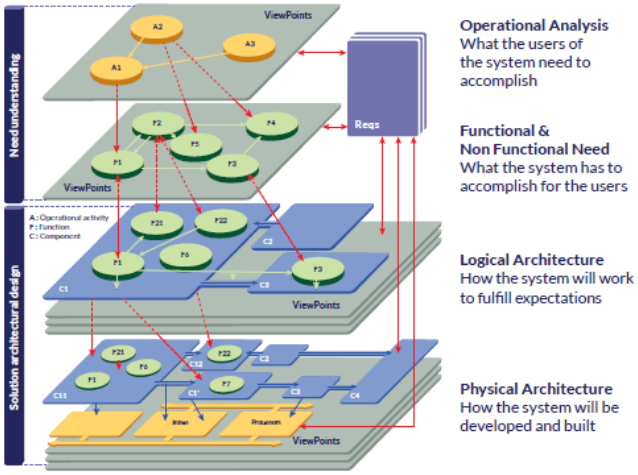


Figure 1. MBSE – The Arcadia Perspectives [11]

Figure 1 illustrates the operational needs and system needs analysis viewpoints against the 'need understanding', and the logical architecture and physical architecture against the solution architecture design. The Federated Enterprise Architecture approach will utilise the Capella open-source tool and Arcadia methodology to validate/justify solutions against operational needs, requirements and impact anaysis.

In reference to the foundation of Architecture Frameworks, in September 1999, the Federal CIO Council published the FEAF, Version 1.1. The purpose of Version 1.1 was for developing an Enterprise Architecture (EA) within any Federal Agency or for a system that transcends multiple inter-agency boundaries. Moreover, common business practices and designs were incorporated while providing standards for developing and documenting architecture descriptions of high-priority areas. Figure 2 shows the FEAF partitions as a given architecture into business, data, applications, and technology architectures. It was also acknowledged that the FEAF included the first three columns of the Zachman framework and the Spewak EA planning methodology [12].

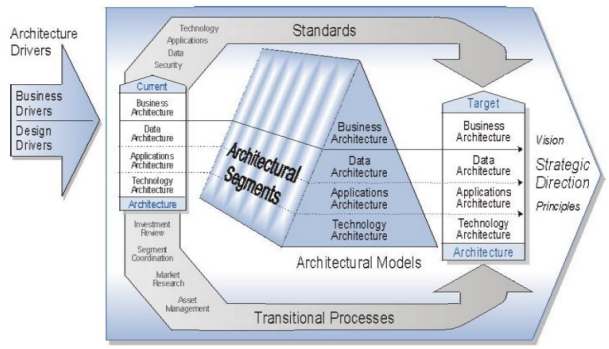


Figure 2. Structure of the FEAF Components

The goal of the Federal CIO Council was to "develop, maintain, and facilitate the implementation of a top-level EA predicated upon the FEAF". The objective of the architecture was to serve as a reference point to facilitate the efficient and effective coordination of common business processes, technology insertion, information flows, systems, and investments among Federal Agencies. Figure 3 illustrates the 3x5 matrix that graphically aligns with the architecture types (Data, Application, and Technology) on one axis of the matrix, and perspectives (Planner, Owner, Designer, Builder, and Subcontractor) on the other.

	Data Architecture	Application Architecture	Technology Architecture
Planner Perspective	List of Business Objects	List of Business Processes	List of Business Locations
Owner Perspective	Semantic Model	Business Process Model	Business Logistics System
Designer Perspective	Logical Data Model	Application Architecture	System Geographic Deployment Architecture
Builder Perspective	Physical Data Model	Systems Design	Technology Architecture
Subcontractor Perspective	Data Dictionary	Programs	Network Architecture

Figure 3. FEAF Architecture Matrix

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

In the authors' opinion, the FEAF cleared the way for future framework excellence such as The Modular Open System Archiecture ecosystem [13] that involves many areas in enegineering across the system life cycle, in particular acquisition program that considers : **plan** – the MOSA strategy with defined objectives ; **modularize** – by decomposing system capabilities into functional modules ; **specify interfaces** – by identifying connections between system building blocks ; **define interfaces** – specifications by capturing how functional modules interact ; and **standardise interfaces** – specifications to allow for opportunities of future mondernisation. It was these considerations that were acknowledged as best practices for this articles **Federated Enterprise Architecture.** The following sections address the overall components of the Federated Enterprise Architecture.

A. Objectives and Ambition

The overall objective is to deliver a more sustainable framework for circular construction and renovation, ensuring adaptability, reuse, and deconstruction of building components while minimizing environmental impact and maximizing stakeholder value. The project proposal leverages cutting-edge digital technologies, advanced methodologies, and innovative materials to align with the European Union's goals for a sustainable, people-centric built environment. The framework adopts a holistic and decentralized approach, integrating federated architectures, life-cycle-based methods, and participatory design processes to address the challenges of resource efficiency, carbon reduction, and material circularity. This vision is realized through a suite of interoperable tools, validated across diverse geographical and climatic contexts. The Strategic Objectives (SO) that will enable the achievement of the framework are described below:

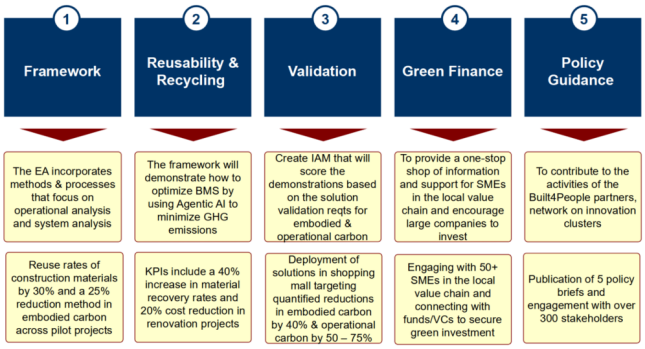


Figure 4. The Requirements and Measurements – 5 Objectives.

a) To analyze and develop a decentralized Federated Framework for construction and renovation processes (objective 1): This objective focuses on creating an interoperable architecture that integrates advanced digital solutions, such as graph technology, Digital Twins (DTs), and blockchain-based systems. The federated framework facilitates interoperability and information sharing between semi-autonomous de-centrally organized Line of Businesses (LOBs), in particular reference to AI & Agents.

b) To optimize the re-usability and recycling of building materials and component (objective 2): The framework will enhance the traceability and performance of sustainable materials, prioritizing bio-based, CO2-storing, and modular solutions such as, the use of waste wood, bio-based insulation, Supervisory Control and Data Acquisition (SCADA) controls, Heating, Ventilation, and Air Conditioning (HVAC) for energy management systems, and product inventories for supply chains. Through advanced Life Cycle Analysis (LCA) and predictive maintenance tools, the framework will establish best practices for disassembly and reuse, enabling a shift towards a circular construction model.

c) To investigate the validation process of the Framework solutions across large-scale pilots in diverse context (objective 3): Two pilot sites across Portugal and Slovenia, and one virtual pilot study (referring to test simulations for example interoperability issues of shared models) between Romania and Denmark, will perform the adaptability and performance analysis of the Federated Framework under real-life and close to real life conditions. These pilots will focus on delivering test solutions in residential and non-residential settings, enhancing adaptability, resource efficiency, and scalability.

d) To promote and implement innovative tools for stakeholder collaboration and green finance integration (objective 4): This objective involves creating a virtual living lab and green finance platform to connect investors with sustainable building initiatives. It integrates digital building logbooks and financial models to enhance the scalability and economic viability of circular practices. The outcome is to provide support and knowledge on investment opportunities, such as 'Growth and Income Fund' [14] and 'Feeder Fund' [15] as investment vehicles to provide capital from different investors and sources.

e) To provide guidance for policy, standardization, and stakeholder adoption (objective 5): The framework will deliver actionable recommendations for regulatory bodies, certification authorities, and industry stakeholders to support the standardization and scaling of circular construction practices. Moreover, the framework will engage in 6 key stewardship of activities such as: standard identification and monitoring; collaboration with standards and policy bodies; gap analysis and recommendations development; advisory policy framework working group; workshops and stakeholder engagement; and policy briefs and contributions, that will increase awareness on best practices for design, adaptability, reuse and deconstruction.

Table I explains the various challenges that will be applied to each objective and its evaluation defined through measurement.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I. REQUIREMENTS & MEASUREMENTS

Evaluation		
Objectives	***Requirements***	***Measurable Key Performance Indicators (KPIs)***
1: Frame-work	The Enterprise Architecture incorporates methods and processes that focus on operational analysis – what the stakeholders need to accomplish and system analysis – what the system has to accomplish for the stakeholders.	KPIs include increased reuse rates of construction materials by 30% and a 25% reduction in embodied carbon across pilot projects.
2: Re-usability & Recycling	The framework will demonstrate how to optimize Building Management Systems (BMS) by using Agentic AI to minimise GHG emissions during the full building operational life cycle that are essential for maintaining high levels of user comfort and well-being, which directly translate into high User QoE (Quality of Experience) KPIs.	The challenges that exist particularly for building systems is that electronic systems or products such as consumer products become obsolete long before the device wears out or fails and are simply discarded and sent to landfill. KPIs include a 40% increase in material recovery rates and a 20% cost reduction in renovation projects.
3: Validation	After the demonstrations the framework will create Impact Assessment Methodology (IAM) that will score the demonstrations based on the solutions validation requirements for both embodied and operational carbon.	KPIs include successful deployment of solutions in a shopping mall in Portugal with a total area of ~200,000 m2 that hosts over 150 stores on 5 floors and a small site 40 m² of built space targeting quantified reductions in embodied carbon by 40% and operational carbon by 50 – 75% respectively.
4: Green Finance	The overall objective for framework is to provide a one-stop shop [16] of information and support for SMEs in the local value chain but to also encourage larger companies to invest in disruptive technologies.	KPIs include engaging 50+ SMEs in local value chains and connecting them with funds/Venture Capitalists (VCs) to secure green investments.
5: Policy Guidance	The framework will contribute to the activities of the Built4People partners and to the Built4People network on innovation clusters through the achievements of the demonstration.	KPIs include the publication of 5 policy briefs and engagement with over 300 stakeholders.

B. *The Scope of the Framework*

The Federated Framework consists of 16 interconnected modules that encompass key enabling tools, processes, and methods. These include advanced IT solutions such as graph technologies, digital twins, distributed ledger systems (blockchains), and Common Data Environments (CDEs), as well as methodologies like LCA, MBSE, and business process mapping. Each module addresses specific aspects of circular construction, such as material traceability, predictive maintenance, and user-centric design.

TABLE II. FRAMEWORK: INNOVATIVE TOOLS, PRODUCTS & TECHNIQUES, PROCESS & METHODS

Framework		
Innovative Tools (INNT)	***Products and Techniques (P&T)***	***Process and Methods (P&M)***
INNT1) AI, IoT and Agents for BMS	P&T1) Digital Building Logbooks including DPP, BRP and MP	P&M1) Buinsess Process Mapping
INNT2) Graph Technology	P&T2) Security Transactions including Data Encryption, and Payments Certs (eIDAS) etc.	P&M2) Life Cycle Analysis including SCBA
INNT3) Large Language Models including LangChain - GenAI, and Chat models	P&T3) Open Source: Open LCA and Open API	P&M3) Federated Architectures including Linked Data CDE
INNT)4 Digital Twins, BIM and GIS Platform	P&T4) Semantic Analysis Techniques, RAG and Indexes	P&M4) MBSE
INNT)5 Distributed Ledger Technologies / Blockchains		P&M5) ETL Data Integration
INNT6) Preventative and Predictive Maintenance - RAM		P&M6) Real-Time Linked Data Space

a. Legends (Digital Product Passport – DPP; Building Renovation Passport – BRP; Material Passport – MP)

The semi-autonomous federated systems architecture will support State of the Art LOBs that uniquely provide solutions and by integrating them with products, techniques, processes and methods they will pioneer co-creation (see Table II). Furthermore, the pilot studies will transition the Use Cases (see Table IV), which are bundle of selected technology advancements, products, techniques and methods & processes from Table II at the demonstrations initial conceptual stage to an advanced operational development stage based on constant evolution and learning. Moreover, using the Federated Architecture will present simulations of the decentralized platforms' abilities to connect with each other in a “common data-space” of open collaboration pooling of information. At pilot level, the framework will encapsulate three very different pilot studies, for which two of them are very much real-life scenarios representing residential (Slovenia) and non-residential projects (Portugal). The other Pilot study comprises of two countries (Denmark and Romania) working on virtual models to test their CDE platforms and interoperability. The impact of these Pilot studies will be evaluated at city council level in the Ukraine to provide added value to the project circular economy results.

III. METHODOLOGY

The methdology section comprises of four sections addressing an exploratory research stage; i) the framework vision, ii) the process, iii) the phases, and iv) the use cases.

A. *The Frameworks Vision*

The framework’s vision is the development of products/materials/services including those that contribute to

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

disassembled and reused, and CO2-storing materials etc. and also the cyphering of materials via Graph Technology. Moreover Graph Technology (ISO/IEC 39075:2024) Information Technology, Database Languages and GQL defines data structures [17]. The framework will provide structured and unstructured data from existing Relational Database and web services such as ECO Building Materials Suppliers Catalogs, deconstruction – reuse warehouse of materials, certified environmental product declaration catalogs and product environmental profiles in compliance with ISO 14025 standard [18], community engagement platforms, European circular economy stakeholder platforms, and environmental monitoring and IoT platforms.The integrating advances of bio-based materials manufacturing technology for example Ceramics and Glass and the use of digital solutions (AI, property and Knowledge Graphs (KG), Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), LangChain) with economy principles, i.e., targeting VCs, Angel funds to provide investment opportunities to enterprises that develop and reuse, deconstruction materials in a life-cycle optimization and circular economy perspective, will offer solutions that not only mitigate environmental impacts, but also drive economic growth and societal well-being.

B. *The Process*

The goal of the framework is to bring a new perspective to circular economy products and processes. However, fostering strategies and roadmaps is not enough and the framework will utilize Federated Enterprise Architecture approach that is traditionally used in the aerospace, automotive, and oil gas industries, for built environment needs.

Figure 5 illustrates the use of MBSE, which is a process offered in the framework to define the solution design. The pilot studies will transition from the initial conceptual stage (pilot study demonstrators original assumptions) to an advanced operational development stage based on constant evolution and learning. In the context of the framework, Figure 5 defines the solutions need analysis, it is an integrated systems open architecture identifying components (e.g., material providers), functions (e.g., supply chain), and exchange items (CO2 carbon content, policies, predictive & preventive maintenance, etc.) pooling of information.

In comparison to traditional architectures that highlight connections of how various components such as ontologies are used and interchanged with semantic platforms and where over ambitious proposals identify lots of connected applications, the framework project builds on a system design approach. It will promote iteration to accommodate the pilot studies' ever changing needs. In addition, new knowledge through the project's evolving research techniques can be implemented and tested to provide better practices and recursion to reach a level to finalize market readiness. Such analysis captures opportunities for disruptive and innovative solutions, processes and methods.

Furthermore, the Market Key Results Investment Platform, New Business Models, and Policy Recommendations will build upon private investments initiatives such as SMEs involvement in the project. Their presence shall extend contribution measures to local and

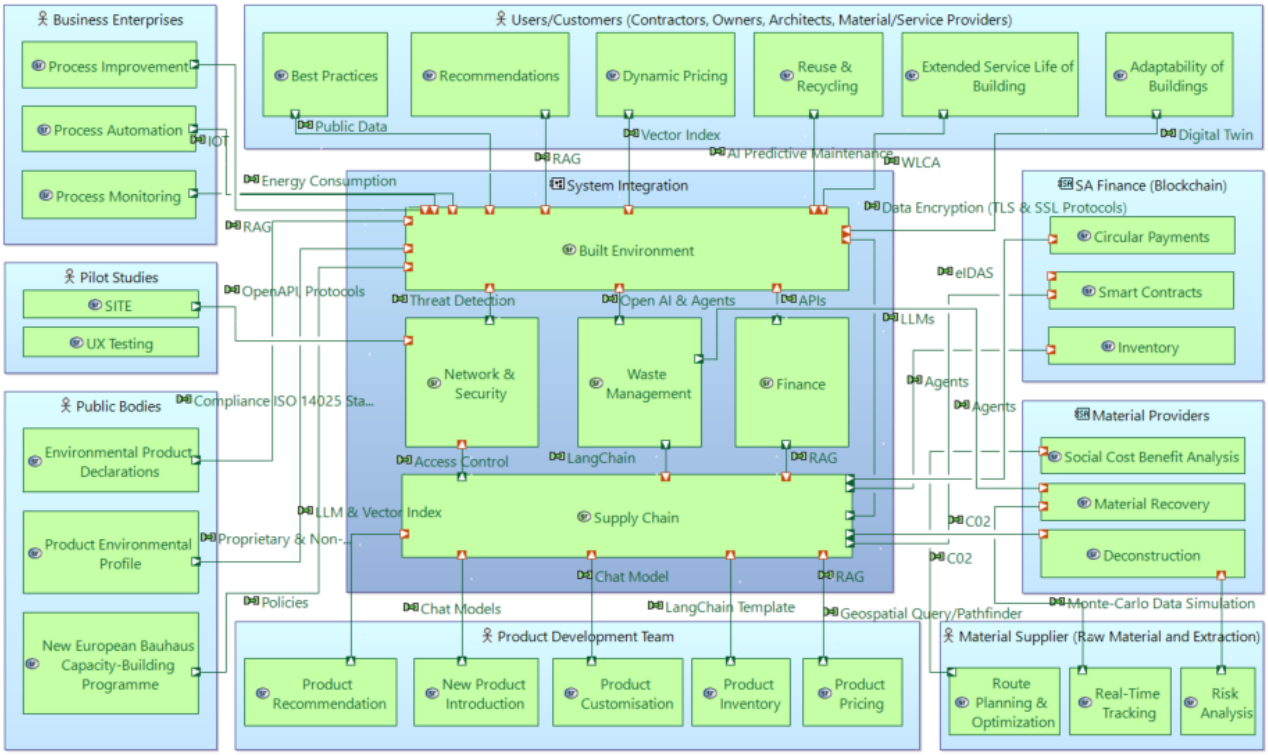


Figure 5. The Framework MBSE for Federated Enterprise Architecture.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

regional value chain approaches, in order to increase innovation buy-in from users. This initiative has driven the proposal to focus on leveraging green finance investment and creating Scale & De-Risk Accelerators for bio-based materials and products (financial simulations). The impact is to not only provide sustainability, standardization, and governance but to also deliver a shared digital connected infrastructure for supporting decision makers with real monetary opportunities 'Bankability' reflecting and helping the circular economy.

C. *The Phases*

To successfully demonstrate this concept and enable a posterior replication of the results obtained within such a project, a clear and well-defined methodology has been defined, which consist in the following 5 phases: Operational Analysis – Define Stakeholder Needs and Environment; Solution Analysis – formalize solutions requirements; logical architecture – develop solutions logical architecture; physical architecture – develop solutions physical architecture; and End Product Breakdown Structure (EPBS) – formalize solutions component requirements. The process of the 5 phases are captured visually in Figure 6 and explained in more detail via the following text.

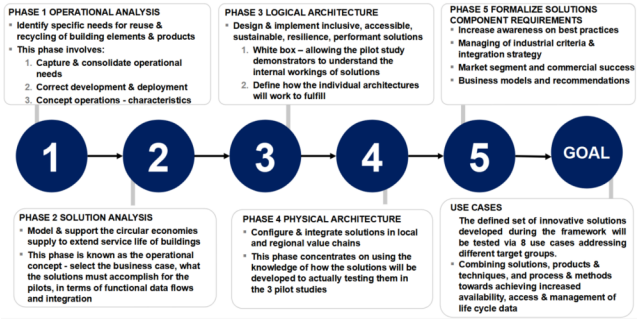


Figure 6. The Five Phases.

a) Phase 1: Identify specific needs for reuse & recycling of building elements and products (Operational Analysis) - this phase involves establishing the project's base-line while capturing and consolidating operational needs from the stakeholders for correct development and later deployment of the solutions. Moreover, it can be related to "Concept operations" where assumptions on the characteristics of the pilot studies have been made.

b) Phase 2: Model and support the circular economies supply to extend service life of buildings (Solution Analysis) - the data from P1 will contribute to P2 development of a Federated Architecture framework enabling the adoption/adaption of the existing methods and processes by identifying the boundary of the solutions, processes and consolidate the requirements. This phase can be referred to as the "operational concept" stage. The progress is more precise for individual pilots such as defining the business case, what the solutions must accomplish for the users, while also modeling functional data flows and dynamic behavior for integration approaches aligned with federated CDE and innovative tools. Whereas the previous phase investigates the overall intent of the pilot studies, OpsCon identifies what they will do (ISO/IEC/IEEE29148) [19].

c) Phase 3: Design and implement inclusive, accessible, sustainable, resilience, performant (energy, cost, etc) solutions for the built environment (Logical Architecture) - P3 represents a "white box" where models, methods are designed to be transparent, allowing the pilot study demonstrators to understand the internal workings of solutions. This phase built upon the information of P1 and the process and architecture of P2, will provide visibility into the pilot studies decision-making process, making it easier to identify the key features and rules contributing to their development. Moreover, P3 will define how the individual architectures will work to fulfill expectations such as the successful deployment of solutions to prove their capability.

d) Phase 4: Configure and integrate solutions in local and regional value chains (Physical Architecture) - P4 concentrates on using the knowledge of how the solutions will be developed to actually testing them in the 3 pilot studies. These 3 pilot studies consist of collecting real data from the field and deploying the practices to be replicated solutions, products, techniques, methods and processes based on platform assessment in the Ukraine (similar to the DAREED platform [20]). All specifications of interfaces deployment configurations, trade-off analysis of the integrated solutions is tested and evaluated.

e) Phase 5: Increase awareness on best practices for design for adaptability, reuse and deconstruction – the managing of the industrial criteria and integration strategy based on the impact outcomes of the pilot studies are assessed for market segment and commercial success. In addition, the outreach of solutions are analysed and considered to propose business models and recommendations to legislators.

D. *Use Cases*

The defined set of innovative solutions developed during the framework will be tested via 8 use cases addressing different target groups. Use cases will represent an already proven concept for the viability of combining these solutions, products & techniques, and process & methods towards achieving increased availability, access, and management of lifecycle data in the built environment. These use cases are one of the cornerstones of the project, as its conception fosters replication, bringing framework to a larger public and set of users.

Table III maps and aligns with Table II Framework: Innovative Tools, Products and Techniques, Process and Methods, including Markets where the 8 Use Cases provide a combination of the frameworks modules for selection

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

during the pilot studies, i.e., several modules can be applied to each UC.

TABLE III. SET OF INNOVATIVE SOLUTIONS DEVELOPED & TESTED

Use Cases		Innovation Tools					Products and Techniques				Process and Methods						Markets		
		1	2	3	4	5	6	P&T1	P&T2	P&T3	P&T4	P&M1	P&M2	P&M3	P&M4	P&M5	P&M6	M1	M2	M3
UC1	Circular Construction and Reuse Framework		X		X						X	X	X						X	X
UC2	Digital Platform		X		X				X					X		X	X			
UC3	Real-Time Linked Data-Space		X		X			X									X			
UC4	RAM Knowledge Platform	X					X	X					X							
UC5	Cybersecurity and Supply Chain Transparency					X			X											
UC6	LCA Models for Renovation Planning and Design		X	X	X		X	X	X	X			X	X	X		X			
UC7	Operational Carbon Management Systems	X	X		X		X	X		X				X	X		X			
UC8	Accelerator for Green Finance Investments		X		X	X		X	X		X	X						X	X	X

According to the American psychology Professor Howard Gardner of Harvard University, the brain functions using mutiple intelligences and as we are living in the Age of Intelligence, in which multiple intelligences rule, in order to perform effectively and productively in life as well as business we need to pay heed to this and develop 'intelligent strategies' [21]. Figure 7 is an all-embracing visual and graphic thinking strategy that is complementary to both Table III and chosen UC Tables V to VII. It expresses the UCs in a chronological order and harnesses the main business practices, problem solutions and strategies addressed in their descriptions. In essence it is a 'visual summary' of these Tables, created as a technique to help the thought process pertaining to the UCs.

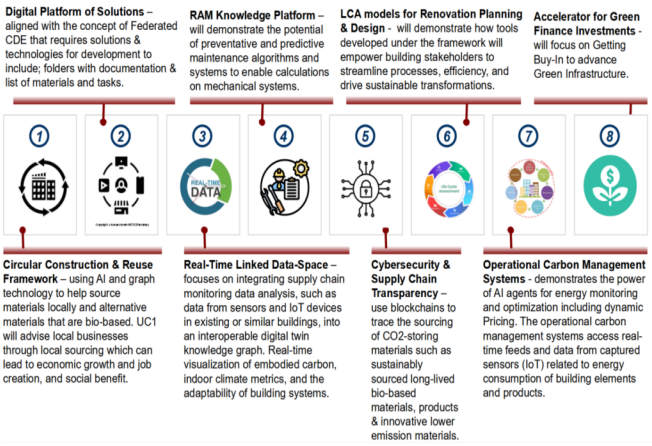


Figure 7. The proposed Solutions, Products & Techniques, and Process & Methods.

Table V provides compartmentalized descriptions of the 8 UCs aligned to Figure 7. The pilot studies will choose the most appropriate UCs to meet their require requirements. The process is similar to an inventory management system but each UC is a unified process containing the proposed Solutions, Products & Techniques, and Process & Methods where the users can bundle the attributes of several UCs in order to achieve the best results for their demonstrations.

TABLE IV. EIGHT USE CASES

A Combination of Framework Modules	
Use Cases	***Proposed Solutions, Products & Techniques, and Process & Methods***
UC1	*Circular Construction and Reuse Framework*: demonstrates different circular and sustainable building solutions to make building and infrastructure better, e.g., following circular and sustainability requirements such as design for adaptability, re-use, and durability. Using Al and graph technology (tracking) UC1 will help source materials locally that travel shorter distances, consuming less fuels and fewer carbon emissions. The preference will be to source reused materials or alternatively materials that are bio-based. UC1 will advise supporting local businesses through local sourcing which, can lead to economic growth and job creation, and social benefit. In addition, the sourcing of financing mechanisms, green insurance and micro-credit for sustainable development via Carbon Platform will provide enterprises with incentive to adapt to greener solutions connected to green finance thus providing a win-win situation.
UC2	*Digital Platform of Solutions*: aligned with CDE to improve collaboration, planning, management and automation within construction projects. The concept of Federated CDE is connected to the project's overall methodology of Federated Architecture Approach. The platform functionality for the framework requires solutions and technologies for development to include: i) folders with documentation, ii) a platform providing a list of materials and tasks, while also mapping to standards where all materials and tasks are awaiting for implementation in the early phases such as selected bills of materials and structures by assemblies. In addition, functions that can provide business intelligence reporting of sales and inventory while also connecting the quantified data with processes that have been identified as a bottleneck.
UC3	*Real-Time Linked Data-Space* [22] [23]: use case focuses on integrating supply chain monitoring data analysis, such as data from sensors and IoT devices in existing or similar buildings, into an interoperable digital twin knowledge graph. This integration supports real-time visualization of embodied carbon, indoor climate metrics, and the adaptability of building systems. The focus of this approach is on optimizing supply chain processes to align with circular economy objectives. Knowledge graphs further enhance this system by linking data on materials, supply chains, environmental performance, and stakeholder roles, thereby enabling informed decision-making. Financial tools, such as agents for circular payments, product pricing, and commercial contracts, ensure that supply chain operations align with circular economy principles.
UC4	*RAM Knowledge Platform* [24]: will demonstrate the potential of preventative and predictive maintenance algorithms and systems to enable calculations on mechanical systems incurred by wear from the moment they are activated. The RAM knowledge platform identifies the useful service life of a system, product, or service by applying real-time monitoring against the preventative and predictive models, extended by proactive, Just-in-Time (JIT) sequence of preventive and corrective maintenance actions and upgrades. The RAM knowledge platform supports information/calculations on system configuration identification elements such as existing, internally developed, reusable components that may consist of Commercial Off-the-Shelf (COTS) products, and Non-Development Items (NDP – hardware and software configuration). Such processes are aligned with early defects detection "Poka-Yoke" [25] and reduce electronic and mechanical materials sent to landfills.
UC5	*Cybersecurity and Supply Chain Transparency*: use case reflects how blockchains will trace the sourcing of CO2-storing materials such as sustainably sourced long-lived bio-based materials and products and innovative lower emission materials/aggregates. In fact each transaction or exchange of

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A Combination of Framework Modules	
Use Cases	***Proposed Solutions, Products & Techniques, and Process & Methods***
	information is recorded in a "block", which is then validated by network members before being added to the existing chain. Once validated, information becomes immutable and traceable. This technology offers unique guarantees in terms of security, transparency and traceability of exchanges, without the need for a centralized trusted intermediary. In UC5 Blockchains will create digital "product passports", containing all the information on a product's composition, manufacture and use. Furthermore according to [26] the concept of "decentralized AI" (DeAI) envisions open source, transparent AI through several blockchain technologies. Decentralized storage and distributed computing networks enhance data integrity, while smart contracts ensure transparent model access and tracking.
UC6	*LCA models for Renovation Planning and Design*: use case will demonstrate how tools developed under the framework will empower building stakeholders to streamline processes, enhance efficiency, and drive sustainable transformations. CO2 studies can be directly imported into openLCA [27] and standard LCA repositories, enabling UC6 users to assess flows and processes in their impact evaluations. UC6 will store all relevant data on a platform incorporating Digital Building Logbooks, which include Building Renovation Passports, BIM-based information, DPP, MP and GIS data. This integrated approach facilitates informed decision-making while ensuring compliance with the Corporate Sustainability Reporting Directive (CSRD). To advance beyond standard LCA principles, UC6 will adopt MBSE. The Carbon platform will generate life cycle inventories and evaluate the environmental impacts of material choices through comprehensive life cycle assessments.
UC7	*Operational Carbon Management Systems*: demonstrates the power of AI agents for energy monitoring and optimization including dynamic Pricing. The operational carbon management systems access real-time feeds and data from captured sensors (IoT) related to energy consumption of building elements and products. The introduction of AI agents and graph technology enhances IoT capabilities and across diverse energy systems. The approach applied to smart buildings is reliant on AI agents for controlling lighting and optimizing energy consumption as they are programmed to learn from their environment and improve over time. The framework will analyse patterns of electricity usage and optimize it such as turning off lights in unoccupied rooms or adjust the HVAC systems based on current occupancy, thus providing Occupant well-being. Furthermore, it will facilitate environmental sustainability by tracking energy performance, carbon emissions, and environmental impact.
UC8	*Accelerator for Green Finance Investments*: Green finance appears to be one of the leading technology solutions which, will further promote the increase in regulating environmental impact activities [28]. Therefore, the asset holders, bondholders, and issuers, among others, will have to refocus their efforts to guarantee that green finance is more useful, significant, inclusive, and environmentally protection oriented. UC8 will focus on Getting Buy-In to advance Green Infrastructure – Creative solutions for green infrastructure are only as viable as those who back them.

IV. DEMONSTRATIONS

To test these innovations and use cases, the demonstrations will cover a spectrum of many items included in the EU policy and market trends regarding data management in the built environment. This broad coverage aims to ensure that the proposed solutions offer high replication potential thanks to a demonstration plan. In this sense, the framework UCs will be tested by scenarios of different building typologies, energy grids and data architectures, via the involvement of a living lab (TRL6) and 2 large-scale pilots (TRL7-8) covering a variety of use cases and target users and 1 virtual pilot that will act as 'Development, Test and Evaluation (DT&E)' before the 2 large-pilots regarding 'Operational, Test & Evaluation' (OT&E) [29]. In addition, ISO 31000 Risk Management Plan will also be used to monitor the project progress [30].

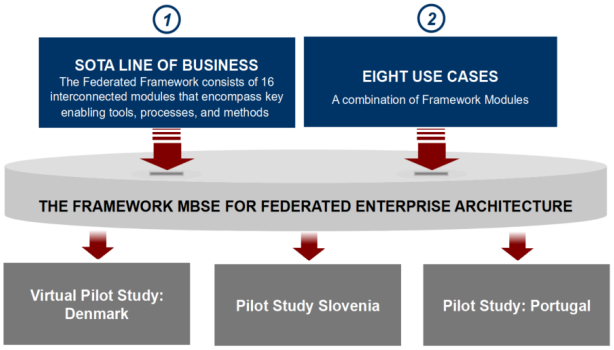


Figure 8. The Demonstrations Architecture

Figure 8 architecture illustrates the process of selecting the best technologies and UCs to be integrated & tested, including UX features, during the project's demonstrations. The three levels of interoperability are **the foundational level,** which refers to the ability of one information system to exchange data with another. The system receiving data must not interpret any information; the data must be system-agnostic. For software tools to be system agnostic, they must operate independently of any specific software environment or platform. This flexibility allows the tool to integrate seamlessly with various systems, making it adaptable to different technological landscapes. **Structural interoperability** is the second level and defines the data structures for information shared between systems. By standardising the message formatting, data is both preserved and unaltered. Data is immediately ready for use in the new system. In a **semantic model**, data is available in cloud formats (storage, platform) for any system to access, regardless of the internal systems in use [31]. Identifying standards for file formats, exchanges, and protocols will support the solutions selected for the project's demonstration. Each of the pilot studies is described in detail in the following sections. The results of the pilot studies will be a major part of providing the data source for widespread dissemination to EU-funded projects, testing facilities addressed and overall knowledge provided from the measured results.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A. *Virtual Pilot Study: Denmark & Romania*

The Danish & Romanian virtual pilot (UCN, Brisk Group & ICI) will focus on a fully digital demonstration, leveraging advanced DT technology and BIM tools to address challenges in interoperability, circular construction, and supply chain integrity. This pilot aligns with the proposals objectives by simulating circular economy principles in construction and renovation processes. The virtual demonstrator will integrate open standards (Industry Foundation Classes ISO 16739-1:2024) [32], Information Container for linked Document Delivery under ISO 21597-2:2020 [33], Cybersecurity measures (courtsey of ICI Bucharest), and lifecycle optimization methodologies to showcase adaptability, reuse, and practices focused on carbon emission reduction. The primary aim of this pilot is to establish effective systems for sourcing materials locally, enabling businesses to develop robust recycling and reuse frameworks. This approach reduces reliance on resource-intensive production processes, minimizes waste generation, and fosters sustainable practices. Local sourcing is encouraged by transparency and decision-making capabilities enabled by the platform. The Carbon platform, equipped with graph technology conforming to ISO/IEC 39075:2024, will enable precise tracking and sourcing of materials, prioritizing reused materials and bio-based alternatives.

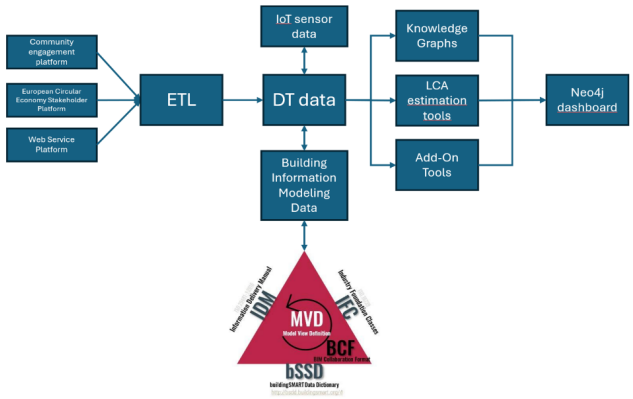


Figure 9. The Virtual Pilot

The Virtual Pilot scenario is focused on pulling data from various data bases via Extract, Transfer, Load (ETL) data processes, into a DT that also feeds from IoT and BIM data (elements of the building). The DT data is queried in real-time by Knowledge Graphs (see UC3) and LCA applied where the results are to be published on the Neo4j dashboards for analysis.

Table V identifies the five chosen UCs for the Virtual Pilot. The outcomes of this demonstration is to achieve the following key elements:

- Interoperability testing – ensuring seamless data integration using BIM, open standards, and federated architecture for decentralised Lines of Business (LOBs);
- Lifecycle optimization – simulating adaptability, disassembly, and reuse scenarios for building elements;
- Cybersecurity enhancements – securing the integrity of supply chains through advanced measures;
- Data-Driven collaboration – utilizing real-time linked data spaces and graph technology to support decision making across stakeholders;
- Real-world data – will be used from material performance and life cycle analysis will be utilized to validate the demonstrator's interoperability and accuracy.

TABLE V. VIRTUAL PILOT CHOSEN USE CASES

N°	Use Cases
UC1	Circular Construction & Reuse Framework (AI enabled tracking on a Carbon platform).
UC2	Digital Platform (interoperability via Digital Twin Integartion).
UC3	Real-Time Linked Data-Space (Environmental Data Integration).
UC5	Cybersecurity & Supply Chain Transparency (through blockchains based products passports).
UC8	Green Financing Accelerator (access to green subsidies & ESG investments).

B. *Pilot Study Slovenia*

The demonstration planned for Cirkulane (Slovenia), is a prefabricated, residential wooden house, addressing the challenges of Build4People topics, including circularity and sustainability. The demonstrator 'GORSKO' aims to focus on:

- Development of ecological modular walls designed for sound and thermal insulation.
- Reduce the carbon footprint and resources by using wood as a main structural material, including the use of waste wood, bio-based (wood, wood fibres, wood chips, straw, clay, sheep wool) insulation, also insulation based on waste textile, and focusing on providing locally produced materials and products.
- Extend the prefabrication processes: In addition to the wooden elements, there is designs to implement a BIM connection to automated production with CNC machines of insulation panels (wood fibre-based).
- Source products inventory and identification to optimize transport and installation.
- Optimisation of preparation and installation process: 3D scanning, e-site, AR/VR use for installation/supervision.
- Developing a smart building (IoT, Digital twin) and setting up smart management and maintenance as a service.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE VI. PILOT STUDY SLOVENIA

N°	Use Cases
UC1	Circular Construction & Reuse Framework
UC2	Digital Platform
UC3	Real-Time Linked Data-Space
UC5	Cybersecurity & Supply Chain Transparency
UC6	LCA models for Renovation and Planning
UC7	Operational Carbon Management Systems
UC8	Green Financing Accelerator

Table VI identifies the seven chosen UCs for the Pilot Study Slovenia. The expected outcomes planned are related to improve:

- Adapability of building design and solutions for different uses;
- Increased reuse of buildings components and increasing the end-of-life value;
- Extend services life by smart maintenance services and;
- Increase awareness and deploy best services in the demonstrators supply chain.

C. *Pilot Study: Portugal*

The Portuguese demonstration will concentrate on the Mechanical and Electrical (M&E) components that contribute to operation costs. The challenges relate to current SCADA systems that collect and manage data from a building's Command & Control infrastructure, which oversees major energy systems such as lighting, HVAC, and power distribution. These systems operate in an event-based manner, meaning that human operators manually analyses data, respond to alarms, and make real-time decisions based on detected anomalies, incidents, or diverse operational needs. While this approach allows for direct human oversight and control, it also is reactive, and carries low efficiency, leading to potential delays in optimizing energy usage and system performance.

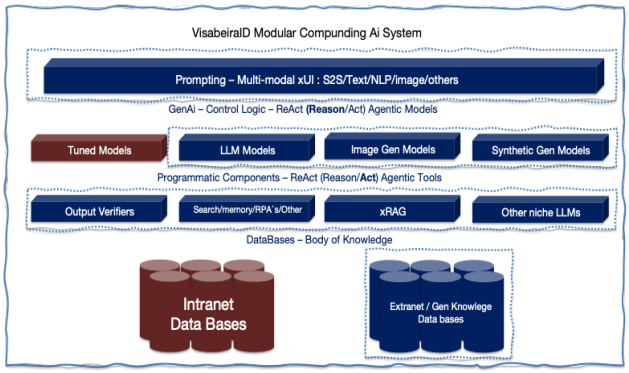


Figure 10. The 'Palacio de Gelo shopping mall' Pilot Study

The architecture of the 'Palacio de Gelo shopping mall' Pilot Study (Portugal) is illustrated in Figure 10. Moreover, the architecture reflects the selected SOTA from the framework and the chosen UCs to support the the scenario. In this To-Be scenario, new energy analytics services will be implemented, such as load forecasting, energy sourcing classifications, anomaly detection, and virtual consumption dis-aggregation to generate meaningful data & insights to be fed into the vertically embedded Agentic AI that will orchestrate building management system. The AI-driven system will utilize endogenous building information extracted from the internal active building management systems, as well as exogenous energy grid information incoming from the local energy Transmission System Operator/Distribution System Operator.

TABLE VII. PILOT STUDY PORTUGAL

N°	Use Cases
UC2	Digital Platform (Building systems' originated data, capturing real-time and historical building operations).
UC3	Real-Time Linked Data-Space (SCADA systems data used as soucring, for creating an integarted digital twin of the building).
UC4	RAM Knowledge Platform (include: I) prompting module, ii) AI ReAct (Reason/Act) Agent Models, a set of GenAI models, following the ReAct paradigm (Reasoning + Acting), will be deployed; iii) Progammatic Agentic Tool Components).
UC6	LCA models for Renovation and Planning (enhance operational processes by integrating a systemic Agentic system approach that will enable energy optimization).
UC7	Operational Carbon Management Systems (environmental impacts and GHG emissions due to large-scale electricity consumption for lighting, climate control, and other energy intensive systems).

Table VII identifies the five chosen UCs for the Pilot Study Portugal. The outcomes of The 'Palacio de Gelo shopping mall' Pilot Study focuses on the following:

- Lifecycle optimization – Agentic AI will be embedded within the building management plan;
- Data-Driven Collaboration – The Portuguese use case focuses on maintaining present user QoE KPIs while lowering the current GHG building emissions footprint;
- The Pilot Study will address circular economy goals while adapting to local/regional sourcing of products and materials to increase replication.

V. TOOLS EVALUATION AND HOLISTIC IMPACT ASSESSMENT

Neo-Eco | Circular Economy Consultancy (France) is the partner assigned to coordinate and manage the "tools evaluation and holistic impact assessment" platform for the Ukrainian test-bed. The overall concept is to pull the best practices, processes, and methods obtained from the three previous demonstrations into a beta stage dashboard. The business analytics of the dashboard will point the city officials to the frameworks best suited tech-powered solutions that can support the Dnipropetrovsk region's sustainable reconstruction needs. The beta stage dashboard will be developed to showcase how the frameworks's suite of solutions were deployed in the previous demonstrations.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

In addition, a scorecard feature will be developed to indicate each tool's level of applicability and usefulness to the selected region based on the needs inputs it provides.

The Ukraine is purposely chosen as the project's validator due to its heavy (re)construction needs and potentially large availability of recyclable rubble and materials, in addition to its energy efficiency investment needs. The result of the project validation exercise will be a diagnostic and prescriptive tool that any city government can use to (1) identify the most relevant and impactful aspects of market building and coordination it should invest in to create a more sustainable construction industry, and (2) become acquainted with the tools that can make this happen. For example; if the selected city administration completes the tool's diagnostic survey and indicates its desire to reduce its built environment's embodied and operational carbon footprint in order to unlock further international donor support, then the dashboard will highlight the solution (tools, products, technique, processes and methods) that engages with its local construction industry accordingly.

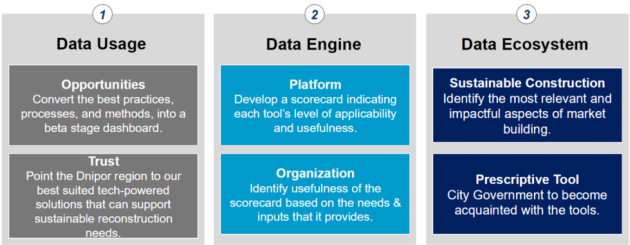


Figure 11. The Toolkit Impact Assessment Platform

Figure 11 addresses the 3 principal components of theToolkit Impact Assessment Platform : (1) Data usage, (2) Data Engine, and (3) Data Ecosystem for the planned validation of the Fedearted Framework for Adaptability, re-use and Deconstruction of Buildings, Aligned with the Principles of Circular Economy.

VI. Conclusion and Future Work

This paper focused on a Federated Enterprise Architecture approach to integrate digital technologies, via advanced processes, with novel materials into a decentralised Framework for (de)construction and renovation. The existing challenges of non-interconnected frameworks associated with monolithic, proprietary system provided the initiative to design an interoperable ecosystem for tools, platforms, and processes that maintained autonomy. The article outlines 5 objectives and their potential requirements and measurable KPIs. The methodology examines in detail the use of MBSE as a process offered in the framework to define the solution design. Furthermore, the methodology describes the evaluation of research methods and their philosophical assumptions in five phases, where each is aligned to MBSE.

The concept of the pilot studies is to facilitate the transition of the selected UCs bundle from the initial conceptual stage to an advanced operational development stage based on constant evolution and learning. The framework's adaptive process will be measurable based on the outcomes of the demonstrations. Furthermore, the anticipated results of the process will provide a long lasting impact on our understanding of how an efficient design for adaptability, re-use and deconstruction of buildings should be approached, as well as to support EU regulation on those issues. Within the context of the paper's output it has identified how to increase adaptability of buildings and the reduction of waste via the Federated Enterprise Architecture, while analyzing supporting structures such as leveraging green finance investment for local and regional economic development, and moreover providing sustainability, standardization, and governance towards the circular economy policy evolution.

Acknowledgment

The author would like to take this opportunity to acknowledge the contribution of F6S (Ireland), InoSens (Serbia), VISABEIRA (Portugal), UCN (Denmark), SGG (Slovenia), ICI and Brisk Group (Romania) & Neo-Eco (France) to the Call: HORIZON-CL5-2024-D4-02 (Efficient, sustainable and inclusive energy use) for which, this article has been based on.

References

[1] A. Redmond "A Framework for Adaptability, re-use and Deconstruction of Buildings, Aligned with the Principles of Circular Economy", Modern Systems 2025, https://www.thinkmind.org/library/MODERN_SYSTEMS/MODERN_SYSTEMS_2025/modern_systems_2025_1_10_10010.html, 26 October 2025.

[2] M. Niemeyer and N. Kozub, "New European Bauhaus Circular Housing in Ukraine (Project B)", New European Bauhaus, beautiful | sustainable | together, https://new-european-bauhaus.europa.eu/system/files/2023-02/221207_NEB_circular_housing.pdf, February 2023.

[3] European Environment Agency, "Building and construction", https://www.eea.europa.eu/en/topics/in-depth/buildings-and-construction?activeTab=07e50b68-8bf2-4641-ba6b-eda1afd544be, modified 20 February 2025

[4] Z. Quist, "Life Cycle Assessment (LCA) – Everything you need to know", Ecochain, https://ecochain.com/blog/life-cycle-assessment-lca-guide/ 03 February 2024.

[5] European Commission: Directorate-General for Research and Innovation, Bioeconomy – The European way to use our natural resources – Action plan 2018, Publications Office, 2018, https://data.europa.eu/doi/10.2777/79401

[6] World Economic Forum, Emerging Technologies, "What is the bioeconomy and how can it drive sustainable development?", https://www.weforum.org/stories/2024/07/bioeconomy-sustainable-development/, 12 July 2024

[7] European Construction, Built Environment and Energy Efficient Building Technology Platform, 'Innovative Built Environment', and Europe Regional Network, 'World Green Building Council', "Built4People Partnership – Strategic Research & Innovation Agenda 2021-2027, European Partnership under Horizon Europe, Built4People –SRIA/Roadmap", https://built4people.eu/wp-content/uploads/2023/10/Built4People-Partnership-SRIA-March-2022.pdf, March 2022.

[8] C. Singam and J. Carter, Model-Based Systems Engineering (MBSE), Guide to the Systems Engineering Body of

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Knowledge (SEBoK), Systems Engineering and Management, https://sebokwiki.org/wiki/Model-Based_Systems_Engineering_(MBSE), accessed 08 May 2026.

[9] A.W. Wymore, Model-Based Systems Engineering (1st ed.). CRC Press. https://doi.org/10.1201/9780203746936, 1993

[10] C. Boudjennah, B. Combemale, D. Exertier, S. Lacrampe1, and M.A Peraldi-Frati, CLARITY: Open-Sourcing the Model-Based Systems Engineering Solution Capella, HAL Id: hal-01186019 https://inria.hal.science/hal-01186019v2, 24 September 2025.

[11] ARCADIA, "A Tool Method To Define, Analyse, Design & Validate System, Software, Hardware Architectures", eclipse.org/capella, https://mbse-capella.org/arcadia.html# [accessed 08 May 2026].

[12] Chief Information Officer Council Version 1.0 February 2001, A practical Guide to CIO Council Federal Enterprise Architecture, pp. 25 - 26, https://www.gao.gov/assets/588407 .pdf.

[13] DoD, Implementing a Modular Open Systems Approach in Department of Defense Programs, Office of Systems Engineering and Architecture Office of the Under Secretary of Defense for Research and Engineering pp. 9 – 11, https://www.cto.mil/wp-content/uploads/2025/03/MOSA-Implementation-Guidebook-27Feb2025-Cleared.pdf, February 2025

[14] J. Chen, "Growth and Income Fund Definition, Investment Mix, Examples", Investopedia (part of the Dotdash Meredith publishing family), https://www.investopedia.com/terms/g/growth-income-fund.asp, last updated 29 April 2026.

[15] FINRA, "Feeder Funds and Retail Investors, Alternative and Emerging Products", https://www.finra.org/investors/insights/feeder-funds#:~:text =A%20feeder%20fund%20is%20an,the%20assets%20and%20directs%20investments, 20 January 2026.

[16] T. Segal, "One-Stop Shop: Definition, Example, History, Pros and Cons", Investopedia (part of the Dotdash Meredith publishing family), https://www.investopedia.com/terms/o/onestopshop.asp, last updated 26 June 2024.

[17] ISO/IEC 39075:2024, Information Technology – Database Language – GQL, Technical Committee: ISO/IEC JTC 1/SC 32 & ICS: 35.060, https://www.iso.org/standard/76120.html, Published (Edition 1, March 2024).

[18] ISO 14025:2006, Environmental Labels and Declarations – Type III environmental declarations – Principles and procedures, Technical Committee: ISO/TC 207/SC 3 & ICS: 13.020.50, https://www.iso.org/standard/38131.html, Published, July 2006.

[19] ISO/IEC/IEEE 29148:2018, Systems and Software Engineering – Life Cycle Processes – Requirements Engineering, Technical Committee: ISO/IEC JTC 1/SC 7 & ICS: 35.080, https://www.iso.org/standard/72089.html, Published November 2018.

[20] H. Wicaksono, B. Yuce, K. McGlinn, and O. Calli, "Linked Data for Smart Energy System", 'Buildings and Semantics, Data Models and Web Technologies for the Built Environment', edited by P. Pauwels and K. McGlinn, Chapter 2. Smart Cities and Buildings", pp.254 – 256, doi: 10.1201/9781003204381, 2022.

[21] T. Buzan, C. Griffiths, and J. Harrison, "Mind Maps for Business, Using the Utimate Thinking Tool to Revolutionise How you Work", 2nd Edition, ISBN: 978-0-273-78435-7, Published 2014.

[22] M. Page and G. Cecconi, "European Data Spaces and the Role of data.europa.eu", Publications Office of the European Union; 2023, available from: doi/10.2830/1603.

[23] E. Curry, "Real-time Linked Dataspaces, Enabling Data Ecosystems for Intelligent Systems", Chapter 1. Real-time Linked Dataspaces: A Data Platform for Intelligent Systems within Internet of Things-Based Smart Environments, pp. 3 – 14, doi: 10.1007/978-3-030-29665-0, published: 18 November 2019.

[24] C. S. Wasson, "System Engineering, Analysis, Design, and Development": Conept, Principles, and Practices, 2nd Edition, Wiley Series in Systems Engineering and Management, A.P. Sage, Series Editor, Chapter 34. System Reliability, Maintainability, and Availability (RMA) pp. 721 – 791, ISBN 978-1-118-44226-5, published December 2015.

[25] H. Obinson, "Case study: Poka Yoke tool in Bohemian Brew Peru, a small business", editor Using Poka-Yoke techniques for early defect detection. Sixth International Conference on Software Testing Analysis and Review; (STAR 1997); Available from: https://www.researchgate.net/publication/373397768_Case_study_Poka_Yoke_tool_in_Bohemian_Brew_Peru_a_small_business[accessed 08 Mai 2025].

[26] M. Li, "Watch decentralized AI in 2025: The convergence of AI and crypto", Forbes, Nov. 12, 2024. Available: https://www.forbes.com/sites/digital-assets/2024/11/12/watch decentralized-ai-in-2025-the-convergence-of-ai-and-crypto, cited in N. Kshetri, "Building Trust in AI: How Blockchain Enhances Data Integrity, Security, and Privacy, Digital Object Identifier 10.1109/MC.2024.3505012, Date of current version: 29 January 2025.

[27] A. Ciroth, J.P. Theret, M. Fliegner, M. Srocka1, V. Bläsig, and Ö. Duyan, "Integrating Life Cycle Assessment tools and information with Product Life Cycle Management / Product Data Management", editor G. Seliger, Proceedings of the 11th Global Conference on Sustainable Manufacturing - Innovative Solutions ISBN 978-3-7983-2609-5, Universitätsverlag der TU Berlin 2013

[28] HM Government, "Greening Finance: A Roadmap to Sustainable Investing", published 18 October 2021, https://www.gov.uk/government/publications/greening-finance-a-roadmap-to-sustainable-investing accessed 08 May 2025.

[29] Defense Acquisition University, 'Developmental Test and Evaluation (DT&E)', AETM 011, DAU Glossary Definition, https://www.dau.edu/acquipedia-article/developmental-test-and-evaluation-dte# accessed 08 May 2025.

[30] ISO 31000:2018, Risk Management – Guidelines, edition 2, International Standard to be Revised [90.92], Technical Committee: ISO/TC 262, ICS: 03.100.01, https://www.iso.org/standard/65694.html#lifecycle, Published February 2018.

[31] Becker's HEALTH IT, "The 3 Levels of Interoperability in Healthcare", https://www.beckershospitalreview.com/health care-information-technology/the-3-levels-of-interoperability-in-healthcare/, 03 June 2015

[32] ISO 16739-1:2024, Industry Foundation Classes (IFC) for Data Sharing in the Construction and Facility Management Industries, Part 1: Data Schema, Technical Committee: ISO/TC 59/SC 13 & ICS : 25.040.40, https://www.iso.org/standard/84123.html, Published March 2024.

[33] ISO 21597-2:2020, Industry Container for Linked Document Delivery – Exchange Specification, Part 2: Linked Types, Technical Committee: *ISO/TC 59/SC 13 &* ICS : 35.240.6791.010.01, https://www.iso.org/standard/74390.html Published November 2020.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Assessment of Sensor Technologies for Detecting Slurry Emissions in Irish Agriculture

Jacqueline Humphries, Abimbola Y. Ikoyi
Technological University of the Shannon (Midwest)
Thurles, Co. Tipperary, Ireland
Email: jacqueline.humphries@tus.ie; abimbola.ikoyi@tus.ie

***Abstract* - There is a rising demand for the reduction of ammonia and greenhouse gases in Ireland to meet European Union targets for climate neutrality by 2050. This has led to considerable interest in the use of various technologies, particularly sensors to aid in mitigation of emissions in slurry life cycle management. In our previous study, a review of publications on the general use of sensors for the detection and determination of ammonia and greenhouse gases was conducted. This extended version builds on the previous findings through validation using responses from Irish farmers. A questionnaire on the use of sensors for slurry management was administered to Irish farmers, and a total of 106 responses were collated. The findings revealed that 22.6% of farmers in this study have adopted sensor technology on farms, and only ~2% utilize it for emission detection. Despite the limited adoption of sensor-integrated technology for the detection of slurry emissions on Irish farms, farmers agree that sensor technologies would be beneficial for slurry management. Some perceived benefits are real-time monitoring and decision support, enhanced detection capabilities, improved emission quantification, spatial and temporal precision, and support for sustainable practices. However, there are currently limitations in their adoption and application, such as high cost and maintenance, limited technical knowledge, inadequate sensor sensitivity and accuracy, operational constraints, limited capability for slurry characterization, and fragmented design. This study revealed a poor correlation between pairs of perceived benefits: real-time monitoring of emissions and emission quantification, and improved detection capabilities and better decision-making. Thus, suggesting that farmer perception of sensor-driven emission detection might be regulatory or compliance-based rather than informing better farm decisions. Therefore, there is a need for a clearer communication strategy that highlights the importance of improved decision-making through sensor-based real-time emission detection. Additionally, this study underscores the importance of subsidies to help reduce costs and encourage the adoption of sensor technologies on Irish farms, supporting the reduction of emissions in slurry management and promoting sustainable agriculture.**

Keywords - Sensors; Precision Agriculture; Ammonia; Greenhouse Gas; Emissions; AgriTech.

I. INTRODUCTION

A preliminary version of this study was presented at the International Conference on Sustainable and Regenerative Farming, 2025 [1]. While the conference paper was limited to a review of articles on the use of sensors for the mitigation of ammonia and greenhouse gases through the slurry life cycle, with a summary of the benefits and challenges of sensor usage, this current study is a substantial extension of those findings. Sections I and II of this current study were adapted from the conference paper.

Irish agriculture is mainly dominated by livestock production, resulting in the generation of a large quantity of animal waste [2]. While livestock slurry is a valuable agricultural resource, its mismanagement poses significant environmental challenges. Slurry contains valuable nutrients like nitrogen and phosphorus, but improper management can lead to significant losses through runoff, leaching, and volatilization. This can cause water pollution (e.g., eutrophication) and air pollution (e.g., ammonia emissions). A recent publication from the Agriculture and Food Development Authority of Ireland (Teagasc) estimates that 40 million tonnes of animal manure is generated and subsequently used annually [2]. It follows therefore, that there is immense pressure on the agricultural sector in Ireland to minimize Ammonia (NH_3) and Greenhouse Gas (GHG) emissions [3][4]. This sector accounts for the majority of Irish national NH_3 (99%) and GHG (37.8%) emissions [3][5][6]. Methane (CH_4) emissions from slurry management represent 10.6% of agricultural GHG in Ireland [4]. It follows therefore, that minimization of Irish national NH_3 and GHG emissions, especially from agriculture, is crucial in meeting emission reduction targets and ensuring the sustainability of the agricultural sector.

Efforts to reduce emissions occur within the many processes involved in the management of slurry, such as removal and storage management, treatment adjustments, slurry application rates, soil uptake, and so on. However, these are not without challenges. For example, the storage of slurry is accompanied by the release of pollutant gases, such as NH_3 and CH_4 emissions [3][7]. In an attempt to minimize this, Ireland developed the Nitrates Action Plan (NAP) in line with the European Union (EU) Nitrates Directives, which specifies guidelines on slurry storage duration and time of application [8]. In addition, Ireland Climate Action Plan (CAP) outlines measures specifically for the agricultural sector with focus on 25% reduction in emissions by 2030, with key target being

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

slurry management to reduce ammonia emissions [4]. Several manure management approaches have been proposed with the possibility of reducing these dangerous gases associated with slurry management. [9] found that the use of additives, which encourage acidification, reduces CH_4 and NH_3 emissions from slurry storage. Guidance from the United Nations Economic Commission for Europe (UNECE) Task Force on Reactive Nitrogen: Ammonia Guidance Document [10] sets out emission abatement measures in the nitrogen lifecycle from livestock feeding strategies, animal housing techniques, manure storage techniques, through manure application techniques. Also, research conducted by [11] in which the impact, potential, and costs associated with abating national NH_3 emissions up to 2030 likewise sets out common mitigation strategies.

Since the UNECE and Teagasc guidance documents [10] [11] were published, there have been exponential advancements in technology. Sensor technology enables the Internet of Things (IoT). Big data is gathered from sensors, hosted on cloud platforms, and analysed using statistical methods or artificial intelligence to enable real-time predictions - driving the Industrial Revolution known as Industry 4.0 [12]. Agriculture 4.0, using the nomenclature of Industry 4.0, promises the same revolution in smart farming. Indeed, many industry consortia, fora and solution providers propose slurry management solutions, which use sensors, and make claims that emissions are reduced [13 - 16]. A rigorous journal review process is necessary to substantiate claims and conclusions made in these channels [1, 17]. Furthermore, whilst the adoption rate of sensor technologies is known to be particularly low in emerging and developing countries [18, 15], it is not clear to what extent those sensors have been adopted in high-income EU contexts such as Ireland, which is primarily comprised of relatively small family-run farms.

In this research, the application of advanced sensor technologies for real-time monitoring and control of emissions in slurry management processes in Irish farms was investigated. The research questions posed are (1) How are sensors used in Irish farms? (2) To what extent are sensors used for the reduction or mitigation of ammonia and greenhouse gas emissions in slurry management on Irish farms? (3) What are the perceived benefits and perceived challenges associated with the use of sensor technologies for slurry management?

To answer these research questions, first a narrative literature review was conducted on sensor utilization for detection of ammonia and greenhouse gases in slurry management, which is set out in Section II. Thereafter, the research method in which Irish Farmers were surveyed is set out in Section III. In Section IV and Section V, the results and discussion on the same are presented. The paper concludes with a summary, limitations, and proposed future research.

II. Sensor Utilisation in the Slurry Lifecycle

The emergence of smart farming and precision agriculture is due to advancements in technology. There has been an increase in the applications of such technologies for sustainable agriculture, and an emerging area is the mitigation of emissions in agriculture. An example is the use of IoT technology for the improvement of slurry management on farms. These field-based IoT sensors record and monitor soil-related conditions targeted at helping farmers make better decisions on best timing for slurry applications to minimize losses and maximize nutrient use. However, these sensors were unable to measure key slurry parameters (such as pH, dry matter, temperature, and nutrient content), perform in situ and online monitoring, or provide data for comprehensive slurry management [19].

Several authors [20 - 25] report on the application of sensors for determination of nutrient components of slurry. However, few reports have been published on the use of sensors for the quantification of gas emissions, such as ammonia and greenhouse gases (methane, nitrous oxide and CO_2). This literature review addresses that gap and covers the quantification of emissions in the three major stages in the traditional management of slurry: slurry production in animal houses, slurry storage and field application.

A. Slurry Production

Livestock production results in the generation of animal waste. Particularly, housing of animals comes with the challenge of handling and management of slurry. Efficient manure management reduces environmental impact, thus maintaining animal health. Environmental sensors measuring factors like air quality and humidity, generate vast amounts of data providing crucial insights into the well-being of the herd and the optimization of the farm environment [12].

Air quality in farmhouses is linked with ammonia, CO_2, Particulate Matter (PM) and Hydrogen Sulphide (H_2S) concentrations. These gases have negative effects on animals and human health in the environment. The quality of air is affected by some other factors, such as frequency of slurry removal and floor type [26]. A 21-day study which utilized an IoT gas and environmental sensors for continuous detection of NH_3, CO_2, H_2S and PM concentrations in two piggeries revealed that housing structures and slurry management systems had a huge impact on the gas emissions in the piggeries. Specifically, slurry management resulted in increased H_2S up to 1.9 ppm and increased NH_3 concentration of 63%. In addition, the structure of housings resulted in accumulation of gases, CO_2 and NH_3 increasing up to 52% and 34% than daily average value respectively [26]. The use of sensors at different times of the day, further confirms the need for advanced technology for the mitigation of environmental impacts of agriculture.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Optimum environmental conditions (temperature, moisture, air quality, etc.) must be maintained in livestock houses. The maintenance of these conditions results in huge electrical energy consumption particularly in poultry houses (broiler houses - 75.5%, laying hen houses - 58.9%) due to the use of various equipment [27]. This is predicted to increase in the future due to technological advancement, which indirectly leads to increased GHG emissions. Consequently, for improved efficiency and sustainability, the prediction of the energy consumption of the indoor environmental condition for intensive poultry farming is expedient [28].

In order to minimize reliance on additional equipment, [28] developed a customized hourly model for the interpretation and analysis of electronically collected data. In this study, gas sensors were utilised for the measurement of CO_2 (Model 336, Huakong Xingye Technology, Beijing, China) and ammonia gas concentration (Model 458, Zhize, Jinan, China) emitted in a poultry house. The average CO_2 and ammonia concentration detected by the sensors were similar to the average predicted data using the developed model [28]. On the other hand, there is need for improvement in the sensitivity levels for the gas sensors to enable accurate detection at extremely low concentrations.

As indicated previously, NH_3 is typically an odorous compound produced from the decomposition of organic nitrogen and is a precursor of secondary inorganic aerosols. Similarly, H_2S, a strong odorous and toxic compound that affects animal and human health, is mainly produced from anaerobic digestion of organic sulphur [29]. These gases are usually at high concentrations in animal houses. A study evaluated the use of Electrochemical (EC) gas sensors for the quantification of odours from ammonia (Model #SO1198 Senko LTD. Korea) and hydrogen sulphide (Model #SO1N8 Senko LTD Korea) in a piggeries' manure treatment facility. Acceptable values were obtained for linearity, accuracy, repeatability, lowest detection limit and response time for the sensors, thus confirming their suitability for on-field testing. However, a longer sampling time of at least 15 minutes might be necessary for ammonia monitoring to reach target concentration point [29].

B. *Slurry Storage*

Upon excretion of faeces in animal houses, the slurry (manure) is usually stored for a specific amount of time. Sensor networks that monitor real-time changes in ammonia concentrations assist in minimizing losses of plant-available nitrogen during manure storage [30]. The duration of storage varies and is affected by several factors, such as time of the year, regulation governing spreading as organic fertilizer, farm slurry storage capacity and so on. Sensors were used in a study for the development of a prediction model for methane and ammonia gas emissions in piggeries with two different types of manure management systems: Long Storage (LS) in deep pits and Short Storage (SS) by daily flushing of a shallow pit with sloped walls and partial manure dilution [31]. The study revealed a positive correlation between calculated and measured CH_4 and NH_3 emissions on an annual basis. This confirms the reduction potential of the studied measures for CH_4 and NH3 emissions from pig houses. In addition, the developed model provides a possibility for the assessment of mitigation measures on CH_4 and NH_3 emissions. This provides a robust basis for assessing the impact of management and housing strategies on CH_4 and NH_3 emissions from pig houses, which in turn, helps support more sustainable practices in pig farming [31].

In a similar study, manure management and sensor location played a huge role in the determination of gas concentration [32]. Higher ammonia concentration was recorded for open slurry pit compared to the slurry management system with daily removal of slurry. Meanwhile, electro-chemical DOL53 ammonia sensors (DOL Sensors, Denmark) located at 1.0m above floor level recorded approximated ammonia concentrations and were more vulnerable to local fluctuations in comparison to those located at 1.8 m above floor level [32].

In contrast to the previous studies where electro-chemical sensors were used for gas concentration determination, a Fourier-Transform Infrared (FTIR) spectroscopy monitor was used to measure gas transport and concentrations of greenhouses gases [methane, carbon dioxide, and nitrous oxide] and ammonia inside manure piles at various depths. Results showed that carbon dioxide dominated the greenhouse gas emissions. An interesting observation in this study was the reduction of gas emissions with increased moisture content in manure with high water holding capacity [33]. Results obtained using FTIR Spectroscopy monitor provided insights into management strategies for emission reduction from solid dairy manure [33].

Drones are used as platforms to carry and deploy sensors, such as RGB cameras, multispectral, hyperspectral, and thermal sensors for aerial imaging and mapping, multispectral or LiDAR sensors for soil and field analysis, and gas sensors (e.g., methane, ammonia), infrared or laser-based detectors sensors to detect and map emission. Drones are effective in counting animal populations and detecting methane leaks in natural gas infrastructure. These techniques have been applied on a small scale to assess and determine livestock-related methane emissions on farms [34].

Electrochemical sensors were found to have several advantages, such as multi-gas non-specific detection, high sensitivity and precision, making them the preferred alternative for emission detection, albeit they have a long response time and short service life. Similarly, FTIR

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

spectroscopy has the advantage of multi-gas non-specific detection but has higher operating cost in comparison with electrochemical sensors [34].

A UAV-based active AirCore system for the estimation of CH_4 emissions from dairy cow farms is outlined in [30]. The inclusion of local wind speed and direction measurement would result in increased accuracy of methane estimation [30]. In addition, there is need for further research in the use of aerial technology for the assessment of emissions from livestock farming.

C. Field Application of Slurry

The application of fertilizers and manure on fields is the largest source of NH_3 in the atmosphere. Ammonia emission from agriculture has negative environmental consequences and is largely controlled by the chemical microenvironment and the respective biological activity of the soil [12]. While gas phase and bulk measurements can describe the emission on a large scale, those measurements fail to unravel the local processes and spatial heterogeneity at the soil-air interface [12].

For better understanding of some of these processes, a two-dimensional (2D) imaging approach which visualized three of the most important chemical parameters associated with NH_3 emission from soil was developed by [35]. Ammonia, O_2 and pH microenvironments were imaged using reversible optodes in real-time with a spatial resolution of <100µm. This NH_3 optode enhanced the understanding of microscale factors influencing NH_3 emissions, allowing for visualization of the soil's chemical microenvironment following manure application [12]. Though there is a surge in the incorporation of precision agriculture tools, these systems often operate in isolation, focusing on specific parameters without providing a holistic view of the agricultural environment [36]. There is a need to bridge this gap by integrating multiple sensors and data sources into a unified monitoring system.

In [36] a comprehensive monitoring system using sensors was developed for the measurement of gases, such as CO_2, methane, and ammonia. This system, known as Agri-Guard, consists of two sets of devices: the IoT based Agri-cones and a centralized camera stand. The Agri-cones consisted of an array of sensors including temperature, humidity, moisture, CO_2 and methane gas sensors. Upon manure application to the soil, substantial increase in sensor readings were observed in the CO_2 and methane gas sensor (MQ9), due to the organic matter decomposition in the manure. Similarly, as microbial decay progressed, the ammonia sensor (MQ135), showed a slight increase, signifying the breakdown of organic nitrogen compounds in the manure [36].

The applications of sensors in slurry management are outlined in Table 1, covering housing, storage, and field use. Their advantages and challenges are summarized in Table 2, highlighting benefits for monitoring and quantification alongside limitations related to sensitivity, cost, and system integration. These advantages and challenges are evaluated from the farmer's perspective in the survey described in the following section.

TABLE 1. SUMMARY OF APPLICATION OF SENSORS FOR THE MITIGATION OF EMISSIONS IN SLURRY MANAGEMENT [1]

	Purpose of Study	***Sensor***	***Monitored animal/slurry source***
1	Evaluation of slurry management in two different housing structures	Environmental Sensor	Pigs
2	Development of energy consumption model for animal houses	Gas Sensors	Pigs
3	Emission monitoring and odour intensity estimation	Electrochemical Sensor	Pigs
4	Development of prediction models for emissions from various slurry storage systems	Gas Sensors	Pigs
5	Effect of manure management and sensor location on emission concentration	Electrochemical	Piggeries
6	Evaluation of compaction effects on emissions from dairy manure	FTIR	Cattle
7	Estimation of emissions from dairy cows' manure	UAV	Cattle
8	Visualization of emissions from soil upon manure application	Optical sensors	Livestock (unspecified)
9	Monitoring of gaseous emissions from manure in farms	Gas sensors	Livestock (Unspecified)

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE 2. ADVANTAGES AND DISAVANTAGES OF SENSORS FOR EMISSION REDUCTION IN SLURRY MANAGEMENT [1]

	Advantages/Opportunities	Disadvantages/Challenges
1	Real time monitoring and decision support [26][36]	Limited capabilities for slurry characterization [19]
2	Enhanced detection capabilities [33][28]	Variation in sensor sensitivity and accuracy [28][29]
3	Improved emission quantification [31]	Operational constraint [32][34]
4	Spatial temporal precisions [12][38]	High cost and maintenance [33]
5	Support and sustainable practices [39]	Fragmented system design [36]

III. MATERIALS & METHOD

According to the 2020 Census of Agriculture, there were just over 135,000 farms in Ireland, of which 75% (approximately 101,250 farms) were classified as cattle rearing, dairy, or other cattle enterprises [37]. These farms constituted the target population for this study. A questionnaire survey was employed as the primary data collection method. The questionnaire served as a structured data collection instrument and was designed to systematically gather information from a sample of participants in order to understand attitudes, opinions, behaviours, and characteristics of the target population.

The questionnaire was designed to capture both quantitative and qualitative data, comprising of a mix of closed-ended questions (including check boxes, Likert-scale ratings, and multiple-choice items) as well as open-ended questions to allow for elaboration on individual experiences. The mixed-question format allowed for both statistical analysis and contextual understanding of farmer perspectives, aligning with best practices in agricultural technology adoption studies [40][41]. The study was not designed to test or operationalize established behavioural adoption models; rather, it aimed to provide a descriptive mapping of usage patterns and perceived advantages and disadvantages. Accordingly, the survey instrument was developed to align with this objective.

The questions were divided into sections covering demographic information (such as age, farm size, location, number of animals, and principal farm activity) and technology-related items, which explored current use of sensor technologies on the farm. Further questions that align with the perceived benefits of the use of sensors, and challenges to adoption as found in the literature review were included for the purposes of validation.

Due to practical constraints, a non-probability sampling approach was employed, and the findings should therefore be interpreted as indicative rather than statistically representative of the national population. Irish farmers were recruited through a range of agricultural networks, including Teagasc advisory services, farmer database contacts associated with the AgriNext and SoilCrates projects, livestock marts, farming networks, social media platforms (e.g., LinkedIn), and third-level Agricultural Science students. Responses were collected anonymously, and participation was voluntary. Data were entered in Microsoft Excel spreadsheet for initial data cleaning and categorization.

With regard to qualitative data, a mix of inductive–deductive coding approaches were used. Initially, deductive coding was conducted using categories informed by the literature review and research objectives. While inductive coding was used to identify emergent concepts and patterns arising directly from respondents' experiences and perspectives. This ensured that various groups and sub-groups generated were data-driven rather than from predetermined themes, thus enabling the identification of novel insight arising from the data.

With regard to quantitative data, exploratory data analysis, which included statistical analysis of the data, examination of descriptive patterns and characteristics, and visualization through frequency distributions and graphical summaries, were performed in Microsoft Excel. Measures of central tendency, including the mean, median, and mode (where appropriate), were calculated to describe typical values, while measures of dispersion such as the range, mean deviation, and semi-interquartile range (SIQR) were used to assess variability within the sample. Data distributions were examined using graphical methods, including histograms, frequency polygons and bar charts, to identify skewness and the presence of outliers. Categorical variables were summarized using frequency counts and percentages; where variables allowed multiple responses, frequencies were reported, and percentages were interpreted accordingly.

Following the Exploratory Data Analysis, associations between binary-coded response themes were examined using Jaccard similarity coefficients and Phi (Φ) coefficients. Jaccard coefficients were used to examine proportional overlap and co-occurrence between selected response themes, and Phi coefficients were calculated to describe the strength of association between the binary-coded variables.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Responses to the multiple-selection question were coded into binary variables (1 = selected, 0 = not selected). Pairwise co-occurrence frequencies were calculated by counting the number of respondents selecting each pair of benefits or challenges, resulting in a co-occurrence matrix. To examine the co-occurrence patterns between perceived challenges/disadvantages and benefits/advantages while accounting for differences in overall selection frequency, Jaccard similarity coefficients were calculated for all pairs of coded themes. The Jaccard coefficient values reflect proportional similarity and co-occurrence between coded response themes rather than statistical dependence or causality between the options selected by the farmers. The resulting coefficients, ranging from 0 (no co-occurrence) to 1 (complete overlap), are illustrated in heatmaps. Generally, the similarity coefficient thresholds between two coded options, and chosen *a priori*, can be interpreted as follows: A value of < 0.20 represents a weak association, a value between 0.20 - 0.40 represents a moderate association, a value between 0.40 - 0.60 represents a strong association, and values above 0.60 represent very strong associations.

In addition to Jaccard similarity coefficients, Phi (Φ) coefficients were calculated as descriptive measures of association for pairwise comparisons of binary-coded variables derived from 2×2 contingency tables. Phi coefficients range from −1 to +1, where values of 0 represent no association, and values approaching ±1 represent stronger inverse or positive associations. For interpretive purposes, coefficient thresholds of approximately ±0.10, ±0.30, and ±0.50 were considered indicative of small, medium, and large associations, respectively, and were chosen *a priori*.

In the analysis, responses selecting 'all of the above' were treated as distinct categories and were not disaggregated into individual response options to avoid overestimation or interpretive bias, and which would have artificially increased joint occurrences and biased Jaccard and Phi estimates upwards.

In the next section, the results are set out.

IV. RESULTS

An exploratory examination of farmers' practices and views is presented in this section. The findings are intended to identify patterns and relationships within the data rather than to provide definitive population-level inference. A total of one hundred and six (n=106) valid responses were collected from participants across the target population. This sample size is in line with, and exceeds, related peer-reviewed studies where questionnaires were used: n=56 [42], n=53 [40], and n=32 [41]. Thus, the achieved sample size is sufficient for exploratory and descriptive analysis of the survey responses.

A. *Respondents and their Farm Enterprises*

The geographical distribution of participants was spread across all four provinces of Ireland, Munster, Connaught, Leinster and Ulster, and from twenty different counties of the possible twenty-six in the Republic of Ireland.

The farmers were from a wide range of age groups, see Fig. 1, with a larger frequency from younger farmers. The use of an online questionnaire partly explains that distribution, as younger people are more likely to be able to access online forms. In fact that was borne out by anecdotal evidence, when one lady was asked would she ask her father to complete the form, she said "I won't ask Dad, he's in his late 70s and wouldn't be able, but I'll ask my brother who mainly looks after the farm anyhow". This joint farm enterprise illustrated in this response is popular in Ireland, when older farmers tend to be supported by younger members of the family in multi-generational farms, and further explains the age-distribution in responses. The sample comprised 75.5% male and 24.5% female respondents, with no participants identifying as non-binary or other.

With respect to Farm Size, respondents had farms within the range of 16 to 1700 acres, indicating substantial heterogeneity among respondents. The farm size distribution is right-skewed and is illustrated in Fig. 2. This heterogeneity arises because the dataset includes both small-scale and large-scale farming operations, which differ considerably in scale, production capacity, and management practices. As a result, measures of central tendency alone may not adequately describe the data, and measures of dispersion such as the range, interquartile range, and mean deviation are required to capture the variability within the sample.

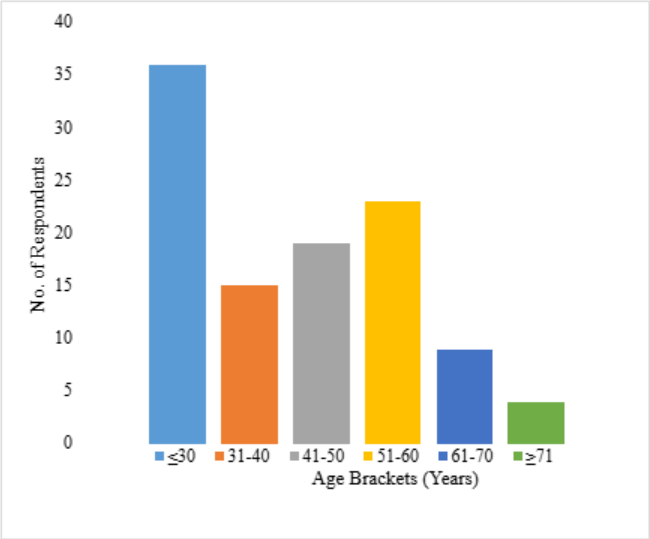


Figure 1. Age Distribution of Respondents

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

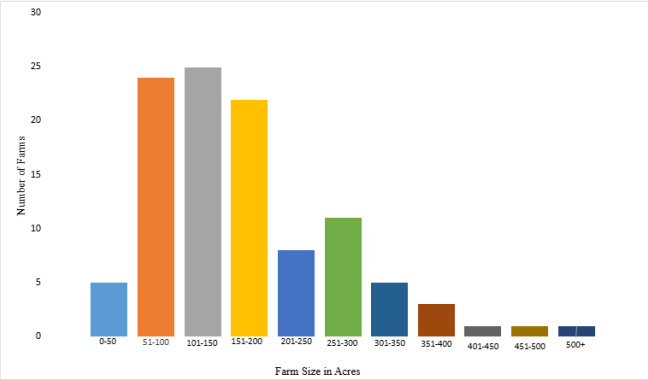


Figure 2. Farm Size Distribution

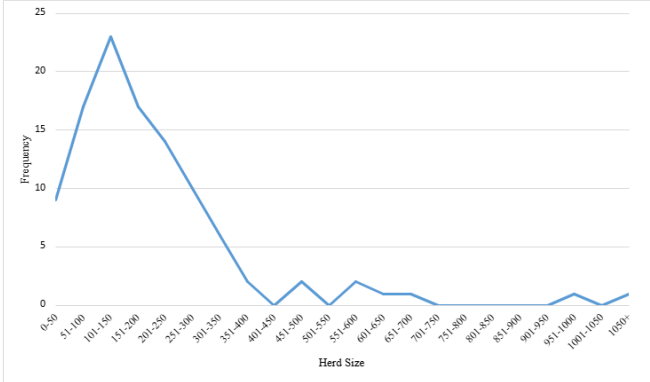


Figure 3. Herd Size Distribution

The mean value is 194, with a median of 150 and a modal value of 100 acres. A semi-interquartile range of 71indicates that the middle 50% of the data is moderately spread out, while the larger mean deviation of 94 indicating that extreme values (range = 1,684 acres) increase the overall variability in farm size.

Fresh water sources were present on or near 90% of respondents' farms of the variety of River, Stream, Pond, Dam, Well or Borehole.

Farm enterprise type was recorded as a multiple-response categorical variable, as farms could engage in more than one enterprise, and in fact 36.89% of farmers indicated more than one principal farming enterprise. Dairy farming was the most frequently reported enterprise (68 farms - 64.15%), followed by cattle rearing (44 farms - 41.51%) and other cattle enterprises (19 farms - 17.92%).

With respect to herd size, a total of 23,132 were reported across all respondents. As expected, the distribution follows a similar pattern to that of the farm size. Respondents reported herd sizes with a minimum herd size of 10 cattle for a young farmer (age ≤ 30 years), up to two farms reporting a herd size of more than 1,000 cattle. The herd size distribution is right-skewed and is illustrated in Fig. 3. As with the farm size, herd size heterogeneity arises because the dataset includes both small-scale and large-scale farming operations. The mean herd size is 220 with a median of 177 and a modal value of 200. The middle 50% of the data shows moderate dispersion (SIQR = 72.65), while overall variability is higher (mean deviation = 122) due to extreme values, reflected in a range of 1,690 animals.

With respect to the type of slurry storage system used on the farm, respondents were provided with a list of categorical labels for which multiple choices were permitted. Underground slatted tanks are used by 96% (n=102) of the farms, overground concrete (or steel) towers (or tanks) are being used by 18% (n=20) farms, lagoons by 7% (n=8) farms, and 0% reported use of slurry bags. More than one method of slurry storage is being used on 24% (n=26) farms.

B. *Use of Sensors on the farm*

Responses regarding the integration of sensors into farm equipment were dominated by negative responses, with 70.7% of respondents indicating that sensors were not integrated. Less than a quarter of respondents (22.6%) reported sensor integration, while a small proportion (6.6%) indicated that integration may be present. This proportion of responses is seen in Fig. 4.

Farmers indicated through an open-ended question where sensor technologies are currently being used. Example responses included 'sat nav and flow meter on slurry tanker', 'sensors on our milking equipment to know when cows are finished milking so the automatic cluster remover can activate', and 'heat detection sensors on cows'. These were grouped into five main categories: Fertilizer Application, Feeding, Fertility Management, Milking and Others, and General Farm maintenance, as seen in Table 3. Dairy Farmers were more likely to employ sensor technologies than farmers of other enterprises, with 58% of all sensors used on farms that identified their main activity as 'dairy'. The types of farm enterprises using sensors are illustrated in Fig. 5.

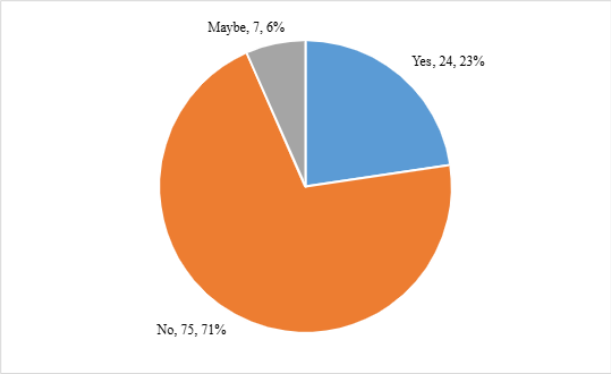


Figure 4. Sensor Integration on Farms

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

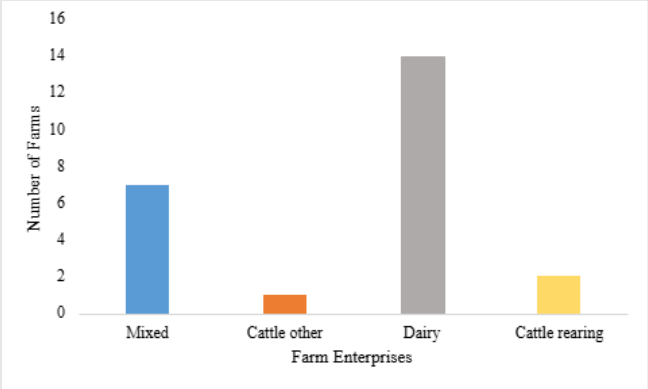


Figure 5. Farm Enterprises with Sensor Devices

When asked is any form of sensor integrated equipment utilized for the measurement of any gas emission in your sheds, only 2 (1.8%) answered yes, another 2 answered maybe (1.8%), with the remaining (96.4%) negative.

When asked do you have slurry sensors for the estimation of slurry level in your storage tank/pit/tower? and do you have any designated sensor for the estimation of emissions in the slurry storage systems? The same two respondents (1.8%) answered yes, with the remaining (98.2%) negative. These two respondents came from different sized farms, one with a 700 acre farm, the other, a younger man, with a farm of 150 acres.

There was slightly higher take-up of sensors in slurry spreading machinery with 12.3% (n=13) of respondents reporting the use of integrated sensor technology in machinery. Of those, 3.8% (n=4) reported that the sensor in the machinery can be used for the estimation of emissions from the field.

C. *Attitude and Perceptions on the Use of Sensors on the farm*

In this section, the results of the attitudinal questions collected using a Likert-scale and multiple-choice questions, are laid out. These measured farmers' perceptions and/or beliefs regarding the benefits and challenges of sensor technologies in supporting emission reduction on farms.

1) Perceived Usefulness of Sensors in Emission Reduction: The respondents indicated how useful they thought sensors were in the fight against greenhouse gas emissions and pollution. Responses were collected using a five-point Likert-type scale, visually represented using star ratings, where one star indicated strong disagreement, and five stars indicated strong agreement. 23% respondents reported not useful (with a star rating of 1 or 2), 28.57% as neutral (star rating of 3), and 48.35% as useful (with a star rating of 4 or 5). The mean response is 3.3, with a mode of 4. With 3 being neutral, this mean response of 3.3 indicated a slight tendency towards agreement, and this is clearly illustrated in the left-skew on the frequency distribution in Fig. 6. The mean score increases to an average of 4.5 for those respondents who have adopted sensor technologies.

TABLE 3. FARMER RESPONSES ON THE USAGE OF SENSORS ON FARMS

Usage of Sensors	Sample verbatim responses
Fertilizer Application	Reversing sensors, on fertiliser spreader, automatic scraper, on sprayer, on loader to prevent it from turning over
	Dribble bar
	Variable rate technology in fertilizer spreader
	Spreading fertiliser GPS
	Sprayer and fertiliser
	Sat nav and flow meter on slurry tanker
Feeding	Automatic calf feeder
	Feed to yield sensors (feeding selection for cows)
	Calf feeder, drafting system for cows
	Cow collars for heat detection Calf tags for auto-feeder
Fertility Management	Heat detection on cows, Feeding cows in milking parlour
	Heat detection sensors on cows
	Health and fertility management
	We have collars on the cows to detect heat, health, activity etc.
Milking and Others	Automatic scraper, Vacuum tanker, Milking parlour, Automatic ration feeders, Automatic milk feeders, Tractors, Water pumps
	We use sensors to know how much water has been pumped out to the drinkers throughout the day and night. There are sensors on our milking equipment to know when cows are finished milking so the automatic cluster remover can activate. We use sensors to know the temperature of the milk on our milk storage.
	Sensors on the milking parlour, that make sure tap on tank is closed, that take off cluster when cow is finished milking, movement sensors on lights
	Robotic milking
General Farm maintenance	Cameras down in the farmyard the light turns on in them when they pick up something
	GPS
	Drafting gate
	Lights seatbelts
	JCB loader
	Diggers
	Balers

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

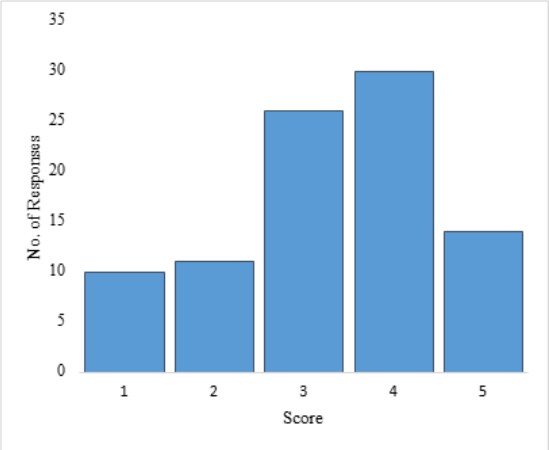


Figure 6. Perceived usefulness of sensors in reducing greenhouse gas emissions and pollution

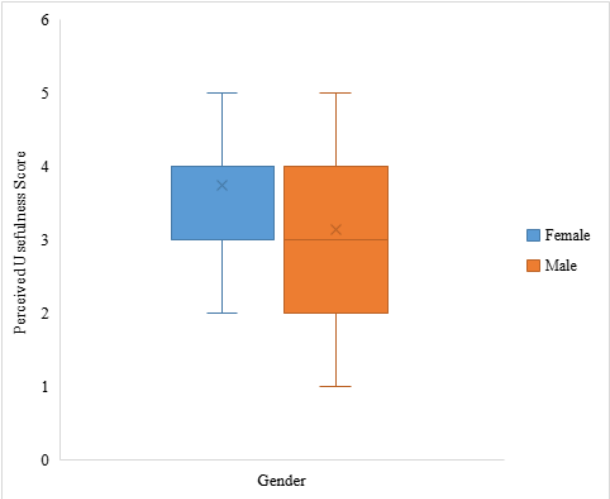


Figure 7. Perceived usefulness of sensors in reducing greenhouse gas emissions and pollution by gender

A visual inspection of the means plot does indicate a U-shape relationship between age (ordered categorical) and perceived usefulness. Younger respondents, ≤30 years, reported a mean score of 3.6, the three groups between 31 and 60 years with a mean score of 3.16, rising again to a mean of 3.65 to those over 60 years of age.

Descriptive analysis revealed gender differences in perceived usefulness, see Fig. 7. Female respondents reported higher perceived usefulness (Mean = 3.74, Mode = 4) compared to male respondents (Mean = 3.15, Mode = 3).

2) Perceived benefits/advantages of the use of sensor technology on the farm: The question posed 'Which of the following sensor benefits would be most useful for your farm?'. This was a multiple-response categorical data point, which was binary in nature (selected / not selected), and allowed more than one option per respondent. The primary descriptive result, which is the frequency of benefits selected by respondents, is illustrated in Fig. 8. These have been sorted in ranking from the highest occurrence to the lowest occurrence.

The most frequently cited benefits are related to decision-making, sustainability, and emissions management, such as 'help us make better decisions' (25.24%, n = 26), 'better for sustainability' (23.3%, n = 24), and 'real-time monitoring of emissions' (22.33%, n = 23) among the highest-ranked responses.

A second tier of benefits related to measurement precision and operational support, including 'supports access to grants and funding' (18.45%, n = 19) and 'accurate emission quantification' (16.5%, n = 17). Less frequently selected benefits were those associated with spatial and longitudinal data granularity, such as 'access to precise data around different locations on the farm' (13.59%, n = 14) and 'access to changes in data over time' (8.74%, n = 9).

Overall, co-occurrence among benefits was moderate rather than strong (Jaccard Similarity Heatmap, Fig. 9). Associations among perceived benefits were generally of moderate magnitude ($\Phi \approx 0.20$-0.40), reflecting consistent but not dominant thematic overlap, as further illustrated in the Phi (Φ) Coefficient Matrix of Co-selected Perceived Benefits Heatmap (Fig. 10).

The strongest association was observed between 'access to changes in data over time' and 'supports access to grants and funding' (Jaccard = 0.40, $\Phi = 0.56$). 'Access to changes in data over time' is the most interconnected variable. It has the highest or near-highest correlations with almost every other benefit grants/funding ($\Phi = 0.56$) and sustainability ($\Phi = 0.48$), and moderate correlations with 'real time monitoring of emissions' ($\Phi = 0.41$), 'help us to reduce our emissions' ($\Phi = 0.41$) and 'help us make better decisions' ($\Phi = 0.37$).

Grants/funding sits in its own cluster. Beyond its strong link to temporal data ($\Phi = 0.56$), it correlates moderately with sustainability ($\Phi = 0.39$) and real-time monitoring ($\Phi = 0.35$), but weakly with accurate quantification ($\Phi = 0.19$) and detection ($\Phi = 0.18$). Two benefit pairings demonstrated weak co-occurrence (Jaccard = 0.14, $\Phi = 0.10$), 'help us to reduce our emissions' and 'real-time monitoring of emissions' were rarely co-selected, as were 'better detection capabilities' and 'help us make better decisions' showed weak co-occurrence. While most pairwise associations have weak or moderate co-occurrence similarity, 'supports access to grants and funding' emerged as the benefit with the most consistent moderate co-occurrence across multiple other benefit categories, with co-occurrence coefficients such as 'real time monitoring of emissions' (Jaccard = 0.31, $\Phi = 0.35$), 'better for sustainability' (Jaccard = 0.34, $\Phi = 0.39$), 'help us make better decisions' (Jaccard = 0.32, $\Phi = 0.36$). Overall, there was absence of consistently strong associations and a limited number of large effect sizes ($\Phi \geq 0.50$).

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

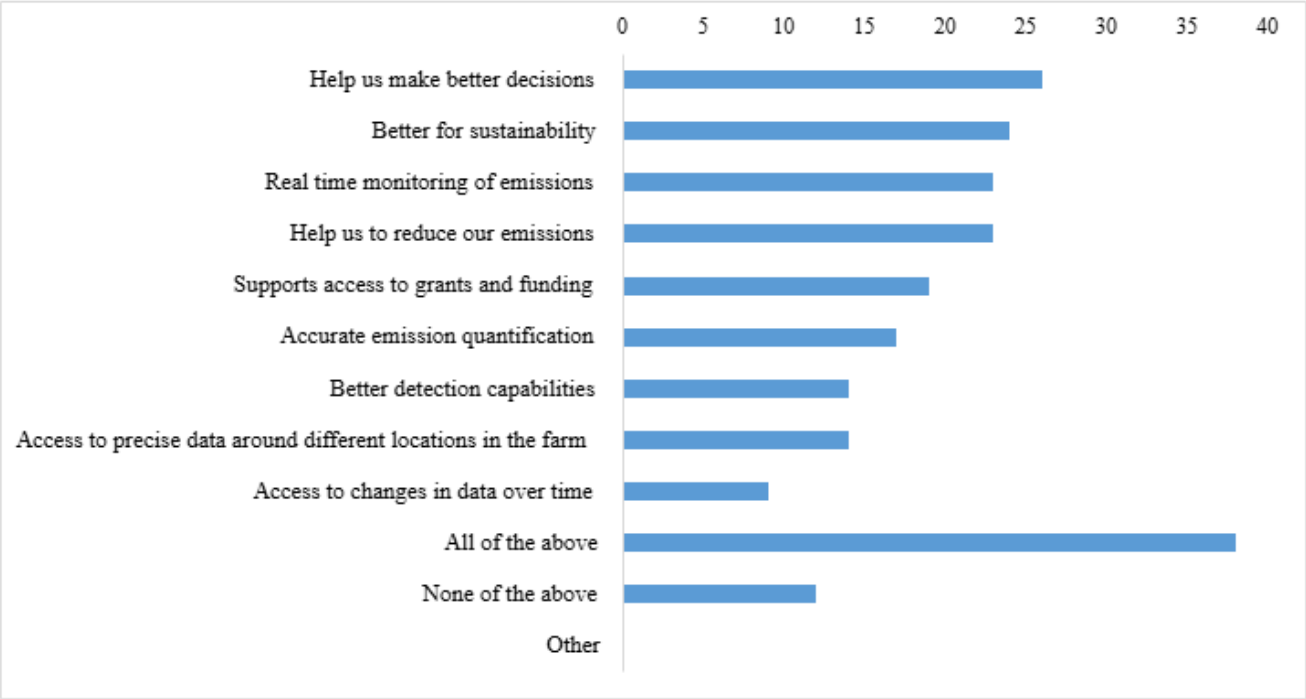


Figure 8. Perceived benefits associated with on-farm sensor adoption

	Help us make better decisions	Better for sustainability	Real time monitoring of emissions	Help us to reduce our emissions	Supports access to grants and funding	Accurate emission quantification	Better detection capabilities	Access to precise data around different locations in the farm	Access to changes in data over time
Help us make better decisions		0.35	0.23	0.32	0.32	0.23	0.14	0.25	0.25
Better for sustainability	0.35		0.27	0.24	0.34	0.24	0.27	0.23	0.32
Real time monitoring of emissions	0.23	0.27		0.18	0.31	0.33	0.16	0.32	0.28
Help us to reduce our emissions	0.32	0.24	0.18		0.24	0.33	0.28	0.23	0.28
Supports access to grants and funding	0.32	0.34	0.31	0.24		0.20	0.18	0.22	0.40
Accurate emission quantification	0.23	0.24	0.33	0.33	0.20		0.29	0.29	0.18
Better detection capabilities	0.14	0.27	0.16	0.28	0.18	0.29		0.27	0.21
Access to precise data around different locations in the farm	0.25	0.23	0.32	0.23	0.22	0.29	0.27		0.21
Access to changes in data over time	0.25	0.32	0.28	0.28	0.40	0.18	0.21	0.21	

Figure 9. Jaccard Similarity Heatmap on Benefits and Advantages of Using Sensors in the Farm.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

	Help us make better decisions	Better for sustainability	Real time monitoring of emissions	Help us to reduce our emissions	Supports access to grants and funding	Accurate emission quantification	Better detection capabilities	Access to precise data around different locations in the farm	Access to changes in data over time
Help us make better decisions		0.37	0.17	0.33	0.36	0.22	0.10	0.29	0.37
Better for sustainability	0.37		0.26	0.20	0.39	0.25	0.32	0.25	0.48
Real time monitoring of emissions	0.17	0.26		0.10	0.35	0.39	0.13	0.40	0.41
Help us to reduce our emissions	0.33	0.20	0.10		0.23	0.39	0.33	0.26	0.41
Supports access to grants and funding	0.36	0.39	0.35	0.23		0.19	0.18	0.25	0.56
Accurate emission quantification	0.22	0.25	0.39	0.39	0.19		0.36	0.36	0.23
Better detection capabilities	0.10	0.32	0.13	0.33	0.18	0.36		0.34	0.28
Access to precise data around different locations in the farm	0.29	0.25	0.40	0.26	0.25	0.36	0.34		0.28
Access to changes in data over time	0.37	0.48	0.41	0.41	0.56	0.23	0.28	0.28	

Figure 10. Phi (Φ) Coefficient of Co-selected Perceived Benefits and Advantages of Using Sensors in the Farm.

3) Perceived challenges/disadvantages of the use of sensor technology on the farm: Respondents were asked to identify the perceived challenges and/or disadvantages associated with sensor use on the farm. This was a multiple-response categorical variable, dichotomous in nature (selected / not selected), permitting more than one selection per respondent. Three null values were removed, leaving 103 valid responses. The frequency of responses is presented in Fig. 11.

Overall, the most frequently cited challenges related to cost, maintenance, and knowledge barriers. 'High initial costs' were the most reported issue by 61.17% of respondents (n=63), followed by 50.49% reporting perceived 'high maintenance effort and associated costs' (n=52). A substantial proportion, 40.78%, of respondents also identified a 'lack of knowledge about sensors' as a challenge (n=42).

A second group of challenges related to practical and operational issues. These included 'operational difficulties' by 19.42% (n=20) and concerns regarding 'inadequate sensor sensitivity and accuracy', 15.53% (n=16). Fewer respondents, 10.68%, cited 'lack of time' as a limiting factor (n =11).

The least frequently selected challenges were associated with data interpretation and system integration, including perceptions that sensors are too fragmented and solve only one problem at a time, 8.74%, (n = 9), 'limited capabilities for slurry characterization', 5.83%, (n = 6), and 'difficulty in reading or interpreting sensor data', 4.85%, (n = 5).

A small proportion of respondents selected 'none of the above' (n=6, 5.83%), indicating that they did not identify with the specific challenges listed in the questionnaire. This response does not necessarily imply the absence of challenges, but rather that respondents did not identify with the challenges listed. Another proportion selected 'don't know the challenges' (n=16, 15.53%), suggesting uncertainty or limited familiarity with sensor technologies among some participants. Half of these also selected 'lack of knowledge about sensors' or 'high costs'.

As with the benefits and advantages, responses to the multiple-selection question were then coded into binary variables (1=selected, 0=not selected). A Jaccard similarity analysis was employed to quantify the proportional overlap between themes, independent of joint absences. The associations are illustrated in Jaccard Similarity Heatmap labelled Fig. 12. Phi coefficients of co-selected perceived

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

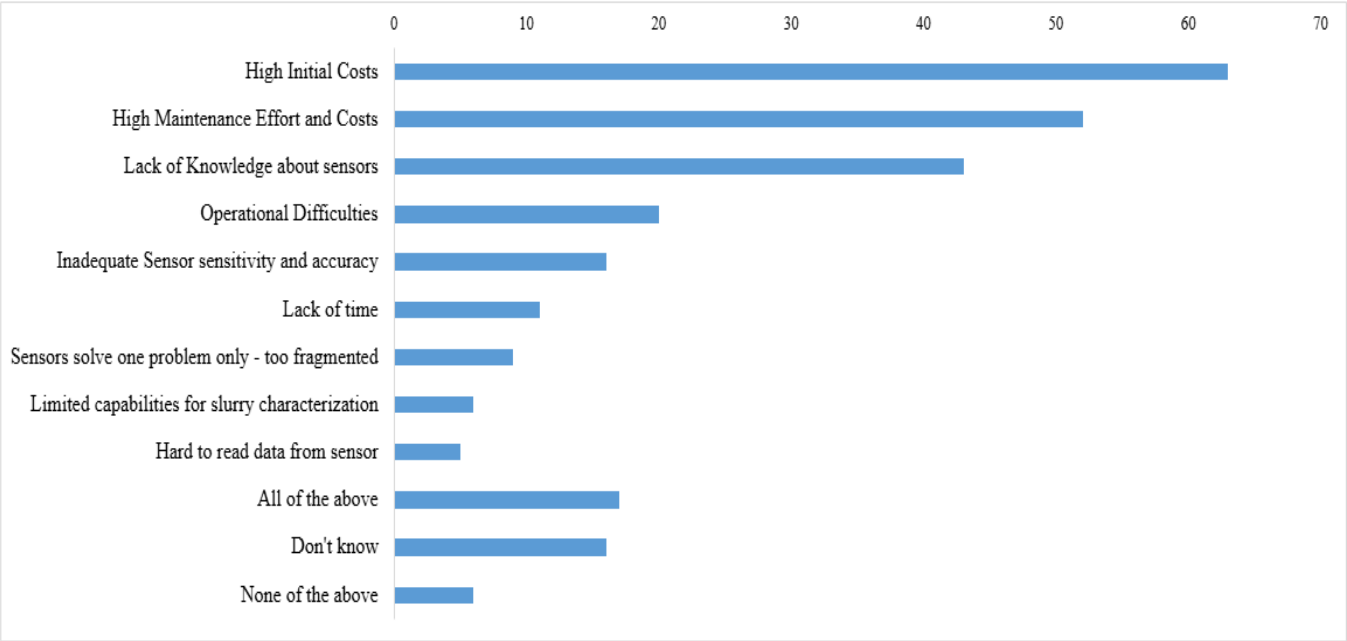


Figure 11. Perceived challenges/disadvantages associated with on-farm sensor adoption

	High Initial Costs	High Maintenance Effort and Costs	Lack of Knowledge about sensors	Operational Difficulties	Inadequate Sensor sensitivity and accuracy	Lack of time	Sensors solve one problem only - too fragmented	Limited capabilities for slurry characterization	Hard to read data from sensor
High Initial Costs		0.67	0.46	0.32	0.25	0.17	0.09	0.08	0.08
High Maintenance Effort and Costs	0.67		0.34	0.31	0.26	0.21	0.11	0.05	0.10
Lack of Knowledge about sensors	0.46	0.34		0.32	0.26	0.13	0.16	0.12	0.12
Operational Difficulties	0.32	0.31	0.32		0.38	0.24	0.16	0.13	0.19
Inadequate Sensor sensitivity and accuracy	0.25	0.26	0.26	0.38		0.17	0.19	0.10	0.11
Lack of time	0.17	0.21	0.13	0.24	0.17		0.05	0.13	0.14
Sensors solve one problem only - too fragmented	0.09	0.11	0.16	0.16	0.19	0.05		0.15	0.17
Limited capabilities for slurry characterization	0.08	0.05	0.12	0.13	0.10	0.13	0.15		0.10
Hard to read data from sensor	0.08	0.10	0.12	0.19	0.11	0.14	0.17	0.10	

Figure 12. Jaccard similarity heatmap on challenges and disadvantages of using sensors in the farm

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

	High Initial Costs	High Maintenance Effort and Costs	Lack of Knowledge about sensors	Operational Difficulties	Inadequate Sensor sensitivity and accuracy	Lack of time	Sensors solve one problem only - too fragmented	Limited capabilities for slurry characterization	Hard to read data from sensor
High Initial Costs		0.57	0.30	0.39	0.34	0.28	0.03	0.11	0.18
High Maintenance Effort and Costs	0.57		0.11	0.34	0.32	0.34	0.10	0.00	0.22
Lack of Knowledge about sensors	0.30	0.11		0.34	0.30	0.10	0.23	0.22	0.27
Operational Difficulties	0.39	0.34	0.34		0.47	0.31	0.20	0.19	0.35
Inadequate Sensor sensitivity and accuracy	0.34	0.32	0.30	0.47		0.20	0.25	0.12	0.15
Lack of time	0.28	0.34	0.10	0.31	0.20		0.00	0.18	0.21
Sensors solve one problem only - too fragmented	0.03	0.10	0.23	0.20	0.25	0.00		0.22	0.25
Limited capabilities for slurry characterization	0.11	0.00	0.22	0.19	0.12	0.18	0.22		0.14
Hard to read data from sensor	0.18	0.22	0.27	0.35	0.15	0.21	0.25	0.14	

Figure 13. Phi (Φ) coefficient heatmap of c0-selected perceived challenges and disadvantages of using sensors in the farm

challenges and disadvantages were calculated and are illustrated in Fig. 13. Jaccard similarity coefficient values range from 0.05 to 0.67, suggesting one particularly strong cluster, several moderate cross-links, and many weak peripheral relationships. The coefficients with distinct high and low values suggest potential clustering among specific themes. Moderate to strong associations were shown between 'high initial costs', 'high maintenance effort and costs', 'lack of knowledge about sensors', 'operational difficulties', 'inadequate sensor sensitivity and accuracy', and 'lack of time'.

The strongest association was observed between 'high initial costs' and 'high maintenance effort and costs' (J=0.67, Φ = 0.57) indicating a large effect size and substantial co-occurrence between financial barriers.

Moderate associations were identified between operational and technical performance-related themes, particularly between 'operational difficulties' and 'inadequate sensor sensitivity and accuracy' (J=0.38, Φ=0.47); and 'operational difficulties' and 'hard to read data' (J=0.19, Φ=0.35), and 'operational difficulties' and 'lack of knowledge' (J=0.32, Φ=0.34).

Lack of time demonstrated generally weak-to-moderate Jaccard similarity coefficients with other barriers (0.05–0.24), with moderate Phi associations observed between lack of time with 'high maintenance effort and costs' (Φ = 0.34), 'operational difficulties' (Φ = 0.31), and 'high initial costs' (Φ = 0.28).

V. Discussion

The primary aim of the study is to evaluate the usage of sensor technologies for the real time monitoring and control of emissions in slurry management processes in Irish farms. The findings are discussed in relation to the three central research questions: firstly, how are sensors used on Irish farms. Secondly, to what extent are sensors integrated for the mitigation or reduction of emissions in slurry management. Thirdly, what are the Irish farmers' perceptions about the benefit/advantages and challenges/disadvantages of use of sensors for monitoring of emissions in slurry management on farms.

A. How are sensors used in Irish farms?

The aim of employing sensors is to minimize negative environmental impacts while optimizing nutrient recovery and beneficial use. Data-driven management facilitated by sensors enables more efficient and environmentally friendly slurry handling.

Findings reported in this study revealed slightly less than a quarter (22.6%) of farmers integrated sensors on the farm, corroborating a report by Teagasc which stated that the adoption of precision agriculture technologies have been slow in Ireland [43]. This is lower than the 25% reported for Northern Europe, and 60% for the Global North overall [44].

In this study, sensor technologies were reported mainly integrated in field fertilizer application, milking, fertility management, water pumps, balers, slurry tank and other general farm operations (Table 3).

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Costs are often reported as the main determinants of adoption of smart farming technologies on Irish farms [41, 43]. This study revealed that farms engaged majorly in dairy enterprises have the highest number of sensor integrated devices. This might be a cost issue, and linked to affordability, as dairy enterprises are deemed to be more profitable than others in Ireland [45].

Previous research found that advanced technologies are mostly integrated on Irish farms which are big in production and employ technologically sophisticated personnel [43]. The findings of the present study provide limited support for this relationship, if a link between farm size and big in production is assumed. In this study, farm size was positively skewed; therefore, farms above the 75th percentile were additionally examined as a 'large farm' subgroup. This threshold approximately corresponded to farms exceeding 242 acres, which is broadly comparable to larger farm classifications reported by the Central Statistics Office. Within the present sample, farmers operating larger farms were not substantially more likely to report the use of advanced sensor technologies than those operating farms closer to the sample median size. Respondents who reported that sensors were integrated into farm equipment or machinery had a median farm size of 150 acres. Similarly, respondents reporting sensor integration within slurry spreading machinery had a median farm size of 150 acres, while those reporting the use of sensor-integrated equipment for gas emission measurement in sheds had a median farm size of 170 acres. However, no sensor use was reported among farmers within the lowest quartile of farm size (<100 acres). Although the findings suggest that very small farms may have lower levels of sensor adoption, the overall low adoption rates observed across the sample indicate that barriers to implementation extend beyond farm size alone.

B. *How are sensors used in the mitigation of ammonia and greenhouse gas emissions in slurry management on Irish farms?*

Exploration of the distribution of the use of sensors throughout the slurry lifecycle, (production, storage and application) revealed significantly low adoption. Integrated sensors for the estimation of slurry levels in storage (tanks) were only reported on two farms of the 106 in the study. The use of sensors in field application of slurry revealed a slightly higher number (13 farms; 12.3%). However, only four (3%) reported that integrated sensors are being used for the estimation of emissions during field application, and only two (~2%) reported the use of sensors for the estimation of emissions from sheds. These data revealed very low adoption and integration of sensors for slurry management and mitigation of emissions on Irish farms overall.

Currently, sensors can be integrated into several aspects of farm activity [43]. Although adoption of specific sensors is dependent on several factors with high cost and maintenance reported as the main factor, farmers' decisions to integrate specific sensors are still mainly affected by perceived usefulness in solving a problem. Agriculture media, farm advisors, farm discussion groups and trade shows in ploughing championships are main influencers for the adoption of various farming technologies by the Irish farmer [46] and could be leveraged. A scheme such as the Knowledge Transfer Program aimed at encouraging knowledge exchange and innovation between advisors, farmers and other stakeholders play an important role in assisting farmers to meet current and future challenges [47]. Schemes such as this would ensure farmers are skilled to utilize different technologies like GPS, computer-based imaging, robotics, etc. [48]. "Targeted Agricultural Modernization Scheme II (TAMS II)" was a scheme from the Irish Government in 2018 to encourage young farmers to use up-to-date technology in their farms by providing grants of 40–60% on the technology used [48]. There is a need for similar schemes which will provide financial support and the required skills to farmers encouraging greater adoption of advanced technologies in the fight against emissions.

C. *What are the perceived benefits and challenges of sensor technologies when used for slurry management?*

Within this study, 'supports access to grants and funding' emerged as the benefit with the most consistent moderate co-occurrence across multiple other benefit categories indicating that respondents frequently selected funding-related benefits alongside a range of other perceived sensor benefits. The strongest association observed between 'access to changes in data over time' and 'supports access to grants and funding' indicates a relatively strong co-selection of these benefits compared to other pairings, suggesting that respondents who value longitudinal data are also more likely to perceive sensor technologies as facilitating access to funding or support schemes. Respondents didn't strongly link technical precision to funding access but did link broader environmental narrative benefits to grants and funding. The fact that perceived support for grant and funding access showed moderate associations with multiple other benefits suggests that farmers view funding-related outcomes as dependent on broader data and monitoring capabilities, rather than as a standalone function of sensor technologies. This pattern indicates an understanding that access to funding is often contingent on the ability to generate, maintain, and demonstrate evidence over time.

Several benefits have been reported from the adoption of sensors for the detection of emissions on farms. Perceived benefits of the use of sensors are that they allow real-time

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

measurement of environmental parameters, such as temperature, humidity, and gas concentrations (e.g., NH_3, CO_2, H_2S), which supports better decision-making regarding optimal slurry application timing to reduce emissions [26][36]. In this study, farmers showed agreement with that, as reflected in the positive score on the Likert scale. This study revealed that several farmers agree with other research findings, that integration of sensor technology can enhance detection capabilities [33][28], improved emission quantification [31], increase spatial and temporal precision [12][38] and support sustainable practices [39]. In this study, higher-frequency responses though tended to emphasise practical and outcome-focused benefits, while lower-frequency responses reflected more technical or analytical uses of sensor data.

Notably, two benefit pairings demonstrated weak co-occurrence suggesting these benefits were perceived as largely independent by respondents. 'Help us to reduce our emissions' and 'real-time monitoring of emissions' were rarely co-selected despite their apparent conceptual proximity, indicating that respondents may have decoupled emissions reduction as an outcome from real-time monitoring as a process. Similarly, 'better detection capabilities' and 'help us make better decisions' showed weak co-occurrence, suggesting respondents may have perceived detection-oriented benefits as serving a different purpose, potentially regulatory or technical compliance, rather than directly informing on-farm decision-making. These distinctions may have implications for how monitoring technologies are framed and communicated to farming stakeholders. In [18] using the Technology Acceptance Model, it was found that attitude, perceived usefulness, and ease of use predicted intention well to adopt sensor technologies in agriculture. Technology Acceptance Models [49] consider acceptance at an individual level, that is usefulness to the person, how easy it is to use, and so on, when considering attitude towards the technology, intention to use technology, or actual use of the technology. Ease of use is found to predict intention to adopt sensor technologies in agriculture [18]. However, with respect to perceived usefulness and attitude, the impact of sensors in tackling the overwhelming environmental and sustainability challenges is quite small [50]. It is arguable that in this study, perceived usefulness in respect of helping to reduce emissions or make better decisions may have limitations where the environmental challenge is perceived as disproportionately large relative to the capabilities of the technology itself.

This study also identified challenges such as high cost and maintenance, fragmented design, and operational constraints, indicating a technology that is far from easy to use. This concurs with the findings in [13] which considered the Theoretical Framework of Acceptability in Smart Agriculture. Their analysis found that opportunity costs and burden are considerable challenges. Similarly, in this study, 65% of all respondents raised financial concerns as challenges, in either or both, of the options 'high initial costs' and 'high maintenance costs and rarely considered acquisition cost in isolation. The financial burden associated with sensors appears to be understood in terms of total cost of ownership, where perceived ongoing maintenance requirements contribute significantly to perceived economic risk. This figure of 65% is higher than reported for a survey of North American farmers in which 52% identified cost as a concern in the adoption of precision agricultural technologies including sensors [44]. Irish farms tend to be smaller than the large commercial intensive scale farms found in North America, and it's possible that the importance of financial concerns is reflective of the amount for investment that smaller Irish farmers have at their disposal. Therefore, this finding suggests that reducing the purchase price of sensors alone may not substantially improve adoption if maintenance complexity and recurring operational costs remain high. The result aligns with broader technology adoption literature in which lifecycle costs are often more influential than initial capital expenditure in determining long-term acceptance and implementation.

Moderate associations identified between operational and technical performance-related themes, suggest that although difficulty interpreting sensor data is not widespread across the sample overall, when it does occur, it is strongly linked to broader operational concerns suggesting that perceptions of technological reliability and are closely linked with practical usability concerns for that cluster. Whilst this association between operational difficulties and difficulty interpreting sensor data was moderately strong, the low frequency of the latter category suggests this relationship should be interpreted cautiously, it does however indicate a clustering of implementation-related challenges. The barriers to adoption may involve both technical performance and data literacy challenges. However, these findings should be interpreted in the context of the very low reported adoption of sensor technologies among respondents. Given the limited direct experience likely present within the sample, these concerns may reflect perceived or anticipated barriers rather than operational difficulties arising from extensive practical use. Indeed, this interpretation is supported by the association between 'operational difficulties' and 'lack of knowledge'. Respondents may associate sensor technologies with unreliable measurements, difficult implementation, or inconsistent performance based on anecdotal information, peer discussions, or broader perceptions of technological complexity. This suggests that uncertainty and lack of trust in sensor reliability may play a significant role in limiting adoption decisions. In this context, perceived implementation

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

challenges may be as important as actual technical challenges, highlighting the importance of demonstration, user training, and evidence of practical reliability under real farm conditions.

Themes related to technological fragmentation and limited slurry characterization capabilities demonstrated weak associations with economic concerns suggesting these represent more independent or specialized barriers. Contrary to initial expectations, 'lack of time' was not typically perceived as a standalone dominant constraint. However, moderate 'associations were observed between 'lack of time' with 'high maintenance effort and costs', 'operational difficulties', and 'high initial costs'. This indicates that time-related concerns may be embedded within wider perceptions of operational burden and resource demand. In practice, respondents may associate sensor adoption with additional workload, setup requirements, maintenance responsibilities, and learning commitments, rather than viewing lack of time as an isolated issue, thus, lack of time may function more as an amplifying contextual barrier than a primary technical concern.

Taken together, the results suggest that farmers primarily perceive barriers to sensor adoption as being driven by economic cost and ongoing maintenance requirements, alongside knowledge and capability constraints, while more technical issues related to data interpretation and system integration were less frequently reported. Overall, the co-occurrence structure suggests three interrelated but distinct domains of barriers: economic constraints, technical/operational challenges, and organizational knowledge limitations.

Whilst this study showed that the farmers perceive several benefits that could be derived from the integration of sensors for mitigation of emission from slurry management, there is low adoption, and therefore the benefits are not being realized. There are perceived challenges identified in the literature, which the farmers in this study agreed with. These include high cost and maintenance [33] and limited technical knowledge about sensors. Others are inadequate sensor sensitivity and accuracy [28][29], operational constraints [32][34], limited capability for slurry characterization [19] and concerns about overly fragmented design [36] (Figs. 11 and 12).

VI. CONCLUSION

Traditional slurry management practices often lead to pollution and greenhouse gas emissions. There is potential within slurry management to reduce these emissions and have a positive impact on national emissions targets. Significant efforts to reduce emissions occur within the lifecycle of slurry, from livestock feed selection through manure spreading or the alternative pathway of biomethane production. In the past ten years there have been exponential developments in technology that have fuelled Smart Agriculture. At the core of these developments are the use of sensors, which capture and, in some instances, analyse data at source. An overview of the various applications of sensors for the monitoring of emissions in slurry management in Irish farms is provided, and as such provides an insight to the reduction of emissions in the slurry lifecycle in livestock farming.

Determinants of technological innovation adoption have been extensively examined through established behavioural frameworks such as the Theory of Reasoned Action, the Theory of Planned Behaviour, the Unified Theory of Acceptance and Use of Technology, the Theoretical Framework of Acceptability, and the Technology Acceptance Model [15][51]. Given the maturity of this literature, the present study did not aim to operationalize or evaluate these established models. Instead, this study adopts a system-oriented and descriptive approach, mapping the current extent of sensor technology utilization and documenting stakeholder-reported advantages and disadvantages. In doing so, it complements rather than replicates existing theory-driven research by characterizing the present implementation landscape.

In this study, Irish farmers were surveyed with respect to their views about, and adoption of, sensor technologies, with a particular emphasis on the use of sensors to measure and combat emissions. Farmers believe that sensor technologies are useful in reducing greenhouse gas emissions and pollution (mean score of 3.3 out of 5). Farmers agree that sensor technologies can provide benefits in slurry management around real-time monitoring and decision-making support, enhanced detection capabilities, improved emission quantification, spatial and temporal precision, and support for sustainable practices. They do not view sensor benefits as a single, unified construct, but rather as a set of distinct yet related functions that address different operational and environmental needs with potential economic and regulatory value, as seen in the most significant correlation between benefits is an understanding that access to data supports access to grants and funding.

Despite the high selection of advantages such as 'help us make better decisions', 'better for sustainability', help us reduce out emissions' and 'real time monitoring of emissions', adoption remains limited, with fewer than one quarter of respondents (22.6%) reporting sensor use and only ~2% utilizing sensors specifically for emission estimation or mitigation. This indicates that, despite technological advancement and policy emphasis on emissions reduction, sensor-based monitoring remains peripheral within current slurry management practice.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Current limitations include high capital and maintenance costs, limited technical knowledge, inadequate sensor sensitivity and accuracy, operational constraints, limited capability for slurry characterization, and fragmented system design.

The co-occurrence of perceived financial barriers and funding-related benefits suggests either a structural misalignment between funding mechanisms and cost realities, or a communication gap regarding available support. These findings highlight the importance of evaluating whether existing grant and incentive schemes adequately offset initial and ongoing costs associated with sensor adoption, and if so, whether these are being adequately communicated to the farmer. Greater integration of economic feasibility modelling within policy design may improve uptake and support climate mitigation objectives.

Overall, this study provides an empirical baseline of current sensor technology utilization in Irish slurry management and identifies key economic and operational barriers to wider adoption. As agriculture faces increasing regulatory and climate mitigation pressures, understanding both technological capability and farmer-level implementation constraints will be critical in aligning innovation pathways with national emissions reduction objectives.

A. Limitations

The survey yielded 106 valid responses. Although the achieved sample size is adequate for exploratory and descriptive analysis, it limits the precision of population-level estimates and the reliability of subgroup comparisons. Consequently, the findings should be interpreted as indicative of emerging patterns rather than as statistically generalizable conclusions. These constraints are reflected in the use of descriptive analytical methods throughout the study. Additionally, the reliance on self-reported data may introduce response bias, particularly in relation to technology adoption and perceived barriers.

B. Further Research

This study has shown that there is limited adoption of sensors technologies for the quantification of greenhouse gases emissions from slurry on Irish farms. Therefore, there is a need for further research to develop, calibrate, and validate robust and reliable sensor systems for measurement of greenhouse gases during all stages of the slurry lifecycle. This includes addressing challenges related to sensor fouling, durability, and data accuracy in harsh, slurry environments.

As highlighted in the literature review [1], future studies should incorporate predictive analytics, which allow forecasting future emissions or simulated scenarios, such as extreme weather events. Predictive analytics are more useful for proactive decision-making and long-term mitigation planning.

This study finds that respondents frequently associate funding-related benefits, such as ‘supports access to grants and funding’, with other perceived advantages of sensor technologies, including ‘real-time monitoring of emissions', ‘better for sustainability’, and ‘helps us make better decisions’. At the same time, financial barriers emerged as a significant concern, with a large effect size and substantial co-occurrence between cost-related factors, particularly ‘high initial costs’ and ‘high maintenance effort and costs'. While costs are therefore perceived as a major barrier, respondents also recognize that sensor adoption may enhance access to funding while delivering sustainability and data-informed decision-making benefits. Future research should incorporate dynamic economic modelling that accounts not only for capital and operational costs but also for potential grant support, environmental incentives, and funding mechanisms that may offset initial investment barriers.

Interesting patterns by age and gender were observed in the data, such as female farmers tending to agree that sensor systems were more useful in reducing greenhouse gas emissions than male famers who were more likely to provide neutral evaluations. However, these patterns should be interpreted cautiously due to the non-random sampling approach and the imbalance in age and gender representation among respondents. The study was not designed to support definitive subgroup comparisons. Future research using larger and more representative samples should explore these patterns further.

Farmers in this study perceived that sensors were useful in the fight against emissions. However, there is low adoption, and there appears to be a disconnect between the detection of emissions and informing their decision making, Further qualitative research is warranted to explore the disconnect. Further research could also integrate established behavioural adoption frameworks (e.g., technology acceptance or diffusion models) with system-level performance metrics to better understand the interaction between technological capability and farmer decision-making. This integration would extend beyond the descriptive scope of the present study and support more targeted intervention strategies.

REFERENCES

[1] A. Y. Ikoyi and J. Humphries, “Evaluation of the Use of Sensors for the Detection of Emissions in Slurry Management”, Second International Conference on Sustainable and Regenerative Farming (CSRF), 2025, Barcelona, Spain. https://personales.upv.es/thinkmind/dl/conferences/csrf/csrf_2025/csrf_2025_1_10_10014.pdf

[2] Teagasc, “Alternative slurry amendments for cattle slurry storage” available at: https://teagasc.ie/news--events/daily/

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

ammonia-emissions-from-pig-production/, 2023. [retrieved: 2026.05.21].

[3] Gov.ie Climate Action Plan. Available at: https://www.gov.ie/en/publication/67104-climate-action-plan [retrieved: 2026.05.22].

[4] Environmental Protection Agency (EPA), "Latest emissions data", available at: https://www.epa.ie/our-services/monitoring-assessment/climate-change/ghg/latest-emissions-data/, 2024c. [retrieved: 2026.05.21].

[5] C. O'Donoghue, M. Ryan, D. Sologon, N. McLoughlin, A. Daxini and K. Daly, "A generalised farmer behaviour model for adoption of environmental measures", Journal of Cleaner Production, 450: 141631, 2024. doi:10.1016/j.jclepro.2024.141631

[6] Environmental Protection Agency (EPA), "Latest emissions data", available at: https://www.epa.ie/our-services/monitoring-assessment/climate-change/ghg/latest-emissions-data/, 2024c. [retrieved: 2026.05.21].

[7] Environmental Protection Agency (EPA) "Ireland's Informative Inventory Report 2023", available at: https://www.epa.ie/publications/monitoring--assessment/climate-change/air-emissions/IIR_Ireland_2024v1.pdf, 2024b. [retrieved: 2026.05.21].

[8] European Commission. Commission Implementing Decision (EU) 2022/696 of 29 April 2022 Granting a Derogation Requested by Ireland Pursuant to Council Directive 91/676/EEC Concerning the Protection of Waters against Pollution Caused by Nitrates from Agricultural Sources; European Commission: Brussels, Belgium, 2022. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022D0696 [retrieved: 2026.05.22].

[9] H. W. Ambrose, F. R. Dalby, A. Feilberg and M. Kofoed, "Additives and methods for the mitigation of methane emission from stored liquid manure", Biosystems Engineering 229: 209–245, 2023. doi:10.1016/j.biosystemseng.2023.03.015

[10] S. Bittman, M. Dedina, C. M. Howard, O. Oenema and M. A. Sutton, "Options for Ammonia Mitigation: Guidance from the UNECE Task Force on Reactive Nitrogen", NERC/Centre for Ecology & Hydrology, 2014.

[11] C. Buckley, D. Krol, G. J. Lanigan, T. Donnellan, J. Spink, K. Hanrahan, A. Boland, P. Forrestal, J. Humphreys, P. Murphy, N. NiFhlatharta, D. O'Brien, T. O'Dwyer, F. O'Mara, K. Richards, L. Shalloo, D. Wall and S. Waters "An analysis of the cost of the abatement of ammonia emissions in Irish agriculture to 2030", available at: https://teagasc.ie/wp-content/uploads/2025/05/NH3-Ammonia-MACC.pdf, 2020. [retrieved: 2026.06.05]

[12] T. Merl and K. Koren, "Visualizing NH_3 emission and the local O_2 and pH microenvironment of soil upon manure application using optical sensors", Environment International, 144, 2020. doi:10.1016/j.envint.2020.106080

[13] R. J. Thomas, G. O'Hare and D. Coyle, "Understanding technology acceptance in smart agriculture: A systematic review of empirical research in crop production", Technological Forecasting and Social Change, 189, 122374, 2023, doi:10.1016/j.techfore.2023.122374

[14] E. I. Soto, A. Barnes, A. Balafoutis, B. Beck, F. B. Sanchez, J. Vangeyte, S. Fountas, D. Van, V. Eory and B. M. Gomez, "The contribution of precision agriculture technologies to farm productivity and the mitigation of greenhouse gas emissions in the EU", EUR 29320 EN, Publications Office of the European Union, Luxembourg, 2019, https://publications.jrc.ec.europa.eu/repository/handle/JRC112505

[15] J. Rosário, L. Madureira, C. Marques and R. Silva, "Understanding Farmers' Adoption of Sustainable Agriculture Innovations: A Systematic Literature Review". Agronomy. 12(11), 2879, 2022, doi:10.3390/agronomy12112879

[16] L. C. McNicol, J. M. Bowen, H. J. Ferguson, J. Bell, R. J. Dewhurst and C. Duthie, "Adoption of precision livestock farming technologies has the potential to mitigate greenhouse gas emissions from beef production" Frontiers in Sustainable Food Systems, 8, 1414858, 2024, doi:10.3389/fsufs.2024.1414858

[17] M. N. Ahmad, "Narrative Literature Reviews in Scientific Research: Pros and Cons", Jordan Journal of Agricultural Sciences, 21(1), 1–4, 2025, doi.org/10.35516/jjas.v21i1.4143

[18] F. Taheri, M. D'Haese, D. Fiems and H. Azadi, "The intentions of agricultural professionals towards diffusing wireless sensor networks: Application of technology acceptance model in Southwest Iran", Technological Forecasting and Social Change,185,122075, 2022, doi:10.1016/j.techfore.2022.122075

[19] J. Hughes, "European Innovation Partnership Using Internet of Things technology to improve slurry management on farms - Year 2 Report", https://businesswales.gov.wales/farmingconnect/sites/farmingconnect/files/documents/EIP%20Wales_IoT%20Slurry%20Management_Final%20Report.pdf, 2023, [retrieved: 2026.05.21].

[20] B. Chugh, S. Thakur, A. K. Singh, R. M. Joany, S. Rajendran and T. A. Nguyen, "Electrochemical sensors for agricultural application", Nanosensors for Smart Agriculture, 147–164, 2022, doi:10.1016/B978-0-12-824554-5.00018-5

[21] X. Feng, R. A. Larson, and M. F. Digman, "Evaluation of Near-Infrared Reflectance and Transflectance Sensing System for Predicting Manure Nutrients", Remote Sensing, 14(4), 2022, doi:10.3390/rs14040963

[22] F. M. Silva, C. Queirós, M. Pereira, T. Pinho, T. Barroso, S. Magalhães, J. Boaventura, F. Santos, M. Cunha and R. C. Martins, "Precision Fertilization: A critical review analysis on sensing technologies for nitrogen, phosphorous and potassium quantification", Computers and Electronics in Agriculture, 224, 109220, 2024, doi:10.1016/j.compag.2024.109220

[23] M. K. Sørensen, O. Jensen, O. N. Bakharev, T. Nyord and N. C. Nielsen, NPK NMR Sensor: Online Monitoring of Nitrogen, Phosphorus, and Potassium in Animal Slurry", Analytical Chemistry, 87(13), 6446–6450, 2015, doi:10.1021/acs.analchem.5b01924

[24] A. Y. Ikoyi and B. A. Younge, "Impact of Sample Volume and Wavelength Region on the Near Infrared Spectroscopy (NIRS) Prediction of Inorganic Nutrient in Equine Faeces", National Symposium on NIR Spectroscopy Proceedings, 2021, https://conf.innorenew.eu/event/2/contributions/74/

[25] A. Y. Ikoyi and B. A. Younge, "Faecal near-infrared reflectance spectroscopy profiling for the prediction of dietary nutritional characteristics for equines", Animal Feed Science and Technology, 290, 2022, doi:10.1016/j.anifeedsci.2022.115363

[26] A. Mautone and A. Finzi, "Air quality monitoring in piggeries through an IoT gas and environmental sensors device", 11th European Conference on Precision Livestock Farming, pp. 1728–1736, 2024.

[27] A. Costantino, E. Fabrizio, A. Biglia, P. Cornale and L. Battaglini, "Energy Use for Climate Control of Animal Houses: The State of the Art in Europe", Energy Procedia, 101, 184–191. 2016, doi:10.1016/j.egypro.2016.11.024

[28] L. Du, L. Yang, C. Yang, C. Hu, C. Yu, M. Qiu, S. Liu, S. Zhu and X. Ye, "Development and Validation of an Energy Consumption Model for Animal Houses Achieving Precision Livestock Farming", Animals, 12(19), 2022, doi:10.3390/ani12192580

[29] H. Joo, S. W. Han, C. S. Lee, H. S. Jang, S. T. Kim and J. S. Han, "Field application of cost-effective sensors for the monitoring of NH_3, H2S, and TVOC in environmental treatment facilities and the estimation of odor intensity",

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Journal of the Air and Waste Management Association, 73(1), 50–64, 2023, doi:10.1080/10962247.2022.2131652

[30] J. K. Whalen, B. W. Thomas and M. Sharifi, "Novel Practices and Smart Technologies to Maximize the Nitrogen Fertilizer Value of Manure for Crop Production in Cold Humid Temperate Regions" 1–85, 2019. doi:10.1016/bs.agron.2018.09.002

[31] P. Sefeedpari, S. H. Pishgar-Komleh and A. J. A. Aarnink, "Model Adaptation and Validation for Estimating Methane and Ammonia Emissions from Fattening Pig Houses: Effect of Manure Management System", Animals, 14(6), 2024, doi:10.3390/ani14060964

[32] J. D. Bus, G. A. Franchi, I. J. Boumans, D. E. Te Beest, L. E. Webb, M. B. Jensen, L. J. Pedersen and E. A. Bokkers, "Short-term associations between ambient ammonia concentrations and growing-finishing pig performance and health", Preventive Veterinary Medicine, 242, 106555. 2025, doi:10.1016/j.prevetmed.2025.106555

[33] F. Chang, E. Fabian-Wheeler, T. L. Richard, and M. Hile, "Compaction effects on greenhouse gas and ammonia emissions from solid dairy manure", Journal of Environmental Management, 332. 117399, 2023, doi:10.1016/j.jenvman.2023.117399

[34] W. Ma, X. Ji, L. Ding, S. X. Yang, K. Guo and Q. Li, "Automatic Monitoring Methods for Greenhouse and Hazardous Gases Emitted from Ruminant Production Systems: A Review". Sensors, 24, 4423, 2024, doi:10.3390/s24134423

[35] J. Humphries, A. Ryan, P. Van de Ven, "A New Model for Quality 4.0" In: Reis, J., et al. Driving Quality Management and Sustainability in VUCA Environments. ICQIS 2023. Springer Proceedings in Business and Economics. Springer, Cham. 2024, doi:10.1007/978-3-031-52723-4_15

[36] K. Singh, K. A. Momin, M. Nishal, C. Sultania, and M. Rao, "Agri-Guard: IoT-Based Network for Agricultural Health Monitoring with Fault Detection", International Conference on Internet of Things, Big Data and Security, IoTBDS - Proceedings, 223–230, 2025, doi:10.5220/0013209500003944

[37] E. Dillon, T. Donnellan, B. Moran, and J. Lennon, "Teagasc National Farm Survey 2024". [online] teagasc.ie, Galway, Ireland: Agricultural Economics and Farm Surveys Department Rural Economy Development Programme, p.86. 2025. Available at: https://teagasc.ie/wp-content/uploads/uploads/media/website/publications/2025/NFSReport2024.pdf [retrieved: 2026.05.22].

[38] K. Vinković, T. Andersen, M. de Vries, B. Kers, S. van Heuven, W. Peters, A. Hensen, P. van den Bulk and H. Chen, "Evaluating the use of an Unmanned Aerial Vehicle-based active AirCore system to quantify methane emissions from dairy cows", Science of the Total Environment, 831, 2022, doi:10.1016/j.scitotenv.2022.154898

[39] S. Neethirajan, "Net Zero Dairy Farming - Advancing Climate Goals with Big Data and Artificial Intelligence", Climate 12(15), 2024, doi:10.3390/cli12020015

[40] F. Otieno, S. R. N. Nalakurthi, M. Raji, A. Tiwari, I. Anton and S. Gharbia, "Farmers' attitudes towards GHG emissions and adoption to low-cost sensor-driven smart farming for mitigation: The case of Ireland tillage and horticultural farmers", Smart Agricultural Technology, 9, 100622, 2024, doi:10.1016/j.atech.2024.100622

[41] V. J. Das, S. Sharma, A. Kaushik, "Views of Irish Farmers on Smart Farming Technologies: An Observational Study". AgriEngineering, 1(2):164-187, 2019, doi:10.3390/agriengineering1020013

[42] M. C. Bianchi, L. Bava, A. Sandrucci, F. M. Tangorra, A. Tamburini, G. Gislon and M. Zucali, "Diffusion of precision livestock farming technologies in dairy cattle farms," Animal, 16(11), 100650, 2022, doi:10.1016/j.animal.2022.100650

[43] Teagasc "Technology Foresight Report 2035" available at: https://teagasc.ie/wp-content/uploads/media/website/publications/2016/Teagasc-Technology-Foresight-Report-2035.pdf [retrieved: 2026.05.22].

[44] E. Saiz., F. Iqbal, J. H. Grant, L. Korir, S. Ullah and A. Radu, "Technological and Socio-Economic Challenges in the Development of Sensors for Precision Agriculture". Agriculture, 16(1), 89, 2025, doi:10.3390/agriculture16010089

[45] Teagasc "Farm income rebound in 2024" available at: https://teagasc.ie/news--events/daily/farm-incomes-rebound-across-all-systems-in-2024/ [retrieved: 2026.05.22]

[46] Irish Farmers Association "Digital Agriculture Technology Adoption Attitudes" available at: https://www.ifa.ie/wp-content/uploads/2020/11/Digital-Ag-Tech-Adoption-Attitudes.pdf [retrieved: 2026.05.22]

[47] Gov.ie "Knowledge Transfer Program" available at: https://www.gov.ie/en/department-of-agriculture-food-and-the-marine/services/knowledge-transfer-programme-2024-2026/ [retrieved: 2026.05.22]

[48] Department of Agriculture, Food and the Marine. TAMS II Support Documents; Department of Agriculture, Food and the Marine: Dublin, Ireland, 2018, Available at: https://www.gov.ie/en/department-of-agriculture-food-and-the-marine/publications/tams-ii-support-documents/ [retrieved: 2026.05.22]

[49] F. D. Davis, "A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results", Doctoral dissertation, Massachusetts Institute of Technology, 1985.

[50] S. Papagiannidis and D. Marikyan, "Environmental sustainability: A technology acceptance perspective", International Journal of Information Management, 63, 102445, 2022, doi:10.1016/j.ijinfomgt.2021.102445

[51] J. Wang, S. Zhang and L. Zhang. "Intelligent Hog Farming Adoption Choices Using the Unified Theory of Acceptance and Use of Technology Model: Perspectives from China's New Agricultural Managers". Agriculture, 13(11), 2067, 2023, doi:10.3390/agriculture13112067

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

From Backscattering Models to Neural Networks: Simulation-Driven Acoustic Identification of Mesopelagic Organisms

Sander Andre Berg Marx[1,*], Babak Khodabandeloo[2,†], Ketil Malde[2,‡]
[1]Department of Informatics, University of Bergen, Bergen, Norway
[2]Acoustics and Observation Methodologies, Institute of Marine Research (IMR), Bergen, Norway
[*]e-mail: sanderamarx@hotmail.com
[†]e-mail: babak.khodabandeloo@hi.no
[‡]e-mail: ketil.malde@hi.no

***Abstract*—Target Strength (TS) is the logarithmic measure of the backscattered acoustic energy reflected toward a sound source when an acoustic wave encounters an organism. It depends on the organism's size, shape, material properties, orientation, and the frequency of the wave, and thus carries information useful for identifying and characterizing marine organisms. Advanced broadband echosounders now allow detailed TS measurements of marine organisms over nearly continuous frequency ranges, which provide valuable information for biomass estimation and ecosystem monitoring. However, interpreting these TS measurements has traditionally relied on manual classification, which makes it difficult to extract biological characteristics for target classification or ecological analysis, especially given the complexity of broadband data. Physics-based backscattering models are versatile tools for modeling the TS frequency response given the shape and material properties of the scatterers. In this study, we employ an exact prolate spheroid model, representative of many marine organisms, to simulate broadband TS spectra for training machine learning models. These models aim to classify and characterize targets based on their TS frequency signatures. A hybrid one-dimensional Convolutional Neural Network (1D-CNN) is proposed for the simultaneous classification (gas- vs. liquid-filled) and regression of geometric properties, and compared against K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF). Furthermore, the model's generalization capability is evaluated against *in situ* broadband measurements of mesopelagic organisms. Results show that while all models achieved perfect classification accuracy, the hybrid 1D-CNN clearly outperformed others in parameter estimation. When applied to real-world data, the model achieved high precision in size estimation (R_{eq}), though challenges remain in predicting elongation. By feeding the 1D-CNN's predictions back into the prolate spheroid model, we observe that the model accurately predicts the frequency of the first resonance peak but struggles with the amplitude and location of the second resonance peak. This discrepancy is attributed to physical damping effects present in real biological tissues but absent in the simulated training data. This demonstrates that simulation-driven machine learning is a promising strategy for overcoming data scarcity, though incorporating realistic biological damping and real-world noise effects remains a critical step toward fully automated acoustic identification.**

Keywords*-Acoustic target classification; machine learning; Convolutional neural network; prolate spheroid backscattering model; simulation-driven learning; in situ validation; real-world acoustic data.

I. Introduction

This article is an extended version of the COCE 2025 conference paper "Machine Learning Using Physics-Based Backscattering Modeling for Acoustic Identification of Mesopelagic Organisms" [1]. Compared with the original conference paper, this journal submission expands the study by adding validation of the model performance on *in situ* real-world acoustic data. Active acoustics is a versatile tool for monitoring marine life, offering unrivaled spatial and temporal resolution compared to other methods, such as net-based biological sampling, optical systems, or video recording. However, interpreting the collected echoes to classify organism size and species remains challenging. Converting acoustic data to biomass is especially difficult when the insonified volume contains mixed species and/or a diverse size distribution within the same species [2][3][4]. Broadband echosounders provide detailed spectra, but these are difficult to interpret manually, making large-scale analysis time-consuming and thereby motivating automated approaches.

When an acoustic wave encounters organisms along its propagation path, the acoustic energy scatters in different patterns depending on the organism's size, shape, material properties, orientation, and the frequency of the wave [5][6]. A portion of this energy is reflected back toward the source, referred to as backscattering. Therefore, the backscattered signal—or its logarithmic measure, known as Target Strength (TS)—carries information that can be used to identify the object from which it originated. Since the TS of an organism varies with frequency, measuring TS over a broad, continuous frequency band provides more detailed information about the scatterers, thereby enhancing the ability to characterize organisms [7]. Broadband fisheries echosounders transmit frequency-modulated pulses, offering two main advantages [8][9]: (1) measurement across a nearly continuous frequency range, and (2) enhanced range resolution through pulse compression, which improves single-target detection. Therefore, the use of broadband echosounders improves the ability to identify organisms and to resolve their spatial distribution within aggregations, both of which are crucial for interpreting and converting acoustic echoes into meaningful biological information.

Backscattering models are an essential tool for interpreting measured backscattered data from marine organisms [10]. Techniques to model acoustic backscattering range from analytical to numerical models. Numerical models can accommodate arbitrary geometries and inhomogeneous material properties,

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

allowing detailed and realistic simulations [11]. However, their high computational cost can limit their practical use, especially when generating massive datasets required to train modern deep learning architectures [12]. To make simulation-driven machine learning feasible, optimizing the execution time of these physics-based models is a critical prerequisite. Although most marine organisms have complex geometries and consist of inhomogeneous materials, backscattering models from canonical geometries with homogeneous material properties are often sufficient to acoustically represent them [13]. For example, spherical models have long been used in fisheries acoustics due to their simplicity [14][15]. However, spherical models cannot capture elongation, which is a critical parameter for accurately representing many marine organisms or their main acoustic reflecting organ, e.g., gas-bladder.

Machine Learning (ML) techniques can be used to classify targets based on trained models that learn to find patterns within the broadband TS spectra. These approaches are generally categorized into unsupervised methods, which group targets based on spectral similarity [16], and supervised methods, which map acoustic signatures to specific biological classes or physical parameters. Supervised learning is typically the most successful approach for precise identification, but it requires large amounts of accurately labeled data to develop robust models [12].

However, acquiring such labels, referred to as "ground truth," is exceptionally challenging for individual fish in the wild. The acoustic scattering environment is highly dynamic; internal structures such as bones, flesh, and the swimbladder create complex impedance contrasts, while constructive and destructive interference across the pulse bandwidth varies significantly as the fish moves. While some studies, such as Kubilius et al. [17], have established ground truth in controlled settings using artificial fish-like targets of known dimensions, obtaining reliable labels for wild organisms remains difficult. Conventional methods rely on invasive biological sampling, such as trawling, which introduces inherent catch biases and provides only a population-level snapshot rather than individual-target validation [17][18].

Furthermore, while real-world broadband acoustic datasets exist, such as the *in situ* measurements used in this study, they are generally limited in scale. These datasets often lack direct physical measurements of the targets and must instead rely on indirect analytical estimations—such as resonance frequency analysis to provide labels. Such estimations, while physically grounded, do not provide the high-confidence ground truth required to train modern, data-intensive deep learning architectures [12].

To overcome these limitations of data scarcity and label uncertainty, we use a physics-based simulation model [19] to generate large labeled synthetic datasets. This paper presents a two-fold contribution: First, we implement and evaluate a computationally optimized version of a prolate spheroid backscattering model to enable efficient data generation. Second, we propose a novel hybrid one-dimensional Convolutional Neural Network (1D-CNN) for the simultaneous classification (gas-filled vs. liquid-filled) and regression of geometric properties such as size, elongation, and orientation.

The performance of the 1D-CNN is then compared against traditional ML models. The method is also tested on a small *in situ* broadband dataset of mesopelagic organisms, specifically small gas-bearing fish, from Khodabandeloo et al. [20]. This allows us to evaluate the model's ability to generalize from idealized simulations to noisier, real-world acoustic measurements.

The remainder of the paper is organized as follows. Section II describes the methods for computational optimization, data generation, and machine learning model training. Section III reports the results on both synthetic and *in situ* data, and Section IV provides a discussion of the findings, including the challenges of real-world generalization. Section V concludes the paper and outlines directions for future work.

II. Methods

A. Physics-Based Acoustic Backscattering Modeling for Synthetic Data Generation

To generate synthetic broadband backscattering data from objects representing mesopelagic marine organisms, with and without gas bladders, we employed an optimized version of the fluid-filled prolate spheroid backscattering model [19]. This physics-based model provides accurate backscattering for liquid- and gas-filled targets over a wide frequency range and for all incident angles, i.e., the relative orientation of the prolate spheroid axis to the direction of wave propagation.

The prolate spheroids are represented by their volume (quantified by the equivalent spherical radius $\boldsymbol{R_{\mathrm{eq}}}$) and elongation (or aspect ratio α) instead of directly using semi-major and semi-minor axes, a and b, respectively. They are related by the following equations:

$$b = \frac{R_{\mathrm{eq}}}{\alpha^{1/3}}, \tag{1}$$

$$a = \alpha \times b. \tag{2}$$

This parameterization facilitates direct control over the size and shape of the modeled organisms, making it easier to explore a wide range of biologically relevant geometries. The range of parameters used to model different organisms are given in Table I.

TABLE I. Parameter ranges used for generating synthetic acoustic data

Parameter	Targets	
	Gas-filled	**Liquid-filled**
Equivalent radius (R_{eq})	0.1–5 mm	5–20 mm
Aspect ratio (α)	1.05–8.0	1.05–8.0
Spheroid density (ρ_s)	2–80 kg/m^3	1.01–1.07 $\times\rho_w$ kg/m^3
Water density (ρ_w)	1027 kg/m^3	1027 kg/m^3
Spheroid sound speed (c_s)	343 m/s	1.01–1.07 $\times c_w$ m/s
Water sound speed (c_w)	1500 m/s	1500 m/s
Incident angle (θ)	0.01–90°	0.01–90°
Frequency range	10–260 Hz (0.5 kHz steps)	

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

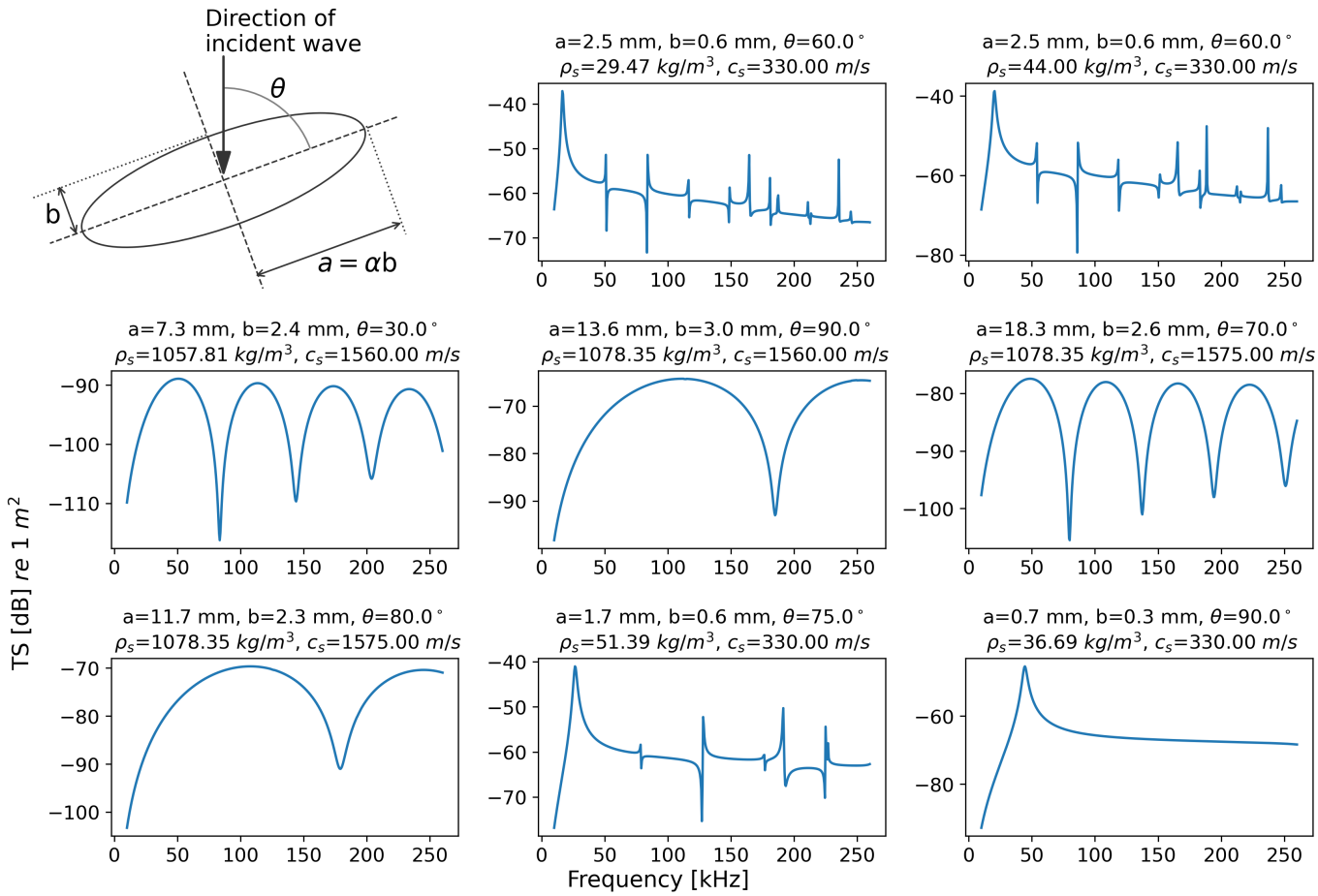


Figure 1. Examples of TS(f) for prolate spheroids with parameters within the ranges in Table I. An illustration of a prolate spheroid with semi-major and semi-minor axes, a and b respectively, along with the incident angle θ, is shown in the upper left corner.

B. Computational Optimization

To enable the generation of large-scale datasets, the original simulation code, a hybrid of Python and Fortran was refactored for efficiency. We performed code profiling to identify computational bottlenecks, which were primarily located in nested Python loops managing frequency-point calculations. To resolve this, we implemented NumPy vectorization, replacing iterative Python loops with array-based operations optimized in C. This approach allowed for the simultaneous calculation of entire frequency bands rather than sequential point-by-point processing. Furthermore, the optimized code was designed to run in parallel across 200 CPU cores on a high-performance computing cluster to maximize data throughput as the frequency-point calculations are independent.

C. Data Preprocessing and Machine Learning Setup

The synthetic acoustic data generated by the prolate spheroid backscattering model underwent several preprocessing steps to prepare it for machine learning. The dataset consists of 30,000 simulated TS spectra, 15,000 liquid-filled and 15,000 gas-filled targets.

Each target is represented by a feature vector with corresponding labels. The input features consist of the broadband TS spectra. Each spectrum is represented as a vector of TS values, sampled at 0.5 kHz steps across the frequency range of 10 to 260 kHz. The models were trained to predict a classification label (liquid-filled or gas-filled) and three regression labels corresponding to the geometric properties of the prolate spheroid: the incident angle (θ), the semi-major axis (a), and the semi-minor axis (b).

Normalization procedures are applied separately to the feature and label columns to ensure consistent scaling and improve model stability [21]. Each frequency TS value is standardized independently using training set statistics, transforming the data to have zero mean and unit variance. The continuous regression labels (θ, a and b) are also normalized to ensure that each parameter contributes equally to model training regardless of original scale.

The complete dataset of 30,000 target instances is then partitioned into training, validation, and test subsets using an 80/10/10 split. The training set (80%) is used for training the ML models and deriving normalization statistics. The validation set (10%) is used for hyperparameter tuning and model selection during training, as well as for monitoring overfitting. Finally, the test set (10%) is reserved for the final, unbiased evaluation of the trained models' performance.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

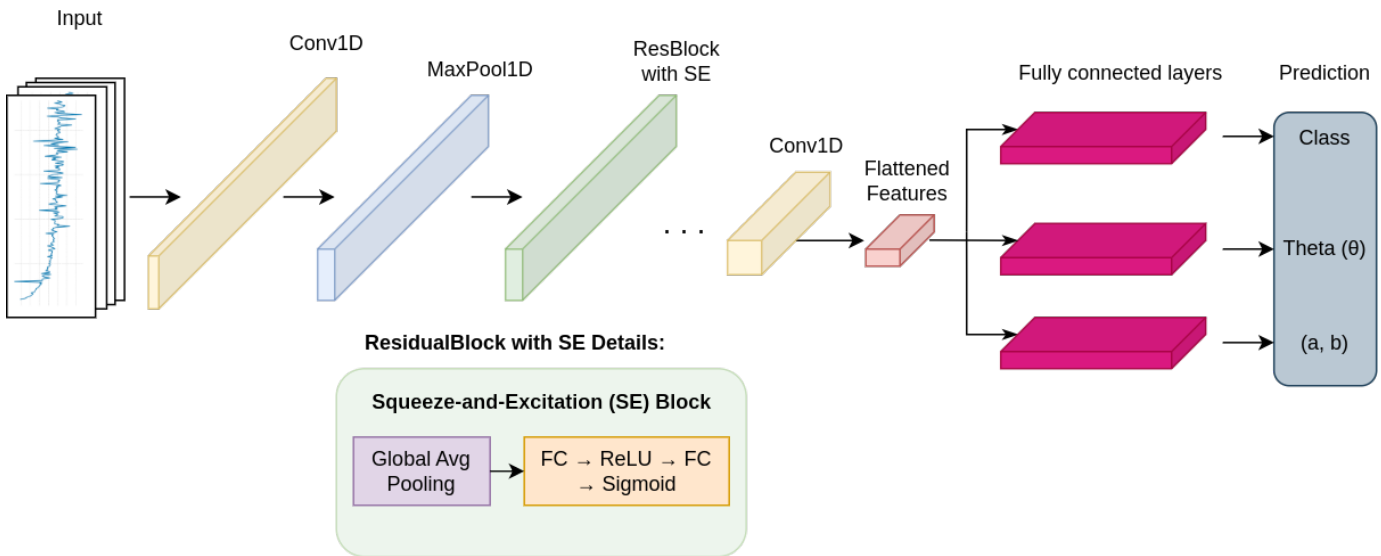


Figure 2. Hybrid one-dimensional convolutional neural network architecture for acoustic target classification.

D. Machine Learning Models

To classify the type and estimate geometric parameters from the TS spectra, a hybrid 1D-CNN was developed. The architecture was chosen for its effectiveness in automatically extracting hierarchical features from sequential data, such as acoustic frequency spectra [22].

The network architecture, illustrated in Figure 2, was designed to process sequential TS data. The architecture incorporates advanced deep learning components, specifically residual blocks [23] and Squeeze-and-Excitation (SE) attention mechanisms [24], to improve feature extraction and training stability. The TS spectra are first processed through an initial 1D convolution layer, followed by a series of residual blocks. The residual blocks contain two 1D convolutional layers, batch normalization [25], and ReLU activations. Squeeze-and-Excitation (SE) attention mechanisms, which perform channel-wise feature weighting by compressing information through global average pooling and then learning channel relationships via two fully connected layers, are also integrated into these residual blocks. The network ends in fully connected layers that lead to a multi-head output structure with dedicated heads for classification and regression: one head for target-type classification, one for regression of the incident angle (θ), and one for regression of the size parameters (a and b).

For comparative analysis of the performance of the 1D-CNN, three additional traditional machine learning models previously used in acoustic target classification research were tested.

- K-Nearest Neighbors (KNN) [26]: An instance-based learner classifying targets based on the majority class of their k nearest neighbors in the feature space, as applied by Cotter et al. [27].
- Support Vector Machine (SVM) [28]: Seeking an optimal hyperplane to separate classes or predict continuous values, utilized by Yang et al. [29] for underwater target recognition.
- Random Forest (RF) [30]: An ensemble method constructing multiple decision trees on random subsets of data and features, aggregating their predictions, as employed by Gugele et al. [31].

These traditional models were not hyperparameter tuned and served as a baseline to evaluate the 1D-CNN.

E. In Situ Generalization Setup

To evaluate real-world performance, a second experiment was designed using an *in situ* broadband dataset of nine single mesopelagic organisms [20], interpreted as swimbladdered mesopelagic fish (gas-bearing targets) based on their characteristic resonance structure. Because the targets were identified acoustically, species-level labels were not available; we therefore treat them as gas-bearing organisms dominated acoustically by a swimbladder-like gas inclusion.

The simulation parameters were adjusted to match the characteristics of these targets. The training dataset was constrained to gas-filled targets with a semi-major axis $a \leq$ 1 mm. To account for the depth-dependent gas compression, the spheroid density (ρ_s) was calculated using the Van der Waals equation [32][33] based on the target depths (140 to 575 meters). Gaussian noise was added to the synthetic training data to replicate environmental and instrumental noise present in the *in situ* data. Finally, the frequency range was adjusted to 38–380 kHz, including gaps to replicate the echosounder channels used during the *in situ* collection.

F. Training and Evaluation

The 1D-CNN was trained using a custom multi-component loss function designed to address both the classification and regression tasks. The total loss ($\mathcal{L}$) is a weighted sum of four components, as shown in (3):

$$\mathcal{L} = \omega_c \, \mathcal{L}_{\mathrm{c}} + \omega_r \left(\mathcal{L}_\theta + \mathcal{L}_{ab}\right) + \omega_{cons} \, \mathcal{L}_{\mathrm{cons}} \tag{3}$$

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

where $\mathcal{L}_c$ is cross-entropy loss for the binary classification task. For regression, Huber loss [34] was used for the angle ($\mathcal{L}_\theta$) and Mean Squared Error (MSE) for the size parameters ($\mathcal{L}_{ab}$). To enforce the geometric relationship between the semi-major and minor axes, a consistency loss ($\mathcal{L}_{\text{cons}}$), also based on Huber loss, was applied by comparing the predicted aspect ratio ($\alpha_{pred} = a_{pred}/b_{pred}$) to its ground truth value. The weights were determined through hyperparameter tuning and set to ω_c = 0.1, ω_r = 2.0, and ω_{cons} = 0.5 to prioritize the more challenging regression task.

Model performance was evaluated on the unseen test set. Classification performance was measured by accuracy, while regression performance was assessed using the coefficient of determination (R^2) and Root Mean Squared Error (RMSE) for θ, a, b, and the derived α. A composite score, averaging the classification accuracy and the R^2 scores of the three primary regression targets, was used for overall model comparison.

III. Results

A. Optimization Results

Before training the ML models, the performance of the NumPy-vectorized scattering model was benchmarked against the original implementation. A "frequency point" is defined as a single calculation of the backscattering cross-section at a specific frequency; a full TS spectrum in this study consists of over 500 such points.

To generate the 39,400 spectra required for this study, the optimized code was executed in parallel across 200 CPU cores on the "Birget" cluster, for a total wall-clock time of 16 days. This equates to approximately 77,000 total CPU hours. Based on the benchmarked 47.6-fold speedup, which reduced the average computation time per frequency point from 334.94 seconds to 7.04 seconds, we can estimate the impact of this optimization on the project timeline.

Using the same 200-core parallel infrastructure, the non-optimized implementation would have required an estimated 760 days of continuous cluster usage to generate the same dataset. This estimate is approximate, as the computational cost varies with target parameters (e.g., size, orientation, and whether the inclusion is gas- or liquid-filled).

The original implementation is used here as a benchmarking reference to quantify relative performance improvements, rather than as an absolute measure of computational efficiency under all conditions. Even allowing for variability, the results clearly indicate that the NumPy-vectorized optimization was a necessary prerequisite for generating the 39,400 spectra within a realistic research timeframe. For more details of the benchmark, see Figure 17 in Khodabandeloo et al. [19].

B. Performance Comparison Across Models (Synthetic data)

Using the simulated backscattering frequency responses, we first classified the targets (i.e., liquid- or gas-filled), then estimated the geometrical properties (semi-major and semi-minor axes) and the incident angle of the prolate spheroids based on their wideband target strength frequency responses, using different ML models. Since the data are simulated, the true model parameters are known, allowing us to quantitatively evaluate the estimates produced by the various ML models. The performance of the models in both classification and morphological parameter estimation is summarized in Table II.

TABLE II. Comparative regression performance (R^2 scores), classification accuracy, and composite score on synthetic test data

Parameter	SVM	KNN	RF	1D-CNN
R^2 for θ	0.742	0.751	0.853	**0.974**
R^2 for a	0.795	0.898	0.902	**0.997**
R^2 for b	0.911	0.931	0.942	**0.996**
Clas. Acc.	1.000	1.000	1.000	**1.000**
Composite score	0.719	0.804	0.867	**0.992**

C. Detailed Performance of Hybrid 1D-CNN

All tested models achieved a perfect classification accuracy of 1.00, correctly identifying every target as either gas-filled or liquid-filled. The primary differences between the models emerged in the regression tasks. Given its superior regression performance, the results of the hybrid 1D-CNN were analyzed in further detail. Figure 3 presents scatter plots of the model's predicted values against the true values for each regression target.

The model demonstrated high precision in predicting the semi-major axis a and semi-minor axis b; data points were tightly clustered along the identity line for both gas-filled and liquid-filled targets, indicating the model's robustness across target types and size ranges. Predictions for the incident angle θ were also strong, though with more visible scatter compared to the size parameters a and b, with notable gas-filled outliers. The derived aspect ratio ($\alpha = a/b$) also showed good performance, but with some increased scatter for targets with higher aspect ratios (α > 4).

D. Analysis of Prediction Outliers

Outlier analysis revealed consistent patterns for both incident angle (θ) and aspect ratio (α).

For incident angle (θ) estimation, the largest errors occurred with small, gas-filled targets with simple TS spectra containing few distinct resonance peaks. The model therefore defaulted toward predicting the training mean (45°), creating the horizontal band of outliers seen in Figure 3. In contrast, liquid-filled outliers showed smaller angle errors and were typically targets with low aspect ratio ($\alpha \approx 1.05$), where the orientation is less defined physically as the target approaches a perfect sphere.

A similar pattern emerged for aspect ratio (α) prediction, where most outliers were again small, gas-filled targets with spectrally simple signatures. This result aligns with the physical principle that accurate estimation of elongation requires at least two distinct resonance peaks [35], a feature these outlier spectra lacked.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

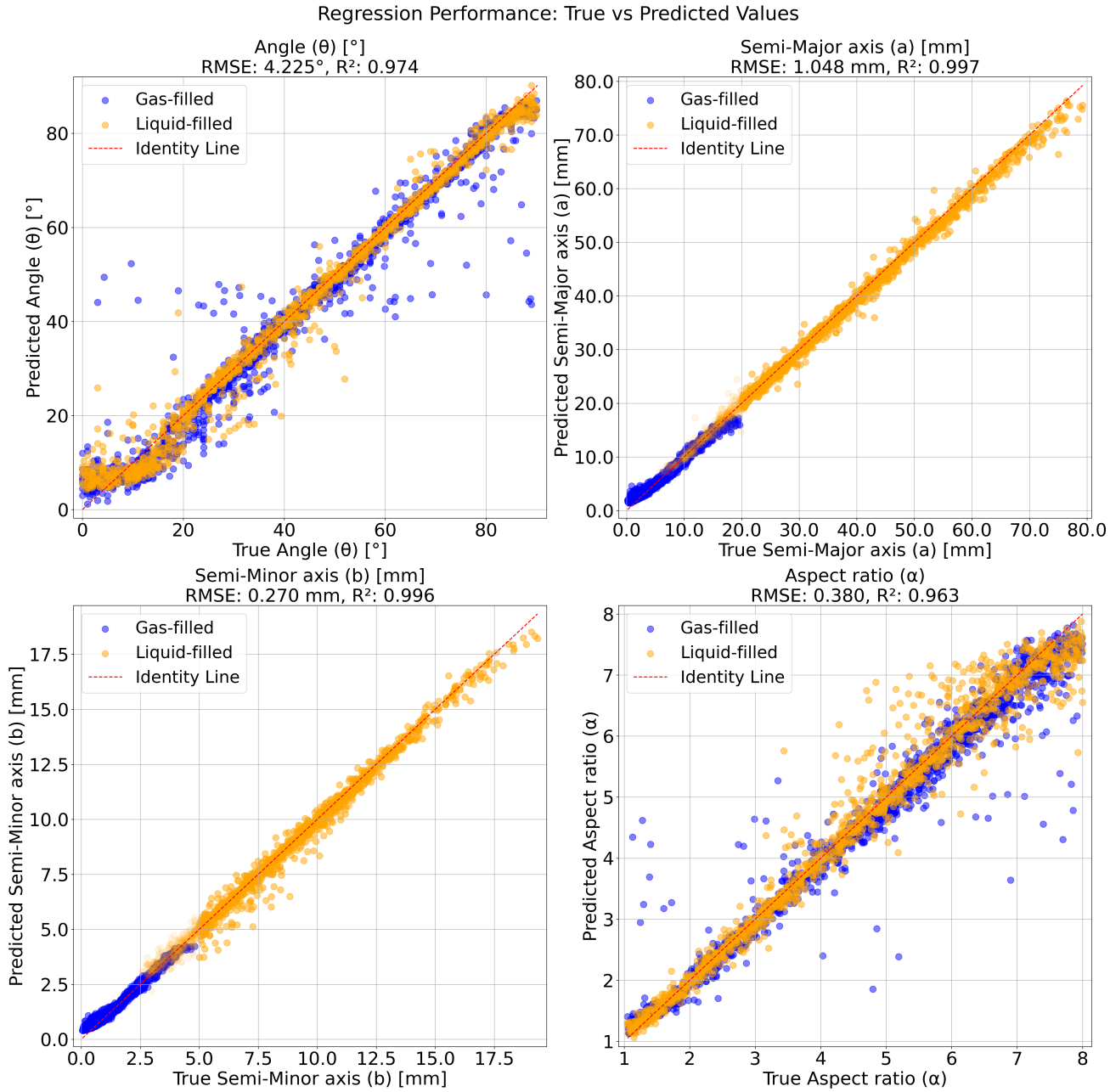


Figure 3. Regression performance of the hybrid 1D-CNN model on test data showing true values against predicted values for (top-left) angle (θ), (top-right) semi-major axis (a), (bottom-left) semi-minor axis (b), and (bottom-right) derived aspect ratio ($\alpha = a/b$). Blue points represent gas-filled targets, and orange points represent liquid targets.

E. Performance of In Situ Measurements

To test real-world generalization, the 1D-CNN trained on simulated data was applied to twelve *in situ* acoustic targets. For the first nine targets, the first and second resonance peaks, as well as corresponding estimates of size and elongation, were obtained from Khodabandeloo et al. [20]. The model's predictions of equivalent radius (R_{eq}) and aspect ratio (α) were then compared against values derived from resonance analysis on targets 1-9 [20], as detailed in Table III.

The model demonstrated high accuracy in predicting the equivalent radius (R_{eq}), achieving a Mean Absolute Error (MAE) of only 0.042 mm. Target 3, for instance, showed a radius error of 0.000 mm, while the largest error (Target 6) was limited to 0.106 mm. Aspect ratio (α) estimation was more variable, with an MAE of 0.94; Target 7 showed the highest accuracy ($|\varepsilon_\alpha| = 0.14$), while Target 9 showed the largest deviation ($|\varepsilon_R| = 3.06$).

To visualize these predictions, Figure 4 compares the measured broadband spectra against two simulated baselines. The black lines represent the actual *in situ* measurements. The red lines represent the spectra generated using the parameters predicted by our 1D-CNN. For benchmarking, the blue dashed

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

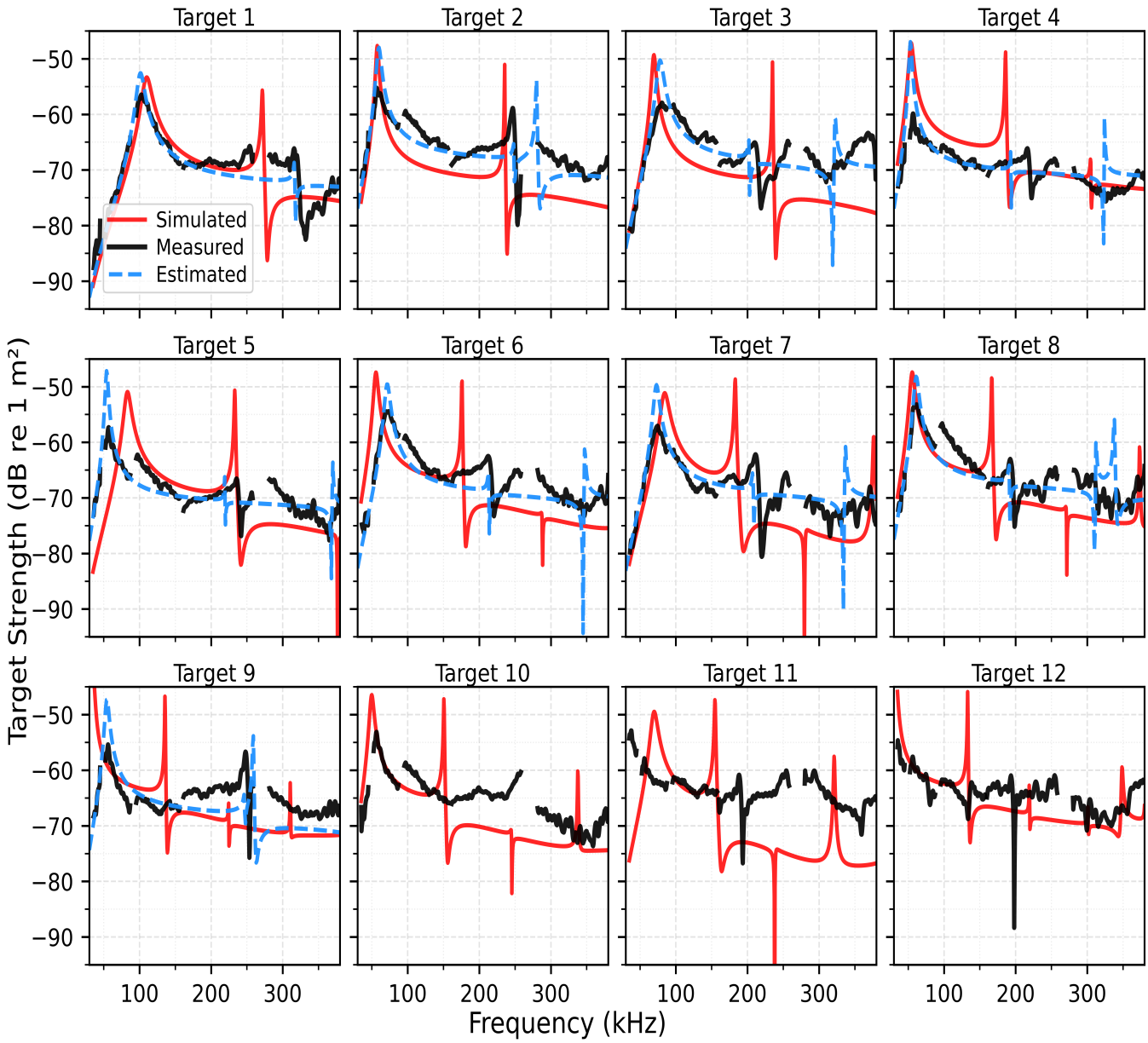


Figure 4. Comparison of measured *in situ* target strength spectra (black) against two simulated models. The red solid line represents spectra generated using the 1D-CNN predicted parameters. The blue dashed line represents spectra generated using the reference estimations from [35] with a fixed tilt angle of $\theta = 70°$.

lines show synthetic spectra generated using the analytical estimations from Khodabandeloo et al. [20], calculated with a fixed tilt angle of $\theta = 70°$. This comparison includes the nine labeled targets and three additional targets (10–12) for which reliable ground truth was not available due to incomplete resonance information.

For the majority of targets (1–8), the model closely predicted the frequency location of the first resonance peak. In several cases (e.g., Targets 2 and 4), the ML prediction aligns closely with the reference estimation (blue), confirming that the neural network has successfully learned to approximate the location of the first resonance peak. However, a systematic discrepancy is evident in the amplitude of the backscatter. As seen in Figure 4, the simulated spectra (red and blue) consistently exhibit sharper and higher-amplitude peaks than the measured *in situ* data (black). This "amplitude gap" indicates that while the model successfully learned the frequency-dependent features related to size, the physics-based training data lacked the damping factors present in real biological tissues. Consequently, the model predicts the correct location of the first resonance peak but the amplitude is exaggerated.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE III. PERFORMANCE OF REGRESSION MODEL PREDICTIONS AGAINST *in situ* MEASUREMENTS. LABELS FOR ELONGATION $(\alpha+\beta)/2$ AND RADIUS R_{ES} ARE TAKEN FROM KHODABANDELOO ET AL. [35]. ERRORS ARE ABSOLUTE DIFFERENCES BETWEEN PREDICTIONS AND LABELS.

	Predictions					Labels		Abs. Errors	
Target	$\boldsymbol{\theta}$ (°)	$\boldsymbol{a}$ (mm)	$\boldsymbol{b}$ (mm)	$\boldsymbol{\alpha} = \frac{a}{b}$	$\boldsymbol{R}_{\mathrm{eq}}$ (mm)	$(\boldsymbol{\alpha}+\boldsymbol{\beta})/2$	$\boldsymbol{R}_{ES}$ (mm)	$\lvert\boldsymbol{\varepsilon}_{\alpha}\rvert$	$\lvert\boldsymbol{\varepsilon}_{R}\rvert$ (mm)
1	36.55	0.54	0.14	3.83	0.220	2.43	0.221	1.40	0.001
2	67.23	0.69	0.42	1.67	0.492	1.16	0.424	0.51	0.068
3	61.21	0.87	0.20	4.32	0.328	2.93	0.328	1.39	0.000
4	41.16	0.54	0.17	3.13	0.250	3.68	0.264	0.55	0.014
5	54.68	0.56	0.14	3.88	0.225	2.90	0.268	0.98	0.043
6	62.17	0.76	0.32	2.36	0.427	2.57	0.321	0.21	0.106
7	37.43	0.60	0.20	2.94	0.290	2.80	0.316	0.14	0.026
8	56.06	0.81	0.36	2.26	0.471	2.06	0.408	0.20	0.063
9	84.09	1.01	0.25	4.12	0.392	1.06	0.446	3.06	0.054
10	71.92	0.89	0.40	2.25	0.520	-	-	-	-
11	82.30	0.94	0.66	1.42	0.742	-	-	-	-
12	78.77	0.94	0.54	1.73	0.649	-	-	-	-

IV. DISCUSSION

A. Analysis of Model Performance

The tested machine learning models, including shallow learners (KNN, SVM, and RF) and a deep learner (1D-CNN), successfully classified gas- versus liquid-filled targets. This is primarily due to the distinct differences in the TS(f) responses between gas- and liquid-filled targets, as observed in Figure 1. In contrast, the models showed varying levels of performance in the regression task. Among them, the 1D-CNN demonstrated the highest accuracy in estimating the model parameters, evidenced by high R^2 scores for a, b, and θ (see Table II).

However, some outliers were observed in the predictions, where the parameters were not correctly estimated. This was especially the case for gas-filled targets in the prediction of angle θ and the computed aspect ratio α. Further investigation of the outliers revealed that they mostly corresponded to small targets (i.e., small $\boldsymbol{R}_{\mathbf{eq}}$) (see [36] for more details). The TS frequency response of small gas-filled targets is known to be insensitive to shape and incident angle near the resonance frequency [37], explaining the model's difficulty in the 10–260 kHz frequency band. An example TS(f) response of such small targets is shown in the lower right panel of Figure 1. For deep learning approaches such as neural networks, large amounts of data are typically required; therefore, the use of optimized code for computing TS(f) was critical. Although the shallow learners achieved lower R^2 scores, it is important to note that less effort was devoted to tuning these models compared to the 1D-CNN. It is possible that more extensive hyperparameter optimization and validation could lead to improved performance for the shallow models.

Despite the strong baseline results, the application of the model to *in situ* measurements (Targets 1–12) revealed the complexities of the "reality gap" between idealized physics-based simulations and real-world acoustic environments.

B. The Reality Gap: Frequency vs. Amplitude

The most significant finding from the *in situ* evaluation (Experiment 2) is the model's high accuracy in predicting the equivalent radius ($\boldsymbol{R}_{\mathbf{eq}}$), achieving an MAE of 0.042 mm. This success is directly linked to the model's ability to identify the frequency location of the first resonance peak (see Figure 4). This suggests that the model successfully learned size-dependent spectral features that are robust enough to transfer from idealized simulations to real-world measurements.

However, as seen in the spectral reconstructions, an "amplitude gap" persists where the model consistently overestimates Target Strength. This discrepancy likely stems from the physical limitations of the fluid-filled prolate spheroid model, which does not account for damping effects caused by, for example, surrounding tissue that absorbs part of the scattering due to viscous mechanisms that affect the peak amplitudes [38] of real marine organisms [37]. Furthermore, while the model identifies the first peak well, it struggles to precisely locate the second resonance peak in most targets (with the exception of Target 5). In contrast, the synthetic spectra generated from manual estimations (blue dashed lines) align more closely with these secondary features.

This observation extends to targets 10, 11, and 12, which lack established ground truth because they do not have a clear second resonance peak. For Target 10, the model accurately predicts the first resonance, and for Target 12, it captures the second peak location, yet both suffer from exaggerated amplitudes. Future iterations should improve accuracy by incorporating more sophisticated material properties into the scattering simulations. In addition, subsequent models should place greater emphasis on invariant features. Specifically, the relative ratio and location of the first and second resonance peaks can be identified independently of the absolute amplitude of the frequency response and are not affected by organism orientation [19].

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

C. Scalability vs. Explainability

There is a clear operational advantage to using the 1D-CNN approach (red lines) over traditional manual resonance estimation (blue lines). Manual estimation is a labor-intensive process requiring significant human expertise and time per target, making it impractical for the large-scale, real-time datasets generated by modern broadband echosounders. Once trained, the ML model can provide near-instantaneous parameter estimations.

However, this efficiency introduces a trade-off in explainability. Manual estimation allows a researcher to justify a classification based on visible physical indicators (e.g., "the distance between these two peaks implies this elongation"). In contrast, deep learning architectures are often viewed as "black boxes" [39]. While the 1D-CNN is faster, the inability to easily verify the logic behind its predictions remains a challenge.

D. Computational Feasibility

Finally, the 47.6-fold speedup achieved through NumPy vectorization was a critical enabler of this research. Deep learning architectures are data-intensive; without the ability to generate tens of thousands of spectra in a reasonable timeframe, the training of the hybrid 1D-CNN would have been unfeasible. This optimization was the result of a systematic evaluation of several techniques. Initial attempts to utilize GPU acceleration through libraries like CuPy [40] and PyTorch [41] provided negligible improvements, as profiling revealed that the overhead of frequent data transfers between the CPU and GPU created a significant performance bottleneck. Similarly, standard Python concurrency strategies yielded minor speed improvements; however, large-scale concurrency proved challenging due to several constraints. The Fortran-based subprocesses created Input/Output bottlenecks, while file-access contention and sequential dependency on externally generated files limited performance gains. Additionally, computational complexity increased with frequency, creating an imbalanced workload distribution. These challenges made further concurrency ineffective without extensive refactoring of the underlying code.

Vectorization with NumPy emerged as the most effective solution by replacing iterative Python loops with array-based operations optimized in C. This approach allowed for the simultaneous calculation of entire frequency bands while maintaining identical numerical precision to the original code. This demonstrates that in simulation-driven ML, technical optimization of the physics engine is just as vital as the neural network architecture itself. While the current implementation was sufficient for this study, future iterations requiring more complex, inhomogeneous geometries may benefit from just-in-time (JIT) compilation or low-level C++ implementations to further bridge the gap between computational cost and model complexity. These methods could offer further speedups beyond standard vectorization, facilitating the modeling of even more complex marine organisms without the prohibitive computational costs.

V. Conclusion and Future Work

This study demonstrates that machine learning, specifically a 1D-CNN trained exclusively on physics-based simulated data, can accurately classify and estimate geometric properties of acoustic targets. We demonstrated that the hybrid 1D-CNN outperforms traditional machine learning methods such as KNN, SVM, and RF. While all models achieved perfect classification of gas- and liquid-filled targets, the 1D-CNN was notably more accurate in the regression task of estimating the targets' geometric properties, including semi-major/minor axes (a, b) and incident angle (θ).

The feasibility of this approach was facilitated by the computational optimization of the scattering model through NumPy vectorization. By achieving a 47.6-fold speed increase, it was possible to generate the massive synthetic datasets required for deep learning within a practical timeframe.

Outlier analysis revealed specific challenges, particularly in estimating geometric parameters of small, gas-filled targets, with simpler spectral features lacking distinct resonance peaks. Addressing these limitations by expanding frequency ranges or incorporating additional acoustic parameters in future research could enhance performance further.

Validation on real-world *in situ* measurements showed that the model can generalize to noisy environments, particularly for size estimation, where it achieved a mean absolute error for the equivalent radius of only 0.042 mm. Addressing current limitations, such as the overestimation of peak amplitudes by incorporating damping effects and more realistic scattering effects in future research could enhance performance further.

We conclude that this simulation-driven approach is a powerful and viable strategy for overcoming data scarcity in marine acoustics. It represents a promising step toward the automated, non-invasive classification of marine organisms, with potential application in real-time classification of mesopelagic organisms during acoustic surveys, supporting biomass estimation and reducing the need for manual analysis.

Acknowledgment

S. M. acknowledges the use of the "Birget" high-performance computing cluster, provided by the Department of Informatics, University of Bergen, which was essential for this research. B. K. and K. M. gratefully acknowledge funding from the Center for Research-based Innovation in Marine Acoustic Abundance Estimation and Backscatter Classification (CRIMAC, Project No. 309512; www.crimac.no). This work is based on the master's thesis of S. M. conducted at the University of Bergen.

References

[1] S. A. B. Marx, B. Khodabandeloo, and K. Malde, "Machine learning using physics-based backscattering modeling for acoustic identification of mesopelagic organisms", in *Proceedings of COCE 2025: The Second International Conference on Technologies for Marine and Coastal Ecosystems*, Barcelona, Spain: IARIA, Oct. 2025, pp. 32–37, ISBN: 978-1-68558-329-3.

[2] C. F. Greenlaw, "Acoustical estimation of zoo-plankton populations", *Limnology and Oceanography*, vol. 24, pp. 226–242, 1979.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[3] A. C. Lavery et al., "Determining dominant scatterers of sound in mixed zooplankton populations", *J. Acoust. Soc. Am.*, vol. 122, pp. 3304–3326, 2007.

[4] R. J. Kloser, T. E. Ryan, J. W. Young, and M. E. Lewis, "Acoustic observations of micronekton fish on the scale of an ocean basin: potential and challenges", *ICES J. Mar. Sci.*, vol. 66, pp. 998–1006, 2009. DOI: 10.1093/icesjms/fsp077

[5] R. Hickling, "Analysis of echoes from a solid elastic sphere in water", *J. Acoust. Soc. Am.*, vol. 34, pp. 1582–1592, 1962. DOI: 10.1121/1.1909055

[6] J. J. Faran, "Sound scattering by solid cylinders and spheres", *J. Acoust. Soc. Am.*, vol. 23, pp. 405–417, 1951. DOI: 10.1121/1.1906780

[7] D. Chu and T. K. Stanton, "Application of pulse compression techniques to broadband acoustic scattering by live individual zooplankton", *J. Acoust. Soc. Am.*, vol. 104, pp. 39–55, 1998. DOI: 10.1121/1.424056

[8] M. E. Zakharia, "Wide band fisheries sounder; From individual echoes analysis to classification of schools at sea", *J. Acoust. Soc. Am.*, vol. 103, pp. 3068–3068, 1998. DOI: 10.1121/1.422849

[9] J. E. Ehrenberg and T. C. Torkelson, "FM slide (chirp) signals: a technique for significantly improving the signal-to-noise performance in hydroacoustics assessment systems", *Fisheries Research*, vol. 47, pp. 193–199, 2000.

[10] A. C. Lavery, T. K. Stanton, D. E. McGehee, and D. Chu, "Three-dimensional modeling of acoustic backscattering from fluid-like zooplankton", *J. Acoust. Soc. Am.*, vol. 111, pp. 1197–1210, 2002. DOI: 10.1121/1.1433813

[11] J. M. Jech et al., "Comparisons among ten models of acoustic backscattering used in aquatic ecosystem research", *J. Acoust. Soc. Am.*, vol. 138, pp. 3742–3764, 2015. DOI: 10.1121/1.4937607

[12] L. Alzubaidi et al., "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications", *Journal of Big Data*, vol. 10, p. 46, 2023, Art. 46. DOI: 10.1186/s40537-023-00727-2

[13] R. Adelman, N. A. Gumerov, and R. Duraiswami, "Semi-analytical computation of acoustic scattering by spheroids and disks", *The Journal of the Acoustical Society of America*, vol. 136, no. 6, EL405–EL410, Nov. 2014, ISSN: 0001-4966. DOI: 10.1121/1.4901318 eprint: https://pubs.aip.org/asa/jasa/article-pdf/136/6/EL405/13131806/el405_1_online.pdf.

[14] H. Medwin, *Sounds in the sea: From ocean acoustics to acoustical oceanography*. Cambridge University Press, 2005.

[15] R. H. Love, "Resonant acoustic scattering by swimbladder-bearing fisha)", *The Journal of the Acoustical Society of America*, vol. 64, no. 2, pp. 571–580, Aug. 1978, ISSN: 0001-4966. DOI: 10.1121/1.382009 eprint: https://pubs.aip.org/asa/jasa/article-pdf/64/2/571/11724662/571_1_online.pdf.

[16] M. D. Agersted, B. Khodabandeloo, Y. Liu, W. Melle, and T. A. Klevjer, "Application of an unsupervised clustering algorithm on in situ broadband acoustic data to identify different mesopelagic target types", *ICES Journal of Marine Science*, vol. 78, no. 8, pp. 2907–2921, Aug. 2021, ISSN: 1054-3139. DOI: 10.1093/icesjms/fsab167 eprint: https://academic.oup.com/icesjms/article-pdf/78/8/2907/41764791/fsab167.pdf.

[17] R. Kubilius, G. J. Macaulay, and E. Ona, "Remote sizing of fish-like targets using broadband acoustics", *Fisheries Research*, vol. 228, p. 105 568, 2020, Art. 105568. DOI: 10.1016/j.fishres.2020.105568

[18] S. Kaartvedt, A. Staby, and D. Aksnes, "Efficient trawl avoidance by mesopelagic fishes causes large underestimation of their biomass", *Marine Ecology Progress Series*, vol. 456, pp. 1–6, Jun. 2012. DOI: 10.3354/meps09785

[19] B. Khodabandeloo, Y. Heggelund, B. Ystad, S. A. B. Marx, and G. Pedersen, "High-precision model and open-source software for acoustic backscattering by liquid- and gas-filled prolate spheroids across a wide frequency range and incident angles: Implications for fisheries acoustics", *Journal of Sound and Vibration*, vol. 616, p. 119 227, 2025, ISSN: 0022-460X. DOI: https://doi.org/10.1016/j.jsv.2025.119227

[20] B. Khodabandeloo et al., "Mesopelagic fish gas bladder elongation, as estimated from wideband acoustic backscattering measurements", *The Journal of the Acoustical Society of America*, vol. 151, no. 6, pp. 4073–4085, Jun. 2022, ISSN: 0001-4966. DOI: 10.1121/10.0011742 eprint: https://pubs.aip.org/asa/jasa/article-pdf/151/6/4073/16529753/4073_1_online.pdf.

[21] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance", *Applied Soft Computing*, vol. 97, p. 105 524, 2020, Art. 105524. DOI: 10.1016/j.asoc.2019.105524

[22] S. Kiranyaz et al., "1D convolutional neural networks and applications: A survey", *Mechanical Systems and Signal Processing*, vol. 151, p. 107 398, 2021, Art. 107398. DOI: 10.1016/j.ymssp.2020.107398

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks", in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745

[25] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift", in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, vol. 37, 2015, pp. 448–456.

[26] E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties", *International Statistical Review / Revue Internationale de Statistique*, vol. 57, pp. 238–247, 1989. DOI: 10.2307/1403797

[27] E. Cotter, C. Bassett, and A. Lavery, "Classification of broadband target spectra in the mesopelagic using physics-informed machine learning", *J. Acoust. Soc. Am.*, vol. 149, pp. 3889–3901, 2021. DOI: 10.1121/10.0005114

[28] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines", *IEEE Intelligent Systems and Their Applications*, vol. 13, pp. 18–28, 1998. DOI: 10.1109/5254.708428

[29] H. Yang et al., "Underwater acoustic target recognition using SVM ensemble via weighted sample and feature selection", in *Proc of the 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2016, pp. 522–527. DOI: 10.1109/IBCAST.2016.7429928

[30] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324

[31] S. M. Gugele et al., "Differentiation of two swim bladdered fish species using next generation wideband hydroacoustics", *Scientific Reports*, vol. 11, p. 10 520, 2021, Art. 10520. DOI: 10.1038/s41598-021-89941-7

[32] B. Khodabandeloo, M. D. Agersted, T. Klevjer, G. J. Macaulay, and W. Melle, "Estimating target strength and physical characteristics of gas-bearing mesopelagic fish from wideband in situ echoes using a viscous-elastic scattering model", *The Journal of the Acoustical Society of America*, vol. 149, no. 1, pp. 673–691, Jan. 2021, ISSN: 0001-4966. DOI: 10.1121/10.0003341 eprint: https://pubs.aip.org/asa/jasa/article-pdf/149/1/673/19745391/673_1_online.pdf.

[33] Y. A. Cengel and M. A. Boles, *Thermodynamics: An Engineering Approach*, 4th ed. Singapore: McGraw-Hill, 2002.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[34] P. J. Huber, "Robust Estimation of a Location Parameter", *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964. DOI: 10.1214/aoms/1177703732

[35] B. Khodabandeloo et al., "Mesopelagic fish gas bladder elongation, as estimated from wideband acoustic backscattering measurements", *J. Acoust. Soc. Am.*, vol. 151, pp. 4073–4085, 2022. DOI: 10.1121/10.0011742

[36] S. A. B. Marx, "Learning acoustic target classification from simulation", Master's Thesis, University of Bergen, 2025.

[37] B. Khodabandeloo, M. D. Agersted, T. A. Klevjer, G. Pedersen, and W. Melle, "Mesopelagic flesh shear viscosity estimation from in situ broadband backscattering measurements by a viscous–elastic model inversion", *ICES Journal of Marine Science*, vol. 78, no. 9, pp. 3147–3161, 2021. DOI: 10.1093/icesjms/fsab183

[38] C. Feuillade and R. W. Nero, "A viscous-elastic swimbladder model for describing enhanced-frequency resonance scattering from fish", *The Journal of the Acoustical Society of America*, vol. 103, no. 6, pp. 3245–3255, Jun. 1998, ISSN: 0001-4966. DOI: 10.1121/1.423076 eprint: https://pubs.aip.org/asa/jasa/article-pdf/103/6/3245/8084717/3245_1_online.pdf.

[39] R. Guidotti et al., "A survey of methods for explaining black box models", *ACM Comput. Surv.*, vol. 51, no. 5, 2018, ISSN: 0360-0300. DOI: 10.1145/3236009

[40] R. Okuta, Y. Unno, D. Nishino, S. Hido, and C. Loomis, "Cupy: A numpy-compatible library for NVIDIA GPU calculations", in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[41] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library", in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035.

[42] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations", *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019. DOI: 10.1016/j.jcp.2018.10.045

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Toward Socially-Aware Explainability for Autonomous Robots

Khatina Sari, Paul Gerhard Plöger, and Alex Mitrevski
Department of Computer Science
Hochschule Bonn-Rhein-Sieg
Sankt Augustin, Germany
e-mail: khatina.sari@gmail.com, paul.ploeger@h-brs.de, alemitr@chalmers.se

***Abstract* - Explainability holds significant importance for autonomous robots deployed in human-centered situations, particularly when errors occur during execution. In the context of robot action, it is important to consider various levels and types of explainability. However, the social dimension of Artificial Intelligence and robotic explanations, which highlights how they affect social interaction, values, and decision-making, has received little to no attention in prior research. Our research addresses the need for multiple types of explainability in relation to robot activity. The aim is to develop a framework that enables the robot to provide explanations at varying levels of detail and complexity, depending on the type of end-user involved. Extending upon previous research on a representation for skill parameterization, we replicate the desired functionality to obtain information about user preferences by introducing simulated explanations that are translated from logical expressions. We explore two potential modalities that could be effectively integrated into the robotic system in the future: heatmaps and natural language explanations. With a particular emphasis on item handover, we hypothesize that users prefer systems with explanations and that explanations in natural language are more appealing than heatmaps. A user study, involving participants from diverse backgrounds and levels of expertise, is conducted to evaluate different levels and preferred types of explainability. The study results support our hypotheses and offer additional valuable information for future system development.**

Keywords - Explainable Artificial Intelligence; Natural Language Processing; Heatmaps; Autonomous Systems; Human-Robot Interaction.

I. INTRODUCTION

This paper builds upon and extends our previously published work in [1]. In addition to the original contributions, we incorporate a more comprehensive review of related literature, provide a detailed description of the research methodology, and present an expanded user study. The user study now extends both quantitative and qualitative analyses, offering deeper insights and a more robust evaluation of the proposed approach.

In recent decades, the fields of robotics and artificial intelligence (AI), which are closely related, have seen significant advancements. Future robotics systems are anticipated to be far more advanced and adaptable as AI and robotics continue to grow. Rule-based systems, also referred to as white-box AI, place an emphasis on transparency, making their logic processes clear and accessible to users. On the other hand, black-box AI, such as neural networks, often does not specify its decision-making process. Therefore, researchers are actively refining the interpretability of black-box AI, which can be used to improve transparency in robot actions, especially when failures occur [2-6].

Some challenges in human-robot interaction (HRI) necessitate transparent communication. Varying user knowledge and expectations pose challenges in maintaining the right level of detail in the explanations. Another challenge is to determine the most effective explanation format for each user [7][8]. It is worth noting that the level of complexity in the explanations provided may need to vary based on the specific users and settings involved. The technical proficiency and domain knowledge of different users may differ. Tailoring the complexity of explanations to the intended users can enhance their comprehensibility and ensure that the information conveyed is meaningful and useful [9]. By providing explanations that reveal the robot's intentions and decision-making processes, users can gain insights into its actions, enabling a higher level of trust and facilitating smoother collaboration. Additionally, to enhance productivity and team effectiveness, it becomes essential for the robots to provide explanations for their actions and decisions. These explanations serve as a means to bridge the gap in understanding between humans and robots, enabling better coordination and shared knowledge.

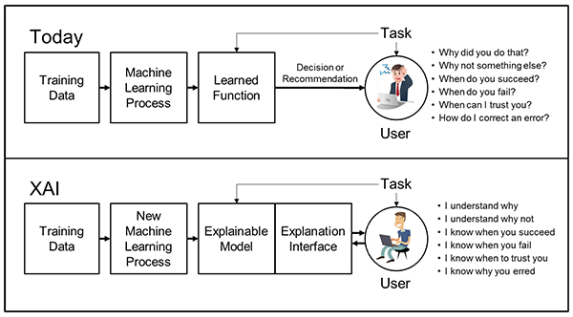


Figure 1. Concept of Explainable Artificial Intelligence (XAI) [6].

Explainability can be classified as local (usually focused on a single input dataset), global (describing how a model behaves generally), model-specific [10] (limited to particular model classes), model-agnostic [11] (may be local or global and independent of machine learning models), and counterfactual [12] (offering an alternate input scenario that

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

would have produced a different model prediction). Meanwhile, there are three common levels of explainability [10]: low-level (includes techniques like linear model coefficients or feature importance scores), medium-level (delves deeper into how specific features impact the model's predictions), and high-level (highlights intricate decision-making processes within the model).

This paper is focusing on robot object handover tasks, with the intention to enhance user understanding and trust in robot actions. The aim is to develop a framework that enables the robot to provide explanations at varying levels of detail and complexity, depending on the type of end-user involved. Essentially, a key element of our work is our user study, which allows us to verify with a diverse and representative set of participants the benefits of offering multiple levels of explainability during item handover. This study aims to encourage innovation in autonomous robotics by providing access to more adaptable, flexible, and user-centered systems.

The remainder of this paper is organized as follows. Section II offers an overview of literature related to the challenging topic this study addresses. Section III describes the general approaches used in our methodology. Section IV outlines our experimental results, both qualitative and quantitative, as well as hypothesis testing. Section V summarizes our findings and includes possible future work.

II. Related work

Transparent or white-box models refer to algorithms that provide users with both the end decision and a summary of the steps used to get there. One of the most common methods used for this is Bayesian network [12][13]. Lomas et al. [13] employed a Bayesian network model to evaluate how well the user understands the robot's actions. The algorithm calculates the amount of detail the robot should provide in its explanations based on a number of variables, including the user's prior knowledge, attention span, and verbal and non-verbal feedback. After determining the user's understanding level, the robot creates an explanation using a combination of pre-defined rules and machine learning methods (decision trees and classification algorithms), as well as graphical representations of its operations. While their approach does provide users with some ways of exploring what actions a robot will take in different scenarios, they do not attempt to carefully select which behaviors to demonstrate. Thus, they require substantial manual effort from users to explore the behavior of the robot and do not ensure that users understand the overall strategy of the robot [14]. Moreover, this method lacks scalability and generalizability because it involves hand-annotating every domain-specific context up front, which hinders application to new circumstances. In contrast, opaque or black-box models are machine learning models that are difficult to explain and understand by experts in practical domains [15][16]. These models include random forest, support vector machine, multilayer neural network, etc. One of the ways to obtain information from such models is to use post-hoc interpretability.

In order to address issues with slow learning and poor interpretability of some learned controllers, [15] proposed a solution that utilizes a two-stage methodology. In the first stage, a genetic algorithm is used to generate a large number of candidate rules. In the second stage, a subset of the most effective and interpretable rules is selected using a set of novel interpretability metrics. Additionally, they applied ant colony optimization (ACO) search algorithm to cut down on learning time. Then they analyzed the interpretability of their proposed approach using several metrics, including the number of rules, the number of fuzzy sets used for each input variable, and the complexity of the rule base. Through a set of simulation experiments where they compared the performance of their proposed approach against other rule-based and model-based approaches, they demonstrated how their approach produced fuzzy controllers that are more interpretable than those generated by other methods, while maintaining comparable or better performance. Nonetheless, they only focused on fuzzy controller design and did not consider non-fuzzy approaches, which may provide effective and interpretable control strategies for mobile robotics applications. As an alternative, we propose the utilization of neural networks and reinforcement learning as part of the machine learning adaptation in robot motion planning. This adjustment highlights our dedication to investigating different and practical approaches that may result in improved responsiveness and flexibility of robotic systems in dynamic settings.

Alvanpour et al. [17] investigated how machine learning can be used to predict robot grasp failure and studied the trade-off between accuracy and interpretability by comparing interpretable ML models that are inherently explainable with black box ML models that tend to be more accurate but do not come with explanations. Their approach consists of two stages: feature extraction and failure mode prediction. Even though they provided explanations for the predictions made by their approach, they did not test their approach on a physical robot, but rather on a simulation environment. Therefore, it remains to be seen how well their approach would perform in a real-world scenario, where sensor noise and other factors could affect the reliability of the system [18]. The evaluation of our proposed approach, on the other hand, spans across both simulated and real-world scenarios, aiming to provide a comprehensive assessment of its applicability and effectiveness across diverse operational contexts. We applied this assessment strategy in order to validate the dependability and applicability of our technique in both the complex and unpredictable settings found in real-world applications, as well as under controlled, simulated situations.

The need for user-centered design practices when creating explanations for AI systems was emphasized by [19]. They suggest involving users in the AI system design process through user studies, interviews, and feedback sessions to understand their needs, mental models, and expectations. Even so, they primarily focused on design practices and guidelines for creating user experiences in explainable AI systems and did not delve deeply into technical solutions or algorithms to achieve explainability. As a result, the technical aspects of implementing the proposed guidelines may require further exploration. Inspired from the user-centric principles they advocated, and given the importance of user trust in human-robot interaction, we conducted a user study to obtain

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

insightful feedback from users. This questionnaire aims to uncover user preferences regarding the different approaches employed in robot motion planning, shedding light on which method resonates more effectively with particular users. Through incorporating user feedback into our assessment framework, we hope to ensure that our method not only satisfies technical requirements but also aligns with user expectations and promotes a constructive and reliable human-robot collaboration.

Based on [9], methods for enabling humans and robots to fully understand each other's internal states should be developed so that robots can communicate with humans without only understanding human instructions at a surface level. In their opinion, a robot needs to have two mechanisms: one that can predict users' thoughts and the other that enables it to convey its own ideas to people in a way that is simple to comprehend. Their survey reveals that the explainability of autonomous robots is a multidisciplinary field that lacks structure. Furthermore, there is currently little to no research on the social dimensions of explanations provided by AI and robots. While a mechanism for creating an interpretable decision-making space has been created, there is currently no way to create a decision-making space that can be used universally in any setting without the need for humans to be involved. Moreover, they make suggestions that it is crucial to model humans who obtain explanations because no approaches have been put out that successfully estimate a human internal model up to this point. But given the interconnection of explanations, it is still a challenging task for the robot to be able to adapt its own knowledge system while delivering explanations. One must take into account the other person, their own knowledge, and the reliability of the information.

Ideally, in human-robot collaboration (HRC), human workers should have the ability to naturally converse with robots, since they are the most crucial members of any HRC team. According to [20], while there are currently few means of communication between human workers and robots, gesture recognition has long been used as an efficient human-computer interaction. They break down the gesture recognition process, which started with gesture raw data collection by sensors. Image-based methods (such as markers, depth sensors, stereo cameras, and single cameras) and non-image-based methods (such as bands, gloves, and non-wearable cameras) are the two main types of data collecting. Since the majority of human motion are dynamic, gesture categorization is the final and most crucial stage in gesture identification. Actuation classification must be completed either in conjunction with or after gesture tracking because a single dynamic gesture is always made up of many frames. In conclusion, they believe that HRC will operate in a safer environment if a depth sensor and body model technique are combined to track human movements.

Beyond the emphasis on trust, our study considers broader implications for user acceptability and cooperation in interaction between humans and robots; incorporating innovative approaches aims to meet users' expectations, preferences, and comfort levels in addition to raising technological criteria. By actively seeking user feedback and insights, our research aspires not only to advance the technical limitations of robotic systems but also to foster an inclusive and user-friendly approach for the comfortable integration of humans and robots in various domains. Additionally, we intend to measure how much these methods add to a generalized increased degree of confidence in the human-robot interaction paradigm.

III. APPROACH

The scope of our study concentrates on the usage of autonomous robots for object handover tasks from robot to human, an important use case that requires an effective explanation strategy. Giving our Toyota Human Support Robot (HSR) a skill set that corresponds to different levels of explainability—or, in some cases, no explainability at all—is the current challenge at hand. Our explainability analysis for skill execution takes into account a number of important factors, one of which is the recognition that explainability in our case is inherently local. Moreover, we investigated the domain of behavior explainability, focusing on cases of execution failures in tasks involving the handover of objects between humans and robots. Although it is obvious how important it is to explain robot behavior, it can be difficult to decide how much information is suitable to provide in these explanations. Our research intends to explore the complicated field of model-agnostic explainability since we acknowledge that various users may have varied expectations regarding the quantity of explanation needed and varying degrees of expertise with autonomous systems.

Within our explainability analysis for skill execution, model agnosticism means taking an approach that goes beyond the complexities of the machine learning models that are used as the foundation for the autonomous robot's decision-making. Model-agnostic approaches offer a general framework that allows us to examine the autonomous robot's decision-making process and behaviors without being limited by the internal model architecture's specific details. These are the methods that we would like to use to our Toyota HSR, which is the physical representation of our research into the problem of robot-human object handover. As we explore the advantages of explainability at several levels, our research methodology highlights the necessity of adaptability and flexibility in giving explanations that are specific to the needs of various users and circumstances.

Our goal in implementing model-agnostic explainability is to guarantee that our understanding of skill execution failures is relevant to all situations, regardless of how complicated the machine learning models are at their core. This methodology is in harmony with the diverse skill sets and demands of users interacting with autonomous robots in human-centered settings, and it eventually leads to a more open, understandable, and intuitive paradigm of interaction.

A. Proposed Approach

The current approach used in our robot to determine the handover position is done by factoring in context-dependent (based on the posture of the detected person) and context-independent (static; based on the context-dependent outcome). However, the handover position in a context-

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

independent approach does not consider any surrounding environment variables; thus, we propose to train a neural network to dynamically set the end-effector position based on the values obtained from the 3D bounding box. By conducting numerous training cycles, the neural network discovers the underlying patterns and connections between various contextual parameters and related handover positions. Through exposure to a broad range of random positions, the network improves its capacity to generalize and make appropriate choices in new scenarios encountered in real-world user interactions. Consequently, the trained neural network gains competency in automatically producing handover positions that properly correspond with user requirements, improving the efficiency and user-friendliness of the robotic system. Regrettably, a prolonged mechanical issue in our Toyota HSR has forced us to delay the implementation of our neural network interpretation. Upon its resumption of operations, we shall resume our work and implement our planned approach.

B. Explainability Setup

One of the primary concerns that drives our research is how to determine the robot's reasoning behind certain decisions, especially why it stops at a specific point in relation to the detected human position during object handover. To carry out this research, an advanced built-in program created by [21] is used, which generates a 3D bounding box to locate the detected person in front of the robot. It follows the right-handed coordinate system, which includes the depth (x-axis), horizontal (y-axis), and vertical (z-axis). Once the person is detected, their position will be determined; in our case, there are three possible positions: standing, sitting, and lying down.

Within our research framework, a number of notations play an important role in influencing how we perceive the spatial connections and decision-making processes associated with handing over objects between humans and robots. Figure 2 shows an illustration of the configuration where W_p represents the location on the robot's effector where the object to be given to the human is held, and W is the robot's base frame. Denoted as B, the bounding box is a three-dimensional structure containing the detected individual. $\bar{B}$ corresponds to the center point of this bounding box, providing a simple reference for the person's overall physical position. The minimum and maximum points of the bounding box, represented as $min(B)$ and $max(B)$ respectively, define the spatial extent of the detected individual within the bounding box. These points serve as key parameters in defining the boundaries of the person's posture and guide the robot's subsequent decision-making.

The relative position between the end effector and the center point of the bounding box is denoted as p. During the handover task, this relative position plays a key role in figuring out the spatial relationship between the identified person and the robot. Additionally, c_p is a threshold parameter that specifies an extra distance allowance in relation to p, enabling the robot's judgment criteria to be flexible and fine-tuned. To account for potential noise and variations in the center position estimate, a small threshold that controls for fluctuations and uncertainties in the center point's determination is introduced as ϵ_p. It serves as a mechanism to mitigate the impact of noise, ensuring a more robust and reliable analysis of the spatial dynamics between the robot and the detected person. Together, these notations improve the robot's accuracy and flexibility in the complex task of object handovers between humans and robots.

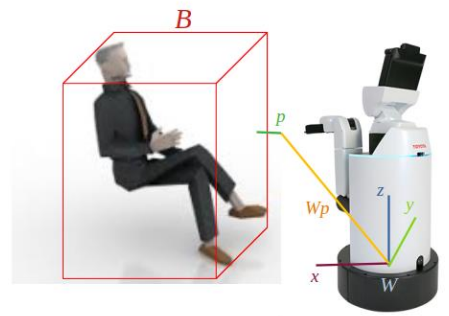


Figure 2. Illustration of the parameters on handover skill.

Logical predicates describing the requirements for a successful handover interaction are adopted from [22] to define the success preconditions. The predicates include $in_front_of_{x,y}(p,B)$, $far_in_front_of_{x,y}(p,B)$, $behind_{x,y}(p,B)$, $far_behind_{x,y}(p,B)$, $above_{x,y}(p,B)$, $below_{x,y}(p,B)$, and $centered_{x,y}(p,B)$. These predicates are made up of a collection of restrictions and conditions that capture the main elements affecting the robot's decision-making. A variety of elements, including spatial connections, postural considerations, and dynamic adaptability, are included in the qualitative representation of success, which is defined by the logical predicates. The following predicates are applied:

$$in_front_of_{x,y}(p,B) \coloneqq p_{x,y} < \min(B)_{x,y} \wedge |p_{x,y} - \min(B)_{x,y}| < C_p$$
$$far_in_front_of_{x,y}(p,B) \coloneqq p_{x,y} < \min(B)_{x,y} \wedge |p_{x,y} - \min(B)_{x,y}| > C_p$$
$$behind_{x,y}(p,B) \coloneqq p_{x,y} > \max(B)_{x,y} \wedge |p_{x,y} - \max(B)_{x,y}| < c_p$$
$$far_behind_{x,y}(p,B) \coloneqq p_{x,y} > \max(B)_{x,y} \wedge |p_{x,y} - \max(B)_{x,y}| > c_p$$
$$above_{x,y}(p,B) \coloneqq p_z > \max(B)_z$$
$$below_{x,y}(p,B) \coloneqq p_z < \min(B)_z$$
$$centered_{x,y,z}(p,B) \coloneqq |p_{x,y,z} - \max(\bar{B})_z| \leq \epsilon_p$$

Overall, these logical predicates include conditions like "the robot maintains a safe distance from the human recipient," or "the handover is executed in accordance with the detected posture". Such conditions provide a qualitative basis for success, allowing us to comprehensively assess the robot's behavior in real-world scenarios. Because these preconditions are explicit, it is easier to analyze the robot's behavior and provides a transparent and interpretable lens through which to assess the handover task's success. Consequently, we present logical predicates that correspond to every posture in order to describe the factors driving the robot's actions. And based on this, we transformed it into an explanation in natural language to give the robot's decision criteria a more understandable and natural flow, as shown in Tables I-III.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

In addition to manual natural language translation, ChatGPT 3.5 [23] was employed to generate automated translation and evaluate the results using the Bilingual Evaluation Understudy (BLEU) score [24]. The first few initial tests did not produce close translations to the manual translation. Therefore, more detailed definitions of each logical expression were provided, as well as separating each predicate that consists of two or more coordinates; for example, $centered_{x,y}(p,B)$ becomes $centered_x(p,B) \wedge centered_x(p,B)$. The outcome of the last iteration was then used for assessment.

TABLE I. PRECONDITIONS FOR STANDING POSITION

Types	Success Preconditions
Logical Predicates	$\boldsymbol{centered_{y,z}(p,B)} \wedge \boldsymbol{in_front_of_x(p,B)} \wedge \neg\boldsymbol{centered_x(p,B)} \wedge \neg\boldsymbol{below_{x,y}(p,B)} \wedge \neg\boldsymbol{behind_{x,y}(p,B)} \wedge \neg\boldsymbol{far_behind_{x,y}(p,B)} \wedge \neg\boldsymbol{above_{x,y}(p,B)} \wedge \neg\, \boldsymbol{in_front_of_y(p,B)} \wedge \neg\boldsymbol{far_in_front_of_y(p,B)}$
Natural Language	The robot's arm should be in front of and centered around a person (corresponding to the person's height and width). It should not be behind, above, beneath, or to the right/left of a human.

TABLE II. PRECONDITIONS FOR SITTING POSITION

Types	Success Preconditions
Logical Predicates	$\boldsymbol{centered_{y,z}(p,B)} \wedge \boldsymbol{in_front_of_x(p,B)} \wedge \neg\boldsymbol{centered_x(p,B)} \wedge \neg\boldsymbol{below_{x,y}(p,B)} \wedge \neg\boldsymbol{behind_{x,y}(p,B)} \wedge \neg\boldsymbol{far_behind_{x,y}(p,B)} \wedge \neg\boldsymbol{above_{x,y}(p,B)} \wedge \neg\, \boldsymbol{in_front_of_y(p,B)} \wedge \neg\boldsymbol{far_in_front_of_{x,y}(p,B)}$
Natural Language	The robot's arm is positioned in front of and around the middle of a sitting person (according to the person's height and width). It is not behind, above, beneath, and to the right or left of the person.

TABLE III. PRECONDITIONS FOR LYING DOWN POSITION

Types	Success Preconditions
Logical Predicates	$\boldsymbol{above_{x,y}(p,B)} \wedge \boldsymbol{centered_y(p,B)} \wedge \neg\boldsymbol{centered_z(p,B)} \wedge \neg\boldsymbol{below_{x,y}(p,B)} \wedge \neg\boldsymbol{behind_{x,y}(p,B)} \wedge \neg\boldsymbol{far_behind_{x,y}(p,B)} \wedge \neg\boldsymbol{in_front_of_y(p,B)} \wedge \neg\boldsymbol{far_in_front_of_{x,y}(p,B)}$
Natural Language	The robot's arm is positioned above and centered around the person's width. It is not below or around their head or feet. It should not extend all the way to the opposite side from where a robot is standing next to.

BLEU provides a quantitative measure by comparing the output of machine translation systems (candidate translation) against reference translations, offering insights into the degree of overlap in n-gram or word sequences with human-generated counterparts [24]. The length of candidate sentences that are shorter than the reference phrases is penalized in the BLEU metric (Brevity Penalty), which is based on the modified n-gram precision measure. The following formula determines the BLEU score:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \cdot logP_n\right), \qquad (1)$$

where BP = Brevity Penalty and P_n= Precision for n-gram.

The Natural Language Toolkit (NLTK) [25] and spaCy [26] are used in our BLEU score computation to provide an unbiased evaluation of machine-generated translations. The translation produced by ChatGPT 3.5 (as a candidate translation) was compared with our original translation (as a reference). The results of the BLEU score for each translation performed by ChatGPT in comparison to the manual translation are presented in Table IV. The final translation output from ChatGPT 3.5 provides a good starting point for future developments. Despite the fact that the translations produced by the first few iterations were not satisfactory, adding further specific information made it generate a translation that was similar to the one that was done manually. The key realization is that it is possible to train models, like ChatGPT, to translate technical terminology into natural languages effectively.

TABLE IV. BLEU SCORE OF CHATGPT 3.5 TRANSLATION

No.	Position	BLEU Score
1	Standing	0.85
2	Sitting	0.81
3	Lying Down	0.88

When it comes to interpreting the neural network's decisions about handover position, Grad-weighted Class Activation Mapping (Grad-CAM) [27] integration shows itself to be an effective tool for insight. It offers a transparent and insightful lens into the decision-making processes of complex models. Grad-CAM fills this gap by giving an illustration of the areas in the input data that have a major impact on a certain outcome. By emphasizing the areas of the input data that have a substantial impact on a certain output, it makes the focus and decision-making processes of the neural network easier to visualize. Grad-CAM allows us to produce heatmaps that show the regions of the 3D bounding box that affected the network's choice of a particular handover position. It primarily serves as a diagnostic tool, revealing the internal mechanisms of the network and helping identify potential areas for improvement or refinement.

Figure 3. Simulated heatmaps on one of the handover scenarios.

Grad-CAM performs best in models with a Global Average Pooling (GAP) layer following the last convolutional layer, which provides a spatial localization of the model's

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

focus. Grad-CAM would be implemented by first loading a neural network model that has already been trained, together with its weights, using TensorFlow, a deep learning framework that is commonly used for this kind of operation. The last classification layer is then eliminated from the model since Grad-CAM requires the output of the last convolutional layer. The next step involves constructing a new model that includes the layers of the pre-trained model up to the final convolutional layer and the GAP layer. This model is intended to produce the model's prediction as well as the output of the final convolutional layer. Unfortunately, the problem with our Toyota HSR prevented us from implementing this method. Despite this obstacle, a previously collected dataset from our research team [28] was leveraged, and the video content was edited to achieve the same heatmap effect (as seen in Figure 3).

This decision allowed us to simulate and observe the intended outcomes, ensuring the continuity of the research despite the technical constraints. The dataset, which includes relevant information but lacks explanations, was then extended by adding explanations in both heatmap and natural language formats. This improvised solution allowed us to proceed with our user study within the designated timeframe, preserved the research objectives, and ensured the timely execution of the study.

C. *Experimental Design*

In our comprehensive user study aimed at investigating user preferences in interacting with AI-based or robotic systems, two distinct hypotheses were formulated to guide our research. The first hypothesis is that users have a preference for systems that offer explanations while they are using them. The second hypothesis is about the preferred explanation format among users; in particular, we hypothesize that people prefer explanations in natural language over alternative visualization techniques like heatmaps. The feedback collected was then analyzed to confirm or deny our hypotheses on users' inclination for technical explanations and their preferred format. Beyond hypothesis testing, the survey plays a crucial role in unraveling the intricate entanglement of users' needs and expectations. For technology to be designed that fundamentally appeals to users, it is extremely important that these demands be understood.

We adopted a mixed-methods strategy to gather quantitative data and qualitative insights through surveys in order to experimentally validate our hypotheses. After being presented with simulated robotic interfaces that include heatmaps and natural language explanations, participants' experiences, preferences, and challenges were carefully examined, allowing us to obtain insights into the factors that contribute to a positive or negative interaction. Additionally, scenarios that are meant to replicate real-world interactions were chosen by giving users experiences that were contextually appropriate and reflected the difficulties and complexities of real-world circumstances. Ten videos and three different explanation varieties were presented to help construct a more comprehensive understanding of user preferences: no explanation, partial explanation using heatmaps, and detailed explanation using natural language. The videos intentionally featured a mix of both successful and unsuccessful handovers to capture a broad spectrum of user experience. In order to prevent any potential biases, eight out of ten videos were purposefully presented in a random order. Following every video, participants were asked to rate how confident they were in their understanding of the robot decision-making process.

IV. EXPERIMENTAL RESULTS

Our user study involved a total of 33 participants, ages ranging from 18 to 40 years old, education ranging from high school to Ph.D., and different academic and professional backgrounds.

The demographic data we were able to obtain from participants via our questionnaire presents a varied view and offers important new perspectives on the traits of our study participants (shown in Figure 4). In terms of age distribution, the vast majority of participants (72.7%) are between the ages of 26 and 40, which indicates that this group is well-represented. In addition, 21.2% of participants are between the ages of 18 and 25, indicating a representation of younger people, while 6.1% are over 40, adding to the age diversity of the study.

Meanwhile, gender representation showcases a predominant presence of males, constituting 66.7% of the participant pool. On the other hand, 27.3% of respondents identify as female and the remaining participants identified as non-binary or opted not to disclose their gender. In terms of education, 45.5% hold a Master's degree, 39.4% have a Bachelor's, 12.1% completed high school, and 3% hold a PhD.

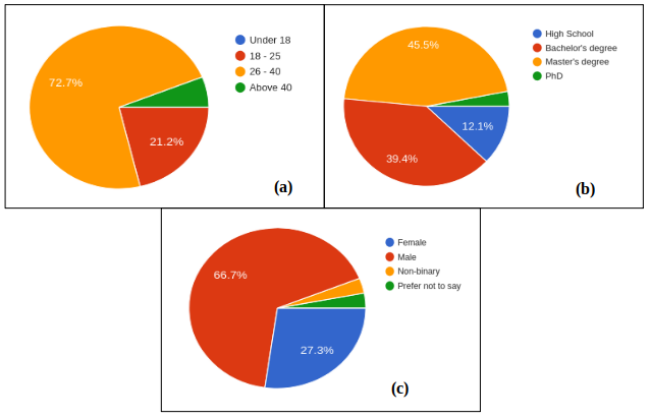


Figure 4. (a) Age range of our user study participants (in years); (b) Gender representation of our user study participants; (c) Education background of our user study participants.

In conclusion, our participants' demographic profiles show a variety of age groups, gender identities, levels of education, and fields of study. This diversity ensures that the responses gathered are representative of a wide cross-section of people, which strengthens the reliability and generality of our user study findings. Based on this diversity, we attempt to determine whether there is any relationship between the preferred explanation technique and the educational background.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A. Quantitative Analysis

Participant profiles in the study reveal that 60.6% of participants have a background in computer science, which is a significant majority. With 18.2% of the total, engineering is the second most common field. Other fields that represent a wide range of academic and professional backgrounds among our participants are finance (9%), natural science (6.1%), social science (3%), and administration (3%). In terms of the participants' experiences and expectations in the realms of robotic systems and AI, 75.8% of them have prior hands-on experience with robotic systems, while an overwhelming 84.8% are familiar with AI or machine learning in their practical lives. In a survey on comfort levels, 72.7% of the respondents said they felt uneasy when AI systems made decisions without providing an explanation, highlighting the significance of transparency.

Table V indicates that the majority of individuals who have experience with robotic systems or AI, including computer science and engineering, prefer natural language over heatmaps. This suggests that, even if they could be considered experts in the topic, they would still rather receive a full explanation than a simplified one. In our qualitative analysis (Subsection C), we will discover more about their reasoning for this preference.

TABLE V. AN OVERVIEW OF PARTICIPANTS' EDUCATIONAL BACKGROUNDS AND THEIR PREFERRED TYPES OF EXPLANATIONS

No.	Educational Background	Preferred Type of Explanation	
		Natural Language	Heatmap
1	Computer Science	8	6
2	Engineering	5	1
3	Natural Science	1	1
4	Administration	1	0
5	Finance	1	2

In our scenario-based questions, two identical videos served as starting points. The first was without explanation, whereas the second included a natural language explanation. The majority indicated that they were unclear about the robot's action in the first video, despite it being a successful object handover scenario. However, the participant's confidence level improved after watching the second video, which revealed a positive beginning. Table VI summarizes participants' confidence levels after eight more videos were shown in a random order.

TABLE VI. AN OVERVIEW OF PARTICIPANTS' CONFIDENCE LEVEL

Video	Outcome	Explanation Type	Confidence Level (%)				
			5	4	3	2	1
3	Succeed	None	9.1	24.2	48.5	18.2	0.0
4	Succeed	Heatmap	24.2	27.3	36.4	12.1	0.0
5	Failed	None	12.1	15.2	30.3	24.2	18.2
6	Failed	Natural Language	24.2	24.2	15.2	15.2	21.2
7	Succeed	None	3.0	18.2	15.2	36.4	27.3
8	Succeed	Natural Language	27.3	57.6	15.2	0.0	0.0
9	Failed	Natural Language	24.2	24.2	18.2	27.3	6.1
10	Failed	Heatmap	18.2	21.2	30.3	27.3	3.0

It reveals that individuals feel more confident when they are given an explanation of how the robot makes decisions. Less than 40% of the participants felt confident about their understanding of the robot decision-making process in the three videos without an explanation, in both successful and unsuccessful handover scenarios. More than 50% of the participants in the two videos where heatmaps were used as an explanation type expressed confidence in the successful handover scenario. However, in the case of an unsuccessful handover, only 34.6% of participants reported feeling confident. With natural language explanations, on the other hand, 48.4% of those surveyed expressed confidence in the unsuccessful scenarios. In the successful scenario, over 80% of the participants expressed confidence and none of them indicated lack of confidence. To conclude, compared to visual explanation (using a heatmap), natural language explanation improves their confidence by over 30% (shown in Figure 5).

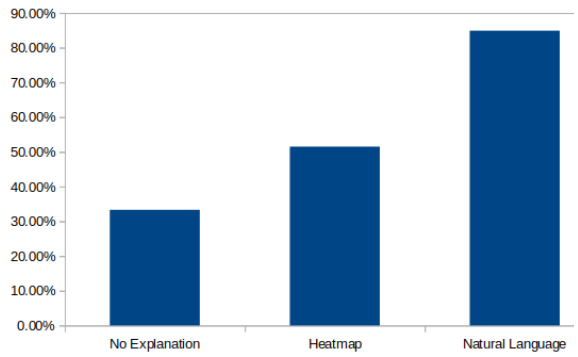


Figure 5. Participants' overall confidence in understanding the robot decision-making process.

B. Hypotheses Testing

We conducted the first hypothesis test to investigate users' preferences regarding the type of videos when seeking information. The hypothesis aimed to determine whether users prefer videos with explanations over videos without explanations. The participants were presented with the question "Which type of video do you prefer when seeking information?" and the response options: videos with explanation, without explanation, and depending on the context.

A chi-square test [29] for independence is employed to analyze the association between the type of video and user preference, where H_0 = no preference difference and H_1 = there is a preference for videos with an explanation. If the p-value of a given dataset is less than 5%, the null hypothesis is rejected because it is assumed that there is a preference difference among the options. To calculate the p-value using chi-square formula (2), the observed value (O) needs to be identified first, which represents the actual counts derived from the sample, and the expected value (E), which represents the values of each category in the event that there was no preference difference between all categories. E is

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

obtained by dividing the total number of observed values by the number of categories. The following calculation can then be used to get its chi-square statistic (χ^2) based on the observed and expected values:

$$X^2 = \sum \frac{(O-E)^2}{E}. \quad (2)$$

The result, along with the degrees of freedom (df), which is a number representing how much variation is involved in the research (n) minus 1,

$$df = n - 1, \quad (3)$$

is used to calculate the p-value from the chi table.

Our observed and expected values based on the survey results are displayed in Table VII. The total observed values—33 in this case—and the number of categories—3 in this case—are then used to compute the expected values, yielding the value E = 11.

TABLE VII. THE OBSERVED AND EXPECTED VALUES

User Preference	O	E	$O-E$	$(O-E)^2$
With Explanation	22	11	11	121
Without Explanation	2	11	-9	81
Depend on the Context	9	11	-2	4

These observed and expected values were used to calculate the chi-square statistic, which was then used to test the hypothesis. The result yielded χ^2= 28.1; with df = 2, the resulting p-value was 0.0000008. Since the p-value is less than α = 5% or 0.05, it is determined that the null hypothesis is rejected.

The second hypothesis is tested based on two identical videos with two distinct explanations—one using a heatmap (video 4) and the other using natural language (video 8). Participants were asked to choose which of the two videos gave them a better understanding of the robot decision-making process. Participants who selected video 8 are considered to prefer the natural language explanation. A one-sample proportion test (Z) [30] is employed to analyze whether the proportion of users who prefer video 8 differs significantly from 50% (no preference). The null hypothesis (H_0) assumed no preference difference, while the alternative hypothesis (H_1) assumed a preference for videos with natural language explanation.

To conduct the test, we need to estimate the proportion $\hat{p}$ as:

$$\hat{p} = \frac{x}{n}, \quad (4)$$

where x is the number of participants who have chosen video 8 and n is the total number of participants. After that, the test statistic can be calculated with the following formula:

$$Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad (5)$$

where p_0 is the pre-specified value; in this case, it is 50% to indicate that if half of the total participants chose video 8, there is no significant preference for that particular video. From there, the calculated Z-value is compared with critical values, which can be obtained from the Z table, from the standard normal distribution. Given that the sampling distribution of our data is a normal distribution with a significant value of 0.05, the critical values are in a range of -1.96 to 1.96. Based on the result of our survey, a one-sample proportion test was calculated with x = 23 and n = 33, which yielded a Z-value of 2.46. Because the Z-value is larger than the maximum critical value, the null hypothesis is rejected.

A post hoc sensitivity analysis [31] was conducted to evaluate the statistical power of our study. Cohen's w,

$$w = \sqrt{\sum \frac{(p_i-p_{0i})^2}{p_{0i}}}, \quad (6)$$

where p_i is the observed value in category i and p_{0i} is the expected value under the null hypothesis in category i, is used to measure the effect size for the chi-square test of the first hypothesis. The thresholds are 0.10 for a small effect, 0.30 for a medium effect, and 0.50 for a large effect. The result yielded w = 0.75, which represents a large effect. Using this result combined with significant value (α= 0.05) and number of participants (N = 33), we achieved statistical power of 0.978 or 97.8%, indicating that the test was adequately powered to detect strong differences in preference.

Furthermore, we assess the effect size for the one-sample proportion test of the second hypothesis with Cohen's h,

$$h = 2\,x\left(\arcsin\left(\sqrt{p_1}\right) - \arcsin\left(\sqrt{p_2}\right)\right), \quad (7)$$

where p_1 and p_2 are the two proportions being compared. The thresholds are 0.20 for a small effect, 0.50 for a medium effect, and 0.80 for a large effect. From our user study result, 23 out of 30 participants preferred video with natural language explanation; thus, p_1 = 69.7%. Then we compare it with p_2 = 50% for the proportion that shows no preference difference. The result yielded h = 0.40, which indicates a moderate effect size. This effect size corresponds to statistical power of 0.369 or 36.9%, which suggests that our study may have been underpowered to reliably detect effects of this magnitude. Therefore, we suggest future studies to be conducted with a larger sample size to confirm our finding.

C. Qualitative Analysis

As proven in our hypothesis 2, natural language explanations are preferable to heatmaps. In order to evaluate it on a qualitative level, the participants were asked why they

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

preferred one type of explanation over the other, and the majority of them responded that they preferred natural language because it is easier to understand and more elaborate. In addition, they believe that natural language explanations can be enhanced by an audio or speech component. In addition to asking the respondents to rank their degree of confidence in each scenario, our scenario-based questions also inquired about any confusion or difficulty the respondents may have had comprehending the robot's decision-making process. This is an outline based on various levels of explanation given:

1) Without Explanation: many wondered why the robot did not come closer to the person and were curious as to how the robot decides when to release the object. In other instances, they questioned what the robot was trying to accomplish and what factors might have played a role when the robot was unable to correctly pass on the item.

2) With Heatmap: even if the heatmap makes the target area easier for them to understand, they believe that more thorough explanation is necessary in order to understand it completely. In their opinion, the heatmap did not help them understand why the object fell or why the robot was unable to detect the user's hand.

3) With Natural Language: Many of them highlighted that the explanation was too lengthy and that they were unable to follow it quickly. Moreover, the goal of the robot was not stated in the provided explanation, and it excluded the reason for the robot's inability to recognize a person's hand.

Given these uncertainties, we have also asked the participants to suggest ways we could improve the overall system. Of the 33 respondents, 8 people said that they have no suggestions regarding how the robot's performance or the explanations it provides could be made better. Most, nonetheless, recommended shorter or simpler sentences for an explanation in natural language or integrating a voice or audio option to improve the explanation process. In order to visualize information more clearly, participants suggested utilizing arrows or targets instead of heatmaps. For a more interesting and educational experience, some participants suggested an interactive graphical explanation that incorporates both written and graphic elements. Other recommendations centered on instructing the robot to recognize what type of item it is handling. To protect delicate objects like glasses or plates, for instance, the robot should handle them with caution and hold them upright.

They were then asked to imagine a situation in which they would favor a different kind of explanation than the one they had previously selected. Those who have chosen natural language say that they prefer heatmaps when a robot performs a simple task, interacts with static objects, or is in a simulation. On the other hand, those who have chosen heatmaps say that they prefer natural language when failure occurs, when the robot is in a dynamic environment, or when the user has no background knowledge about the system.

When asked to imagine a situation in which they would prefer to have no explanation at all, the majority of respondents believe that in a straightforward or routine task that is repeated, there is no need for an explanation because the rationale is obvious. While some claim that they cannot think of any situation in which it is preferable not to have an explanation, others highlight this point by stating that, even in tasks that appear straightforward, having an explanation is desirable since it provides a clear reasoning behind the robot's chosen action.

Last but not least, the following are some further comments, suggestions, or feedback about the application of AI systems with various levels of explainability: in order to give users the flexibility to select the level of explanation they are looking for and when it should be provided, the majority of them recommended making explanations available upon request; a few more said that the whole experience would be enhanced by increasing the level of engagement: to make that happen, they recommended implementing a voice-activated technology; one other noteworthy feedback is that they recommend providing additional context or reference data, such as the user's height or the robot's expected reach, to help non-expert users who are less familiar with the system understand it better.

V. CONCLUSION AND FUTURE WORK

Effective communication of intent is crucial for successful interaction between humans and robots in applications involving HRI. Humans need to understand the robot's internal state, including its purposes and intentions, to engage in collaborative tasks effectively. This requirement has fueled the development of explanation techniques that make the robot's internal workings transparent to people. Given that explanations are essential for building user confidence and comprehension of robotic systems, our research emphasizes how important it is to include them in HRI. The capacity to explain robot decision-making becomes crucial for users to interact with these intelligent machines in a meaningful way as they become more and more integrated into diverse sectors. In order to fulfill this demand, our method entails offering two different kinds of explanations inside the robot handover framework. First, we added natural language explanation, translating complex logical expressions into easily understandable statements. This improves the interpretability of the robotic system's activities and makes them more consistent with human cognitive processes, which makes it easier for people to understand the robot's behavior more naturally. Second, we introduced a heatmap-like explanation in the style of Grad-CAM that could be applied to a trained neural network. This visual representation provides a supplementary and visual explanation by enabling users to identify the regions that are relevant within the input data that impacted the robot's decision-making.

Our user study results supported our hypotheses, offering statistical evidence that users do, in fact, prefer explanations when interacting with robotic systems. These findings highlight that providing explanations improves users' trust and understanding of robot systems. Although the study demonstrates a clear preference for explanations in natural language as opposed to heatmap visualizations, respondents express a preference for heatmaps or no explanations at all when the robot is performing regular or routine tasks. This

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

tendency implies that, in situations they are familiar with, participants think that the visual representations of the heatmaps are sufficient or that perhaps they prefer them more when the tasks are simple and require no extra information. Due to the wide range of participant preferences, flexible communication strategies that take into account varying user expectations and levels of experience with certain robotic tasks are necessary.

Even though the results suggest that users prefer systems that provide explanations over those that do not, it is important to acknowledge a potential bias in how this hypothesis was tested. The question itself highlights the presence or absence of an explanation, which might have led participants to gravitate toward the condition with explanations, independent of their actual utility in decision making. Future studies should aim to mitigate this bias by embedding explanations in more naturalistic tasks where the usefulness of the explanation emerges organically rather than being made explicit to participants.

While our findings indicate that participants preferred natural language explanations, it is important to recognize that this result may partly reflect differences in interpretability between formats. Natural language requires little effort to process, whereas heatmaps demand additional interpretation and prior familiarity. This asymmetry may have disadvantaged the heatmap condition. To address this imbalance, future studies should explore providing training or familiarization with visual explanations, refining visualization design to reduce cognitive effort, or presenting hybrid formats that combine textual and visual elements for complementary strengths.

Our research lays the groundwork for compelling avenues of future work, particularly in the domain of automating the translation of scientific terms into natural language. Building on the success of employing ChatGPT 3.5 to translate logical expressions, the next logical step involves expanding this method to encompass a broader spectrum of scientific terminology. By refining and extending the capabilities of such model, we envision a system that can seamlessly convert intricate scientific terms into clear and accessible natural language, catering to both expert and non-expert users. One prospective approach for this development involves fine-tuning language models on specialized scientific domains. The model could be adapted to recognize and translate specific scientific terms, ensuring accurate and contextually relevant conversions. Additionally, incorporating domain-specific ontologies or knowledge graphs could enhance the model's understanding of relationships between scientific terms, contributing to more coherent translations.

Further studies could explore automating the translation of scientific terms into natural language to provide explanations for nonexpert users. To implement audio explanations effectively, future work may explore the integration of speech synthesis technologies or Natural Language Processing (NLP) models specialized in generating spoken content. Additionally, exploring the potential of machine learning techniques, such as reinforcement learning, could contribute to optimizing explanation selection. This way, the system could learn over time which combination of explanation modalities yields the most positive user responses or facilitates optimal task performance.

ACKNOWLEDGMENT

The authors would like to express gratitude to all of the participants who willingly participated in the research, contributing their time and insights. Without their cooperation, this study would not have been possible.

REFERENCES

[1] K. Sari, P. G. Plöger, and A.Mitrevski, “Explainability Analysis for Skill Execution,” in EXPLAINABILITY 2025 The Second International Conference on Systems Explainability, pp. 14-20, 2025.

[2] O. Loyola-Gonzalez, “Black-box vs. white-box: understanding their advantages and weaknesses from a practical point of view,” IEEE Access, vol. 7, pp. 154096–154113, 2019.

[3] W. Samek, T. Wiegand, and K.-R. Mueller, “Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models,” arXiv preprint arXiv:1708.08296, 2017.

[4] A. Adadi and M. Berrada, “Peeking inside the black-box: a survey on explainable artificial intelligence (xai),” IEEE Access, vol. 6, pp. 52 138–52 160, 2018.

[5] R. Guidotti et al., “A survey of methods for explaining black box models,” ACM Computing Surveys (CSUR), vol. 51, no. 5, pp. 1–42, 2018.

[6] D. Gunning and D. Aha, “Darpa’s explainable artificial intelligence (xai) program,” AI Magazine, vol. 40, no. 2, pp. 44–58, 2019.

[7] A. Holzinger, “From machine learning to explainable ai,” in 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA). IEEE, pp. 55–66, 2018.

[8] N. Burkart and M. F. Huber, “A survey on the explainability of supervised machine learning,” Journal of Artificial Intelligence Research, vol. 70, pp. 245–317, 2021.

[9] T. Sakai and T. Nagai, “Explainable autonomous robots: A survey and perspective,” Advanced Robotics, vol. 36, no. 5-6, pp. 219-238, 2022.

[10] A. Das and P. Rad, “Opportunities and challenges in explainable artificial intelligence (xai): A survey,” arXiv preprint arXiv:2006.11371, 2020.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, “Model-agnostic interpretability of machine learning,” arXiv preprint arXiv:1606.05386, 2016.

[12] R. M. Byrne, “Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning,” in IJCAI, pp. 6276–6282, 2019.

[13] M. Lomas et al., “Explaining robot actions,” in Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot, pp. 187-188, 2012.

[14] D. Das, S. Banerjee, and S. Chernova, “Explainable ai for robot failures: generating explanations that improve user assistance in fault recovery,” in Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, pp. 351–360, 2021.

[15] O. Amir, F. Doshi-Velez, and D. Sarne, “Summarizing agent strategies,” Autonomous Agents and Multi-Agent Systems, vol. 33, pp. 628–644, 2019.

[16] M. Mucientes and J. Casillas, “Quick design of fuzzy controllers with good interpretability in mobile robotics,” IEEE

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Transactions on Fuzzy Systems, vol. 15, no. 4, pp. 636–651, 2007.

[17] A. Alvanpour, S. K. Das, C. K. Robinson, O. Nasraoui, and D. Popa, "Robot failure mode prediction with explainable machine learning," in 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE). IEEE, pp. 61–66, 2020.

[18] S. B. Remman and A. M. Lekkas, "Robotic lever manipulation using hindsight experience replay and shapley additive explanations," in 2021 European Control Conference (ECC), pp. 586-593, 2021.

[19] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: informing design practices for explainable ai user experiences," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–15, 2020.

[20] H. Liu and L. Wang, "Gesture recognition for human-robot collaboration: a review," International Journal of Industrial Ergonomics, vol. 68, pp. 355–367, 2018.

[21] A. F. Abdelrahman, "Incorporating contextual knowledge into human-robot collaborative task execution," Technical Report, H-BRS Sankt Augustin, 2020.

[22] A. Mitrevski, "Skill generalisation and experience acquisition for predicting and avoiding execution failures," Ph.D. Dissertation, RWTH Aachen, 2023.

[23] OpenAI. (2023). ChatGPT (Oct 16 version) [Large language model]. [Online]. Available from: https://chat.openai.com/chat. Retrieved: June 2024.

[24] M. Evtikhiev, E. Bogomolov, Y. Sokolov, and T. Bryksin, "Out of the bleu: how should we assess quality of the code generation models?" Journal of Systems and Software, vol. 203, p. 111741, 2023.

[25] S. Bird, "NLTK: the natural language toolkit," in Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pp. 69-72. 2006.

[26] Y. Vasiliev, "Natural language processing with Python and spaCy: A practical introduction," No Starch Press, 2020.

[27] R. R. Selvaraju et al., "Grad-cam: visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626, 2017.

[28] IROS 2022 HEART-MET Handover Failure Detection Challenge. Available from: https://codalab.lisn.upsaclay.fr/competitions/6757#learn_the_details-evaluation. Retrieved: September 2025.

[29] Pandis, N., "The chi-square test," American Journal of Orthodontics and Dentofacial Orthopedics, vol. 150, no. 5, pp. 898–899, 2016.

[30] David I, Adubisi O, Farouk B, and Adehi M., "Assessing MSMEs growth through rosca involvement using paired t-test and one sample proportion test," J Soc Econ Stat, vol. 9, no.2, pp. 30–42, 2020.

[31] Cohen, D. "Culture, social organization, and patterns of violence," Journal of Personality and Social Psychology, vol. 75, no. 2, pp. 408–419, 1998.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Energy Management of Stand-Alone Rural Micro-Grid Application based PV – Batteries–System in Guinea

Mohamed Lamine Touré[1,2], Mamadou-Baïlo Camara[1,2,3], Fahad Maoulida[1], Alireza Payman[1], and Brayima Dakyo[1]

[1] GREAH Laboratory, University of Le Havre Normandie, 75 Rue Bellot, 76600 Le Havre, France
[2] Laboratoire d'Enseignement et de Recherche en Énergétique et Automatique (LENA), Université Gamal Abdel Nasser Conakry, Guinea
[3] Laboratoire de Recherche en Sciences Appliquées de Mamou (LaReSA), Institut Supérieur de Technologie, Telico, Mamou 063, Guinea
[4] LIED Laboratory, UMR 8236 CNRS, Université Paris Cité, 35 rue Hélène Brion, 75006 Paris, France

e-mail: mohamed-lamine.toure@etu.univ-lehavre.fr; mamadou-bailo.camara@univ-lehavre.fr; maoulidafahad@gmail.com; alireza.payman@univ-lehavre.fr; brayima.dakyo@univ-lehavre.fr

Abstract — **Using photovoltaic (PV) energy in stand-alone micro-grid systems represents an effective solution for improving access to electricity in remote rural areas while reducing dependence on conventional power grids. This paper presents an energy management and control strategy for a PV–batteries micro-grid supplying an AC load through a three-phase voltage source inverter. The proposed system aims to maximize the solar energy production while ensuring reliable and stable operation of the micro-grid. The PV generator is modeled using a voltage model that considers the variation of the series resistance as a function of solar irradiance and operating temperature. To ensure optimal power extraction, a Maximum Power Point Tracking (MPPT) algorithm based on the incremental conductance method is implemented. A rule-based energy management system coordinates the power exchange between the PV generators and the batteries energy storage system. In addition, the DC-bus voltage is regulated using a proportional–integral controller through the bidirectional converter in batteries side, ensuring system stability under varying operating conditions. The DC power is converted into AC power by a two-level three-phase voltage source controlled using the space vector pulse width modulation (SVPWM) technique to provide high-quality output voltage with reduced harmonic distortion. The performances of the proposed system are evaluated through simulations using Matlab/Simulink software under different solar irradiance levels and load conditions. Simulation results demonstrate the effectiveness of the proposed control strategy in stabilizing the DC-bus voltage, maximizing PV power extraction, and mitigating power fluctuations through the batteries energy storage, thereby ensuring reliable power supply for rural micro-grid applications.**

Keywords — Photovoltaic system; battery energy storage; energy management system; DC-bus voltage regulation; space vector PWM (SVPWM); voltage source inverter; stand-alone micro-grid.

I. INTRODUCTION

Climate change worldwide, combined with the gradual depletion of fossil fuel resources, necessitates the implementation of sustainable energy solutions aimed at reducing our dependence on these polluting energy sources. In this context, the integration of renewable energies into stand-alone micro-grids emerges as an effective alternative for limiting greenhouse gas emissions while enhancing energy autonomy and reliability.

Among these sources, solar photovoltaic (PV) energy stands out particularly in the sub-Saharan region of West Africa. This case of study is focused to Republic of Guinea, which benefits from an average daily sunshine duration of 8 to 9 hours, representing a significant advantage for rural communities and areas located far from major national grid interconnections. Therefore, the development of solar integration solutions within these stand-alone micro-grids will facilitate access to electricity and promote the creation of income-generating activities.

However, the optimal operation of a photovoltaic generator remains complex due to its inherently variable and non-linear behavior, as well as the phenomena related to the penetration of solar radiation through the Earth's atmosphere. Indeed, the power output is highly dependent on environmental conditions, particularly solar irradiance and temperature. This makes the use of Maximum Power Point Tracking (MPPT) techniques essential to ensure optimal energy efficiency.

Several MPPT approaches have been developed, including the Perturb and Observe (P&O) method, the Incremental Conductance (InC) technique, and adaptive step-size methods, as presented in [1], [2], [3], [4]. Among these, the InC algorithm stands out for its ability to more effectively track the maximum power point, especially under conditions of high solar irradiance variability.

Furthermore, energy management of photovoltaic systems coupled with batteries Energy Storage Systems (ESS) requires the implementation of an Energy Management Strategy (EMS) capable of responding rapidly to efficiently direct energy either to the micro-grid or to the batteries. Various studies-based EMS approaches have proposed based on fuzzy logic, decision rules, or hierarchical architectures [4-13].

For instance, Swetha et al. [5] developed a centralized model-based strategy adapted to electric vehicle applications in a micro-grid environment. Similarly, Oukkacha et al. [6] proposed an energy management method based on frequency sharing among a fuel cell, lithium-ion batteries, and supercapacitors, all interfaced through DC-DC converters. Badawy and Sozer [7], on the other hand, designed and implemented an Energy Management Strategy Control (EMSC) for a micro-grid integrating renewable energy sources and batteries.

Moreover, Baqar et al. [8] conducted a comparative analysis of various EMSC strategies applied to a hybrid system combining a fuel cell, supercapacitors, and batteries, highlighting the importance of effective coordination to extend battery lifespan. Finally, Bonkile and Ramadesigan [9] proposed an autonomous energy management approach for a PV-battery hybrid system aimed at minimizing overload constraints using the Runge-Kutta method with high-stability time steps.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

In terms of DC-bus voltage regulation, which is a key element in stand-alone micro-grid architectures, whether embedded or stationary, conventional controllers such as PI and PID remain widely used to ensure system voltage stability [10], [11]. In this context, Yaqoob et al. [12] proposed an energy management strategy based on plateau control for a stand-alone photovoltaic-battery hybrid system, with the objective of stabilizing the DC-bus voltage while optimizing power distribution among the different sources.

Similarly, an Energy Management Strategy and Control (EMSC) designed for a DC micro-grid integrating a photovoltaic module, batteries, and a load was developed in [13], aiming to optimize energy flows while extending battery lifespan. Furthermore, Benzaouia et al. [14] conducted an experimental evaluation of a control strategy for a PV/battery system dedicated to water pumping applications, using neural networks to maximize power extraction on the PV side and fuzzy logic on the battery side to maintain a balance between energy supply and demand.

Despite these advances, only a limited number of studies have combined an Incremental Conductance (InC)-based MPPT strategy with a PI controller dedicated to DC-bus voltage regulation within a PV/batteries/inverter architecture intended for stand-alone AC micro-grids. Furthermore, few works consider a photovoltaic system voltage model that incorporates a variable series resistance while also accounting for realistic weather conditions.

In this context, the present work proposes a hybrid PV/batteries/inverter system designed for stand-alone micro-grid applications. The developed energy management strategy is based on Incremental Conductance (InC) MPPT algorithm, enabling optimal extraction of photovoltaic power under variable irradiance conditions while accounting for the effect of the PV generator's series resistance.

This system is coupled with two PI controllers: the first ensures the DC-bus voltage regulation by effectively controlling the batteries charging and discharging through a bidirectional DC-DC converter, while the second, associated with a unidirectional DC-DC Boost converter, guarantees power transfer from the photovoltaic system to the DC-bus.

A two-level three-phase inverter is connected to the DC- bus and controlled using the SVPWM algorithm to ensure a reliable AC power supply for domestic loads.

The system is modeled and simulated in Matlab/Simulink using realistic climatic data, particularly solar irradiance and temperature measured at the Dialakoro pilot site in Republic of Guinea. The study aims to evaluate the performances of the micro-grid under various production and consumption scenarios, with a particular focus on the DC-bus voltage stability, the batteries safety, and the inverter operation to ensure an appropriate voltage supply with a frequency of 50 Hz.

The obtained results highlight effective regulation, efficient utilization of solar energy, optimal batteries energy management, and high-quality power delivered to consumers, thereby confirming the relevance of this approach for stand-alone micro-grid applications.

The paper is organized as follows: Section II presents micro-grid architecture, emphasizing the modeling of photovoltaic sources, batteries storage systems, and DC/AC conversion through a three-phase inverter. Section III is dedicated to energy management strategies which introduce the principle of MPPT control and adopted energy management strategy. Section IV presents simulation results and provides an in-depth analysis. Section V concludes the study by summarizing the main findings.

II. Description of Micro-grid

Figure 1 presents the overall architecture of the studied hybrid system, which includes a photovoltaic (PV) source characterized by a variable series resistance, a module of the batteries, a three-phase inverter, and a load represented by a household. The photovoltaic generator is connected to a boost converter, controlled by the Maximum Power Point Tracking (MPPT) strategy. This strategy relies on an algorithm that optimizes in real time the power extracted from the PV array by continuously adjusting the operating point.

The energy storage system consists of batteries connected to DC-bus through a bidirectional DC-DC converter, controlled by a dual PWM signal generated by the energy management system. This device enables batteries charging when photovoltaic production exceeds demand and discharging to supply the load when solar generation is insufficient.

Furthermore, a three-phase inverter, controlled using the SVPWM technique, ensures the conversion of energy from the DC-bus to the AC bus. It consists of three legs, each comprising two IGBT switches that are driven in a complementary manner to generate an alternating voltage or current at the desired frequency, set to 50 Hz in this case.

The considered micro-grid as the main load, is supplied through the AC bus, whose voltage is regulated via coordinated power management across all converters. This architecture enables both intelligent and adaptive energy management, facilitating the integration of the photovoltaic system into a stand-alone micro-grid, while enhancing energy flexibility.

A. Behavior Modeling of Solar PV

$$\begin{cases} V_{pv} = \alpha \cdot \dfrac{n \cdot k \cdot T}{q} \cdot \ln\left(\dfrac{I_{ph} - \frac{I_{pv}}{N_{p,ch}} \cdot k_{\lim i}}{N_{sc} \cdot I_o} \cdot \beta + 1 \right) - I_{pv} \cdot R_s \cdot \lambda \\ \alpha = N_{s,ch} \cdot N_{sc} \\ \beta = \left(\dfrac{T}{T_{ref}} \right)^3 \cdot \exp\left(\dfrac{q \cdot E_g \left(\frac{1}{T_{ref}} - \frac{1}{T} \right)}{n \cdot k} \right) \\ \lambda = \dfrac{N_{s,ch} \cdot N_{sc} \cdot k_{\lim v}}{N_{p,ch}} \\ k_{\lim i} = \dfrac{I_{sc} - I_{mp}}{I_{mp}} \\ k_{\lim v} = \dfrac{V_{oc} - V_{mp}}{V_{mp}} \end{cases} \quad (1)$$

Several mathematical models of PV panels have been developed with the aim of accurately representing their electrical properties and operation, which derive directly from the physical structure of PV cells as in [15], [16]. In this study, the used PV model is based on the methods described in [3],[1]. The PV

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

voltage model with variable series resistance is connected directly to the DC-bus via a boost converter, thus guaranteeing power transfer from the PV string to the DC-bus. Table I presents the main technical characteristics of the PV module and DC-DC boost converter. In addition, the voltage-current (I-V) properties of the PV panel, and the links with power output, can be explained using equation (1), which summarizes the electrostatic behavior of the module under standard irradiance and temperature conditions, where I_{pv}, V_{pv}, I_o and I_{ph} present respectively the PV current, voltage, saturation current and photocurrent. The series and parallel resistances are R_p and R_s. N_s and N_p are the total number of solar modules connected in series and parallel respectively. T is the PV surface temperature. n, E_g and k are the ideality factors, gap energy and Boltzmann constant, respectively.

TABLE I. CHARACTERISTICS OF PV PANEL AND BOOST CONVERTER

Electrical Parameters	Values
Solar PV source	
Maximum power P_{mpp}	*345 W*
Maximum Current I_{mpp}	*9.05 A*
Short-circuit current I_{sc}	*9.52 A*
Maximum voltage V_{mpp}	*38.14 V*
Open-circuit Voltage Voc	*46.52 V*
Temperature coefficient of current k_{sc}	+0.049 % /°C
Temperature coefficient of voltage k_{oc}	-0.315 % /°C
Series cells *Nsc*	72
Number of modules in series in a PV string $Ns_{,ch}$	7
Number of parallel strings in a PV string $N_{p,ch}$	80

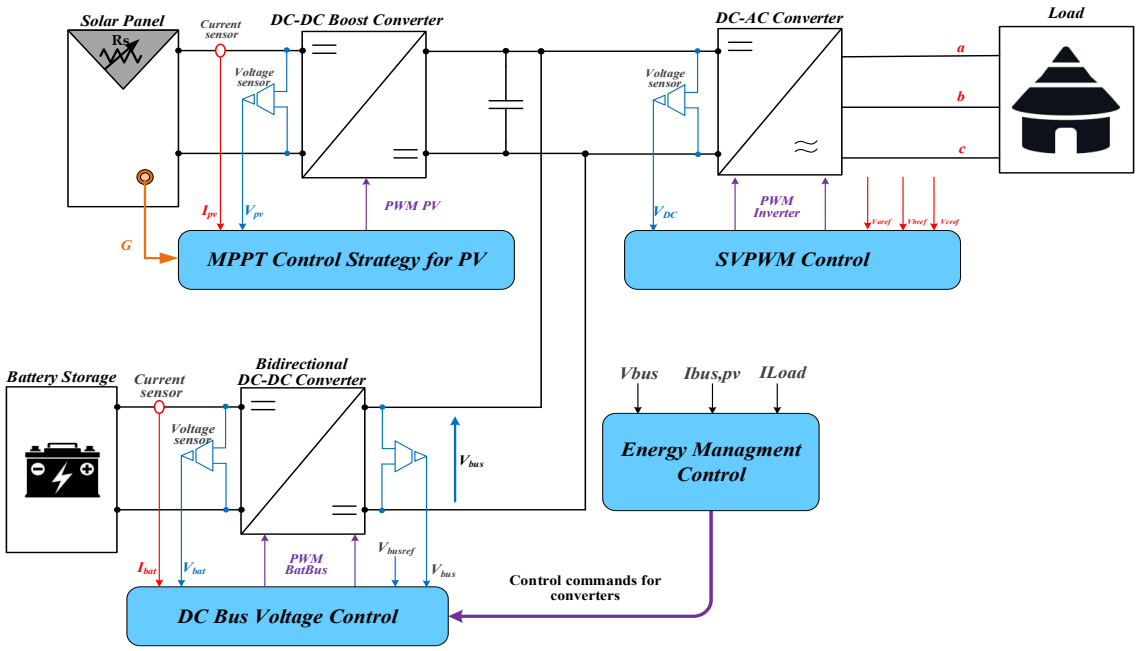


Fig. 1. Proposed power system topoloy.

B. Behavior Modeling of the Batteries

The batteries are high energy density, quick dynamic response, and low rate of self-discharge make it a promising technology for storing renewable energy in hybrid systems. As depicted in Fig. 1, the batteries are connected to the DC-bus via a bidirectional DC-DC converter, enabling efficient DC power supply to the load. The considered model of the batteries is presented in equation (2), [17], [18].

$$\begin{cases} V_{bat} = E_0 - K \cdot \dfrac{Q}{Q - it}\, it - R_b \cdot i_{bat} + A_b\, e^{-B \cdot it} - K \cdot \dfrac{Q}{Q - it}\, i^* \text{(Discharge)} \\ V_{bat} = E_0 - K \cdot \dfrac{Q}{Q - it}\, it - R_b \cdot i_{bat} + A_b\, e^{-B \cdot it} - K \cdot \dfrac{Q}{it + 0.1Q}\, i^* \text{(Charge)} \end{cases} \quad (2)$$

$$\text{With } it = \int_0^t i_{bat} dt \text{ and } i^* = LFP(i_{bat})$$

In equation (2), the open-circuit voltage is denoted by E_0. Q is the capacity (Ah) of a typical battery. i_t is the battery's current charge (Ah). The polarization constant is denoted by K. The exponential zone amplitude (in V) is shown by A_b. B represents the exponential zonetime constant inverse in the exponential zone $(\text{Ah})^{-1}$. The internal resistance (in Ω) is denoted by R_b. The battery's current is denoted by i_{bat}, and the filtered current (in A) by i^*.

The battery design was modeled using simulation based on the technical specifications listed in Table II, which also includes the parameters of the bidirectional DC-DC converter. The model describing the battery's state of charge SoC_{bat}, is evaluated using equation (3) [18], [19], where SoC_{bat} is the battery state of charge (%), Q_{bat} is the maximum battery capacity (Ah).

$$SOC_{bat} = 100 \cdot \left(1 + \frac{1}{Q_{bat}} \cdot \int_0^t i_{bat}(t)\, dt \right) \quad (3)$$

The battery's charge-discharge cycle is mainly determined by the amount of power available and the power level required by the system. This cycle depends on both available energy capacity and demand dynamics. The battery's State of Charge (SoC) limits are used to define operational constraints, framing

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

minimum and maximum operating thresholds. These limits are used to set safety limits and optimize battery use according to the actual power capacities that can be supplied or absorbed at any given time, as described in equation (4), where SoC_{batmin} and SoC_{batmax} are respectively the state of charge minimum and maximum.

$$SOC_{bat\,\min} \le SOC_{bat} \le SOC_{bat\,\max} \quad (4)$$

TABLE II. CHARACTERISTICS OF BATTERY AND BIDIRECTIONAL CONVERTER

Electrical Parameters	Values
Battery source	
Battery capacity: Q	648 Ah
Constant voltage: E_0	273.2 V
Polarization constant: *K*	0.0029 (V/Ah)
Battery internal resistance R_b	0.0038 (Ω)
Exponential zone amplitude A_b	21.16 (V)
Exponential zone time constant inverse B	0.094 $(Ah)^{-1}$
SOCbat	50%

C. Modeling of the DC–AC Converter

The considered DC-AC converter is based on two-level three-phase inverter designed to convert the energy from the DC-bus into alternating current (AC) suitable for supplying domestic loads. It consists of three legs corresponding to the three phases, each including two power switches based on IGBT technology equipped with antiparallel diodes. Inverter is controlled using the Space Vector Pulse Width Modulation (SVPWM) technique, which enables the generation of a quasi-sinusoidal three-phase voltage, thereby improving power quality while reducing harmonic distortion.

Furthermore, an LC filter is installed at the inverter output to smooth the voltage and current waveforms by attenuating the harmonic components generated by the switching of the power devices. This filtering significantly improves the quality of the supplied energy by ensuring a voltage waveform closer to a pure sinusoid at a frequency of 50 Hz, making it suitable for stand-alone micro-grid applications (Fig. 2).

III. ENERGY MANAGEMENT STRATEGIES

This section provides an in-depth explanation of the MPPT control principle, and the energy management strategy proposed in this paper.

A. Principle of MPPT Control for PV

The photovoltaic (PV) panel, as a generator of electrical energy, converts solar irradiance into electricity through a direct conversion process. Nevertheless, this conversion is subject to the intrinsic variability of solar resources, which induces intermittent and non-linear behavior in energy production. This characteristic adversely affects the stability and performance of the PV system. To mitigate these effects and optimize energy capture, the integration of a dedicated control system is essential.

To do this operation, the MPPT algorithms play a crucial role, adapting the operating point of the PV system in real time to extract the maximum available power. Several MPPT strategies have been proposed in the literature [3], [20], [21], aiming to autonomously control the voltage of each PV module, in order to maximize energy efficiency under varying sunlight and temperature conditions.

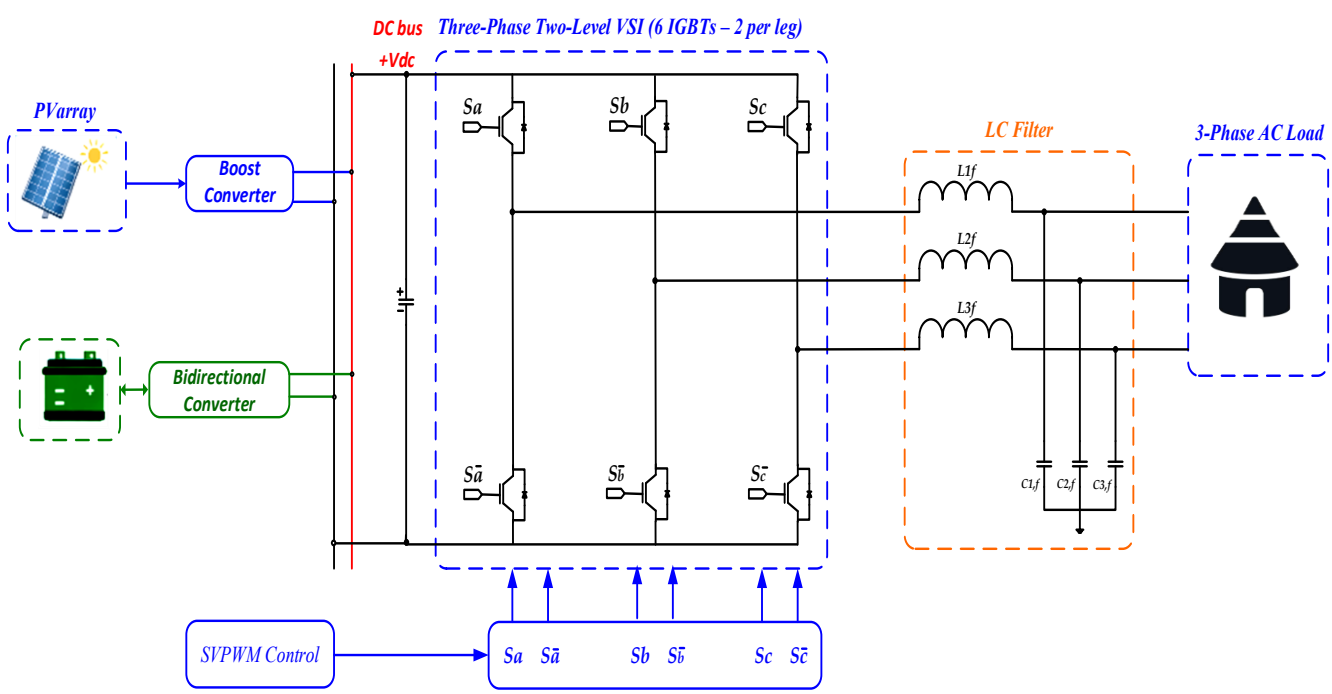


Fig. 2. DC–AC inverter structure and control.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

In the present work, the authors have opted for the control strategy with the integrated Incremental Conductance (InC) algorithm, whose principle is illustrated in Fig. 3, as the method for tracking the maximum power point. The InC algorithm is based on an analysis of the power gradient as a function of voltage (*dP/dV*), enabling precise estimation of the optimum operating point. In combination with the irradiance sensor, it provides the reference current that is calculated using equation (5). Figure 4 presents the Incremental Conductance (InC) MPPT algorithm.

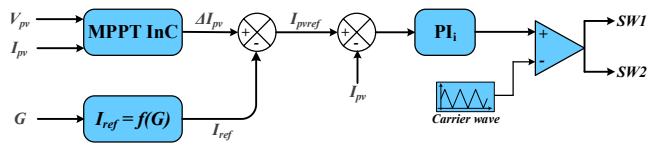


Fig. 3. MPPT Control Strategy.

$$I_{ref} = \left(I_{sc} + k_{sc} \cdot (T - 298.15)\right) \cdot \left(\frac{G}{1000}\right) \quad (5)$$

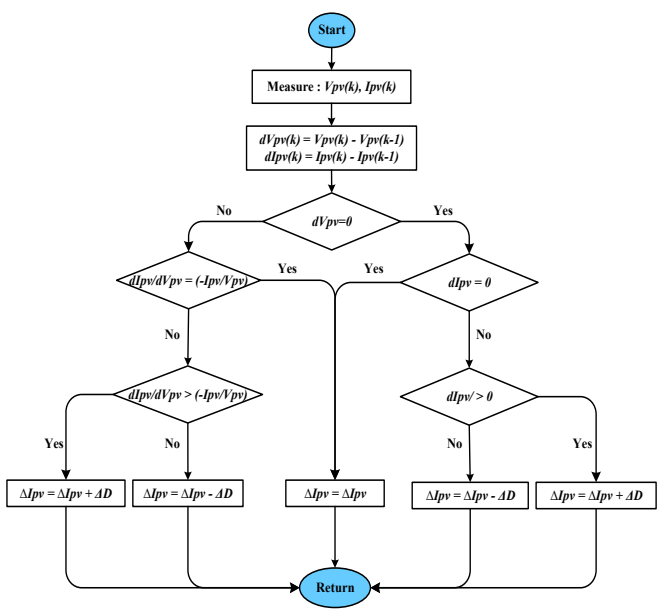


Fig. 4. Diagram of the InC MPPT.

B. Energy Management Strategy and Control (EMSC)

The objective of this study is to develop an energy management strategy within a micro-grid based on the use of a Proportional-Integral (PI) controller, to optimize energy exchanges between the different components of the system. A key aspect of this approach is to ensure the stability of the DC-bus voltage, which must be maintained at a predefined reference level despite external disturbances and inherent system uncertainties. Indeed, controlling this voltage is essential to guarantee the proper operation, reliability, and safety of the micro-grid.

The proposed method is based on minimizing the deviation between the measured DC-bus voltage V_{bus} and its reference value V_{busref} through a robust and structured control strategy. The operation of the micro-grid requires continuous adaptation to fluctuations in photovoltaic power generation. In this context, the batteries energy management plays a crucial role in preventing system instabilities caused by excess or insufficient production, while ensuring a continuous balance between energy supply and demand.

Since the batteries module is connected to the DC-bus through a bidirectional DC-DC converter, the main role of the PI controller is to operate this converter in either buck or boost mode, depending on the energy requirements. The reference current, generated by the PI voltage controller, is then used to determine the duty cycle applied to the pulse-width modulation (PWM) module, thereby ensuring effective control of the converter.

As illustrated in Fig. 2, photovoltaic energy is first injected into the DC-bus of the micro-grid and then delivered to the load through an inverter. In the event of excess production, the surplus energy is stored in the batteries. The power transfer from the DC-bus to the load is ensured by the inverter, whose control, combined with an LC filter at its output, improves the quality of the energy supplied to the consumer.

This strategy ensures accurate regulation of both current and voltage while maintaining an overall energy balance. It promotes optimal using of solar energy due to batteries assistance.

The structure of the control illustrated in Fig. 5, is based on a PI control loop that generates the appropriate duty cycle to ensure tracking of the reference voltage V_{busref}, set at 1000V. This approach enables the batteries charging and discharging processes without the degree of freedom limit due to cascade control architectures.

The k_p and k_i parameters of the proportional-integral controllers applied to the photovoltaic current and the DC-bus voltage were determined using equation (6) to optimize the dynamic response of the system. The optimal values obtained for these coefficients are presented in Table III. The system performance is subsequently tested through simulations.

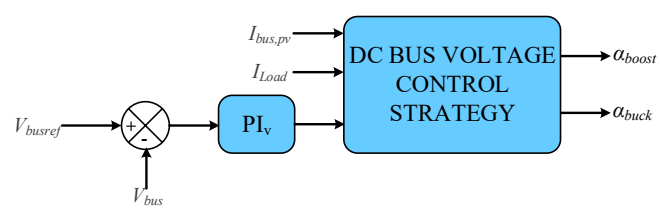


Fig. 5. DC-bus voltage control.

$$\begin{cases} k_{pi} = L \cdot \omega_n^2 \\ k_{ii} = 2 \cdot \varepsilon . L \cdot \omega_n^2 \\ k_{pv} = C \cdot \omega_n^2 \\ k_{iv} = 2 \cdot \varepsilon . C \cdot \omega_n^2 \end{cases} \quad \begin{cases} \omega_n = 2 \cdot \pi \cdot \frac{f}{10} \\ \varepsilon = \frac{\sqrt{2}}{2} \\ \tau_{(i,v)} = \frac{k_{p(i,v)}}{k_{i(i,v)}} \end{cases} \quad (6)$$

The current and voltage control bandwidth ω_n is fixed to 10% of that the converters control frequency [5], [7].

Figure 6 presents the flowchart of the proposed control strategy. The batteries energy storage system performs a dual function, both supplying energy when required and absorbing excess production. In stand-alone mode, when generation exceeds demand, the algorithm controls the batteries charging

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

process while respecting its safety constraints, particularly the maximum allowable current.

TABLE III. PI CONTROLLER PARAMETERS

Coefficients	Values
k_{pi}	0.1
k_{ii}	100
k_{pv}	0.1
k_{iv}	1
f	10kHz
L	1mH
C	3300µF

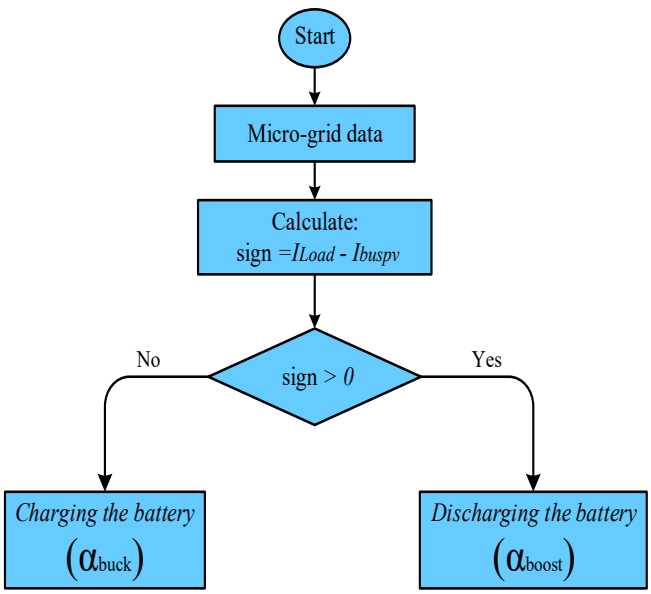


Fig. 6. Voltage control strategy.

Conversely, in the event of insufficient photovoltaic generation, the batteries supply the required energy to meet the load's demand until its discharge limit is reached. Furthermore, the energy management system continuously monitors the batteries state of charge, thereby ensuring reliable and secure operation of the micro-grid.

C. *Control of Inverter Using SVPWM*

Grid-connected PV systems represent a key solution for sustainable energy production; however, their efficient integration requires advanced power conversion and control strategies. In [22], a new three-phase three-level NPC (Neutral Point Clamped) inverter structure combined with the SVPWM (Space Vector Pulse Width Modulation) technique is proposed for injecting PV energy into the grid without storage, using a high-gain interleaved Boost converter. The results demonstrate a high efficiency of 96.1%, although the output voltage THD (Total Harmonic Distortion) remains relatively high (22.33%).

In [23], an MPPT control strategy based on the incremental conductance algorithm combined with a dual closed-loop technique is developed for a hybrid PV-battery microgrid, improving power tracking, energy flow management, and DC-bus voltage regulation in accordance with the EN50530 standard. Similarly, [24] proposes an improved MPPT approach coupled with *d-q* current regulation to stabilize the DC-link voltage and reduce harmonics; however, the absence of THD results limits a comprehensive evaluation of its performance.

Furthermore, [25] introduces an advanced Fractional Order Integral Sliding Mode Control (FOISMC) strategy optimized using the Particle Swarm Optimization (PSO) algorithm, applied to a PV system including a DC-DC converter and a three-phase inverter. This approach significantly enhances system dynamic performance, achieving notable reductions in voltage, active and reactive power, as well as overshoot and response time, highlighting the effectiveness of intelligent control techniques.

Finally, [26] emphasizes the importance of multilevel inverters in large-scale PV plants, showing that the choice of topology (CHB – Cascaded H-Bridge, NPC – Neutral Point Clamped, FC – Flying Capacitor) has a significant impact on overall system performance. While the CHB topology offers better voltage tolerance, NPC and FC structures simplify the system by eliminating the need for multiple DC sources. Overall, these studies demonstrate that improving PV system performance relies on a synergy between inverter topologies, control strategies, and modulation techniques, paving the way for more efficient, stable, and grid-compliant energy systems.

In this study, the authors propose SVPWM control as an inverter management strategy for the grid. SVPWM algorithm presented in Fig. 7 first converts the three-phase reference voltages into the stationary $\alpha\beta$ reference frame using the Clarke transforming.

The reference voltage vector is then located within one of the six sectors of the space vector diagram. The dwell times of the adjacent active vectors and zero vectors are calculated to synthesize the reference voltage during each switching period. Finally, the corresponding duty cycles are used to generate the gating signals for the six IGBT switches of the two-level inverter.

The sector identification is performed using a decision-based algorithm that avoids the computation of the arctangent function. The algorithm determines the sector of the reference voltage vector in the $\alpha\beta$ plane by analyzing the signs of V_α and V_β and comparing the ratio V_β/V_α with $\sqrt{3}$ as presented in Fig. 8. This method significantly reduces the computational complexity and is suitable for real-time implementation of the SVPWM strategy.

The proposed inverter is a two-level three-phase voltage source inverter connected to DC-bus and designed to deliver 150 kW/ 400 V AC/50 Hz three-phase AC load. A switching frequency of 10 kHz is adopted for the SVPWM signal generation. To improve the output voltage quality, an LC filter composed of a 1 mH phase inductance and a 25 µF phase capacitor is used. The DC-bus capacitor in inverter input is fixed to 2 mF to ensure DC-bus voltage stability.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

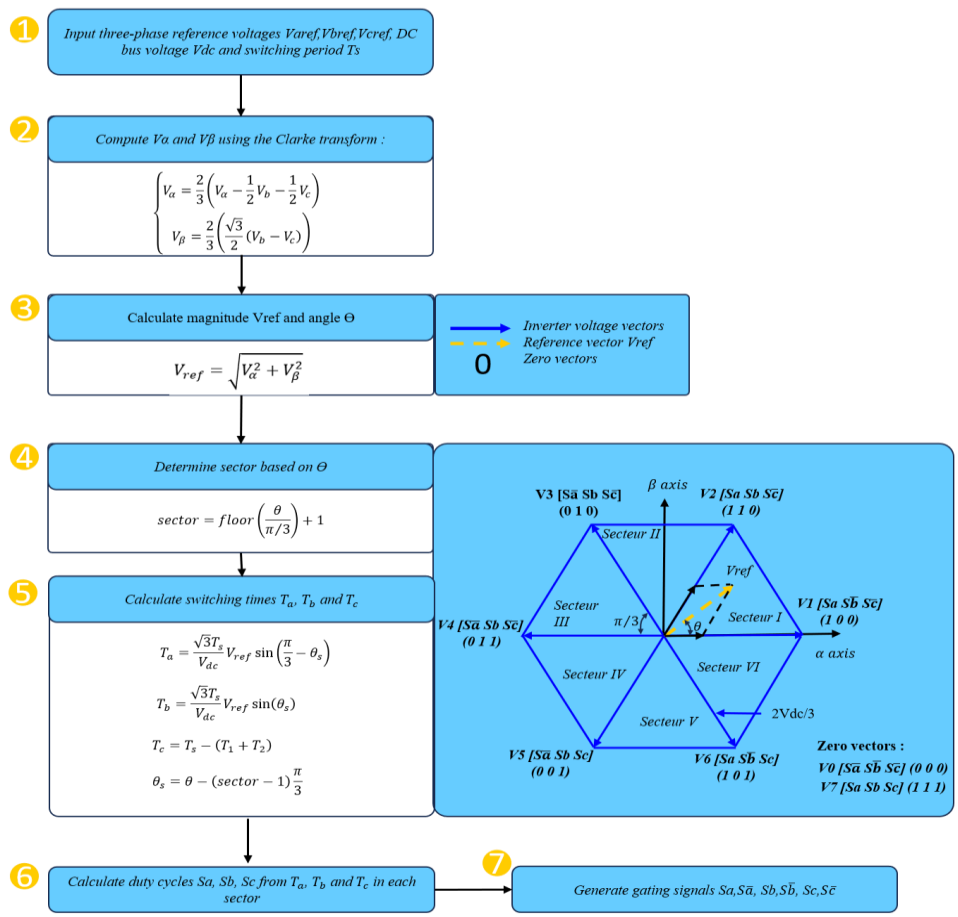


Fig. 7. Space Vector PWM Algorithm.

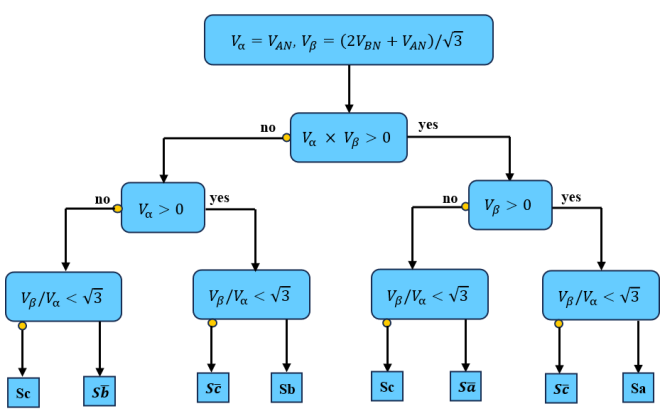


Fig. 8. Decision algorithm for sector identification in the SVPWM control strategy.

In order to accurately reproduce the dynamic behavior of the load, a time-varying current profile is imposed on the three-phase system. The studied system consists of a three-phase voltage source inverter supplied by a DC-bus voltage V_{dc} of $1000V$ .The inverter generates a three-phase AC voltage with a line-to-line RMS value is given in equation (7),

$$V_{LL} = \frac{m * V_{DC}}{\sqrt{3}} \quad (7)$$

at a fundamental frequency of 50 Hz.

For a balanced three-phase system, the phase voltage is given in equation (8),

$$V_{ph} = \frac{V_{LL}}{\sqrt{3}} \quad (8)$$

which yields $V_{ph} \approx 231$ V for $V_{LL} = 400V$ and $m = 0.69$.

TABLE IV. INVERTER PARAMETERS

Electrical Parameters	Values
Inverter	
Switching frequency fsw	*10 kHz*
Modulation index (SVPWM)	*0.69*
Output frequency	*50Hz*
Sampling time	*10^{-5} s*
SOCbat	*50%*
LC Filter at output of the inverter	
Filter inductance L_{1f}, L_{2f}, L_{3f}	*1 mH*
Filter capacitance C_{1f}, C_{2f}, C_{3f}	*25 μF*

The natural frequency of the filter is given in equation (9).

$$f_c = \frac{1}{2\pi\sqrt{L1f.C1f}} \quad (9)$$

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Thus, for the selected filter values in our system, the cutoff frequency f_c, after calculation, is approximately 1006Hz.

IV. Simulation Results and Discussion

After modeling the PV system, the energy storage device, the inverter, and the various control strategies, Matlab/Simulink software was used to implement and simulate the hybrid energy system to evaluate the performances of the proposed energy management strategy.

The DC-bus side is characterized by the interaction between the photovoltaic system and the batteries through the main DC-bus, via the respective Boost and Buck-Boost converters, which enable continuous power flow management through their control algorithms. The simulated climatic conditions presented in Fig. 9 include the variations of solar irradiance (in W/m²), and ambient temperature (in °C) as well as load's demand over a period of 1000 hours. These parameters affect the energy production of photovoltaic modules.

The load current profile I_{Load}, shown in Fig. 10, is interpolated and applied at the inverter output in a stepwise manner. This implementation is carried out using the *Three-Phase Breaker* and *Three-Phase Series RLC Branch* blocks in MATLAB/Simulink, configured in resistive mode, in order to accurately reproduce the real dynamic behavior of the measured load.

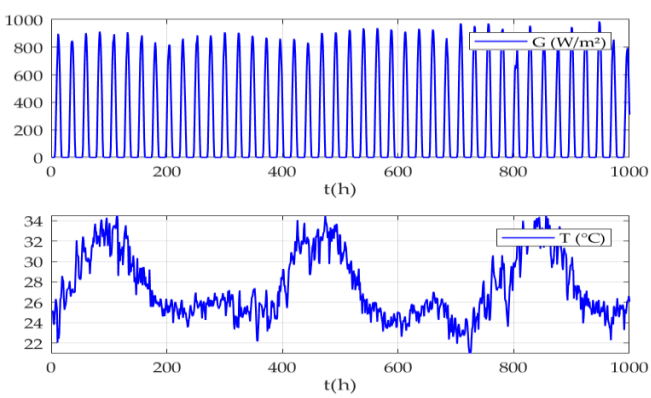


Fig. 9. Variations in experimental weather conditions, solar radiation, temperature.

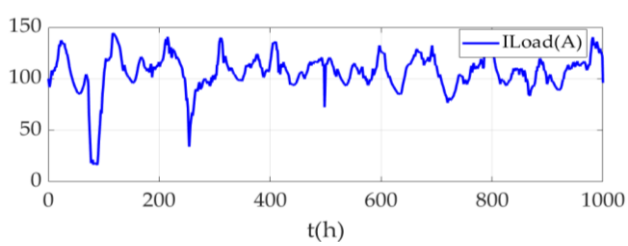


Fig. 10. Measured Load Current Profile

The curves presented in Fig. 11 show the temporal evolution of the PV modules voltage *Vpv* and its current I_{pv}. It can be observed that the system effectively responds to variations in irradiance and temperature, thereby ensuring optimal energy production. These results highlight the ability of the MPPT algorithm, along with the Boost converter control strategy, to dynamically adjust the operating point of the photovoltaic generator.

The waveforms illustrated in Fig. 12 highlight the measured current at the output of the PV array and the reference current generated by the MPPT controller through irradiance data provided by a pyranometer. The strong agreement between these two signals demonstrates the effectiveness of the incremental conductance algorithm, which is capable of rapidly adapting the reference current to variations in solar irradiance, thereby ensuring optimal energy production efficiency.

The DC-bus voltage is maintained around a reference value set at 1000 V. Its evolution during the simulations is presented in Fig. 13. Thanks to the action of the PI controller combined with the adopted control strategy, effective DC-bus regulation is achieved even in the presence of disturbances arising from the load connected to the inverter or from PV generation. Maintaining this voltage is essential to ensure the proper operation of all supplied devices, particularly the micro-grid loads.

The waveforms illustrated in Fig. 14 highlight the batteries behavior in response to system demands. The variations of the current correspond to the charging and discharging phases, depending on the availability of solar energy and the consumption requirements. The voltage remains within nominal ranges, reflecting the proper sizing of the bidirectional converter as well as the reliability of the energy management strategy.

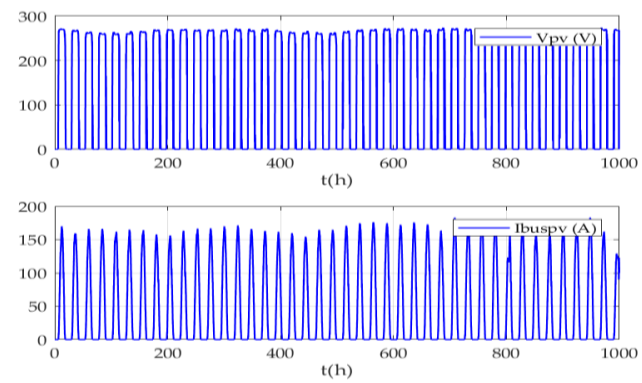


Fig. 11. Waveform of the current and voltage on the PV side.

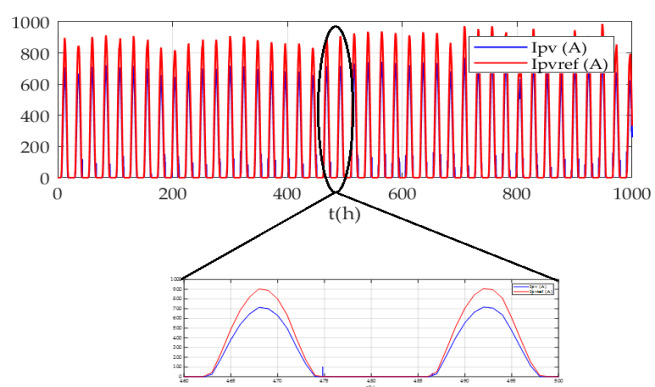


Fig. 12. Waveform of the current and its reference on the PV side.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

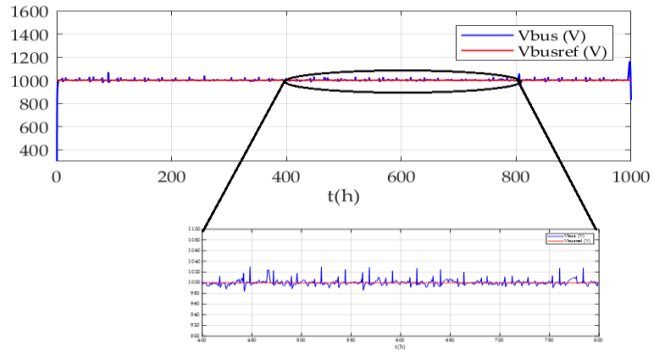


Fig. 13. DC-bus voltage.

The voltage across the batteries module terminals presented in Fig. 15 fluctuates around 269V, which is very close to its nominal value of 270 V. This slight decrease in voltage is due to the voltage drop caused by its internal resistance. The power supplied by the PV and batteries system is illustrated in Fig. 16. This result is obtained by emulating the real load profile on the DC-bus. The power compensation provided by the batteries when the PV power is zerocorresponds in practice to the period between sunset and sunrise.

Figure 17 shows the evolution of the three-phase currents i_a, i_b, i_c and the three-phase phase-to-neutral voltages at the inverter output v_a, v_b, v_c are presented in Fig. 18. Over the entire study period, amplitude variations are observed in both currents and voltages, reflecting the system's adaptation to fluctuations of the PV power and load demand. The zoom over the interval $t = 464.4\,s$ to $t = 464.44\,s$ highlights perfectly sinusoidal waveforms, balanced and phase-shifted by 120° between the three phases, which is characteristic of proper three-phase operation. It is important to emphasize that this current amplitude enables the simultaneous supply of multiple domestic loads or households in rural areas.

The similarity between the voltage waveforms v_a, v_b, v_c and the current waveforms i_a, i_b, i_c is explained by the resistive nature of the load, for which the current is proportional to the voltage. Furthermore, the regularity of the signals and the absence of significant distortion confirm the effectiveness of the SVPWM control strategy as well as the LC filter at the inverter output. These results demonstrate the high quality of the energy delivered to the micro-grid, in agreement with the low total harmonic distortion values.

Figure 19 presents the evolution of the duty cycles T_a, T_b, and T_c, normalized with respect to the switching period T_s, as well as the switching sector generated by the SVPWM control strategy applied to the three-phase inverter. Over the entire simulation period, the duty cycles vary continuously according to the system operating conditions, particularly fluctuations in photovoltaic generation and load's demand, while the sector variable takes inter values between 1 and 6 as illustrated in Fig. 20, corresponding to the different regions of the $\alpha\beta$ vector plane. The zoom over the interval $t = 464.4\,s$to $t = 464.44\,s$highlights periodic waveforms of the duty cycles, bounded between 0 and 1 and phase-shifted among the three phases, reflecting the time distribution of active and zero voltage vectors.

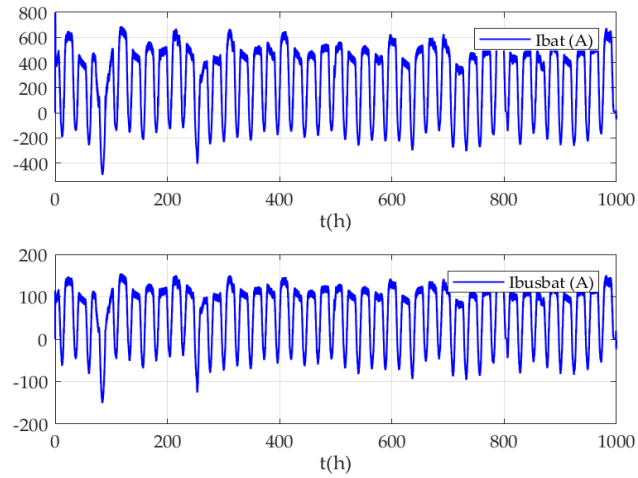


Fig. 14. Battery current input and output Waveform

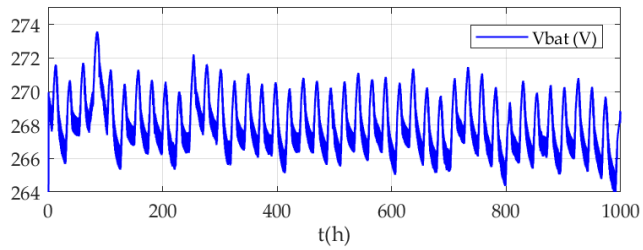


Fig. 15. Waveform of the voltage and current on the battery side.

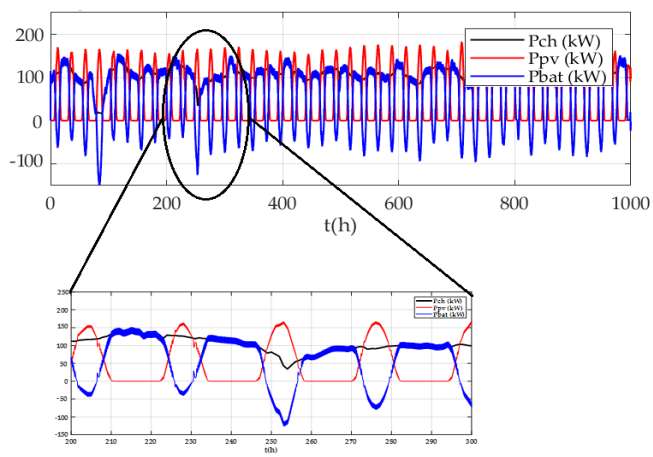


Fig. 16. Waveform of the power.

At the same time, the ordered succession of sectors reflects the continuous rotation of the reference voltage vector in the complex plane. Each sector transition corresponds to a new combination of active vectors, ensuring an accurate synthesis of the desired three-phase voltage. This coordination between duty cycles and sector selection

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

confirms the correct operation of the SVPWM algorithm. It ensures optimal utilization of the DC-bus voltage, efficient switching of the power devices, and a significant reduction of harmonic content. Consequently, these results demonstrate the capability of the control strategy to generate balanced, high-quality three-phase voltages, contributing to the overall performance improvement of the system.

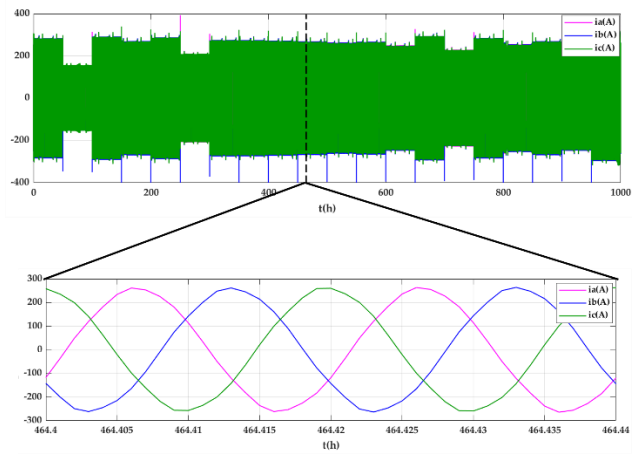


Fig. 17. Inverter output current.

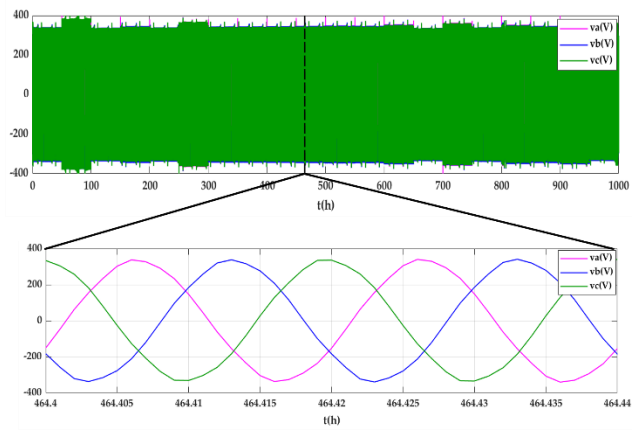


Fig. 18. Inverter output voltage.

Figures 21-23 present the PWM control signals of the inverter switches (S_1, S_4 , S_3, S_6 , S_5, S_2) along with the associated reference voltages ($V_{a,ref}$, $V_{b,ref}$, $V_{c,ref}$). Over the entire simulations period, the PWM signals appear as high-frequency rectangular waveforms, reflecting the rapid switching of the power devices, while the reference voltages maintain a sinusoidal shape. The zoom over the interval $t = 464.4$ *sto* $t = 464.44$ *s*highlights the fundamental principle of SVPWM modulation: the comparison between an implicit triangular carrier waveform and the sinusoidal reference signals, resulting in time-varying duty cycles. It can be observed that the pulse widths continuously vary according to the instantaneous amplitude of the reference voltages, enabling the reconstruction, after filtering, of balanced sinusoidal three-phase voltages.

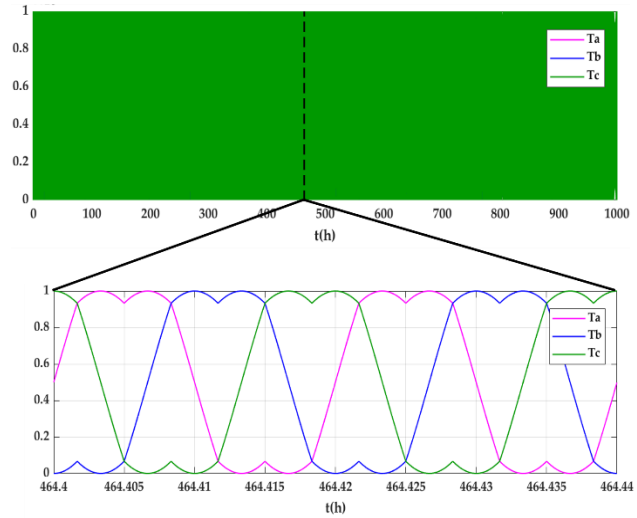


Fig. 19. Space vector PWM principle and switching sequence.

The complementary switching signals (S_1/S_4 , S_3/S_6 , S_5/S_2) ensure alternate control of the switches in each inverter leg, guaranteeing the proper synthesis of voltage levels. This pulse-width modulation accurately reflects the trajectory of the reference voltage vector in the $\alpha\beta$ plane, in accordance with the SVPWM principle.

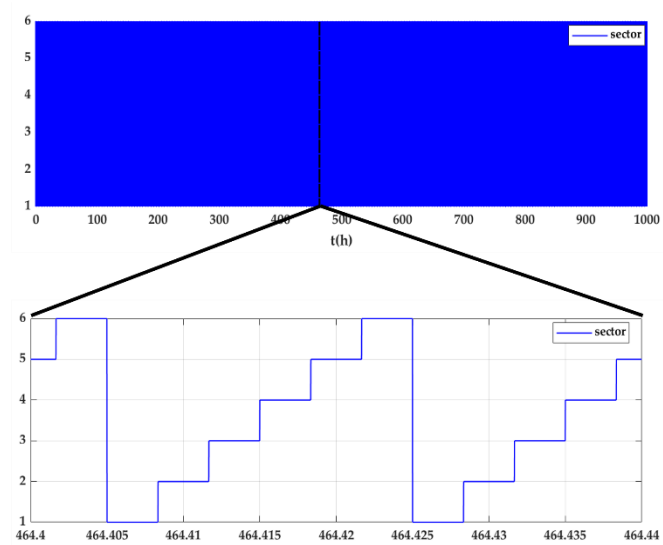


Fig. 20. Waveform sector

The consistency between the PWM signals and the reference voltages confirms the correct operation of the control algorithm. It ensures optimal utilization of the DC-bus voltage, efficient switching of the devices, and a significant reduction in harmonic contents. Consequently, these results demonstrate the capability of the SVPWM strategy to generate high-quality three-phase voltages, contributing to the overall performance improvement of the system.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

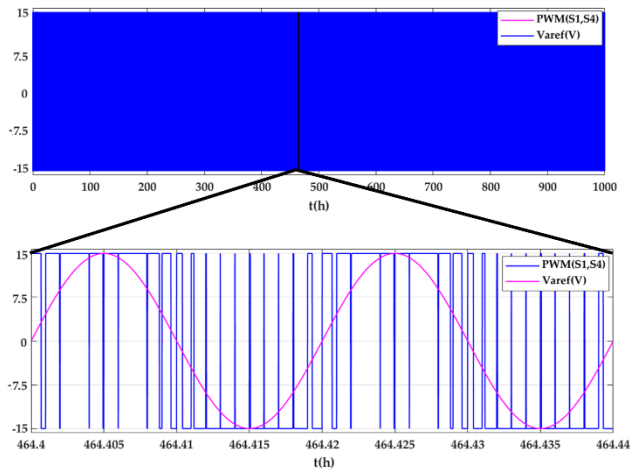


Fig. 21. PWM waveform and reference voltage Varef

Figure 24 shows the temporal evolution of the active power P supplied to the system over a period of 1000 hours on the DC-bus. It can be observed that the power varies dynamically according to operating conditions, particularly due to the power fluctuations of the PV and load in. The maximum power reached is approximately 150kW, which corresponds to the system's maximum capacity. The power variations reflect changes in operating conditions, particularly transitions between different levels of solar irradiance and load demand. The rapid oscillations observed around the power levels are associated with converter switching phenomena and control actions, without significantly affecting the overall stability of the system. Despite these fluctuations, the power remains generally well-regulated around consistent levels, highlighting the effectiveness of the energy management strategy. The system ensures an appropriate distribution of the power between the PV, the batteries storage, and the load, while respecting the maximum power limit and maintaining stable operation of the micro-grid.

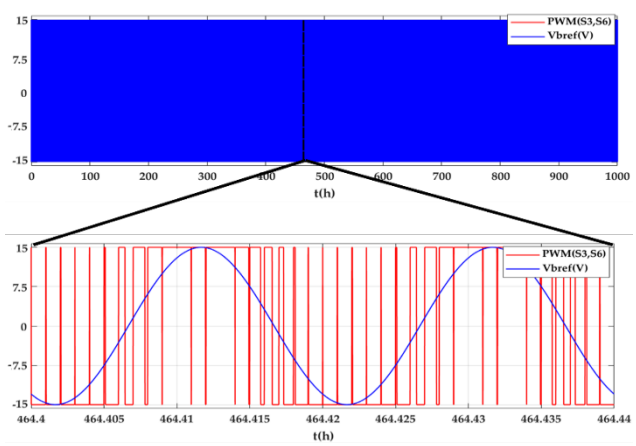


Fig. 22. PWM waveform and reference voltage Vbref

Figures 25 and 26 present the frequency analysis of the currents i_a, i_b, i_c as well as the voltages v_a, v_b, v_c at the inverter output, obtained using Fourier transform (FFT). It can be observed that the spectrum is dominated by the fundamental component at 50 Hz, corresponding to the grid frequency, while the harmonic components are very low. The total harmonic distortion (THD) values are around 2.4% for all phases, for both currents and voltages.

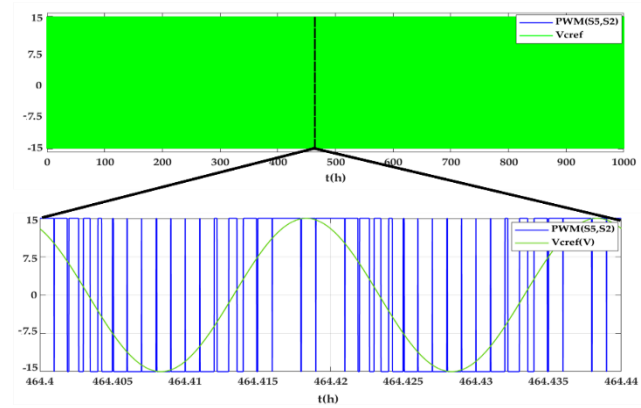


Fig. 23. PWM waveform and reference voltage Vcref

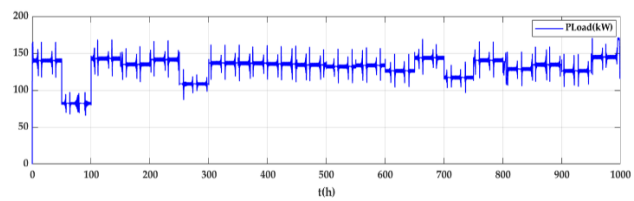


Fig. 24. Dynamic variation of load power.

The similarity of THD values between currents and voltages is explained by the resistive nature of the load, for which the current is proportional to the voltage, resulting in similar waveforms. Furthermore, the low amplitude of harmonics highlights the effectiveness of the SVPWM control strategy as well as the LC filter at the inverter output, which significantly attenuate the high-frequency components generated by switching. These results demonstrate a good power quality delivered to the micro-grid, meeting power quality standards, and confirm the adequacy of the filter design as well as the adopted control strategies.

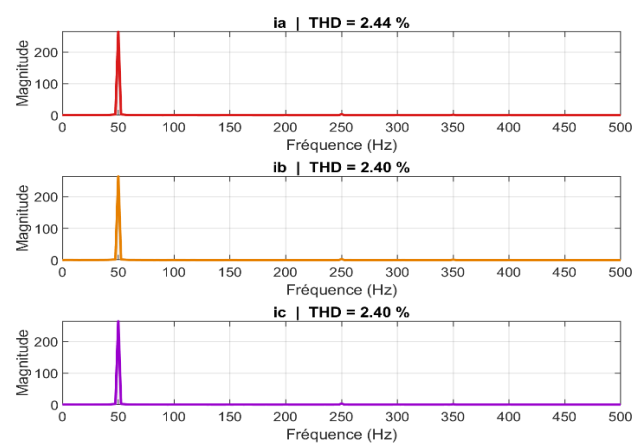


Fig. 25. Frequency analysis of the currents ia,ib and ic

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

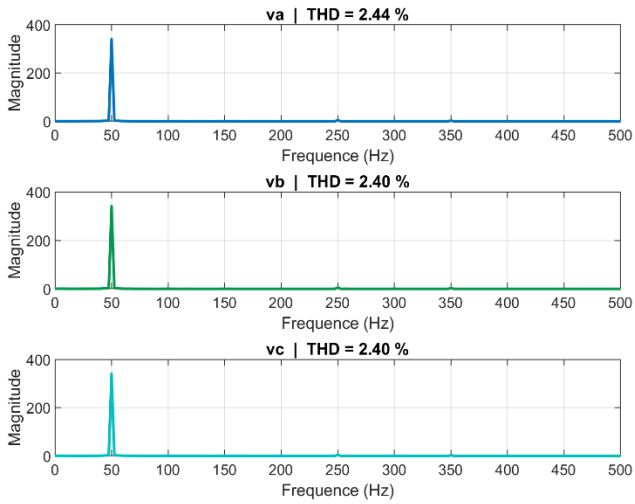


Fig. 26. Frequency analysis of the voltages va,vb and vc

V. Conclusion and Future Work

Furthermore, the DC-bus voltage is stably regulated around 1000 V despite variations in solar irradiance and load conditions. The batteries storage system plays a key role in energy balancing through the batteries charging and discharging operation according to production and consumption conditions. The observed power flows confirm an efficient distribution of energy between the PV, the batteries, and the load.

The three-phase inverter, controlled using the SVPWM technique and combined with an LC filter at its output, enables the generation of balanced and sinusoidal three-phase voltages and currents. The obtained results show a strong agreement between the reference signals and the actual quantities, as well as a good power balance. Frequency analysis highlights a low total harmonic distortion (THD) of 2.44% for both voltages and currents, confirming the high quality of the energy supplied to the micro-grid. This performance is mainly attributed to the effectiveness of the SVPWM strategy and the proper design of the LC filter.

The simulations are done under realistic conditions, demonstrate the relevance of the proposed strategy. The system ensures high power quality (low THD), stable operation, and optimized energy management, making it particularly suitable for stand-alone micro-grid applications in rural areas.

Acknowledgment

This work is done at the GREAH laboratory, Université Le Havre Normandie, France, by the Renewable Energies and Storage Systems (MERS) research team, as part of the 1000PhD and 5000 Masters train-the-trainer project initiated by the Ministry of Higher Education, Scientific Research and Innovation of the Republic of Guinea (MESRSI-GUINEE). The authors would like to thank the Guinean authorities for their financial support, and the University of Le Havre Normandie for the research activity support to ensuring the success of this project.

References

[1] M. L. Touré, M.-B. Camara, C. B. Nzoundja Fapi, A. Payman, and B. Dakyo, Eds., “Energy Management of PV-Batteries System for a Rural Micro-Grid Application in Guinea,” presented at the International Conference on Green Communications, Computing and Technologies, Wilmington, DE, USA: IARIA, 2025.

[2] A. Chellakhi and S. El Beid, “A Real-Time Investigation of an Enhanced Variable Step PO MPPT Controller for Photovoltaic Systems Using dSPACE 1104 Board,” *Energies*, vol. 18, no. 13, p. 3343, Jun. 2025, doi: 10.3390/en18133343.

[3] M. L. Touré, M. B. Camara, and B. Dakyo, “Symmetrical Multilevel High Voltage-Gain Boost Converter Control Strategy for Photovoltaic Systems Applications,” *Electronics*, vol. 13, no. 13, p. 2565, Jun. 2024, doi: 10.3390/electronics13132565.

[4] C. B. Nzoundja Fapi *et al.*, “Fuzzy Logic-Based Maximum Power Point Tracking Control for Photovoltaic Systems: A Review and Experimental Applications,” *Arch Computat Methods Eng*, vol. 32, no. 4, pp. 2405–2428, May 2025, doi: 10.1007/s11831-024-10210-7.

[5] C. H. Swetha, N. S. Jayalakshmi, K. M. Bhargavi and P. B. Nempu, “Control strategies for power management of PV/battery system with electric vehicle,” *In 2019 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER) (pp. 1-6). IEEE.*

[6] I. Oukkacha; M. B. Camara and B. Dakyo, “Energy Management in Electric Vehicle based on Frequency sharing approach, using Fuel cells, Lithium batteries and Supercapacitors,” *in Inter. Conf. on Renewable Energy Research and Applications (ICRERA2018), Paris, France, (pp. 986-992), IEEE.*

[7] M. O. Badawy and Y. Sozer, “Power flow management of a grid tied PV-battery system for electric vehicles charging”, IEEE Transactions on Industry Applications,” vol. 53, no. 2, pp. 1347–1357, 2017.

[8] A. Baqar, M. B. Camara, and B. Dakyo, “Supercapacitors Fast Ageing Control in Residential Microgrid Based Photovoltaic/Fuel Cell/Electric Vehicle Charging Station,” *Energies*, vol. 16, no. 13, p. 5084, Jun. 2023, doi: 10.3390/en16135084.

[9] M. P. Bonkile and V. Ramadesigan, “Power management control strategy using physics-based battery models in standalone PV-battery hybrid systems,” *Journal of Energy Storage*, vol. 23, pp. 258–268, Jun. 2019, doi: 10.1016/j.est.2019.03.016.

[10] S. Ahmad, M. Shafiullah, C. B. Ahmed, and M. Alowaifeer, “A Review of Microgrid Energy Management and Control Strategies,” *IEEE Access*, vol. 11, pp. 21729–21757, 2023, doi: 10.1109/access.2023.3248511.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[11] J. Hu, Y. Shan, K. W. Cheng and S. Islam, "Overview of power converter control in microgrids—Challenges, advances, and future trends", IEEE Transactions on Power Electronics," vol. 37, no. 8, p. pp. 9907–9922, 2022.

[12] S. J. Yaqoob, H. Arnoos, M. A. Qasim, E. B. Agyekum, A. Alzahrani, and S. Kamel, "An optimal energy management strategy for a photovoltaic/li-ion battery power system for DC microgrid application," *Front. Energy Res.*, vol. 10, Jan. 2023, doi: 10.3389/fenrg.2022.1066231.

[13] Y. Alidrissi, R. Ouladsine, A. Elmouatamid, and M. Bakhouya, "An Energy Management Strategy for DC Microgrids with PV/Battery Systems," *J. Electr. Eng. Technol.*, vol. 16, no. 3, pp. 1285–1296, May 2021, doi: 10.1007/s42835-021-00675-y.

[14] M. Benzaouia, A. Rabhi, B. Hajji, S. Benzaouia, H. Midavaine, and B. K. Oubbati, "Real-Time Control and Power Management Strategies of PV/Battery Standalone System," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 9135–9140, 2023, doi: 10.1016/j.ifacol.2023.10.151.

[15] C. B. Nzoundja Fapi, M. L. Touré, M.-B. Camara, and B. Dakyo, "Control Strategy for DC Micro-Grids in Heat Pump Applications with Renewable Integration," *Electronics*, vol. 14, no. 1, p. 150, Jan. 2025, doi: 10.3390/electronics14010150.

[16] S. R. Fahim, H. M. Hasanien, R. A. Turky, S. H. E. A. Aleem, and M. Ćalasan, "A Comprehensive Review of Photovoltaic Modules Models and Algorithms Used in Parameter Extraction," *Energies*, vol. 15, no. 23, p. 8941, Nov. 2022, doi: 10.3390/en15238941.

[17] M. B. Camara, H. Gualous, F. Gustin, A. Berthon, and B. Dakyo, "DC/DC Converter Design for Supercapacitor and Battery Power Management in Hybrid Vehicle Applications—Polynomial Control Strategy," *IEEE Trans. Ind. Electron.*, vol. 57, no. 2, pp. 587–597, Feb. 2010, doi: 10.1109/tie.2009.2025283.

[18] H. Rezk, A. M. Nassef, M. A. Abdelkareem, A. H. Alami, and A. Fathy, "Comparison among various energy management strategies for reducing hydrogen consumption in a hybrid fuel cell/supercapacitor/battery system," *International Journal of Hydrogen Energy*, vol. 46, no. 8, pp. 6110–6126, Jan. 2021, doi: 10.1016/j.ijhydene.2019.11.195.

[19] B. Benlahbib *et al.*, "Experimental investigation of power management and control of a PV/wind/fuel cell/battery hybrid energy system microgrid," *International Journal of Hydrogen Energy*, vol. 45, no. 53, pp. 29110–29122, Oct. 2020, doi: 10.1016/j.ijhydene.2020.07.251.

[20] M. A. Danandeh and S. M. Mousavi G., "Comparative and comprehensive review of maximum power point tracking methods for PV cells," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 2743–2767, Feb. 2018, doi: 10.1016/j.rser.2017.10.009.

[21] M. L. Katche, A. B. Makokha, S. O. Zachary, and M. S. Adaramola, "A Comprehensive Review of Maximum Power Point Tracking (MPPT) Techniques Used in Solar PV Systems," *Energies*, vol. 16, no. 5, p. 2206, Feb. 2023, doi: 10.3390/en16052206.

[22] R. Subbulakshmy and R. Palanisamy, "SVPWM control strategy for Novel Interleaved High Gain DC converter fed 3-level NPC Inverter for Renewable Energy Applications," *ISA Transactions*, vol. 140, pp. 426–437, Sep. 2023, doi: 10.1016/j.isatra.2023.05.019.

[23] S. Prajapati, R. Garg, and P. Mahajan, "Modified control approach for MPP tracking and DC bus voltage regulation in a hybrid standalone microgrid," *Electric Power Systems Research*, vol. 236, p. 110935, Nov. 2024, doi: 10.1016/j.epsr.2024.110935.

[24] I. Al-Wesabi, Z. Fang, H. M. Hussein Farh, A. A. Al-Shamma'a, and A. M. Al-Shaalan, "Comprehensive comparisons of improved incremental conductance with the state-of-the-art MPPT Techniques for extracting global peak and regulating dc-link voltage," *Energy Reports*, vol. 11, pp. 1590–1610, Jun. 2024, doi: 10.1016/j.egyr.2024.01.020.

[25] K. Boutaghane, N. Bennecib, M. Benidir, H. Benbouhenni, and I. Colak, "Performance enhancement of a three-phase grid-connected PV inverter system using fractional-order integral sliding mode controls," *Energy Reports*, vol. 11, pp. 3976–3994, Jun. 2024, doi: 10.1016/j.egyr.2024.03.049.

[26] A. Azizi, M. Akhbari, S. Danyali, Z. Tohidinejad, and M. Shirkhani, "A review on topology and control strategies of high-power inverters in large- scale photovoltaic power plants," *Heliyon*, vol. 11, no. 3, p. e42334, Feb. 2025, doi: 10.1016/j.heliyon.2025.e42334.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Investigation of Trap-Assisted Tunneling Currents in Capacitive Micromachined Ultrasonic Transducers (CMUT) with Silicon Nitride Membranes

Chukwuebuka Okoro, Sazzadur Chowdhury
Department of Electrical and Computer Engineering
University of Windsor
Windsor, Ontario, Canada
E-mail: okoro3@uwindsor.ca, Sazzadur@uwindsor.ca

***Abstract*—The results of a theoretical investigation of Poole-Frenkel and 3D Hartke models to predict tunneling currents due to electron emissions from trap centers in silicon nitride membranes in Capacitive Micromachined Ultrasonic Transducers (CMUTs) are presented. A CMUT, with a 0.5 micrometer thick silicon nitride membrane suspended over a 600 nanometer thick vacuum gap was used in the study. The study revealed that the Poole-Frenkel model overestimated electron current densities caused by emissions from silicon nitride trap centers by more than 99% in comparison to the Hartke model prediction. This was observed for a DC bias voltage range of 11-77 volts. This significant variation in predicted current densities is primarily due to the inclusion of multi-directional emission paths in the Hartke model. In contrast, the Poole-Frenkel model considers emissions only in one direction. The study provides valuable insights into the dielectric charging mechanisms in CMUTs, which can aid in the development of solutions to enhance their reliability by minimizing dielectric charging.**

Keywords- Trap-assisted Tunneling; Current Density; Hartke; Poole-Frenkel; CMUT; Silicon nitride; Dielectric Charging.

I. INTRODUCTION

Capacitive Micromachined Ultrasonic Transducers (CMUTs) rely on electrostatically or acoustically forced vibration of a microfabricated membrane suspended over a sealed vacuum cavity to transmit or receive ultrasound [1]. The CMUTs offer attractive functional characteristics, such as higher fractional bandwidth, lower mechanical impedance, lower internal loss, better thermal dissipation, low cost batch fabrication capability, and easier integration with microelectronic integrated circuits to achieve better signal-to-noise ratio that are superior to the conventional piezoelectric transducers [2]–[7].

The CMUTs are particularly envisioned to overcome the limitation of piezoelectric transducers for high-resolution 2D and 3D non-ionizing multi-modal biomedical and Non-Destructive Evaluation (NDE) applications [8]–[14]. Leveraging cost-effective advanced lithography along with precise micro and nanoscale deposition and etching techniques, it is now feasible to fabricate high-frequency linear, planar, non-planar, and flexible CMUT arrays [1].

As the CMUTs and microelectronic Integrated Circuits (ICs) are fabricated using similar microfabrication processes, the CMUTs can easily be integrated with microelectronic ICs to realize 3D-integrated ultrasound microsystems [13]. Such microsystems, by incorporating both the transducers and processing electronics as chiplets in a multifunctional System-In-Package (SIP), can provide high-quality data with a better Signal-to-Noise Ratio (SNR) due to shorter vertical interconnection paths and smaller in-plane routing distance. Examples of such advanced capabilities include dual-mode operation [14], e.g., the same array performing imaging and High-Intensity Focused Ultrasound (HIFU) for conventional therapy, histotripsy (non-thermal tissue ablation), ultrasound elastography, catheter-based invasive high frequency probing of internal body organs, and transparent arrays for 2D and 3D photoacoustic imaging [8]–[15]. Thus, the CMUTs unfold a new set of advanced functionalities that are not achievable with conventional piezoelectric transducers.

Despite these excellent competitive edges, approximately a quarter of a century after their invention in 1996, the CMUTs are still struggling to become the choice of technology in the industries due to some serious reliability issues, of which the electrostatic instability [1], [16]–[18], specifically the dielectric charging phenomena, is considered to be the most serious.

The dielectric charging phenomenon in CMUTs refers to the gradual accumulation of tunneled or trapped charges within the dielectric structural layer of the CMUT membrane, which limits the long-term reliability of the CMUTs. The studies in [1], [16]–[18] categorically noted that the dielectric charging problem in CMUTs must be resolved for the main stream adoption of CMUTs in the biomedical and NDT fields.

The authors in [18] categorized the dielectric charging mechanism in CMUTs into two primary modes: contact charging and non-contact charging. The contact charging mode is triggered when the CMUT membrane collapses on to the backplate as the electrostatic force originating from the applied voltages exceeds the collapse threshold. This leads to a high electric field near the contact interface; promoting aggressive charge injection and dipole alignment (two equal and opposite charges orient themselves due to the applied electric field). This accelerates performance degradation.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

The non-contact charging occurs when the CMUT membrane remains suspended without collapsing into the dielectric layer. In this case, charges arise due to internal defects, migration of free charges, and dipole reorientation within the dielectric materials.

In [18], the authors investigated the dominant electrical transport mechanism in typical dielectric materials, silicon nitride and silicon dioxide, used to fabricate CMUT membrane and support posts. The investigation utilized established charge transport models and employed parametric analysis in addition to a dual mode classification of charging mechanism in CMUT materials and geometries. The analysis revealed that membrane thickness, cavity height, thickness of the bottom insulating layer, and the operating temperature, significantly contribute to the dielectric charging phenomenon in CMUTs, providing a better understanding of CMUT charging behavior, aiding in improved design strategies.

Recent studies suggest that the quantum tunneling effect and internal Schottky effect are significant contributors to this charge injection and trapping under high electric fields [19][20]. The Poole-Frenkel (PF) [21]–[23] and Hartke [24][25] models, are two widely used theoretical frameworks to predict trap-assisted charging current density in CMUTs with silicon nitride membranes. On the other hand, the Fowler-Nordheim model [26] is commonly used to predict the charging current density due to the quantum tunneling [19] of charges from the metal electrode to the silicon nitride structural layer in a CMUT membrane.

Silicon nitride is a common dielectric structural material used to fabricate CMUT membranes due to its highly attractive mechanical properties and ease of fabrication. However, the silicon nitride membranes inherently contain structural defects, known as *K*-center defects [21]–[23]. These defects arise during thinfilm deposition processes or from high electric field-induced stress, and they act as charge-trapping sites capable of capturing both electrons and holes [21]–[23]. Under a strong electric field, these trapped charges can be thermally or field-activated to escape into the conduction band, a process known as field-enhanced thermal emission or the Poole-Frenkel (PF) effect [21]–[23].

In the PF model, the applied electric field reduces the Coulombic potential barrier of the trap, thereby enhancing the likelihood of charge emission along the field direction [20]. However, the PF model considers the field-induced potential barrier lowering only in one dimension and does not account for multi-directional emission paths.

On the other hand, the Hartke model extends this concept into three dimensions by averaging the field-lowered barrier over all possible emission angles, thereby offering a more accurate depiction of emission behavior in real materials. This is particularly significant at submicron dielectric thicknesses, which are typical in CMUTs, where charge injection and release occur over a distributed field profile [27].

Various design strategies involving variations in materials, geometries, and fabrication methods have been reported to minimize the dielectric charging phenomena in CMUTs. The authors in [28] presented post CMUT device as another solution to the dielectric charging. Post CMUT is composed of isolation posts of dielectric materials instead of a whole layer. In the design, the post isolation was implemented to reduce the surface area of the insulation layer, which significantly reduces the dielectric charging and its adverse effects. Unlike the conventional CMUTs, which showed shifts and shape changes in the current-voltage characteristics curve due to charge movement and trapping under high electric field, the post CMUTs exhibited stable CV characteristics with no noticeable charging or hysteresis even after high voltage cycling/sweep and spiking tests. This stability is attributed to the limited contact area provided by the posts, effectively isolating the dielectric layer and preventing charge accumulation.

In another technique as reported in [29][30], an enhanced insulation design for CMUTs has been presented that employs a deep trench oxidation method to form a centrally located, thermally grown silicon dioxide layer extending approximately 30 μm into the substrate. This effectively minimized the parasitic charging and device capacitance while also improving the breakdown voltage characteristics.

The charging behavior of various dielectric materials was also investigated in [17] with a focus on comparing plasma-enhanced silicon dioxide and plasma-enhanced silicon nitride using CV measurements. The results showed that charging began once the applied stress voltage reached the CMUT's pull-in voltage, and the plasma-enhanced silicon nitride exhibited a higher susceptibility to charge accumulation than the plasma-enhanced silicon dioxide. Further, the presence of both DC and AC voltage during operation led to more significant charging effects than the use of only DC voltage.

In [31], the authors presented experimental measurement data on dielectric charging in SOI CMUTs. The data showed a clear deviation of experimental values from the analytically estimated values of the capacitance changes as a function of AC and DC excitations. This deviation was attributed to the trapped charges in the silicon-on-insulator layer induced by the strong electric field. The measured capacitance values, which were 4–10 times higher than the estimated value, showed clear evidence of the charging in the dielectric layer. Additionally, the charging effect intensified with an increasing bias voltage. This approach presents a sound interpretation of dielectric charging mechanisms and advances CMUT design by introducing structural pathways for improving the device reliability.

Another highly reliable hexagonal shape CMUT cell structure with a diameter of 60 μm and cavity height of 100 nm was studied in [17], to address the issue of dielectric charging during high voltage operation. The study employed a 200 nm thick PECVD silicon nitride structural layer to form the CMUT membrane. Mechanical spacers or bumps were strategically positioned below the vibrating membrane, along with partial openings in the top membrane directly above these spacers. The primary function of the spacers is to physically limit the membrane's downward displacement, thereby preventing full contact with the bottom cavity surface even when the applied driving voltage exceeds the device collapse voltage. By constraining the membrane to

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

substrate contact area, the structure effectively minimizes current leakage through the dielectric during transient contact events, enhancing dielectric integrity. Furthermore, the introduction of partial electrode openings serves to reduce local electric field intensities at the regions of minimal electrode separation, particularly above each spacer. This redistribution of the field significantly improves the dielectric reliability of the CMUT cells compared to conventional configurations. Long-term experimental measurements as reported in [17] demonstrated that the proposed structure maintained stable receiving sensitivity, with fluctuations remaining below 1dB after more than 6×10^{11} membrane vibration cycles at voltages beyond collapse. Additionally, the design enabled a three times increase in transmitted acoustic pressure under the same driving conditions, indicating substantial performance gains. Collectively, these enhancements make the structure highly suitable for demanding medical imaging applications, where both high output and long-term reliability are critical.

In this context, this paper investigates both Poole-Frenkel and Hartke models predicted current densities in the context of CMUTs with silicon nitride membranes. The investigation results can be used for more accurate design of CMUTs to improve their reliability by decreasing the trap-assisted dielectric charging effects.

The remainder of this paper is structured as follows: Section II provides the operational and material aspects of a CMUT, Section III outlines the theoretical models governing field-induced barrier lowering and electron emission in CMUTs with silicon nitride membranes, Section IV presents the results of MATLAB-based analytical modeling, Section V provides a brief analysis of charging current density variation based on nitrogen composition, and finally Section VI provides the concluding remarks.

II. CMUT Operation and Material Considerations

The CMUT is a reciprocal electrostatic transducer. The cross-section of a typical CMUT geometry is shown in Figure 1. As Figure 1 shows, a CMUT geometry is constructed to have a flexible membrane at the top, which is anchored around its edges by dielectric support posts and suspended over a vacuum cavity formed between the top membrane and a fixed bottom plate [1]. The geometry essentially constituents a variable capacitor where the top membrane acts as a deformable upper electrode and the fixed plate at the bottom acts as the fixed bottom electrode. In some implementations, a thin insulating layer is added on the top of the bottom electrode to prevent device damage during a collapse mode operation. The top membrane is constructed with a highly doped single crystal silicon layer, or a functionalized dielectric structural layer such as silicon nitride or a suitable polymer.

A. 2.1 CMUT Operation

Approximating the CMUT geometry as a parallel plate variable capacitor as shown in Figure 2, if the upper plate is displaced by a distance x, the resulting capacitance of the structure can be expressed as

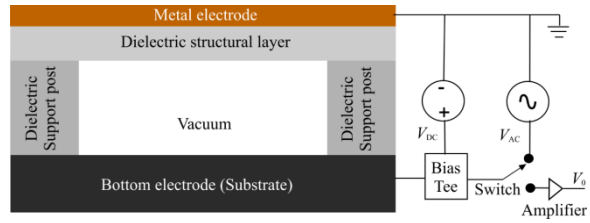


Figure 1. A cross-section of a CMUT and its typical operational setup.

$$C = \varepsilon_0 \varepsilon_r \frac{A}{(g_{eff} - x)} \tag{1}$$

and the effective plate gap height g_{eff} can be expressed as

$$g_{eff} = \frac{t_i}{\varepsilon_r} + \frac{t_m}{\varepsilon_m} + g_0 \tag{2}$$

In (1) and (2), A is coupling area of the capacitor plates, ε_0, ε_r ,and ε_m are the permittivities of free space, the bottom insulating layer, and the membrane structural material, respectively, g_0 is the initial gap between capacitor plates, x is the membrane deflection, t_m, and t_i are the membrane and insulator thickness, respectively.

The mechanical behavior of this configuration can be described using Newton's equation of motion as [32]

$$m\frac{d^2x}{dt^2} + b\frac{dx}{dt} + kx = F \tag{3}$$

In (3), F is the actuating force (electrostatic or acoustic) m is the mass of the movable plate, k is stiffness of the movable plate, and b represents the damping parameter of the system.

The CMUT operates both as a transmitter to emit ultrasound in the medium and also as a receiver to generate an electric signal in response to an incident ultrasound wave.

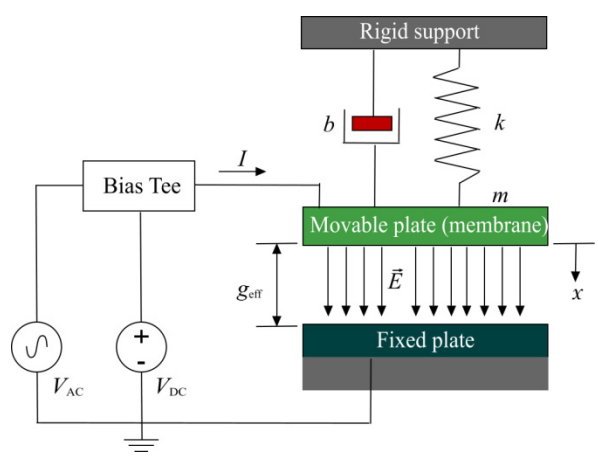


Figure 2. Electromechanical model of a CMUT.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A control signal operates a switch to enable mode switching as shown in Figure 1.

In the transmit mode, an AC voltage of desired frequency and amplitude, superposed with a suitable DC bias voltage, is applied across the CMUT geometry as shown in Figure 1. The resulting time-varying electrostatic attraction force between the CMUT electrodes (top membrane and the fixed bottom electrode) causes the membrane to vibrate and generate ultrasound waves in the medium.

During the receiving mode, an incident ultrasound wave forces the membrane to vibrate at the incident wave frequency. This vibration of the membrane dynamically alters the gap between the membrane and the bottom electrode, which affects a change in CMUT capacitance. This capacitance change is converted to an output voltage using a suitable microelectronic circuit.

Due to the fundamentally different physical principles of operation, ability to detect nano and picoscale precision variations of operational parameters (e.g., membrane displacement), and high precision microfabrication capabilities, CMUT transmit and receive performances are much superior compared to their piezoelectric counter parts. A comparison of some of the key performance parameters of CMUTs and piezoelectric transducers are provided in Table I.

B. CMUT Membrane Material

The mechanical and electrical properties of the CMUT membrane material are crucial to obtain the necessary output pressure (transmit operation) and generated voltage (receive operation). Common materials to fabricate the CMUT membranes are listed below with their relative advantages and disadvantages.

TABLE I. CMUT AND PZT PERFORMANCE COMPARISON [33].

S/N	Parameter	CMUT	PZT
1	Fabrication Method	Wafer bonding, surface micromachining, bulk micromachining	DF (Dice and Fill), Laminating
2	Array fabrication	Low cost	Expensive
3	Frequency range	1MHz - >100MHz	Comparatively small 50 kHz to 20MHz
4	Bandwidth	Wider fractionnal bandwidth (>100%)	Narrower Bandwidth
5	Output pressure for a similar voltage level	Comparatively low but improving	High
6	IC integration	Can be integrated easily with front-end electronics	Comparatively difficult to integrate with front-end electronics
7	Thermal stability	High thermal stability	Low thermal stability
8	Frequency use	Easier high-frequency fabrication	Difficult high-frequency fabrication
9	Sensitivity (mV/kPa)	22.57	4.28
10	DC bias	Required	N/A
11	SNR (dB)	22–87	18–22

Single crystal silicon: This method uses a thin layer of single-crystal silicon, typically the device layer of an SOI (silicon-on-insulator) wafer or a bulk micromachined silicon wafer, to realize the CMUT membrane. The objective is to achieve a desired level of electrical conductivity by doping and mechanical strengths for the membrane to attain target vibrational characteristics, optimal membrane deflection, and electrostatic actuation forces. However, the SOI wafers are expensive, and the thickness uniformity of the device layer remains a challenging issue.

Polycrystalline silicon (polysilicon): Polysilicon offers similar benefits and issues as the single-crystal silicon. However, as the suspended polysilicon membranes are realized using a Low-Pressure Chemical Vapor Deposition (LPCVD) method in a surface micromachining process, the realization of a suspended membrane using a sacrificial layer creates fabrication complexity, non-perfect geometry, and involves additional manufacturing cost.

Polymers and Metals: Various types of polymers such as Benzocyclobutene (BCB), SU-8, or Parylene-C are becoming popular to realize CMUT membranes using a wafer-bonded process. The polymer layers are functionalized with a thin metal layer, typically gold, deposited at the top or at the bottom of the membrane. Polymers contribute flexibility and structural resilience, while the metal layer as a high-conductivity electrode. Due to the high flexibility of the polymer membranes, it is possible to obtain high deflection during acoustical or electrical actuation. The polymer membranes can be realized using a simple spin deposition technique at a low cost.

Ceramics and metals: Ceramics (e.g., Silicon nitride, 3H-SiC) offer high mechanical strength and durability, facilitating reliable membrane vibration over a longer period [31]. Microfabricated (both surface micromachined and wafer bonded) silicon nitride thinfilms, specifically stoichiometric silicon nitride (Si_3N_4) functionalized with gold thinfilms, are widely used to fabricate CMUT membranes due to their high mechanical strength and low residual stress.

A comparative review in [34] revealed that ceramic materials exhibit superior performance when optimizing both diaphragm deflection and resonant frequency simultaneously, significantly outperforming metallic and polymer alternatives. The analysis in [34] that used a wide set of materials indicated that among all materials evaluated, only 3H-SiC, silicon nitride, and silicon with (110) crystal orientation surpass the performance characteristics of conventional silicon membranes. While silicon nitride membranes demonstrate the highest load-deflection characteristics, the dielectric charging phenomena in silicon nitride membranes is a critical issue that seriously compromises the operational reliability of CMUTs. Most importantly, the dielectric charging phenomena in silicon nitride membranes causes a drift of their resonant frequency

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

from the designed value and vibrational characteristics due to a spring softening effect [32]. While the issue of dielectric charging in silicon nitride MEMS switches is being investigated by various researchers [21], the dielectric charging issues in CMUTs with silicon nitride membranes is a much less explored area.

C. Silicon Nitride as CMUT Membrane Material

Silicon nitride is a hard material from the family of ceramics consisting of silicon and nitrogen [35]. Bulk silicon nitride is produced from materials compositions such as silicon powder (starting raw material), beta silicon nitride powder (diluting agent), high purity nitrogen, sodium chloride (molten salt) as listed in [35]. It has been observed in [35] that the nitridation of silicon with sodium chloride (NaCl) additives yielded two distinct product layers, which include the upper fibrous layer (impurities) and lower blocky layer (pure Si_3N_4). Also at 1250°C-1350°C a mixture composition of α-Si_3N_4, β-Si_3N_4 and silicon was seen but the silicon vanished at 1350°C making α-Si_3N_4 the dominant constituent.

Microfabricated silicon nitride thinfilms are typically deposited using two primary methods: Low Pressure Chemical Vapor Deposition (LPCVD) and Plasma-Enhanced Chemical Vapor Deposition (PECVD) [36][37]. While both techniques are extensively utilized across various applications, they differ in deposition rates and film quality. LPCVD operates at a slower pace, whereas PECVD enables faster film formation [37]. Further, as the LPCVD method realized silicon nitride films are stoichiometric (Si_3N_4) in nature with significantly low residual stress compared to PECVD method realized nitrides (SiN), the stoichiometric silicon nitride (Si_3N_4) is preferred to realize CMUT membranes. A comparison of physical properties of both types of thinfilm silicon nitrides are provided in Table II [37].

TABLE II. COMPARISON OF SELECTED PHYSICAL PROPERTIES OF LPCVD AND PECVD SILICON NITRIDE THINFILMS [37].

Property	LPCVD	PECVD	Unit
Deposition temperature	973–1073	473–673	K
Particles	Few	Many	
Step coverage	Conformable	Poor	
Dielectric constant	6–7	6–9	
Residual stress	1T	2C-5T	10^8 N/m
Young's modulus	270	300	GPa
Film quality	Excellent	Poor	
Dielectric strength	10	5	10^8 V/m
Resistivity	10^{14}	$10\text{-}10^{13}$	Ω-m

Note: C = compressive; T = tensile.

III. PHYSICS OF DIELECTRIC CHARGING IN SILICON NITRIDE THINFILMS

The fundamental physical laws of dielectric charging in silicon nitride thinfilms have been investigated by many researchers [3][21][22][29][30].

A. K-center Defects in Silicon Nitride Thinfilms

The dielectric charging phenomenon in both N-rich PECVD and stoichiometric silicon nitride (Si_3N_4) thinfilms mainly is attributed to the $\mathrm{Si} \equiv \mathrm{N}$ dangling bonds (defined as an unpaired electron of a silicon atom that is bonded to three nitrogen atoms [21][22][38]). These $\mathrm{Si} \equiv \mathrm{N}$ dangling bonds form energy states that can trap either holes or electrons as shown conceptually in Figure 3 [21]. Known as *K*-center defects, these traps are characterized as amphoteric trap centers (can capture both electrons and holes) with a trap energy state of approximately 4.2eV for a silicon dangling bond just under the conduction band minimum and -2eV for nitrogen dangling bond near the valence band minimum in silicon nitride [23]. This suggests that the nitrogen dangling bond makes no significant contribution to charge trapping, as it forms a state outside the band gap.

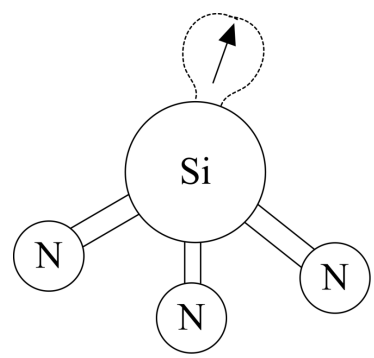


Figure 3. Formation of silicon nitride dangling bond as a silicon atom bonds with 3 nitrogen atom with one dangling bond ($\mathrm{Si} \equiv \mathrm{N}$) [22].

Following [22], due to strong negative electron-electron correlation energy, *K*-center with two spins (↑↓) is energetically more stable than the neutral *K*-centers (K^0) with single spin of electron (↑), making the neutral *K*-centers (K^0) more stable. As electrons are injected into either N-rich PECVD or stoichiometric silicon nitride thinfilms, they can be captured by neutral *K*-center (K^0) and become trapped there to make the *K*-center negatively charged ($K^0 + e^- = K^-$).

Similarly, when holes are injected into N-rich PECVD or stoichiometric silicon nitride thinfilms, the neutral *K*-centers (K^0) become positively charged ($K^0 + h^+ = K^+$). The phenomenon is illustrated conceptually in Figure 4. It is to be noted here that while the neutral *K*-centers (K^0) exhibit paramagnetic behavior with an unpaired electron in silicon atom, both positive and negatively charged *K*-center exhibit diamagnetic behavior without any unpaired electron (for hole capture, the unpaired electron is annihilated and for electron capture, two electrons of opposite spin occupies the energy level).

In [38][39], the *K*-center defect densities were measured and quantified using Electron Spin Resonance (ESR) along with Capacitive-Voltage (C-V) measurement technique, and it was discovered that SiN_x films illuminated with high

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

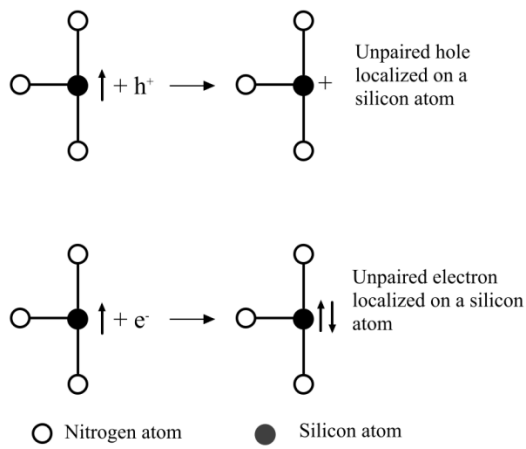


Figure 4. Schematic diagram of the *K*-center showing an unpaired electron localized on a silicon atom and bonded to three nitrogen.

energy UV(~4.9 eV) radiation increased the paramagnetic signal. Based on the measurement, it was concluded that the UV light was able to annihilate all the charges present in the film (K^+ or K^-) and convert them to neutral paramagnetic K^0 states. Further, the authors in [38][40] used an Electron Nuclear Double Resonance (ENDOR) technique to determine the structure of the defects in silicon nitride thinfilms. The investigation discovered that the silicon dangling bond with three nitrogen atoms is the primary defect responsible for charge trapping in the silicon nitride thinfilms. The detailed studies of the charge trapping due to silicon or nitrogen dangling bond is found in [20]–[22],[38].

These trap sites, due to different silicon and nitride composition, exhibit varying defect densities and thereby influencing the charge trapping behaviors, and ultimately the operational reliability of the CMUTs. Therefore, understanding the dielectric properties of silicon nitride and its compositions is essential in evaluating the extent of charge injection and retention during CMUT operation.

The movement of trapped charge carriers that are thermally emitted from the *K*-centers (basically coulombic trap states) into the conduction or valence band of the dielectric material gives rise to a current density. At the same time, charges are injected from the metal electrode to the silicon nitride layer through a tunneling process, contributing to a tunneling current density. Both current densities contribute to dynamically changing charge densities in the CMUT membrane, which forces a dynamic shift in the operating point and resonant frequency of CMUT membrane vibration, which affects the output pressure, vibrational characteristics, and receiving sensitivity [20]. Due to limited scope, this research work is focused on the trapped charge contributed current density in CMUTs with silicon nitride membranes.

B. Field-Induced Barrier Lowering In a CMUT

The potential energy distribution in a typical field emission system comprised of a metal electrode and a dielectric material after applying an electric field is shown in Figure 5 [20].

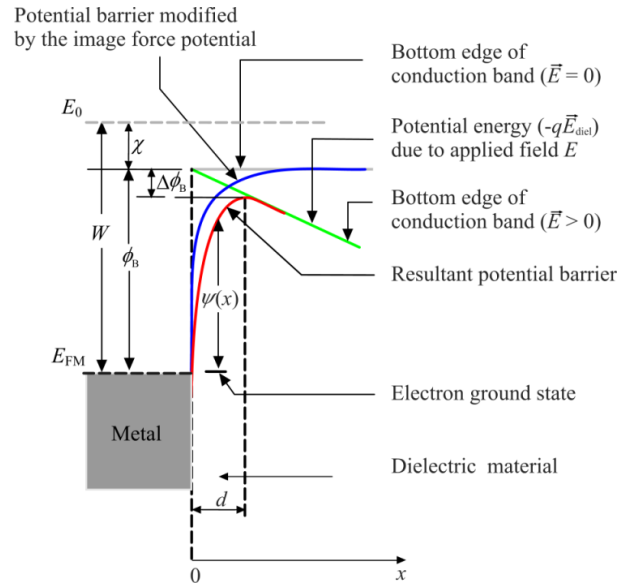


Figure 5. Potential energy barrier lowering in a typical metal-dielectric emission system due to an applied electric field [42].

In Figure 5, E_{FM} is the metal Fermi level, W is metal work function, χ is insulator electron affinity, $\vec{E}$ is applied electric field, d is dielectric thickness, ϕ_B is the potential barrier height, $\Delta\phi_B$ is decrease in potential barrier height, and $\psi(x)$ is the resultant potential barrier.

The potential energy barrier from metal to the dielectric material can be expressed as (4) below [19][21].

$$\psi(x)=\begin{cases}0 & x<0\\ V_0-\chi-q\vec{E}_{\text{diel}}x & 0\le x<d\\ V_0+q(\vec{E}-\vec{E}_{\text{diel}})d-q\vec{E}x & x\ge d\end{cases} \quad (4)$$

This potential energy barrier lowering due to an applied electric field is known as the Schottky effect.

In a dielectric material, such as silicon nitride with trap centers, the electric field due to an applied bias voltage also decreases the height of the potential energy barrier of the trap center (trap depth) in a similar fashion and is known as the internal Schottky effect [20]–[22]. If the potential energy barrier lowering due to the internal Schottky effect is sufficient, a trapped electron can overcome the barrier and be transported to the conduction band with little thermal agitation. As the trapped charges are transported to the conduction band, they contribute to dielectric charging and in turn, a leakage current.

This theoretical framework forms the basis of investigating the dielectric charging of silicon nitride membranes in CMUT due to the *K*-center defects. When a bias voltage is applied across the CMUT electrodes during operation, the internal Schottky effect associated with the electric field lowers the potential barrier of the *K*-center defects in silico nitride membranes, enabling the electrons to

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

overcome the potential barrier related to the traps, and move to the conduction band. A conceptual depiction of this internal Schottky effect-induced barrier lowering in the context of a CMUT is shown in Figure 6, assuming the barrier to have a triangular profile. In Figure 6, the field emission occurs from the top metal electrode surface. The interface between the metal and the stoichiometric Si_3N_4 structural layer of the membrane is positioned at $x = 0$ where d is the thickness of the Si_3N_4 layer.

The metal is characterized by a Fermi level E_{FM}, work function W, and electron affinity χ and ε_r is the relative permittivity of the Si_3N_4 layer.

The initial energy E of an electron at a specific location x along the x-axis is determined by its position x. The electric field $\vec{E}$ due to the bias voltage works across the vacuum gap and the $\vec{E}_{diel}$ represents the electric field in the silicon nitride layer.

For the CMUT geometry in Figure 6, the Poole-Frenkel emission model predicted current density J_{PF} can be calculated using (5) below [21].

$$\left.\begin{aligned} J_{PF} &= C\vec{E}\exp\left(\frac{-\left(\phi-\beta_{PF}\sqrt{\vec{E}}\right)}{k_B T}\right) \\ \beta_{PF} &= \sqrt{\frac{q^3}{\pi\varepsilon_0\varepsilon_r}} \end{aligned}\right\} \quad (5)$$

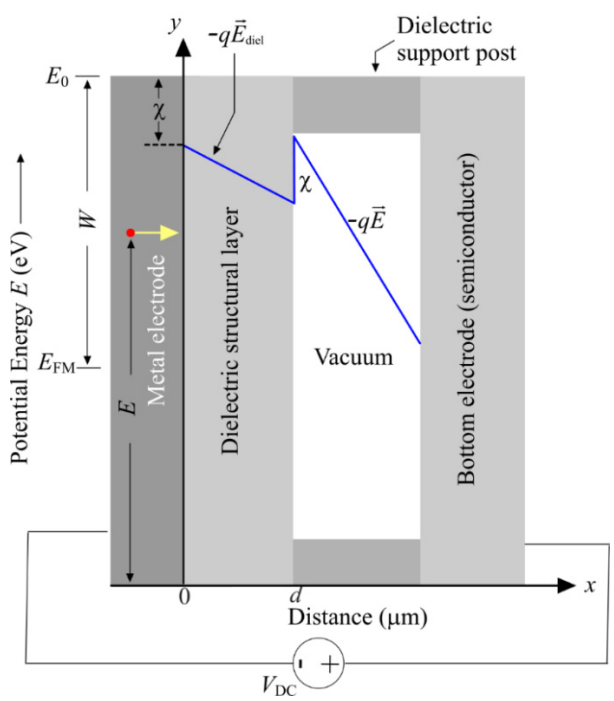


Figure 6. CMUT model showing the triangular barrier lowering due to applied potential energy.

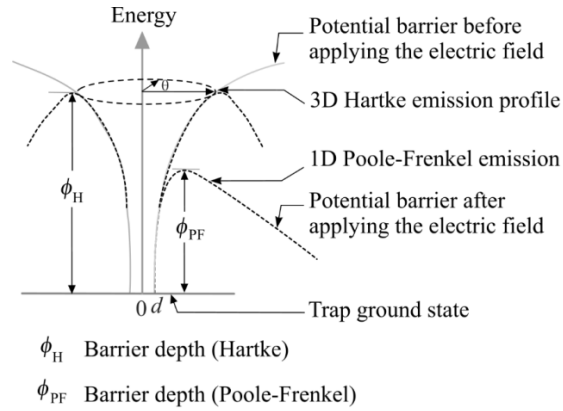


Figure 7. Conceptual diagram showing unidirectional Poole-Frenkel and 3D Hatke models of electron emission due to potential barrier lowering.

In (5), the parameter ϕ is the energy level of the charge trap; C is a pre-exponential factor, β_{PF} is the Poole-Frenkel potential barrier lowering coefficient; ε_0 is the permittivity of free space, ε_r is the relative permittivity of the nitride layer; q is the electronic charge; k_B is the Boltzmann constant and T is the temperature in kelvin. In [42], the value of the pre-exponential factor C has been reported as 0.1 A/V-m for Si_3N_4.

One important feature of the Poole-Frenkel emission model as expressed in (5) is that the model predicted emission current density is unidirectional (e.g., along the x-direction as in Figures 5 and 6). However, in reality, charge emissions from the trap centers due to the internal Schottky effect are not unidirectional but occur over a range of directions in a 3D space, as shown conceptually in Figure 7.

The major difference between these two models lies on the fact that unlike the unidirectional Poole-Frenkel emission model, the Hartke emission model accounts for the internal Schottky effect in a 3D space by introducing angular dependencies where the barrier lowering varies with an angle θ between the direction of the applied field and the emission path, described as $\Delta\phi\cos\theta$ [24][25]. By integrating this relationship over all possible emission angles, the effective average barrier can be calculated with a higher accuracy. Accordingly, the emission current density resulting from the electric field due to the CMUT bias voltage following the Hartke model can be expressed as [24][25]

$$\left.\begin{aligned} J_H &= \exp\left(\frac{-\phi}{k_B T}\right)\left(\frac{k_B T}{\Delta\phi\sqrt{\vec{E}}}\right)^2 \\ &\quad\times\left(1+\left(\left(\frac{\Delta\phi\sqrt{\vec{E}}}{k_B T}\right)-1\right)\exp\left(\frac{\Delta\phi\sqrt{\vec{E}}}{k_B T}\right)\right)+\frac{1}{2} \\ \Delta\phi &= \sqrt{\frac{q^3}{\pi\varepsilon_0\varepsilon_r}} \end{aligned}\right\} \quad (6)$$

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

where $\Delta\phi$ is the barrier-lowering coefficient and other parameters remains the same as in (5).

The authors in [24][25] reported that the 3D Hartke model predicted barrier-lowering amount is less than the maximum values predicted by the 1D Poole-Frenkel model. Thus, the unidirectional (1D) Poole-Frenkel model calculated emission current in a CMUT with a silicon nitride membrane is expected to deviate from the values calculated using the more realistic 3D Hartke model. This deviation is particularly significant at submicron dielectric thicknesses, which are typical in CMUTs, where charge injection and release occur over a distributed field profile [25].

IV. RESULTS

Matlab-based analytical modeling was conducted to compare the 1D Poole-Frenkel model with the 3D Hartke model. A CMUT geometry with a gold-coated silicon nitride membrane has been used for the comparison. The cross-section of the CMUT geometry is the same as in Figure 1. The CMUT specifications are listed in Table III.

TABLE III. CMUT SPECIFICATIONS.

Parameters	Dimension	Unit
Top electrode thickness (gold)	0.4	µm
Silicon nitride (Si_3N_4) structural layer thickness	0.5	µm
Vacuum gap thickness	0.6	µm
Bottom electrode thickness (silicon)	0.5	µm
Gap between electrodes	1.1	µm
Temperature	300	K

Figure 8 shows the Poole-Frenkel model predicted current density J_{PF} as a function of the electric field $\vec{E}$ due to an applied DC bias voltage. As Figure 8 reveals, the Poole-Frenkel current density J_{PF} increases rapidly as the effective barrier height decreases with an increasing electric field.

Figure 9 shows the Hartke model predicted current density J_H as a function of the electric field $\vec{E}$ due to an applied DC bias voltage.

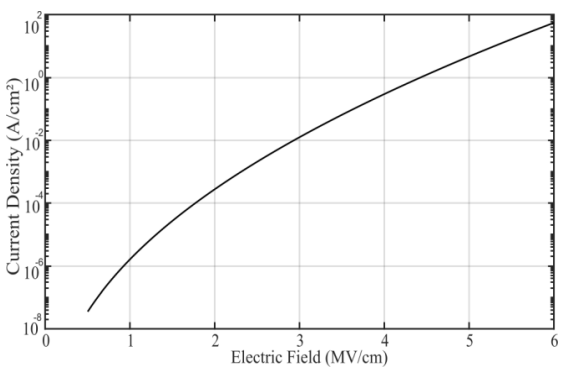


Figure 8. Poole-Frenkel emission model predicted current density due to the trap centers in the CMUT silicon nitride membrane as a function of the electric field associated with the CMUT bias voltage.

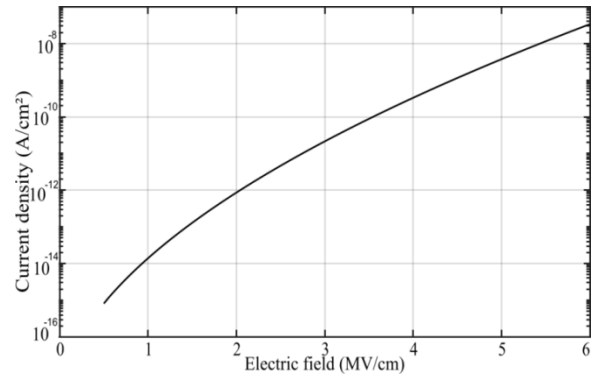


Figure 9. 3D Hartke model predicted current density due to the trap centers in the CMUT silicon nitride membrane as a function of the electric field associated with the CMUT bias voltage.

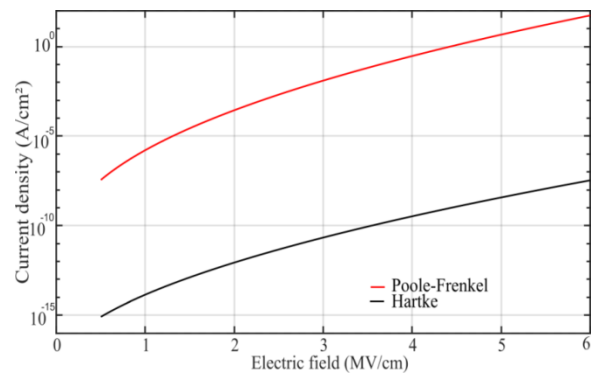


Figure 10. Comparison of Hartke and Poole-Frenkel emission models predicted current densities.

As expected, the Hartke model predicted current density also increases rapidly as the electric field due to the CMUT bias voltage increases due to a lowering of the barrier height. Figure 10 compares the current densities predicted by both the models for the same CMUT geometry subject to the same bias voltage variations. The percent variations in the current densities predicted by both models are tabulated in Table IV.

As Table IV shows, the Poole-Frenkel model predicted current density consistently predicts higher current densities compared to those predicted by the Hartke model by more than 99% for a bias voltage range of 11V-77V.

TABLE IV. CURRENT DENSITY COMPARISON.

Bias Voltage (V)	Hartke model current density, J_H (A/cm^2)	Poole-Frenkel model current density J_{PF} (A/cm^2)	$\Delta\%$ $\frac{J_{PF}-J_H}{J_{PF}}\times 100$
11	2.689×10^{-12}	1.077×10^{-10}	97.5
22	8.154×10^{-12}	8.771×10^{-10}	99.0
33	2.016×10^{-11}	3.864×10^{-9}	99.47
44	4.425×10^{-11}	1.277×10^{-8}	99.65
55	8.950×10^{-11}	3.555×10^{-8}	99.83
66	1.704×10^{-10}	8.796×10^{-8}	99.80
77	3.099×10^{-10}	1.996×10^{-7}	99.84

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

V. Current Density Variation Based on Nitrogen Composition

As the trap density in a silicon nitride layer depends on the actual composition of silicon and nitrogen in N-rich PECVD or stoichiometric LPCVD silicon nitride thinfilms, it is worth investigating the current densities as a function of silicon and nitrogen composition. To conduct this investigation, three different silicon nitride layers with varying silicon and nitrogen compositions as listed in Table V were used [42]. Corresponding current densities as a function of bias voltage associated electric field are plotted in Figure 11. As it can be seen in Figure 11, the layer with the lowest nitrogen content has lower trap depth and, consequently, higher emission current density for the same electric field.

TABLE V. Different LPCVD Silicon Nitride Compositions With Corresponding Trap Depths [42].

Composition	Trap depth/ionization potential ϕ (eV)
$SiN_{0.75}$	0.56
$SiN_{1.17}$	0.92
$SiN_{1.22}$	1.13

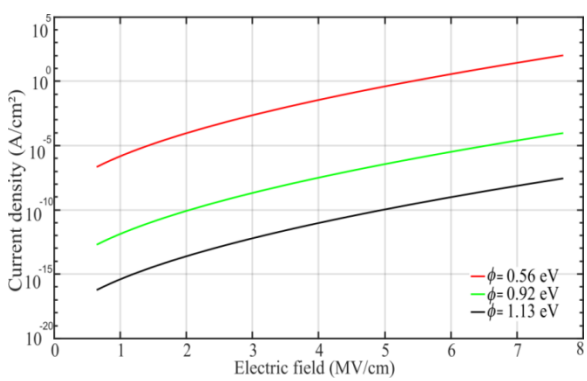


Figure 11. The Hartke model predicted current densities for three different silicon nitride layers with varying composition of silicon and nitrogen and trap depth energy levels as listed in Table V.

VI. Conclusions

The investigation revealed that in a CMUT geometry with silicon nitride membranes, the Poole-Frenkel model consistently predicts emission current densities that are more than 99% higher than those predicted by the Hartke model within a bias voltage range of 11–77 volts. To provide a context, a theoretical study published in [25] concluded that the unidirectional (1D) Poole-Frenkel model consistently predicted a 66% higher emission rate compared to the 3D Hartke model for two plane parallel electrodes separated by a dielectric medium. Obviously, the Poole-Frenkel model overestimates the emission current.

Apparently, this is due to its assumption of a one-dimensional potential barrier lowering. However, the Hartke model, which takes into account the three-dimensional nature of the *K*-center traps, aligns more closely with actual trap behavior. It can be concluded that, in comparison to the Poole-Frenkel model, the Hartke model is a more accurate method for predicting the emission current in 3D microstructures, such as a CMUT that uses a vibrating silicon nitride membrane to transmit or receive ultrasound.

Furthermore, the study found that the trap depths (potential barrier height) in a silicon nitride layer that in turn contributes to the extent of the emission current density in a CMUT, is influenced by the specific composition of silicon and nitrogen within the layer. The parameters of the deposition process, such as plasma-enhanced chemical vapor deposition (PECVD) or low-pressure chemical vapor deposition (LPCVD), greatly influence the composition of the deposited materials in a thinfilm. Investigation revealed that as the nitrogen content increases, the energy trap depth also increases. Accordingly, among the 3 investigated silicon nitride layers, the one with the lowest nitrogen concentration (and therefore the lowest energy trap depth of 0.56 eV) results in the highest emission current density at the same electric field strength.

Furthermore, during vibration, the silicon nitride membrane of a CMUT deforms causing distortion of the energy levels of the charge traps located within the membrane. It is thus necessary to evaluate the Hartke model predicted current densities in dynamically vibrating CMUT membranes against experimentally measured values to better understand the dynamic behavior of the physics of trap-assisted emission currents to mitigate their effects.

Acknowledgment

This research work is supported by Natural Sciences and Engineering Research Council of Canada (NSERC)'s Discovery Grant RGPIN-2025-04794 and IntelliSense Corp, 220 Broadway, Suite 102, Lynnfield, MA 01940, USA.

References

[1] C. Okoro and S. Chowdhury, "Comparison of Poole-Frenkel and Hartke Emission Currents in a Capacitive Micromachined Ultrasonic Transducer (CMUT)," in *Proc. of The Sixteenth International Conference on Sensor Device Technologies and Applications* (*SENSORDEVICES 2025*), Barcelona, Spain, 2025, pp. 10-14.

[2] A. Ergun, G. Yaralioglu, O. Oralkan, and B. T. Khuri-Yakub, "Techniques and Applications of Capacitive Micromachined Ultrasonic Transducers," in *MEMS/NEMS Handbook Techniques and Applications,* Vol. 1, Design Methods, C. Leondes, Ed. _ New York: Springer, 2006, pp. 553-615.

[3] R. Manwar, T. Simpson, A. Bakhtazad, and S. Chowdhury, "Fabrication and characterization of a high frequency and high coupling coefficient CMUT array," *Microsystem Technologies*, vol. 23, pp. 4965–4977, Oct. 2017.

[4] M. Klemm, "Acoustic Simulation and Characterization of Capacitive Micromachined Ultrasonic Transducers," Ph.D. Dissertation, Faculty of Electrical Engineering and Information Technology, Technical University of Dresden, Dresden, Germany, 2017.

[5] S. Berg, "Capacitive Micromachined Ultrasonic Transducers–Acoustic Challenges and Proposed Solutions," Ph.D. Dissertation, Department of Electronics and Telecommunications, Norwegian University of Science and Technology, Trondheim, Norway, 2012.

[6] X. Zhang, "Capacitive Micromachined Ultrasonic Transducers (CMUTs) on Glass Substrates for Next–

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Generation Medical Imaging and Beyond," Ph.D. Dissertation, Dept. of Electrical and Computer Engineering, North Carolina State University, Raleigh, North Carolina, 2017.

[7] M. Vallet et al., "Quantitative Comparison of PZT and CMUT Probes for Photoacoustic Imaging: Experimental validation," *Photoacoustics*, vol. 8, pp. 48–58, Dec. 2017.

[8] B. Heiles et al., "Ultrafast 3D Ultrasound Localization Microscopy using a 32×32 Matrix Array," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 2005-2015, Apr. 2019.

[9] Z. Xu, "Histotripsy: the First noninvasive, Non-ionizing, Non-thermal Ablation Technique based on Ultrasound," *Int. J Hyperthermia*, vol. 38, pp. 561–575, Apr. 2021.

[10] Focused Ultrasound Foundation, "Focused Ultrasound for Alzheimer's Disease," Charlottesville, VA, 2022.

[11] J. Jang et al., "Dual-Mode Integrated Circuit for Imaging and HIFU With 2-D CMUT Arrays," in *Proc. of 2015 IEEE International Ultrasonics Symposium* (*IUS*), 2015, pp. 1-14.

[12] C. Arvanitis and F. Degertekin, "Multi-Functional Sparse Phased Arrays for Guiding Focused Ultrasound Therapies," US Patent 12 427 344, Sept. 30, 2025.

[13] T. Matéo et al., "A 1-D CMUT transducer with front-end ASIC in a 9 French catheter for intracardiac echocardiography: Acoustic and Imaging evaluation," in *Proc. of 2020 IEEE International Ultrasonics Symposium* (*IUS*), 2020, pp. 1-8.

[14] M. Engholm, "Capacitive Micromachined Ultrasonic Transducers for 3–D Imaging," Ph.D. Disssertation, Department of Micro- and Nanotechnology, Technical University of Denmark, Kongens Lyngby, Denmark, 2018.

[15] M. Schiffer, "Function-Driven Innovation: How Technology Follows the Needs of Chiplet Systems," in *Proc. of Electronic Packaging Day 2025 Conference*, Fraunhofer IZM, Berlin, Germany, November 6-7, 2025, pp. 1-16.

[16] J. Munir, Q. Ain, and H. Lee, "Reliability issue related to dielectric charging in capacitive micromachined ultrasonic transducers: A review," *Microelectronics Reliability*, vol. 92, pp. 155–167, Jan. 2019.

[17] S. Machida, T. Takezaki, T. Kobayashi, H. Tanaka, and T. Nagata, "Highly reliable CMUT cell structure with reduced dielectric charging effect," in Proc. of *2015 IEEE International Ultrasonics Symposium* (*IUS*), 2015, pp. 1-4.

[18] J. Li, Y. Li, and P. Zhang, "Modelling and Analysis of Dielectric Charge of CMUTs," *IOP Conf. Ser.: Mater. Sci*. Eng., vol. 768, pp. 062106, Mar. 2020.

[19] S. Kasap, *Principles of electronic materials and devices,* Fourth edition. New York, NY: McGraw-Hill, 2018, pp. 221-230.

[20] D. Choi, "Numerical Modeling of Space Charge Dynamics And Electrical Breakdown in Solid Dielectrics," Ph.D. dissertation, Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA, 2013.

[21] A. Dogan, "The Reliability of the Silicon Nitride Dielectric in Capacitive MEMS Switches," Master's thesis, Department of Materials Science and Engineering, The Pennsylvania State University, University park, PA, 2005.

[22] V. Sharma, "Study of Charges Present in Silicon Nitride Thin Films and Their Effect on Silicon Solar Cell Efficiencies," Ph.D. dissertation, Department of Electrical Engineering, Arizona State University, Tempe, AZ, 2013.

[23] J. Simmons, "Poole-Frenkel Effect and Schottky Effect in Metal-Insulator-Metal Systems," *Phys. Rev*., vol. 155, pp. 657-660, Mar. 1967.

[24] J. Hartke, "The Three-Dimensional Poole-Frenkel Effect," *Journal of Applied Physics*, vol. 39, pp. 4871–4873, Sept. 1968.

[25] P. Murgatroyd, "Theory of space-charge-limited current enhanced by Frenkel effect," *J. Phys. D: Appl. Phys*., vol. 3, pp. 151–156, Feb. 1970.

[26] R. Fowler and L Nordheim, "Electron Emission in Intense Electric Fields," *Proc. of the Royal Society of London*, *Series A*, vol. 119, pp. 173-181, 1928.

[27] M. Zubair, Y. Ang, and L. Ang, "Fractional Fowler-Nordheim Law for Field Emission from Rough Surface with Non-parabolic Energy Dispersion," *IEEE Trans. Electron Devices*, vol. 65, pp. 2089–2095, Jun. 2018.

[28] Y. Huang et al., "A solution to the charging problems in capacitive micromachined ultrasonic transducers," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control,* vol. 52, pp. 578–580, 2005.

[29] M. Kupnik A. Ergun, H. Huang, and B. T. Khuri-Yakub, "6F-2 extended insulation layer structure for CMUTs," in *Proc. of 2007 IEEE International Ultrasonics Symposium* (*IUS*), 2007, pp. 511-514.

[30] M. Kupnik, M. Ho, S. Vaithilingam, and B. T. Khuri-Yakub, "CMUTs for air coupled ultrasound with improved bandwidth," in *Proc. of* 2011 *IEEE International Ultrasonics Symposium* (*IUS*), 2011, pp. 592–595.

[31] J. Hernandez, T. Zure, and S. Chowdhury, "Capacitance measurements of an SOI based CMUT," in *Proc. of IEEE Fourth Latin American Symposium on Circuits and Systems (LASCAS),* 2013, pp. 1-4.

[32] S. Senturia, Microsystems Design; Kluwer Academic Publisher: Norwell, MA, USA, 2000.

[33] S. Upadhyay, "A MEMS Planar Fresnel Lens for CMUT Array," Master's thesis, Electrical and computer Engineering Department, University of Windsor, Windsor, ON, Canada, 2023.

[34] Z. Mehmood, I. Hanif, and F. Udrea, "Material selection for optimum design of MEMS pressure sensors," *Microsystem Technologies*, vol. 26, pp. 2751–2766, Sept. 2020.

[35] N. Andraskar, G. Tiwari, and M. Goel, "Impact response of ceramic structures - A review," *Ceramics International*, vol. 48, pp. 27262–27279, Oct. 2022.

[36] B. Joshi et al., "LPCVD and PECVD silicon nitride for microelectronics technology," *Indian Journal of Engineering and Materials Sciences*, vol. 7, pp. 303-309, Oct. 2000.

[37] M. Madou, Fundamentals of Microfabrication, CRC Press, Boca raton, FL, 1997.

[38] L. Warren, F. Rong, E. Poindexter, G. Gerardi, and J. Kanicki, "Structural identification of the silicon and nitrogen dangling-bond centers in amorphous silicon nitride," *Journal of Applied Physics,* vol. 70, pp. 346–354, Jul. 1991.

[39] D. Krick, P. Lenahan, and J. Kanicki, "Stable photoinduced paramagnetic defects in hydrogenated amorphous silicon nitride," *Applied Physics. Letters*, vol. 51, pp. 608-610, Aug. 1987

[40] P. Lenahan and S. Curry, "First observation of the Si hyperfine spectra of silicon dangling bond centers in silicon nitride," *Applied Physics Letters*, vol. 56, pp. 157, Jan. 1990.

[41] Y. Zhou and P. Zhang, "Theory of field emission from dielectric coated surfaces," *Phys. Rev. Research*, vol. 2, p. 043439, Dec. 2020,

[42] R. Apodaca, "Electrical Characterization of Silicon-Rich Nitride and Silicon Oxynitride Films Deposited by Low-Pressure Chemical Vapor Deposition," Master's Thesis, Department of Electrical Engineering, University of New Mexico, Albuquerque, NM, 2005.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Adaptive Forward Error Correcting Decoders: A Survey of Architectures and Implementations

Sergei Sawitzki
FH Wedel (University of Applied Sciences)
Wedel, Germany
e-mail: Sergei.Sawitzki@fh-wedel.de

***Abstract*—This work describes adaptive designs of decoders for several forward error correction codes. After a brief introduction to the corresponding code family and basic decoder architecture, several possibilities of multi-standard implementation of such decoders are discussed. In the first scenario, decoders capable of decoding the codes with different parameters belonging to the same family are introduced. The results suggest, that the hardware overhead caused by the additional flexibility is – in most cases – as low as 5 – 10% additional silicon footprint while the throughput is reduced by 2 – 26 % compared to the implementation based on the fixed code parameters. In the second scenario, reconfigurable designs of multi-family decoders are discussed. Since some parts of the data-path and internal memory can be reused for different decoders, silicon area savings of more than 40 % are achievable compared to the overall chip area costs of the individual decoder implementation per code family. The overall usefulness of such reconfigurable decoder designs depends on the application case, for instance, the throughput requirements or the necessity to process data streams encoded with different codes in parallel. The figures reported herein summarize and extend some previous work known from literature, as well as research carried out by the author.**

Keywords-forward error correction; FEC decoders; adaptive decoder architectures; convolutional codes; Viterbi algorithm; Reed-Solomon codes; turbo codes; LDPC codes; multi-standard FEC architectures.

I. Introduction and Related Work

Most of the modern digital communication and data storage standards enforce the use of different error correcting codes. Since in most cases the error correction process is carried out by the receiver – without resending or rereading the data – these codes are usually referred to as Forward Error Correcting (FEC) codes and the respective hard- or software units are called FEC encoders and decoders. Since the code used in particular storage or communication system is usually selected in such a way that it addresses the specific noise characteristics of the channel (memoryless vs. with memory, single vs. burst errors etc.), the more general term *channel code* is often used in the same context synonymously.

The FEC codes most widely used nowadays are belonging to one of the following code families: convolutional codes, Reed-Solomon (RS) codes, Low-Density Parity-Check (LDPC) codes, and turbo codes. Some standards specify varying codes for different data rates or combine codes from different families within one run to address more of the specific channel properties like cross-talk and fading improving the overall bit error rate.

Over the last few years, mobile devices got significantly more flexible. Modern mobile phones and other wearable devices support numerous communication interfaces providing the user with almost unlimited connectivity. Even the low-price devices nowadays support wireless LAN, different mobile internet standards, and Bluetooth, just to name a few. At least some of them rely on the same kind of FEC algorithms, although the specific parameters like code generator polynomials, coding rates, constraint length, block sizes, and so on may vary even within the same standard or application. To address this issue, adaptive Viterbi and RS decoders will be discussed below and compared to the straight-forward implementation for the single code instance. Such comparison will provide a glimpse of the overhead introduced by the additional flexibility.

On the other side, some computations carried out during the decoding process of different code families are quite similar. In addition, some data need to be temporarily stored either throughout the complete decoding procedure or, in case of LDPC and turbo codes, between adjacent decoding iterations resulting in specific memory requirements. This poses the question, if at least some parts of the silicon dedicated to the data-path and memory can be reused among different standards reducing the silicon footprint and improving the power balance compared to the straight-forward implementation of every single decoder. This question will be addressed in the discussion of combined Viterbi-turbo and LDPC-turbo decoders.

The applicability and advantages of the adaptive FEC decoder implementation depend on the use case, e.g., the necessity to process several data streams encoded with the same code or with different codes in parallel or the option to switch between different data rates. This should be kept in mind during the discussion of the different design options as will be addressed below in more detail. In general, this contribution is the revised and significantly extended version of the conference paper [1] published previously by the same author.

In principle, FEC decoding algorithms can be implemented both in software and hardware. However, given the fact that most of them are quite computationally intensive, software-based solutions are usually lagging behind the high throughput requirements imposed by the majority of contemporary communication and data storage standards. For instance, iterative algorithms involved in the decoding process of turbo and LDPC codes may require as many as 1 500 machine instructions per bit (or even more depending on the number of decoding iterations) [2]. Running such decoders at the data rates of several hundreds of megabits per second requires computational power, which is

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

beyond the capabilities of even most powerful microprocessors. For this reason, the scope of this paper is limited to the dedicated hardware architectures, i.e., direct mapping of the FEC decoders to silicon. Alternative approaches, which are less flexible than the software solution but still offer at least some properties of the architectures built around the instruction set paradigm, are Application-Specific Instruction-set Processors (ASIP) [3][4] and Networks-on-Chip (NoC) [5]. These are beyond the scope of this contribution. A comprehensive overview concerning the scalability issues and multi-standard capabilities of different FEC decoders is provided in [2][6][7]. Based upon these studies, some of the results are re-evaluated, supported by new findings and extended herein.

The rest of this paper is organized as follows. Section II introduces an adaptive implementation of the Viterbi decoder and compares several designs known from literature. In Section III, the additional costs of a reconfigurable design for a Reed-Solomon decoder compared to the appropriate single-code implementation are discussed. Section IV summarizes the general options of combining decoders for different code families within one design and analyzes the architectures of reconfigurable Viterbi-turbo and LDPC-turbo decoders in more detail. In Section V, architectures of all presented decoder families are briefly reviewed. Finally, Section VI provides a summary of this work and draws some conclusions concerning adaptive and multi-family FEC decoder implementation showing some directions for future research.

The discussion of every code family starts with the description of the key aspects relevant for the decoder implementation. For the readers interested in more detailed theoretical background, some references are provided where appropriate.

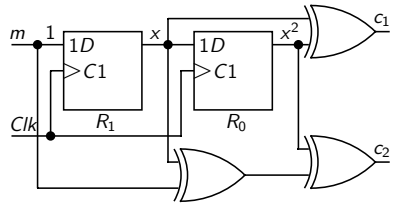
II. Viterbi Decoders

Convolutional codes were introduced in 1955 [8] and became one the most widely applied FEC code family. The following subsections describe the structure and the parameters of the code as well as the mapping of the decoding algorithm to hardware finishing with the discussion of the additional costs imposed by an adaptive implementation.

A. Code Structure

Convolutional encoder passes the incoming data stream through a shift register. The input of the first storage element as well as the outputs of several storage elements of this shift register are combined using an XOR operator. Particular XOR operators are selected to produce the output data streams. The number of delay stages in the shift register including the input of the first delay stage is referred to as *constraint length* K. Larger K values provide better error correction capabilities (at the cost of the increasing decoder complexity). Constraint lengths between 5 and 11 are common in real-world applications.

If the encoder does not have any feedback structures, i.e., no outputs of the delay stages are fed back to the input via XOR operators, it is called non-recursive. Such convolutional encoder can be seen as Finite Impulse Response (FIR) filter. The output



$g_1(x) = x + x^2 \cong 011_2 = 3_8$
$g_2(x) = 1 + x + x^2 \cong 111_2 = 7_8$

Figure 1. Simple convolutional encoder (non-recursive, non-systematic).

is the discrete convolution of the input data with the impulse responses (one for every output stream) of this FIR in the time domain, hence the name of the code [9, pp. 455 – 456]. In almost all applications of the non-recursive codes, the original input does not appear in the encoded message. In terms of the coding theory such code is classified as non-systematic. Recursive systematic convolutional codes will be discussed in the Subsection IV-A in the context of the turbo codes.

The number n of the output data streams defines the code rate $R = 1/n$. For the codes used in most communication standards 2 to 4 output bits are produced for every input bit, thus the code rate R lies between 1/2 and 1/4. Depending on the required error correction performance, it is possible to adapt the code rate $1/n$ of the initial code (called *mother code*) to different values by puncturing, i.e., removing some encoded bits before transmission and replacing them afterwards with dummy bits at the receiver before decoding. Higher code rates reduce the performance of the code.

The particular codes differ depending on which bits from the sequence stored in the shift register are involved in the XOR calculation. This information is provided in the form of the so-called generator polynomials g, one for every output. Figure 1 shows an example of a simple convolutional encoder with $K = 3$, $R = 1/2$, $g_1 = 011$ and $g_2 = 111$. As can be seen, the generator polynomials are often specified by the binary or octal representation of their coefficients. The encoder is a Finite-State Machine (FSM) whose output depends on g. Recording all possible state transitions of this FSM over time creates a directed graph called *trellis diagram*. As illustrated in Figure 2 (for the code produced by the encoder in Figure 1), every input sequence corresponds to the unique path through this graph, while every state transition on this path generates one output symbol. Since the decoder only sees the output sequence, while the input sequence and the corresponding state transitions remains inaccessible, the underlying abstraction is called Hidden Markov Model (HMM).

B. Decoding Algorithm

The decoding process is based on the maximum-likelihood (MLL) principle. The decoder's task is to find the state transition sequence which the encoder FSM most likely underwent, based on its output sequence, which in most cases is corrupted by noise during the transmission. Afterwards, the state transitions can be mapped to the corresponding input data sequence restoring the original message. In graph theory, this

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

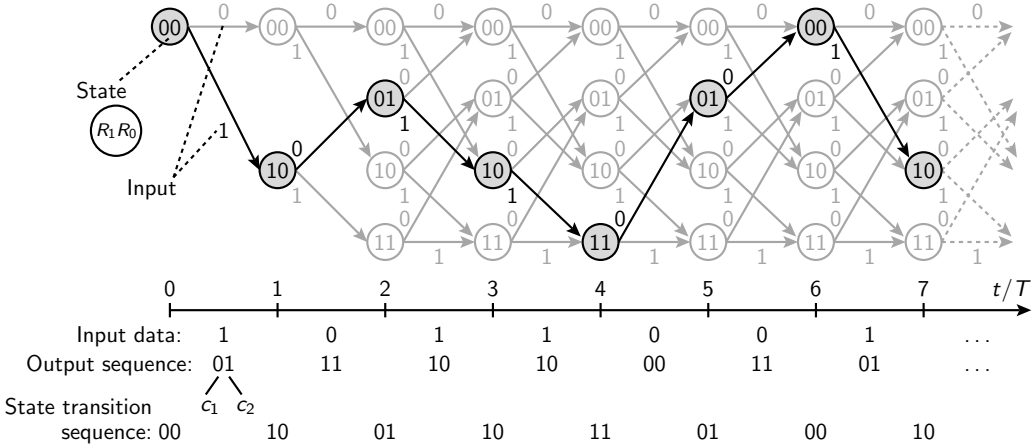


Figure 2. Trellis diagram.

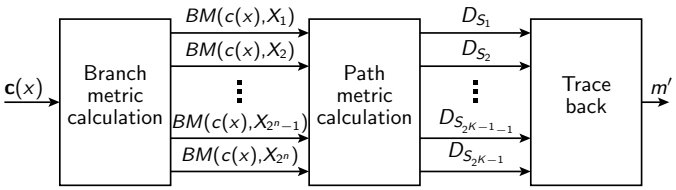


Figure 3. Top view architecture of the Viterbi decoder.

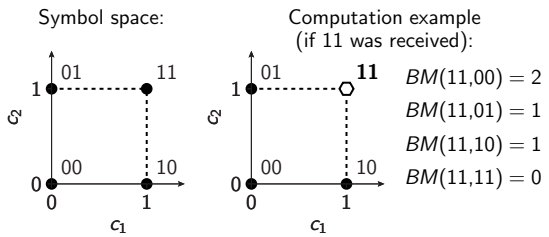
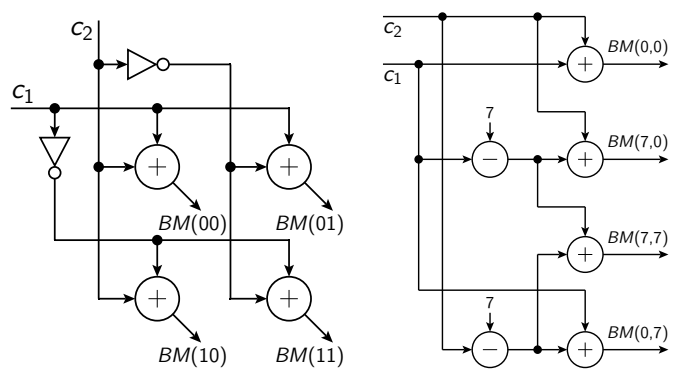


Figure 4. Branch metric example.

TABLE I. INPUT FORMATS FOR SOFT-INPUT DECODING WITH $b = 3$.

Interpretation	Two's complement	Signed magnitude	Unsigned magnitude
Strongest 0	011	011	000
	010	010	001
	001	001	010
Weakest 0	000	000	011
Weakest 1	111	100	100
	110	101	101
	101	110	110
Strongest 1	100	111	111



(a) Hard-input (Hamming distance) [11]

(b) Soft-input for $b = 3$ (Manhattan distance, unsigned magnitude)

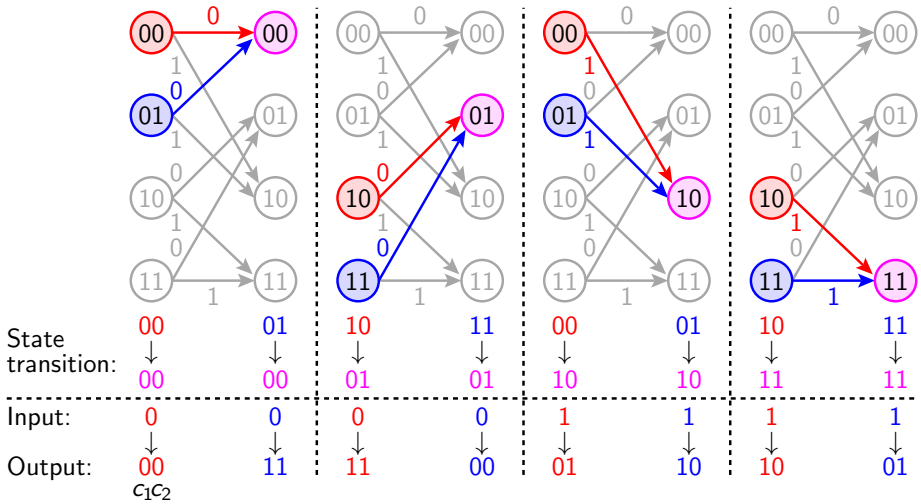
Figure 5. Branch metric computation.

problem can be framed as finding the path through the trellis diagram that minimizes the cumulative error metric.

The most widely used decoding algorithm for convolutional codes was introduces in 1967 by Andrew Viterbi and is based on dynamic programming [10]. As shown in Figure 3, a Viterbi decoder consists of the three major parts, which are executed consecutively and are called branch metric calculation, path metric calculation, and trace-back.

In the first stage of the decoding process, every received symbol is compared to all legal output symbols of the encoder. The result is called Branch Metric (BM). In the simplest case of the so-called hard-input decoding, where every input symbol to the decoder is considered to be n bits wide, BM is the Hamming distance as shown in Figure 4 for the case $n = 2$. However, after the modulation and analog-digital conversion every bit composing any code symbol is mapped to an integer number represented by several bits. This case is referred to as soft-input decoding with the input bit-width b (resulting in $b \times n$ bits per code symbol). Table I shows different representation conventions for soft-bits. Note that independent of the format, hard-input symbols can be generated from the soft-inputs by keeping the Most Significant Bit (MSB) and removing the rest. Soft-input decoding provides better error correction performance at the cost of implementation complexity. For example, other measures like Manhattan or (squared) Euclidean distance are needed to represent the branch metrics. The number of BM values to be calculated per received symbol depends on the code rate, since every additional output bit at the encoder doubles the number of legal output symbols, and the decoder must compare each received symbol against all 2^n possibilities.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



$$
\begin{aligned}
PM_{t+1}(00) &= \min(PM_t(00) + BM(x, \overbrace{(0,0)}^{c_1,c_2}), PM_t(01) + BM(x, \overbrace{(1,1)}^{c_1,c_2})) \\
PM_{t+1}(01) &= \min(PM_t(10) + BM(x, (1,1)), PM_t(11) + BM(x, (0,0))) \\
PM_{t+1}(10) &= \min(PM_t(00) + BM(x, (0,1)), PM_t(01) + BM(x, (1,0))) \\
PM_{t+1}(11) &= \min(PM_t(10) + BM(x, (1,0)), PM_t(11) + BM(x, (0,1)))
\end{aligned}
$$

Figure 6. Path metric calculation.

Figure 5 shows the branch metric calculation for hard- and soft-input decoding.

In the second stage, branch metrics are used to calculate the Path Metrics (PM). One PM needs to be calculated for every state of the encoder. All path metrics are set to zero in the beginning and updated with every processed symbol. The corresponding operation is called Add-Compare-Select (ACS) and is the computational core of the decoding process. Every ACS unit consists of two adders, one comparator and one multiplexer. In addition, one register is required per unit to store the actual PM value. As the number of ACS units is equal to the number of states of the encoder, it can be calculated as 2^{K-1}, i.e., it increases exponentially with the constraint length of the code. The result of the ACS calculation is the new path metric and the binary decision, which state transition the encoder underwent while the corresponding output symbol was generated. This decision is stored for every processed symbol and every ACS unit in form of a single bit called *decision bit*. The decision bits accumulated over time represent several paths through the possible state transition sequences of the encoder with every path corresponding to the certain input sequence. At the same time, the actual path metric of every state stores the cumulative error of the current path leading to this state as the sum of single BM values calculated in the previous stage. Figure 6 illustrates the path metric calculation for the code produced by the encoder in Figure 1.

With every processed symbol, new BM values are added to the actual PM values, thus path metrics are monotonically increasing, requiring normalization since the bit-width of the registers storing the actual values is limited. This can be achieved by shifting every PM to the right by one or more

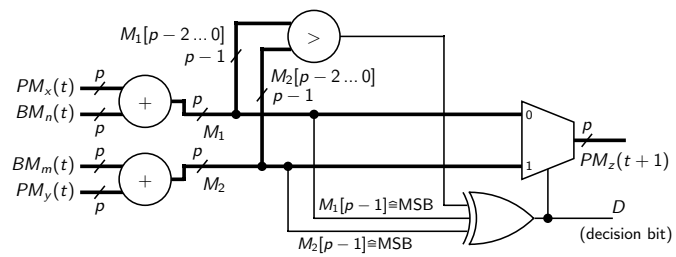


Figure 7. ACS unit with modulo-based normalization (adapted from [13]).

bits once any PM reaches the maximum value representable within the reserved bit-width, thus avoiding overflow. The more efficient approach, however, is based on modulo normalization. The comparison operation within the ACS unit calculates the difference between two sums of a path metric with a branch metric. It can be proven, that this difference has the upper bound determined by the input bit-width, the constraint length, and the coding rate [12][13]:

$$P_{\max} = \left(2^b - 1\right) \cdot \frac{K}{R}. \tag{1}$$

The essence of the modulo normalization is setting the bit-width p for the PM representation to $\lceil \log_2(2 \cdot P_{\max}) \rceil = \lceil \log_2(P_{\max}) + 1 \rceil$. By doing this, the overflow can be ignored. Instead, the result of the comparison is inverted, if the MSBs of the compared values differ (which means that the smaller value underwent an overflow and is actually greater). Hardware implementation of the ACS unit with modulo-based normalization logic is shown in Figure 7.

The final stage of the Viterbi decoding algorithm is called trace-back. During this stage the decision bits produced by the

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

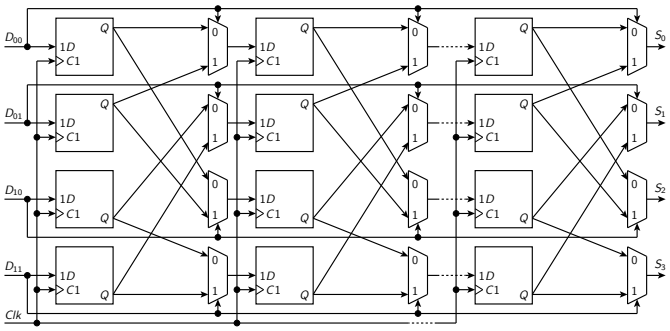


Figure 8. Trace-back by register exchange.

preceding PM calculation are used to follow the path with the smallest cumulative error from the current state back to the initial state of the decoder. It can be shown that under normal conditions all paths through the trellis merge to a single one after a certain number of steps which is called Trace-Back Depth (TBD). This path represents the most probable sequence of the encoder's state transitions. As a consequence, decoder needs to store the decision bits for at least the number of symbols equal to TBD. The exact value of the TBD depends on the channel conditions. For Additive White Gaussian Noise (AGWN) channels TBD values of $5 \times K$ are sufficient, but they may be higher than $20 \times K$ for channels with fading and multi-path propagation.

The most obvious implementation of the trace-back step is to map the trellis diagram to hardware directly in form of the network of storage elements connected by multiplexers. This approach is called register exchange and is depicted in the Figure 8 for the trellis diagram shown in the Figure 2. If the number of trace-back stages is large enough and the channel conditions do not overstress the error correction performance of the code, all S_x outputs of the last stage should be the same after at most TBD clock cycles, so any of these outputs can be used as the decoded message.

The major disadvantage of the register exchange approach is its power inefficiency for higher values of the constraint length. In the extreme case of $K = 11$ on a multi-path channel, decoding of one single bit would require more than 200 000 accesses to the storage elements. Even for smaller K values and under the assumption that not every flipflop changes its value at every clock cycle, the power consumption remains quite high. For this reason, memory-based trace-back techniques are used in most real-world Viterbi decoders for $K > 3$.

Given the fact that one decision bit is produced by every ACS unit, the memory capacity for a single trace-back run equals to $\mathrm{TBD} \times 2^{K-1}$. As the decision bits cannot be overwritten during the trace-back, this amount needs to be at least doubled. Many Viterbi decoder implementations use three memory banks: one for storing the actual decision bits, one for merging the different paths to the single one determining the starting point for the decoding, and one for reading the decoded bits starting with the state from the merging step. As soon as one iteration is completed, i.e., one memory bank was completely read of

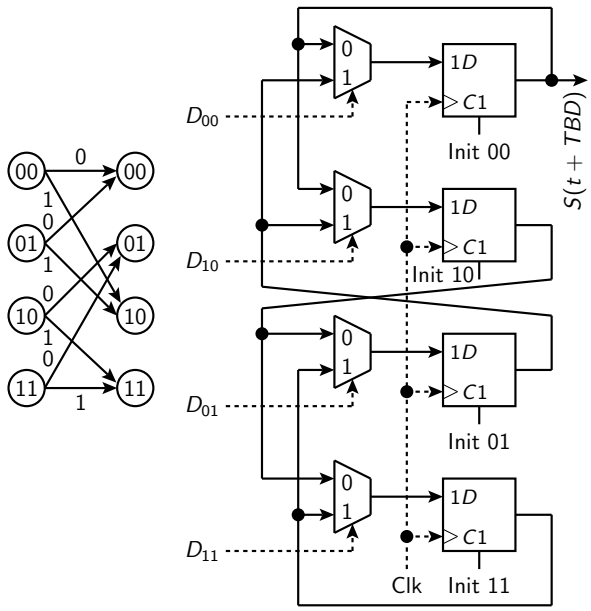


Figure 9. Trace forward unit (adapted from [14]).

written, the banks change their roles. This approach requires the memory capacity of $3 \times \mathrm{TBD} \times 2^{K-1}$. However, it has been shown that the merging step can be replaced by the single stage using 2^{K-1} registers of $K-1$ bits connected by a simple multiplexer network. This technique is called *trace forward* and allows to save one memory bank reducing the overall memory capacity to $2 \times \mathrm{TBD} \times 2^{K-1}$ [14]. Figure 9 shows the trace-forward hardware for $K = 3$ (note that every register is $K - 1 = 3 - 1 = 2$ bits wide). After TBD decoding steps the output $S(t+\mathrm{TBD})$ provides the bit pattern, which encodes the starting state for the actual trace-back run.

As already mentioned, the decision bits corresponding to the most probable path are traversed in the reverse order during the trace-back. Thus, the corresponding decoded input bits must be stored in a Last-In First-Out (LIFO) buffer. This LIFO memory is finally read to produce the decoded input sequence in the right order (which, in optimal case, will be the same as the original input to the encoder).

C. Adaptive Implementation

Table II summarizes some applications of convolutional codes in different communication standards. As already mentioned, even within a single standard code parameters may vary. While a Viterbi decoder for a specific convolutional code has fixed design parameters like the number of the BM and the ACS units, TBD, and so on, it is quite different in the case of the adaptive decoder. Concerning the most settings, the decoder must be dimensioned for the worst case, for instance, the largest value of the trace-back depth or the number of quantization levels for the soft-bit input.

As the decoder should be capable of processing both hard- and soft-input bits, the first design decision is the maximum input bit-width. As has been shown by previous research, increasing the soft-bit resolution beyond 5 bits has almost

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE II. CONVOLUTIONAL CODES IN DIGITAL COMMUNICATION SYSTEMS.

Standard	Constraint length K	Generator polynomials	Code rate R (mother code)
IEEE 802.11a/g/n (wireless LAN)	7	$g_1 = 133_8,\ g_2 = 171_8$	1/2
ECMA-368 (wireless USB)	7	$g_1 = 133_8,\ g_2 = 165_8,\ g_3 = 171_8$	1/3
DAB (Digital Audio Broadcasting)	7	$g_1 = g_4 = 133_8,\ g_2 = 171_8,\ g_3 = 145_8$	1/4
DVB (Digital Video Broadcasting), CCSDS 131.0-B (Space Data Systems, telemetry)	7	$g_1 = 171_8,\ g_2 = 133_8$	1/2
UMTS (Universal Mobile Telecommunications System)	9	$g_1 = 561_8,\ g_2 = 753_8$	1/2
		$g_1 = 557_8,\ g_2 = 663_8,\ g_3 = 771_8$	1/3
WiMAX (Worldwide interoperability for Microwave Access)	7	$g_1 = 171_8,\ g_2 = 133_8$	1/2
GSM (Global System for Mobile Communications)	5	$g_1 = 31_8, g_2 = 33_8$	1/2
		$g_1 = 33_8, g_2 = 25_8,\ g_3 = 37_8$	1/3
CDMA2000 (Code Division Multiple Access)	11	$g_1 = 4656_8,\ g_2 = 4726_8,\ g_3 = 5562_8,\ g_4 = 6372_8$	1/4
GMR-1 (GEO Mobile Radio interface), satellite phone	5	$g_1 = 31_8, g_2 = 27_8$	1/2
		$g_1 = 25_8,\ g_2 = 33_8,\ g_3 = 37_8$	1/3
		$g_1 = 31_8, g_2 = 27_8,\ g_3 = 25_8,\ g_4 = 37_8$	1/4
		$g_1 = 25_8,\ g_2 = 33_8,\ g_3 = 37_8,\ g_4 = 35_8,\ g_5 = 27_8$	1/5
	7	$g_1 = 133_8,\ g_2 = 171_8$	1/2
GMPRS (GEO Mobile Packet Radio Service), in addition to GMR-1	6	$g_1 = 45_8,\ g_2 = 55_8,\ g_3 = 73_8,\ g_4 = 77_8$	1/4
	9	$g_1 = 435_8,\ g_2 = 657_8$	1/2
		$g_1 = 755_8,\ g_2 = 633_8,\ g_3 = 443_8$	1/3
		$g_1 = 671_8,\ g_2 = 645_8,\ g_3 = 473_8,\ g_4 = 537_8$	1/4

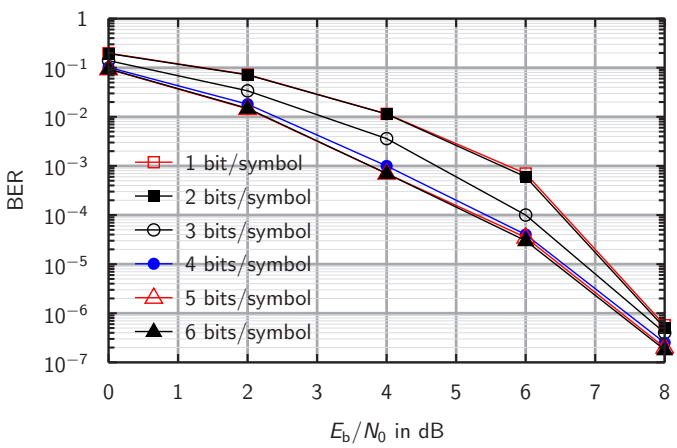
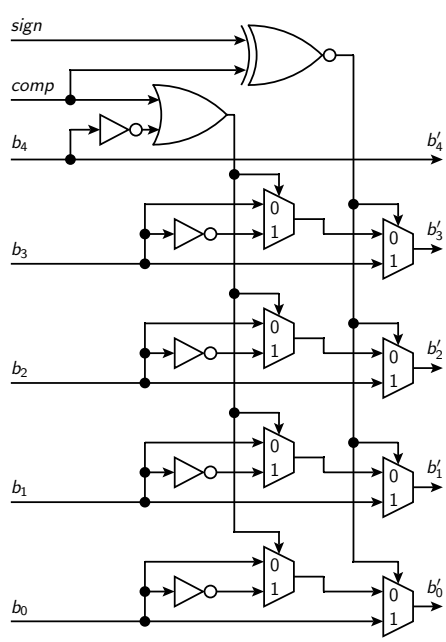


Figure 10. Effect of increasing the input bit width of a Viterbi decoder on BER (adapted from [11]).

no impact on the Bit Error Ratio (BER) [11]. Figure 10 summarizes the corresponding dependencies. According to these findings, the b parameter will be limited to the range from 1 (hard-input) to 5. To unify the BM calculation, the input is converted to the unsigned magnitude format. The corresponding circuit is shown in Figure 11. Depending on the value of the *sing* and *comp* inputs, the soft-bits are considered to be in the two's complement ($sign = 0$, $comp = 1$), signed magnitude ($sign = 1$, $comp = 0$) or unsigned magnitude ($sign = comp$) format (see Table I for explanation). The output always represents a corresponding unsigned magnitude value. Note that the size of the converter circuit does not depend on any other code parameter except the input bit-width, so for the fixed b value it remains constant and introduces almost no overhead to the overall silicon footprint. If the input bit-width is less than 5, then the corresponding bits are padded with zeros.



(a) Circuit

b	$c_x^{(i)} \mapsto b_y$
1	$c_x^{(0)} \mapsto b_4$
2	$c_x^{(1)} \mapsto b_4$
	$c_x^{(0)} \mapsto b_0$
3	$c_x^{(2)} \mapsto b_4$
	$c_x^{(1)} \mapsto b_1$
	$c_x^{(0)} \mapsto b_0$
4	$c_x^{(3)} \mapsto b_4$
	$c_x^{(2)} \mapsto b_2$
	$c_x^{(1)} \mapsto b_1$
	$c_x^{(0)} \mapsto b_0$
5	$c_x^{(4)} \mapsto b_4$
	$c_x^{(3)} \mapsto b_3$
	$c_x^{(2)} \mapsto b_2$
	$c_x^{(1)} \mapsto b_1$
	$c_x^{(0)} \mapsto b_0$

(b) Input mapping

Figure 11. Unsigned magnitude input converter ($c_x^{(i)}$ ist the ith bit of the input symbol c_x)

The number of BM values depends on the code rate R and is equal to 2^n. For instance, if the decoder should be able to decode any code up to $R = 1/5$, 32 branch metrics need to be calculated. For this purpose the circuit shown in Figure 5b can be scaled accordingly, while the constant values have to be adapted to the input bit-width b, e.g., 16 in case of $b = 4$ or 32 in case of $b = 5$. For other metrics than Manhattan distance, the data path has to be slightly extended. For instance, calculating squared Euclidean distance require an additional multiplier or

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

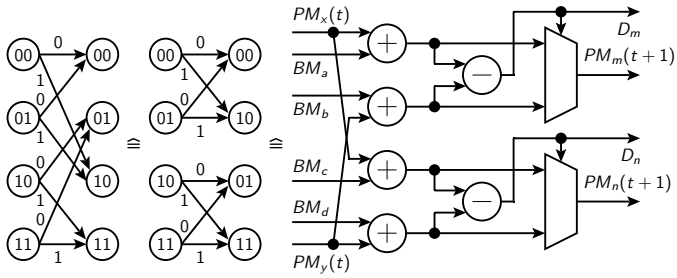


Figure 12. Butterfly structure of the path metric calculation. Comparators are represented by subtractors.

TABLE III. OUTPUT OF DIFFERENT CONVOLUTIONAL ENCODERS WITH $K = 3$.

State transition	Output c_1c_2 for $g_2 = 111$ and	
	$g_1 = 011$	$g_1 = 101$
$00 \rightarrow 00$	00	00
$00 \rightarrow 10$	01	11
$01 \rightarrow 00$	11	11
$01 \rightarrow 10$	10	00
$10 \rightarrow 01$	11	01
$10 \rightarrow 11$	10	10
$11 \rightarrow 01$	00	10
$11 \rightarrow 11$	01	01

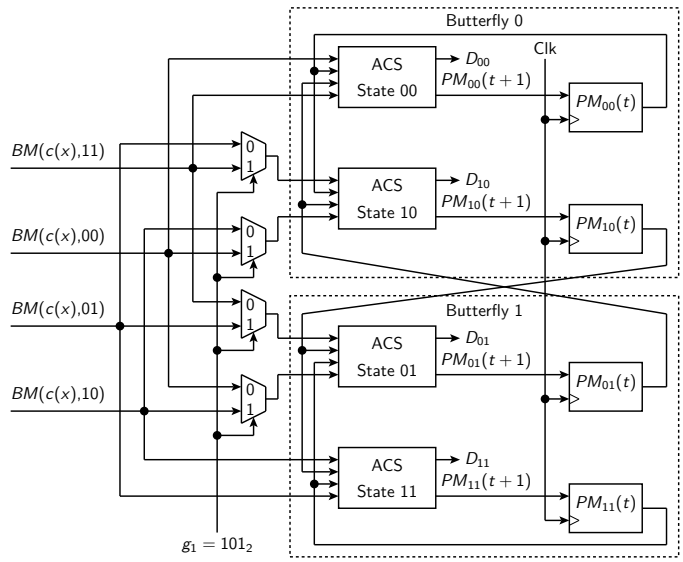


Figure 13. Path metric calculation for a reconfigurable Viterbi decoder.

a Look-Up-Table (LUT) per symbol. Again, once dimensioned for the lowest code rate (the largest value of n), the silicon area required to implement the adaptable branch metric unit does not change with any other code parameter.

The design of an adaptable ACS unit is determined be the number of codes with different generator polynomials as well as the maximal constraint length. For $K = 11$, the number of ACS calculations per decoded bit is equal to 1024. Mapping all of them to hardware so that the decoder will be able to achieve maximal degree of parallelism would require far too much silicon area. A reasonable approach is to implement 256 ACS units, so that all codes up to $K = 9$ can be decoded fully parallel, while for $K = 10$ and $K = 11$ two or four clock cycles are required to decode a single bit, respectively. Considering current CMOS technology nodes, even using this time-multiplexed mode for higher constraint lengths provides throughputs of several hundreds of megabits per second, which should be enough for most applications.

Changing the generator polynomials of the code does not affect the stage transitions of the encoder. In fact, due to the binary nature of the input, every state has two predecessors and two successor states while the edges connecting these states are following the same pattern, independent of the number of states. Figure 12 shows how these patterns can be used to combine pairs of states within the so-called *butterfly* structure. If generator polynomials are replaced by the new ones, the encoder, however, produces different output, since different bits (compared to the initial code) are passed to the XOR gates. Table III indicates how the output symbols produced by the same state transition change, if the generator polynomial $g_1 = 011$ of the encoder represented in Figure 1 is replaced by $g_1 = 101$ (while g_2 remains the same). Consequently, every new set of generator polynomials requires a different interconnect pattern between the branch metrics and the inputs of the butterfly units.

Figure 13 shows the ACS calculation unit for a decoder capable of decoding the output produced by the encoder in Figure 1, as well as the code generated by $g_1 = 101$ and $g_2 = 111$. While for each particular code the interconnect between the BM calculation outputs and PM calculation inputs is fixed, the adaptive decoder needs to switch between two different patterns depending on the code to be decoded. This functionality is realized by the multiplexers placed at the inputs of the particular ACS units. If the decoder has to process codes with different constraint lengths, some additional flexibility is required in the connections between the registers storing the actual PM values and the inputs of the ACS units as well. However, if the ACS units are grouped to the butterflies as shown above, the interconnect between them show a very regular pattern, so the overall area overhead is limited. Still, with larger number of different polynomials and larger values of the constraint length more and larger multiplexers at the inputs of the adaptive PM calculation unit are required reducing the clock frequency and throughput. A detailed design of the interconnect network for an adaptable Viterbi decoder including the particular interconnect patterns for different settings is described in [15].

In general, varying the constraint length and/or the generator polynomials has the strongest impact on the additional chip area and the critical path of the decoder. For the trace-back stage, the memory capacity has to be calculated for the worst case, i.e., the largest K and TBD values. In addition, the addressing scheme has to be adapted to every single case, which imposes the need for programmable address counters. Compared to other modifications this overhead is almost negligible. Figure 14 shows the trace-back unit of the reconfigurable Viterbi decoder

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

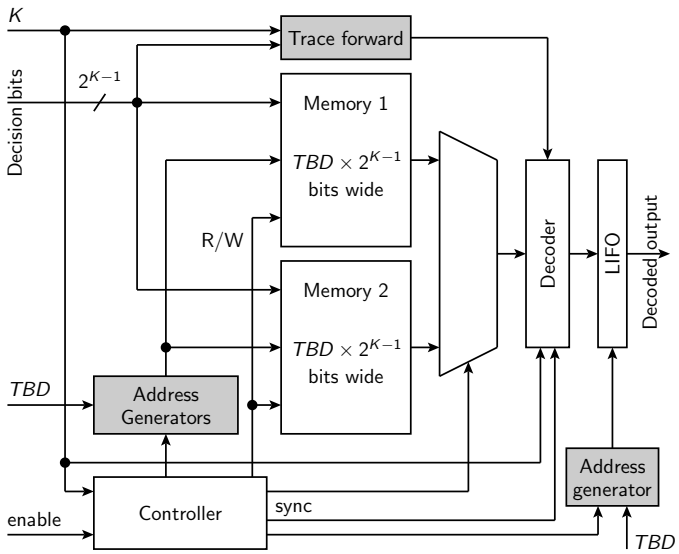


Figure 14. Trace-back unit for a reconfigurable Viterbi decoder with clock and reset signals omitted for the sake of clarity.

for variable values of the constraint length. The building blocks affected by the reconfigurable nature of the design are shaded. In principle, the address generators could be dimensioned for the worst case, just like the memories. However, larger TBD increases the latency of the decoder, which would mean an unnecessary overhead for smaller K values. At the same time, making the address generators pre-loadable and scalable adds almost nothing to the overall silicon footprint of the decoder.

The trace forward block discussed in the previous subsection needs to be dimensioned for the largest K value, while superfluous registers are bypassed when decoding codes with smaller constraint lengths. Again, some additional multiplexers are needed to implement this feature. Since trace forward unit is not on the critical path of the decoder, it does not introduce a major burden. One memory block is used to store the actual decision bits (write access) while the other is used to trace back the decision bits from the previous cycle (read access). The memories swap their roles after one cycle is completed, which explains the need for the multiplexer at their read ports.

Table IV summarizes the area and timing figures of different adaptable Viterbi decoder designs known from literature [16][17][18]. The major findings can be generalized as follows:

- The overhead increases with the degree of flexibility. The Viterbi decoders designed to support a (small) fixed amount of codes introduce the least – in some cases even negligible – overhead. In contrast, the full flexibility regarding the constraint length, the code rate as well as the generator polynomials requires up to 50 % more silicon area and reduces the throughput by about 26 %. Although looking diminishing at the first glance, it may still be quite a good price to pay considering the alternative of implementing a dedicated decoder for every single code.
- Varying the constraint length K has the highest impact on the area overhead of the adaptive Viterbi decoder implementation. Two factors can be identified as being responsible for this matter. On one hand, the number of ACS units increases exponentially with K, which means, the number and the size of the multiplexers required to implement different interconnect patterns increase accordingly. On the other hand, the TBD value grows linearly with K, so the memory depth of the trace-back grows accordingly. At the same time, the memory width grows exponentially with K. These consequences become crucial for higher values of K: the area overhead for $K = 9$ is about 9 %, while for $K = 11$ it is more than 50 %. For Viterbi decoders with smaller constraint lengths the area overhead is almost negligible.

TABLE IV. COMPARISON OF ADAPTABLE AND SINGLE-CODE VITERBI DECODERS.

Architecture	Metric	Reference	Result
[16] $K = 3 \ldots 7$, g variable, b fixed	silicon area	$K = 7$, g fixed	$+2,9\,\%$
	throughput	$K = 7$, g fixed	$-1,5\,\%$
VITURBO [17] $K = 3 \ldots 9$, g variable, b fixed	silicon area	$K = 9$, g variable	$+9\,\%$
	max. clock frequency	$K = 9$, g variable	$-30\,\%$
[18] $K \in \{7, 9\}$, g according to EDGE, CMDA2000 and WCDMA, b fixed	silicon area	$K = 9$, g fixed (CDMA2000)	overhead is negligible
	critical path	$K = 9$, g fixed (CDMA2000)	$+4\,\%$
[7] $K = 3 \ldots 11$, g variable, $n = 2 \ldots 4$, $b = 1 \ldots 5$	silicon area	$K = 11$, $n = 4$, $b = 5$, g fixed	$+50,1\,\%$
	throughput	$K = 11$, $n = 4$, $b = 5$, g fixed	$-26,1\,\%$

TABLE V. ELEMENTS OF THE FINITE FIELD $\mathrm{GF}(2^3)$ GENERATED BY $\mathbf{f}(x) = 1 + x + x^3$ IN DIFFERENT REPRESENTATIONS.

Symbolic	Binary	Polynomial	α-polynomial
0	000	0	$0 = 0 + 0 \cdot \alpha + 0 \cdot \alpha^2$
α^0	100	1	$1 = 1 + 0 \cdot \alpha + 0 \cdot \alpha^2$
α^1	010	x	$\alpha = 0 + 1 \cdot \alpha + 0 \cdot \alpha^2$
α^2	001	x^2	$\alpha^2 = 0 + 0 \cdot \alpha + 1 \cdot \alpha^2$
α^3	110	$1 + x$	$1 + \alpha = 1 + 1 \cdot \alpha + 0 \cdot \alpha^2$
α^4	011	$x + x^2$	$\alpha + \alpha^2 = 0 + 1 \cdot \alpha + 1 \cdot \alpha^2$
α^5	111	$1 + x + x^2$	$1 + \alpha + \alpha^2 = 1 + 1 \cdot \alpha + 1 \cdot \alpha^2$
α^6	101	$1 + x^2$	$1 + \alpha^2 = 1 + 0 \cdot \alpha + 1 \cdot \alpha^2$

III. REED-SOLOMON DECODERS

The basic idea of the error-correcting linear cyclic codes, nowadays known as the Reed-Solomon codes, can be traced back to 1952 [19][20]. In their current form, the codes were introduced in the seminal paper by Irving S. Reed and Gustave Solomon, hence the name [21].

A. Code Structure

RS codes are defined over a finite field $\mathrm{GF}(q^m)$, which is an extension field of $\mathrm{GF}(q)$, where q is a prime number and m is a natural number different from zero. All practically applied codes are based upon $q = 2$, which implies that the field size is a power of two. Every finite field is generated by an irreducible polynomial $\mathbf{f}(x)$ of the degree m, whereby the elements of the field can be themselves represented as

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

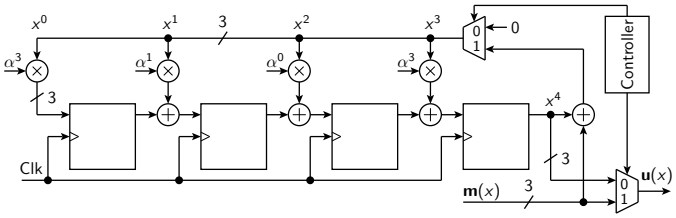


Figure 15. Reed-Solomon encoder for a (7,3) code over $\mathrm{GF}(2^3)$.

polynomials of the degree $m-1$ with the coefficients from $\mathrm{GF}(q)$ as shown in Table V. Choosing a different polynomial as the generator of the field results in different representation of its elements, however, it can be shown, that all finite fields for the fixed q und m values are isomorphic to each other [20]. If the generator polynomial of the field is primitive, its roots (called primitive elements) can be used as the generators of the multiplicative group of the field, i.e., every non-zero element can be derived from any root α by raising it to consecutive powers $\alpha^0, \alpha^1, \alpha^2, \ldots, \alpha^{q^m-2}$.

Binary representation of the elements of $\mathrm{GF}(2^m)$ requires m bits per element. This is the reason, why almost all RS codes used in communication and storage systems are defined over $\mathrm{GF}(2^8)$: each symbol can be encoded by exactly one byte. In contrast to the most other code families, the error correcting properties of the RS codes are defined symbol-wise, i.e., it does not matter, if a single bit or any other number of bits within one symbol are corrupted – the decoder treats it as a single symbol error. This explains, why RS codes are very powerful at correcting burst errors. A (255,239) RS code has the block size of 255 bytes, with 239 bytes of original message data and $2t = 255 - 239 = 16$ bytes of the checksum added during encoding. $t = 16/2 = 8$ is the number of corrupted symbols which can be corrected by the code. If 8 adjacent symbols (64 adjacent bits in the worst case) would be corrupted during the transmission, the decoder would still be able to correct all of them.

The general notation used to specify RS codes is (n, k), where n is the block size and k is the original message size before encoding (both specified in symbols of m bits width). As stated in the example above $n - k = 2t$, where t is the number of correctable symbol errors and $R = k/n$ is the code rate. The t parameter can be chosen depending on the required error correction capacity of the code. Once fixed, it defines the generator polynomial for the code, which has the degree $2t$:

$$\mathbf{g}(x) = g_0 + g_1 x + g_2 x^2 + \ldots + g_{2t-1} x^{2t-1} + x^{2t}. \quad (2)$$

The coefficients $g_0, \ldots, g_{2t-1}$ are all elements of $\mathrm{GF}(2^m)$ and the polynomial itself is defined in such a way, that its $2t$ roots correspond to the $2t$ consecutive powers of the single element of $\mathrm{GF}(2^m)$:

$$\mathbf{g}(x) = (x - \alpha) \cdot (x - \alpha^2) \cdot (x - \alpha^3) \cdot \ldots \cdot (x - \alpha^{2t}). \quad (3)$$

Although in principle any $2t$ consecutive powers can be chosen to define the generator polynomial, for practical reasons, the primitive element $\alpha = \alpha^1$ of the extension field $\mathrm{GF}(q^m)$ is chosen as the first root in most cases (being the generator of the multiplicative group of this field). Equation (3) follows this convention as well.

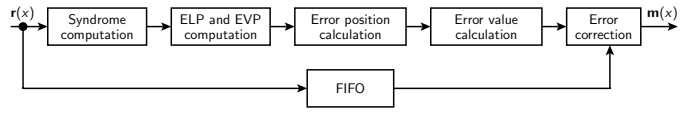


Figure 16. Top view architecture of the Reed-Solomon decoder.

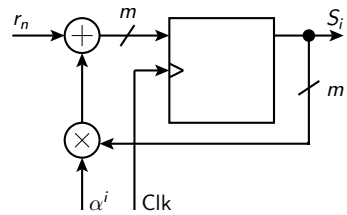


Figure 17. Basic cell for the syndrome computation.

RS codes are block-codes, i.e., they extend messages of the fixed size k by adding $n-k$ redundant symbols, so the encoded information block consists of n symbols. It is possible to vary the block size without changing the error correction capability of the code by replacing a part of the original message by z zeros before encoding. This z zero symbols are not transmitted. Instead, the receiver inserts the corresponding number of zeros before decoding, so effectively, the message size is reduced to $k - z$, while the difference $(n - z) - (k - z) = n - k = 2t$ determining the number of errors correctable by the code remains the same. This procedure is called shortening and the resulting code shortened, while the original code is referred to as mother code, just like in case of puncturing discussed in Subsection II-A.

The encoding procedure appends the rest $\mathbf{p}(x)$ of the division of the polynomial representing the original message $\mathbf{m}(x)$ shifted to the left by $n-k$ symbols by the generator polynomial:

$$x^{n-k}\mathbf{m}(x) = \mathbf{q}(x)\mathbf{g}(x) + \mathbf{p}(x), \quad (4)$$

which is equivalent to

$$\mathbf{p}(x) = x^{n-k}\mathbf{m}(x) \bmod \mathbf{g}(x). \quad (5)$$

This operation can be easily carried out using feedback shift-register as illustrated in Figure 15 for a very simple (7,3) RS code. Note that addition and multiplication operations are defined within the corresponding finite field (i.e., they differ from standard arithmetic). Since the original message is not changed by the encoding procedure, RS code is classified as systematic in coding theory (non-systematic RS codes exist as well but are not as widely used as their systematic counterparts).

B. Decoding Algorithm

Figure 16 shows the global architecture of the RS decoder ($\mathbf{r}(x)$ and $\mathbf{m}(x)$ stand for received and decoded message respectively). In a nutshell, the decoding process consists of five consecutive steps called syndrome computation, Error Locator Polynomial (ELP) and Error Value Polynomial (EVP)

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

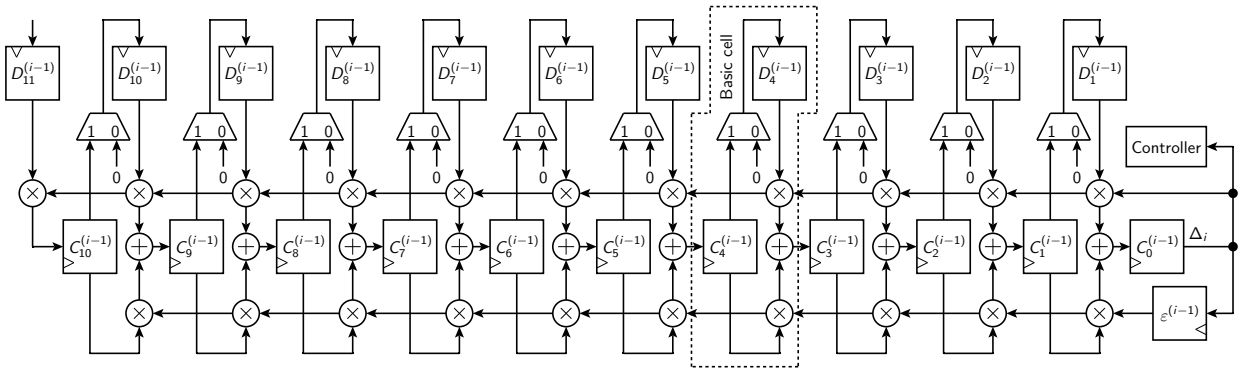


Figure 18. Dual-line implementation of the Berlekamp-Massey algorithm for $t = 5$ in hardware with clock and reset signals as well as multiplexer selector inputs omitted for the sake of clarity (adapted from [22]).

computation, error position calculation, error value calculation and error correction. Compared to the rest of the decoding algorithm, the first and the last step are rather trivial. Since decoding requires a number of clock cycles to accomplish, the First-In First-Out (FIFO) memory block is used to shift the received message accordingly, so the symbols arrive at the error correction stage at the same time as the corresponding correction values are provided.

Given that the encoding process can be represented by polynomial multiplication over $\mathrm{GF}(q^m)$, it becomes clear how transmission errors can be detected (and corrected). Since the multiplication operation preserves the roots of the generator polynomial, the decoder can verify the received message by checking whether $2t$ consecutive powers of the primitive element of $\mathrm{GF}(q^m)$ are roots of the polynomial $\mathbf{r}(x)$. This is the first step of the decoding process called syndrome computation. If at least one of the syndromes is not equal to zero, then the message was corrupted. Depending of the latency requirements, the syndromes can be calculated sequentially or in parallel by inserting the consecutive powers of α (elements of the field $\mathrm{GF}(2^m)$) into $\mathbf{r}(x)$. Figure 17 (adapted from [23]) shows the hardware used for the syndrome computation with r_n representing the nth symbol of the received message and S_i standing for the ith syndrome value.

After the syndromes are known, they are used in the next step of the decoding process to compute the number and the positions of the corrupted symbols. The most intuitive way to accomplish this is – in the first iteration – to assume the number of errors to be one and try to solve the corresponding system of equations $\mathbf{\Lambda}_l \cdot \mathbf{S}^\top = 0$ based on the vector of syndrome values $\mathbf{S}$ and the error locator polynomial

$$\begin{aligned}\mathbf{\Lambda}_l(x) &= x^l \cdot \prod_{i \in I_{\mathrm{FP}}} \left(x - \alpha^i\right) \\ &= x^l \cdot \left(\lambda_0 + \lambda_1 x + \ldots + \lambda_{r-1} x^{r-1} + x^r\right)\end{aligned} \tag{6}$$

with l as the (assumed) number of corrupted symbols, and $I_{\mathrm{FP}} = \{i : e_i \neq 0, 0 \leq i < n\}$ as the set of positions of all error values e_i. For $i \in I_{\mathrm{FP}}$, $x = \alpha^i$ is a root of ELP meaning that $e_i \neq 0$ (ith symbol of the message is corrupted).

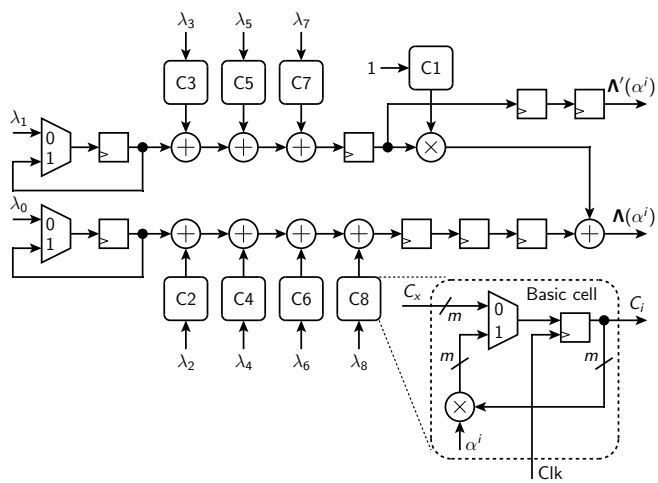


Figure 19. Hardware implementation of the Chien search for a (255,239) RS decoder. Clock and reset signals are omitted for the sake of clarity (adapted from [24]).

Accordingly if $i \notin I_{\mathrm{FP}}$, then $x = \alpha^i$ is not a root of ELP meaning that $e_i = 0$ (ith symbol of the message is correct).

The solution for $l = 1$ indicates the position of the corrupted symbol. In case the system of equations for a single error is not solvable, the assumption of two corrupted symbols ($l = 2$) is made resulting in a different system of equations. The process is repeated until a solvable system of equations is found and solved delivering the ELP coefficients $\lambda_0 \ldots \lambda_{l-1}$. This iterative procedure of finding the number of the corrupted symbols is called Peterson-Zierler-Gorenstein algorithm [25][26].

When ELP coefficients are known, finding the positions of the erroneous symbols can be easily accomplished by Chien search [27]. It checks which symbols of $\mathrm{GF}(2^m)$ (except zero) are roots of the $\mathbf{\Lambda}_l(x)$. As already mentioned, ELP is constructed in such a way, that the equation $\mathbf{\Lambda}_l(\alpha^i) = 0$ is valid, if ith symbol of the received message is corrupted. Consequently, the power i of every root found by the Chien search indicates the position of one corrupted symbol.

Since RS codes work on symbols of m bits width, knowing the error position of the symbol is not enough. An additional

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

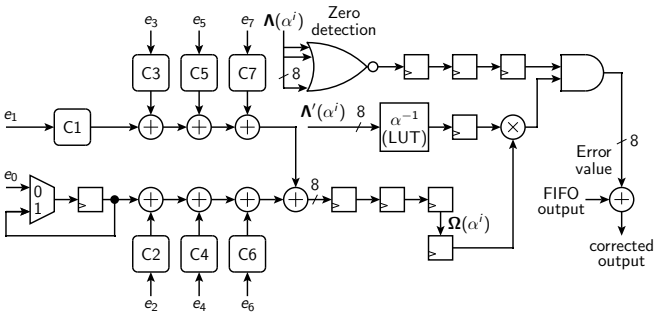


Figure 20. Hardware implementation of the Forney algorithm for a (255,239) RS decoder. Basic cells are the same as ones used for the Chien search. Clock and reset signals are omitted for the sake of clarity (adapted from [24]).

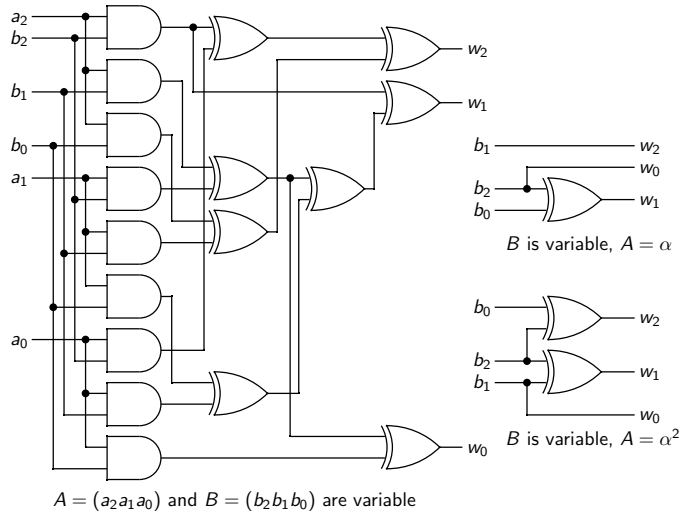


Figure 21. Variable-variable versus constant-variable multiplier for $\mathrm{GF}(2^3)$.

step is required to determine the error values. This can be done by using the syndromes computed in the first step, since every set of syndromes can be mapped to certain errors. Alternatively, the error value polynomial

$$\mathbf{\Omega}(x) = 1 + e_1 x + e_2 x^2 + \ldots + e_{e-1} x^{e-1} \tag{7}$$

with e_x representing the error values can be used [28]. EVP stands in relation to the ELP and the syndrome vector via the so-called *key equation*

$$\mathbf{\Lambda}(x)\mathbf{S}(x) \bmod x^{2t} = \mathbf{\Omega}(x). \tag{8}$$

Since Peterson-Zierler-Gorenstein algorithm does not scale very well, Berlekamp-Massey algorithm is more frequently used for the hardware implementation of the RS decoders [29]. It can be parallelized and modified to reuse the same hardware to calculate both error location and error value polynomials by solving the key equation. Figure 18 shows the hardware implementation of the Berlekamp-Massey algorithm using the so-called dual-line architecture [22]. Its detailed description, however, would go far beyond the scope of this paper. Again, after the coefficients of the ELP and EVP are calculated, Chien search must be carried out to determine the error locations. Figure 19 shows the corresponding circuit, where the notation Cx within the basic cell means, that α^x value is used for the calculation. In parallel to the error locations the so called *formal derivative* $\mathbf{\Lambda}'$ of the ELP is calculated:

$$\mathbf{\Lambda}'(x) = \sum_{i=1}^{t} i \cdot \lambda_i x^{i-1} \;\text{ with }\; i \cdot \lambda_i = \underbrace{\lambda_i + \lambda_i + \cdots + \lambda_i}_{i \text{ times}}. \tag{9}$$

Assuming that the sequence of the $2t$ consequtive powers of the GF elements used as the roots of the generator polynomial $\mathbf{g}(x)$ starts with $\alpha^1 = \alpha$, the error values can be calculated with Lagrange interpolation using the formal derivative and the (already caluculated) EVP [30][31]:

$$e_i = -\frac{\mathbf{\Omega}(\alpha^{-i})}{\mathbf{\Lambda}'(\alpha^{-i})} \;\text{ with }\; \alpha^{-i} = \frac{1}{\alpha^i} \;\equiv\; \alpha^i \cdot \alpha^{-i} = 1, \tag{10}$$

whereby the same basic cell as used for the Chien search can be applied and the multiplicative inverse values α^{-i} can be pre-calculated and stored in a LUT. This method of determining the error values are known as Forney algorithm [28]. As can be seen from Figure 20, which shows the corresponding hardware, the data path resembles some of the structure used for the Chien search. The inputs of the basic cells are now the coefficients of the EVP (instead of ELP) and both ELP $\mathbf{\Lambda}(x)$ and its formal derivative $\mathbf{\Lambda}'(x)$ are used as inputs.

In the final decoding step, the error values are added to the received symbols at the calculated error positions restoring the original message. As already mentioned, a FIFO memory is used to make sure the right values appear at the output of the decoder at the right time. The finite filed addition corresponds to a bitwise XOR operation, which scales well and is trivial from the implementation point of view.

C. Adaptive Implementation

Table VI summarizes some applications of RS codes in different communication standards. Depending on the required error correction performance, various finite fields, primitive polynomials and t values are used in practice.

As can be seen from the previous subsection, decoding the RS codes requires numerous calculations based on the finite field arithmetic. While the finite field addition is trivial and scales seamlessly with the bit-width m, efficient implementation of the finite field multiplication has some issues. In an RS decoder designed for a specific code, many of the GF-multipliers have one constant input (variable-constant multipliers). For adaptive decoding of different codes, a variable-variable multiplier is required. The corresponding overhead is quite high as Figure 21 illustrates for a very simple case of $\mathrm{GF}(2^3)$. This becomes even more critical with larger values of m. Luckily, only the multipliers involved in the syndrome and error location computation are affected, so the impact on the area of the whole RS decoder is limited, though not negligible.

The traditional way to implement a variable-variable GF-multiplier is based on the polynomial modulo reduction. If the

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE VI. REED-SOLOMON CODES IN DIGITAL COMMUNICATION SYSTEMS.

Standard	Galois-Field	Primitive polynomial $\mathbf{f}(x)$	(Mother) code
ITU-T J.83 (digital cable TV, USA)	$\mathrm{GF}(2^7)$	x^7+x^3+1	(128,122)
	$\mathrm{GF}(2^8)$	$x^8+x^4+x^3+x^2+1$	(204,188), (207,187)
ECMA-368 (wireless USB)	$\mathrm{GF}(2^8)$	$x^8+x^4+x^3+x^2+1$	(23,17), shortened from (255,249)
DAB (Digital Audio Broadcasting)	$\mathrm{GF}(2^8)$	$x^8+x^4+x^3+x^2+1$	(204,188) shortened from (255,239)
DVB (Digital Video Broadcasting)	$\mathrm{GF}(2^8)$	$x^8+x^4+x^3+x^2+1$	(204,188) shortened from (255,239)
DOCSIS 3.0 and 3.1 (Data Over Cable Service Interface Specifications)	$\mathrm{GF}(2^8)$	$x^8+x^4+x^3+x^2+1$	$(255,k)$ with $k=223,225,227,\ldots,251,253$, also shortened
WiMAX (Worldwide interoperability for Microwave access)	$\mathrm{GF}(2^8)$	$x^8+x^4+x^3+x^2+1$	(255,239)
VDSL (Very high speed Digital Subscriber Line)	$\mathrm{GF}(2^8)$	$x^8+x^4+x^3+x^2+1$	(144,128), (240,224), and others
GSM (Global System for Mobile communications)	$\mathrm{GF}(2^8)$	$x^8+x^4+x^3+x^2+1$	(85,73) shortened from (255,243)
IEEE802.3bp (automotive Ethernet)	$\mathrm{GF}(2^9)$	x^9+x^4+1	(450,406)
IEEE802.3bj (100Gb/s Ethernet)	$\mathrm{GF}(2^{10})$	$x^{10}+x^3+1$	(528,514), (544,514)
IEEE802.11ad (directional multi-gigabit wireless LAN)	$\mathrm{GF}(2^8)$	$x^8+x^4+x^3+x^2+1$	(224,208) shortened from (255,239)
GMR-1 (GEO Mobile Radio interface), satellite phone GMPRS (GEO Mobile Packet Radio Service)	$\mathrm{GF}(2^4)$	x^4+x+1	(15,9)
CCSDS 131.0-B (Space Data Systems, telemetry)	$\mathrm{GF}(2^8)$	$x^8+x^7+x^2+x+1$	(255,223), (255,239)

elements of the finite field are written in the polynomial form, their product can be calculated as follows [24]:

$$W = A\cdot B = \sum_{k=m}^{2m-2} d_k h^k + \sum_{k=0}^{m-1} d_k h^k \ \text{ with } \ d_k = \sum_{i=0}^{k} a_i b_{k-i}, \tag{11}$$

where a_x and b_x are the corresponding coefficients of the polynomial representation of A and B. Since a_x and b_x are both elements of $\mathrm{GF}(2)$, $d_k \in \mathrm{GF}(2)$ is valid as well. Given $h^k \in \mathrm{GF}(2^m)$ and $m \le l \le 2m-2$, h^k can be rewritten as

$$h^k = \sum_{i=0}^{m-1} g_i^k h^i, \tag{12}$$

while g_i^k with $0 \le i \le m-1$ is uniquely defined. Using the substitution $w_k = d_k + \sum_{j=m}^{2m-2} d_j g_k^j$, the product can be rewritten as

$$W = \sum_{k=0}^{m-1} w_k h^k, \ \text{ i.e., } \ W(x) = A(x)\cdot B(x) \bmod \mathbf{f}(x) \tag{13}$$

with $\mathbf{f}(x)$ as the generator polynomial of the finite field (not of the RS code) [24]. Thus, finite field multiplication can be divided into two steps: polynomial multiplication according to the equation (11) and modulo reduction according to the equation (13).

For $\mathrm{GF}(2^3)$, the polynomial representation of the elements looks like $A = a_2\alpha^2 + a_1\alpha + a_0$ and $B = b_2\alpha^2 + b_1\alpha + b_0$, so the first step delivers

$$\begin{aligned} A\cdot B &= (\alpha A b_2 + A b_1)\alpha + A b_0 = (\alpha(a_2\alpha^2 + a_1\alpha + a_0)b_2 \\ &\quad + (a_2\alpha^2 + a_1\alpha + a_0)b_1)\alpha + (a_2\alpha^2 + a_1\alpha + a_0)b_0 \\ &= a_2b_2\alpha^4 + (a_1b_2 + a_2b_1)\alpha^3 \\ &\quad + (a_0b_2 + a_1b_1 + a_2b_0)\alpha^2 + (a_0b_1 + a_1b_0)\alpha + a_0b_0. \end{aligned} \tag{14}$$

The higher powers $\alpha^4 = a^2 + \alpha$ and $\alpha^3 = \alpha + 1$ are reduced modulo 3 in the second step

$$\begin{aligned} A\cdot B &= a_2b_2\alpha^4 + (a_1b_2 + a_2b_1)\alpha^3 \\ &\quad + (a_0b_2 + a_1b_1 + a_2b_0)\alpha^2 + (a_0b_1 + a_1b_0)\alpha + a_0b_0 \\ &= a_2b_2(\alpha^2 + \alpha) + (a_1b_2 + a_2b_1)(\alpha + 1) \\ &\quad + (a_0b_2 + a_1b_1 + a_2b_0)\alpha^2 + (a_0b_1 + a_1b_0)\alpha + a_0b_0 \\ &= (a_2b_2 + a_0b_2 + a_1b_1 + a_2b_0)\alpha^2 \\ &\quad + (a_2b_2 + a_1b_2 + a_2b_1 + a_0b_1 + a_1b_0)\alpha \\ &\quad + a_1b_2 + a_2b_1 + a_0b_0. \end{aligned} \tag{15}$$

The expression (15) reflects exactly the structure of the variable-variable multiplier depicted in the Figure 21, where nine AND gates represent the multiplication and eight XOR gates the addition over $\mathrm{GF}(2)$. Since the polynomial multiplication and modulo reduction have the same logic depth, a pipeline stage can be added between them for larger values of m to improve the clock frequency.

Although providing reasonable performance for a fixed m value, the polynomial modulo reduction approach cannot be applied for an adaptive architecture. Some solutions exists for specific cases, e.g., if $\mathrm{GF}(2^l)$ is a sub-field of $\mathrm{GF}(2^k)$ while l is an integer divider of k, a multiplication in $\mathrm{GF}(2^k)$ can be implemented by several multiplications in $\mathrm{GF}(2^l)$ [9][32].

A better approach is based on the Montgomery reduction [33], which uses two identical multiplier circuits to implement a multiplication over $\mathrm{GF}(2^m)$, while the same hardware can be reused for the multiplication over $\mathrm{GF}(2^l)$ for $l \le m$. Figure 22 (adapted from [34]) shows the top-view architecture of the Montgomery multiplier with an example implementation for $\mathrm{GF}(2^4)$. Note that m_i are the (binary) coefficients of the generator polynomial $\mathbf{f}(x)$ of the finite field and $K(x)$ is a constant for every fixed $\mathbf{f}(x)$. The correspondence $m_0 = 1$ can be used to simplify the sub-circuits which use m_0 as an input.

Having a suitable GF multiplier in place, the implementation of an adaptable RS decoder is quite straightforward. The hardware components introduced in the Subsection III-B must be dimensioned for the largest envisioned values of t and m.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

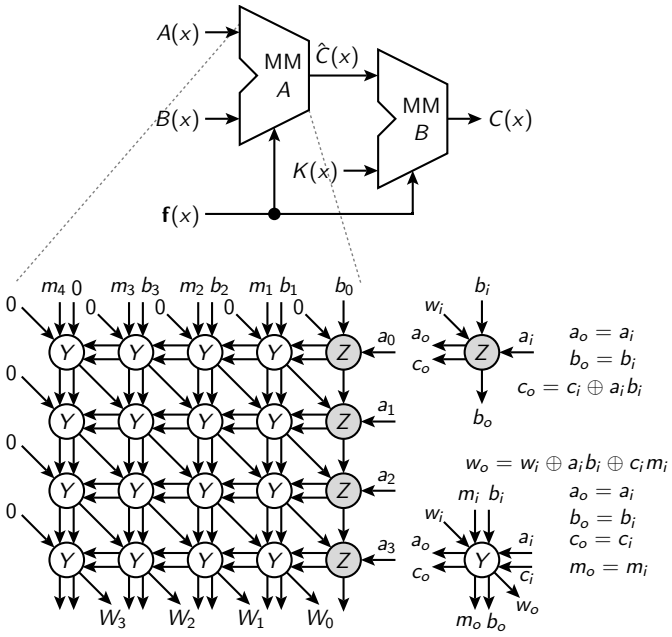


Figure 22. Montgomery multiplier for $\mathrm{GF}(2^m)$ with $m \leq 4$.

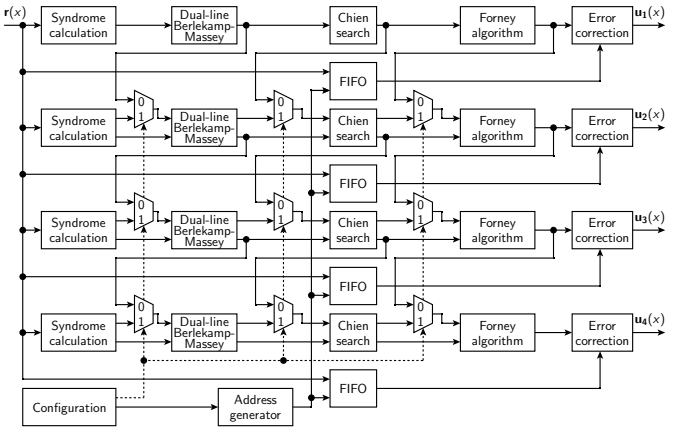


Figure 23. Top-level architecture of a multi-standard multi-stream Reed-Solomon decoder.

For smaller values of m, corresponding bits in the symbolic representation of the finite field elements must be filled with zeros. The same applies to the syndrome computation when decoding codes with smaller t parameters, i.e., the "nonexistent" roots are set to zero in advance. Alternatively, the inactive hardware can be used to decode one or more additional data streams in parallel. This can be done as long as the condition

$$\sum_{i=1}^{w} t_i \leq t_{\max} \tag{16}$$

holds (t_i represents the error correction capacity of the code for the data stream i). The corresponding architecture is shown in Figure 23 and can, for instance, be used to decode a single stream with $t = 16$ or up to four data streams encoded using a code with $t = 4$. The configuration block stores the control information to direct the data stream(s). In addition, the Montgomery multipliers must be configured depending on the chosen generator polynomials.

TABLE VII. COMPARISON OF ADAPTABLE AND SINGLE-CODE REED-SOLOMON DECODERS.

Comparison metric	Decoder architecture				
	[37]	[38]	[39]	[40] 1st variant	[40] 2nd variant
Code parameter	fixed	n and t variable	Universal decoder: n, t, m and $\mathbf{f}(x)$ variable		
m	8	8	$1 \ldots 8$	$1 \ldots 10$	$1 \ldots 8$
t	8	$1 \ldots 8$	$1 \ldots 8$	$1 \ldots 8$	$1 \ldots 16$
Erasure correction	no	no	no	≤ 16	≤ 16
Throughput in Gb/s	1,6	0,8	0,048	2,2	2,4
Gate equivalents	21 000	34 000	44 000	75 000 + 35 Kbit RAM	39 000 + 15 Kbit RAM

The last element affected by the adaptable implementation is the FIFO memory. Since its capacity does not depend on t, it has to be dimensioned for the largest envisioned block size and use programmable address generators to decode smaller blocks. The resulting overhead is almost negligible given the fact, that – depending on the implementation – the FIFO buffer can consume up to 75 % – 90 % of the overall chip area [35][36].

Table VII summarizes the area and throughput figures of different RS decoder designs known from literature [37][38][39][40]. For the adaptive decoding, n, t, and m parameters can be changed, though in practice m is usually kept constant, since almost all real-world RS codes are based on $m = 8$ (ITU-T J.83, IEEE802.3bp, and IEEE802.3bj standards being an exception by specifying codes based on $\mathrm{GF}(2^7)$, $\mathrm{GF}(2^9)$, and $\mathrm{GF}(2^{10})$ correspondingly). In addition, different polynomials can be chosen as the finite field generators. As mentioned in Subsection III-A, varying the GF generator polynomial $\mathbf{f}(x)$ results in a different encoding of the particular elements of the finite field, however, all finite fields for the fixed values of q and m are isomorphic to each other.

The conclusions from the study of the different RS decoder designs are as follows:

- The area overhead of the adaptive RS decoder implementation can be quite significant. Compared to the fixed-code decoder with $m = t = 8$, a fully reconfigurable design can consume about 85 % more silicon area in terms of gate equivalents in addition to 15 KBit more SRAM. Just as in the case of an adaptive Viterbi decoder this is still quite a low price to pay, since the error correction capabilities are much better and the additional erasure correcting feature is useful in many application scenarios (erasures are errors whose position is known in advance).
- Throughput is usually not an issue with RS decoding, since the symbol-per-second decoding rate is multiplied by m to obtain the bit-per-second values. For adaptive decoding, this means that the slower clock rates, caused by the additional overhead of the adaptive implementation, are not significant in most cases.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

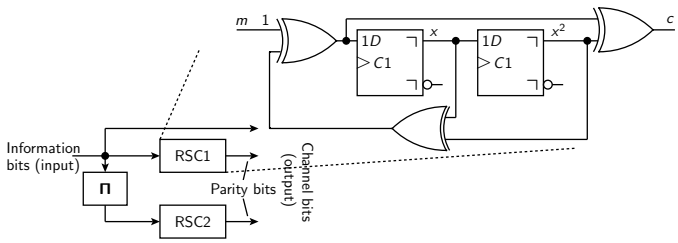


Figure 24. Simple turbo encoder.

TABLE VIII. TURBO CODES IN DIGITAL COMMUNICATION SYSTEMS.

Standard	Maximum constraint length K	Maximum block size (bits)	Code rate (mother code)
LTE (Long Term Evolution)	4	6 144	1/3
WiMAX (Worldwide interoperability for Microwave access)	4	4 800	1/3
UMTS (universal mobile telecommunications system)	4	5 144	1/3
CDMA2000 (Code division multiple access)	4	20 730	1/5
EV-DO (Evolution data optimized)	4 4	4 096 1024	1/3 1/5
DVB-RCS (digital video broadcasting return channel satellite)	4	1 728	1/3
DVB-RCS2 (digital video broadcasting return channel satellite, second generation)	5	4 792	1/3
CCSDS 131.0-B (Space Data Systems, telemetry)	5	8 920	1/4

IV. MULTI-FAMILY DECODERS

As can be seen from the previous sections, the arithmetic involved in the decoding of convolutional and RS codes is quite different, which limits the possibilities of meaningful hardware reuse. However, it may be an option to combine decoders for the other code families within one design with Viterbi decoder or with each other.

A. Turbo Codes

Turbo codes were introduced in 1993 [41]. The basic idea is to use several (usually two) recursive convolutional encoders to process the same input data. First encoder receives the data directly, the second one gets an interleaved data stream as illustrated in Figure 24 (adapted from [42]). The size of the interleaver is usually in the order of several kilobits. Pseudo-random address generators are commonly used to implement the interleaving operation, which ensures that the same data appears as several statistically independent messages.

In contrast to the non-recursive non-systematic encoders discussed in the Subsection II-A, turbo encoders are usually built based upon Recursive Systematic Codes (RSC) due to their ability – especially if several encoders are concatenated – to produce codewords of higher weight, which has advantages concerning the recursive decoding procedure. The outputs of the encoders are transmitted as parity information together with the original message, so the resulting code is systematic. Due to the feedback within the data path, the RSC can be seen as an Infinite Impulse Response (IIR) filter [9, pp. 453–454]. Just like in case of conventional convolutional codes, puncturing can be applied to turbo codes to change the code rate.

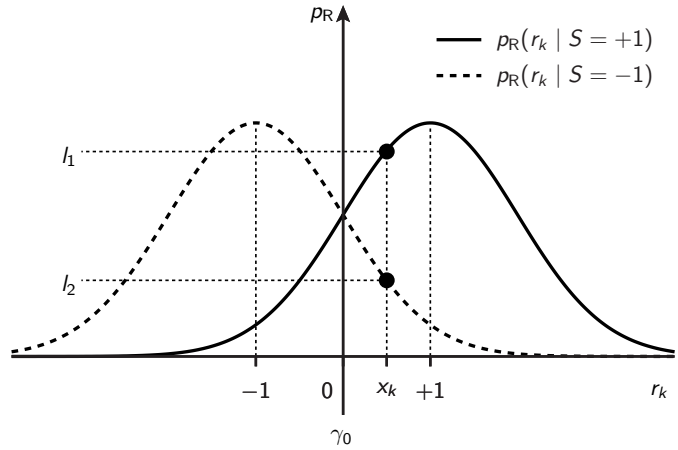


Figure 25. Likelihood functions for a BPSK symbols transmitted over AWGN channel [43].

Table VIII summarizes some applications of the turbo codes in different communication standards with their mother code rates (some standards specify additional code rates, which are not included in the table). The probably simplest explanation of the decoding principles assumes the Binary Phase Shift Keying (BPSK) signal s_k transmitted over an AGWN channel with noise specified as n_k [2][43, Section 8.1]. The received signal is the sum of the transmitted signal and noise

$$r_k = s_k + n_k \;\text{ with }\; p_{\mathrm{N}}(n_k) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{n_k}{\sigma}\right)^2} \tag{17}$$

as probability density with variance σ^2. The likelihood of the received signal to have a certain value under the assumption that s_k was sent can be calculated as follows:

$$\begin{aligned} p_{\mathrm{R}}(r_k \mid S = s_k) &= p_{\mathrm{N}}(r_k - s_k \mid S = s_k) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{r_k - s_k}{\sigma}\right)^2}. \end{aligned} \tag{18}$$

Since for a BPSK signal the only possible values for s_k are $+1$ and -1, these likelihood functions can be easily visualised as shown in Figure 25. A hard-decision maximum-likelihood decoder would compare the received value x_k against the likelihoods l_1 and l_2 and produce the value with the higher likelihood as the output. In other terms, such a decoder returns $+1$, if x_k lies on the right side of the decision threshold γ_0 and -1, if x_k lies to the left of γ_0 [43, Section 8.1.1.2].

Turbo decoder, however, works with soft decisions (soft output), i.e., instead of ± 1, the decoded value is the likelihood for a bit to be zero or one. A common way to represent this likelihood is to encode the bit decision as a sign bit (MSB) and the reliability of this decision as a magnitude. Log-Likelihood-Ratio (LLR) is an example of a metric constructed this way [44]:

$$LLR_{\mathrm{R}} = \ln \frac{p_{\mathrm{R}}(r_k \mid S = +1)}{p_{\mathrm{R}}(r_k \mid S = -1)}, \tag{19}$$

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

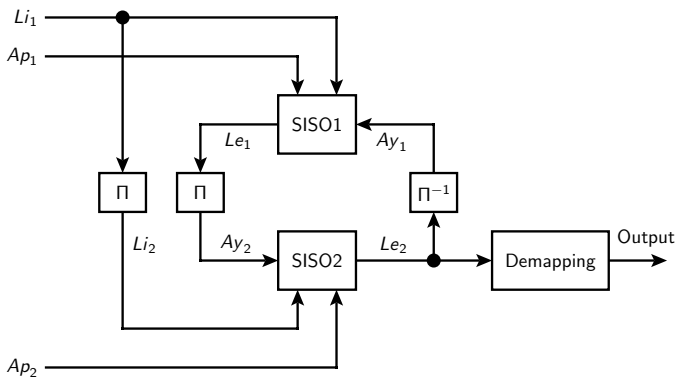


Figure 26. Top view architecture of the turbo decoder [2]. Li_x represent systematic bits (original message).

which is frequently used in soft-output decoding due to its numerical advantages (no floating-point multiplicaion required).

Figure 26 shows the global structure of the turbo decoder. The decoding is done iteratively using two decoders that process the received data and pass their outputs to each other, with interleaving and deinterleaving steps between them. Since each decoder uses the soft metrics produced by the other (called a-priori information Ay_x) as an own input and produces soft output decisions (called extrinsic bits Le_x) itself, the corresponding process is called Soft-Input Soft-Output (SISO) decoding.

In contrast to the hard-decision MLL decoding, each SISO module uses the a-priori information to improve its own decisions, which are then passed back to the first decoder and so on. During this process, the initial likelihood values associated with x_k are improved and shifted towards a more reliable decision after each iteration, so that the decoding process usually can be stopped after several iterations [43, Subsection 8.2.5]. This feedback loop gave the name to the code family in analogy to the turbocharger principle used in combustion engines. In the systems with fixed maximal latency requirements, the decoding process is stopped after several iterations. In addition, the decoding process can be stopped, if the results of subsequent iterations do not change (so-called *early stopping*), which can be easily implemented in LLR representation by comparing the MSBs of the corresponding soft-decisions. In the final step, a block called a demapper converts soft decisions into hard bits based on their probability metrics. Since both SISO modules are identical, common implementations of the turbo decoders often have only one SISO block in place and reuse it in a time-multiplexing manner as SISO1 and SISO2 (Figure 26) interchangeably to save the silicon area.

The error-correcting capabilities of the turbo codes depend on the constraint length and the size of the interleaver. Increasing both parameters improve the performance but has negative impact on the silicon footprint. As can be seen from Table VIII, the common range for K is between 3 and 5 and for block size between 1 and 21 kbit.

TABLE IX. LDPC CODES IN DIGITAL COMMUNICATION SYSTEMS.

Standard	Maximum block size (bits)	Code rate
IEEE 802.11n (wireless LAN)	1 944	1/2, 2/3, 3/4, 5/6
IEEE 802.11ad (WiGig, 60 GHz WLAN)	672	1/2, 3/4, 5/8, 13/16
IEEE 802.11.3c (WiGig, 60 GHz WLAN)	672	1/2, 3/4, 5/8, 7/8
IEEE 802.11ay (WiGig, 60 GHz WLAN)	1 344	1/2, 2/3, 3/4, 5/8, 5/6, 7/8, 13/16
WiMAX (Worldwide interoperability for Microwave Access)	2 304	1/2, 2/3, 3/4, 5/6
DVB-T2 and -S2 (Digital Video Broadcasting Terrestrial and Satellite)	64 800	1/2, 3/5, 2/3, 3/4, 4/5 5/6 (both), 1/4, 1/3, 2/5, 8/9, 9/10 (S2 only)
GMR (GEO mobile radio interface, satellite phone)	11 136	1/2, 2/3, 3/4, 4/5, 9/10
GMPRS (GEO Mobile Packet Radio Service)	8 880	1/2, 2/3, 3/4, 4/5, 9/10
IEEE 802.3an (10 Gigabit Ethernet)	2 048	7/8
IEEE 802.22 (WRAN, Wireless Regional Area Networks)	480	1/2, 2/3, 3/4, 5/6
CCSDS 131.0-B (Space Data Systems, telemetry)	8 160	1/2, 2/3, 4/5, 7/8
CCSDS 231.0-B (Space Data Systems, telecommand)	512	1/2

B. LDPC Codes

LDPC codes were described in detail in the dissertation of Robert Gallager in 1963 [45]. Due to the high computational requirements of the decoding process, the first practical applications emerged more than 30 years later, when LDPC codes were proposed as channel codes for several communication standards (see Table IX for examples). In a nutshell, LDPC codes are linear block codes with very large (up to several hundreds rows and columns), very sparsely filled generator matrix. As for other block codes, the encoding process can be described by the vector-matrix multiplication of the message block $\mathbf{m}$ by the generator matrix $\mathbf{G}$ producing the encoded message $\mathbf{c}$:

$$\mathbf{c} = \mathbf{m} \times \mathbf{G}. \tag{20}$$

The major difference to the "normal" block codes is the large block size (in the order of $10^4 \ldots 10^5$ bits) with the small number of parity bits (i.e., number of ones in the generator matrix), hence the name. For a systematic block code, $\mathbf{G}$ has the following form

$$\mathbf{G} = \left(\mathbf{P} \vdots \mathbf{I}_k \right) \tag{21}$$

with $\mathbf{P}$ as the parity matrix and $\mathbf{I}_k$ as the identity matrix of size k (the length of the message block $\mathbf{m}$). Since the parity matrix determines which message bits are involved in the corresponding parity equations, the correctness of the received information block of length n can be checked using the parity check matrix $\mathbf{H}^\top$ defined as

$$\mathbf{G} \times \mathbf{H}^\top = \mathbf{0} \text{ with } \mathbf{H}^\top = \begin{pmatrix} \mathbf{I}_{n-k} \\ \cdots \\ \mathbf{P} \end{pmatrix}. \tag{22}$$

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

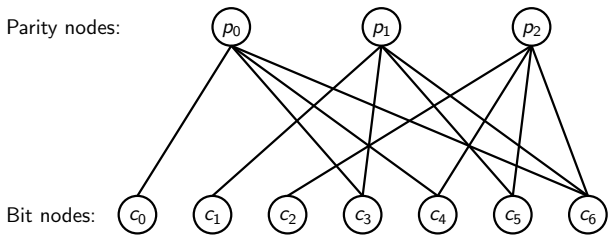


Figure 27. Tanner graph for a (7,4) Hamming code.

If the received codeword $\mathbf{c}'$ was affected by the error vector $\mathbf{e}$, the syndrome computation (multiplying $\mathbf{c}'$ by $\mathbf{H}^\top$) will deliver a non-zero value:

$$\begin{aligned}\mathbf{s} &= \mathbf{c}' \times \mathbf{H}^\top = (\mathbf{c}+\mathbf{e}) \times \mathbf{H}^\top = (\mathbf{m} \times \mathbf{G} + \mathbf{e}) \times \mathbf{H}^\top \\ &= \mathbf{m} \times \underbrace{\mathbf{G} \times \mathbf{H}^\top}_{=\mathbf{0} \text{ see } (22)} + \mathbf{e} \times \mathbf{H}^\top = \mathbf{e} \times \mathbf{H}^\top. \end{aligned} \tag{23}$$

According to the equation (23), the syndrome vector depends on the error vector, so it can be used to determine the bits affected by noise and restore the original message. The same principle applies to the non-systematic codes as well, though their generator and parity check matrix do not include the identity matrix. LDPC codes can be defined either as systematic or non-systematic with latter being used much more often in practice. The code rate of an LDPC code can easily be changed by adjusting the size of $\mathbf{G}$ as well as by applying puncturing.

In case of linear codes, if $\mathbf{e}$ is a valid codeword itself, its addition to the submitted message would result in a valid codeword, so the decoder would not notice the problem. However, due to the sparse filling of the generator matrix with ones, LDPC codes have a very large code distance making the latter case extremely unlikely to appear.

Due to the large block size, direct decoding of the received codeword by inspecting the syndrome vector is not practical. Instead, just like in case of turbo codes, the decoding process is carried out in an iterative manner. Bipartite graphs known as Tanner graphs are used to represent the parity check matrix as shown in Figure 27 for a simple Hamming code defined by the parity check matrix

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & \vdots & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & \vdots & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & \vdots & 0 & 1 & 1 & 1 \end{pmatrix}. \tag{24}$$

For a $m \times n$ generator matrix, Tanner graphs consists of two disjoint set of nodes: $C = \{c_0, \ldots, c_{n-1}\}$ representing the bits of the encoded message and $P = \{p_0, \ldots, c_{m-1}\}$ representing the parity equations. A node $c_x \in C$ is connected to a node $p_y \in P$ if xth bit is involved in the yth parity check. An LDPC code is called regular, if the weight w_r of every row as well as the weight w_c of every column of its parity check matrix is the same, i.e., the sum of ones in every row is constant and the sum of ones in every column is constant (though w_r and w_c do not have to be equal). Such a code can be characterized by a triple (n, w_c, w_r).

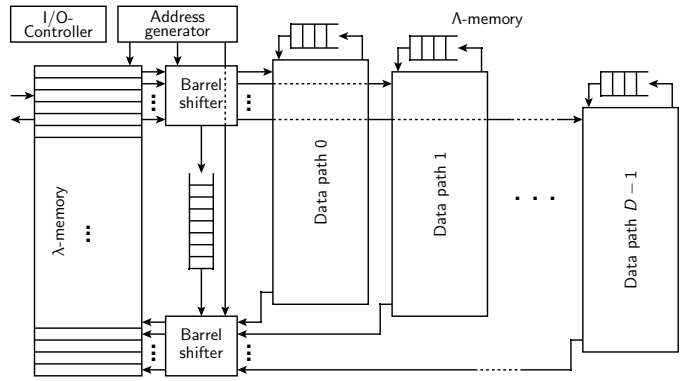


Figure 28. Top view architecture of the LDPC decoder [2].

The simplest way to decode the LDPC codes is to apply Gallager's hard decision algorithm. In the first iteration, parities are checked to determine, which bits of the information block are responsible for most unsatisfied parity equations. These bits are inverted and the next iteration starts. The decoding ends, as soon as all equations are satisfied or a certain number of iterations is reached. For a very simple regular (16,3,4) LDPC code with the parity check matrix

$$\mathbf{H} = \begin{pmatrix} 1\,1\,1\,1 & \vdots & 0\,0\,0\,0 & \vdots & 0\,0\,0\,0 & \vdots & 0\,0\,0\,0 \\ 0\,0\,0\,0 & \vdots & 1\,1\,1\,1 & \vdots & 0\,0\,0\,0 & \vdots & 0\,0\,0\,0 \\ 0\,0\,0\,0 & \vdots & 0\,0\,0\,0 & \vdots & 1\,1\,1\,1 & \vdots & 0\,0\,0\,0 \\ 0\,0\,0\,0 & \vdots & 0\,0\,0\,0 & \vdots & 0\,0\,0\,0 & \vdots & 1\,1\,1\,1 \\ \hdashline 1\,0\,0\,0 & \vdots & 1\,0\,0\,0 & \vdots & 1\,0\,0\,0 & \vdots & 1\,0\,0\,0 \\ 0\,1\,0\,0 & \vdots & 0\,1\,0\,0 & \vdots & 0\,1\,0\,0 & \vdots & 0\,1\,0\,0 \\ 0\,0\,1\,0 & \vdots & 0\,0\,1\,0 & \vdots & 0\,0\,1\,0 & \vdots & 0\,0\,1\,0 \\ 0\,0\,0\,1 & \vdots & 0\,0\,0\,1 & \vdots & 0\,0\,0\,1 & \vdots & 0\,0\,0\,1 \\ \hdashline 1\,0\,0\,0 & \vdots & 0\,1\,0\,0 & \vdots & 0\,0\,1\,0 & \vdots & 0\,0\,0\,1 \\ 0\,1\,0\,0 & \vdots & 0\,0\,1\,0 & \vdots & 0\,0\,0\,1 & \vdots & 1\,0\,0\,0 \\ 0\,0\,1\,0 & \vdots & 0\,0\,0\,1 & \vdots & 1\,0\,0\,0 & \vdots & 0\,1\,0\,0 \\ 0\,0\,0\,1 & \vdots & 1\,0\,0\,0 & \vdots & 0\,1\,0\,0 & \vdots & 0\,0\,1\,0 \end{pmatrix} \tag{25}$$

this process looks as follows [20, Section 15.6].

Assume, the codeword

$$\mathbf{c}' = (1,1,0,0,1,0,1,0,0,0,0,0,0,0,0,0) \tag{26}$$

is received (with bits of $\mathbf{c}'$ numbered in ascending order from left to right and rows of $\mathbf{H}$ numbered in ascending order from above starting with 1). After calculating $\mathbf{c}' \times \mathbf{H}^\top = (0,0,0,0,0,1,1,0,1,0,0,1)$ it becomes clear, that the parity equations 6, 7, 9, and 12 are violated while the bits 6, 10, 11, and 15 are involved in the most (respectively two) violated equations. Flipping these bits results in

$$\mathbf{c}'' = (1,1,0,0,1,1,1,0,0,1,1,0,0,0,1,0). \tag{27}$$

After the product $\mathbf{c}'' \times \mathbf{H}^\top = (0,1,0,1,0,1,1,0,1,0,0,1)$ is calculated once again, the parity equations 2, 4, 6, 7, 9, and 12 appear violated, with bits 6 and 15 involved in the most

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE X. COMBINATION OPTIONS OF DIFFERENT DECODERS (BASED ON DATA FROM [2]).

Decoder	REED-SOLOMON	VITERBI	Turbo	LDPC
REED-SOLOMON	=	−	−	−
VITERBI	−	=	+	−
Turbo	−	+	=	+
LDPC	−	−	+	=

equations (three each). Finally, flipping these bits results in the codeword

$$\mathbf{c} = (1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0), \qquad (28)$$

which, according to the condition $\mathbf{c} \times \mathbf{H}^\top = \mathbf{0}$, was the original message [20, p. 600].

This process can be visualized using the Tanner graph of the code: Every parity node passes the modulo-2 sum (XOR) of the bits involved in the corresponding parity equation to every bit node within its network (the bit receiving the sum is not included in its calculation). The bit nodes with the largest amount of received ones are those violating the most parity equations and must be flipped [43, Subsection 8.2.4.2]. Better error correction performance is achieved by using the soft-decision approach, which – instead of flipping bits – propagates probabilities through the Tanner graph trying to maximize the a-posteriori probability for every bit to satisfy all parity equations [9, Section 15.5]. The bit probabilities based on the soft decision are assigned to the corresponding bit nodes and sent to the parity nodes, which use them to calculate the probabilities that the corresponding parity equations are satisfied. These probabilities are then sent back to the bit nodes according to their involvement in the respective parity equations and used to refine the initial assumptions. This procedure called *belief propagation* is repeated several times and can be seen as a kind of message passing algorithm [43, Subsection 8.2.4.4]. At the end, either the codeword is decoded or the decoding process stops after a fixed number of iterations without producing a solution. As in the case of turbo codes, logarithmic metrics like LLR can be used to represent the likelihood values, which are converted to the hard bit decisions at the end.

Figure 28 shows the global architecture of the LDPC decoder with soft-decision memories referred as λ, and parity node message memories referred as Λ. The number of data paths can be scaled according to the throughput requirements, while the barrel shifter take care of providing the correct values for the computation.

C. Combining Different Decoders

The conclusion from the discussion of the different decoding algorithms in the previous sections is that advantages of multi-family FEC decoder implementation result from memory and data-path reuse. Table X summarizes the potential of different combinations. As already mentioned, RS decoders have quite different arithmetic requirements, which explains, why the first row and the first column of the table are filled with "−" limiting the options to Viterbi, turbo, and LDPC decoders.

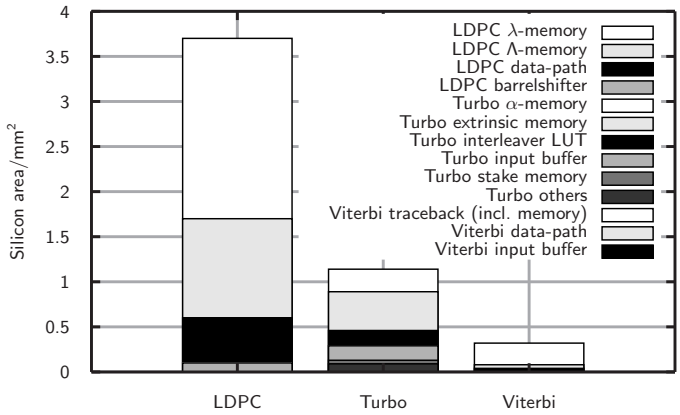


Figure 29. Chip area distribution for different FEC decoder designs (adapted from [2]).

According to the brief description of the iterative decoding, it should be clear that a lot of local memory is required to store the information bits, soft decisions, and intermediate results. Figure 29 shows the silicon area distribution between different parts of the design for LDPC, turbo, and Viterbi decoders. The extreme disproportion in the required silicon area between LDPC and Viterbi decoders suggests, that a combination of both within the same design with the goal to reuse some of the resources does not make sense. Even if the complete Viterbi decoder area could be reused for LDPC (which is virtually impossible), the overall gain would sum up to less than 10 % of the silicon footprint [2].

The SISO module required for turbo decoding usually implements the Bahl-Cocke-Jelinek-Raviv (BCJR) algorithm, which ensures the minimal (a-posteriori) symbol error probability [46]. BCJR is often used as a synonym for Maximum-a-Posteriori (MAP) approach. Depending on the probability metrics, modified versions of the same algorithm are known as Log-MAP and Max-Log-MAP (ML-MAP) [47]. An alternative SISO implementation is based on the Soft-Output Viterbi-Algorithm (SOVA), which can also be used to decode the convolutional codes. In contrast to the hard decision Viterbi decoder, SOVA uses a different path metric, which is based on the a-priory probabilities of the input symbols and produces a reliability metric instead of the hard bit decision as an output [9, Subsection 14.3.17]. The reliability could, for instance, be based on the difference between the branch metrics of the kept and discarded branches accumulated over the length of the TBD. SOVA minimizes the cumulative error for a sequence of symbols, which – unlike the BCJR approach – does not necessary result in a minimal symbol error probability.

The difference between two algorithms is that BCJR calculates both forward and backward iterations for every state transition (following the path to the actual and succeeding state from different directions), while Viterbi follows only one path [48]. The BCJR algorithm assigns three values to every state transition. For the transition from x to y (assuming x is reached at time t), the probability of the sequence up to time t ending at state x is first calculated by the forward iteration

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

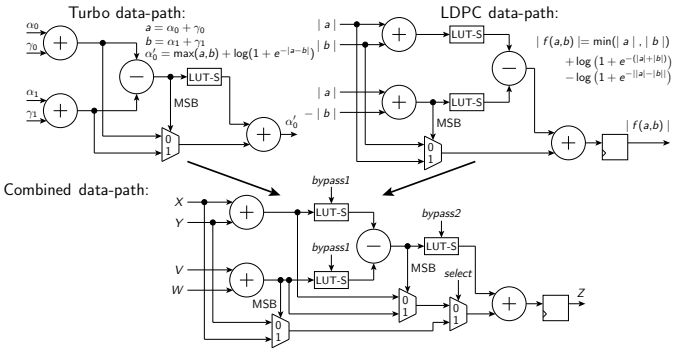


Figure 30. Combining Turbo and LDPC data-paths within a single FFU (adapted from [50], LUT-S are Look-Up-Tables storing the approximated logarithmic values).

of the BCJR algorithm and is commonly designated as $\alpha_t(x)$. Next, the probability of the future sequence from time $t+1$ to the end, given the state y at time $t+1$, is calculated by the backward iteration of the BCJR algorithm and is usually designated as $\beta_{t+1}(y)$. Finally, the branch metric of the state transition between x and y is calculated as $\gamma(x, y)$. Since the values of the previous iteration are required to compute the next one, they need to be stored. The corresponding memories are called α, β (or stake), and γ (or extrinsic), which explains the chart legend related to the turbo decoder in Figure 29.

Using SOVA instead of BCJR increases the potential for hardware reuse in a combined Viterbi-turbo decoder at the cost of slightly reduced bit error ratio for the turbo decoder compared to the SISO module based on the Max-log-MAP algorithm. In the case of an ML-MAP BCJR implementation, at least the adders of the Viterbi's ACS units can be reused in the LLR metric calculation. The same hardware can be reused for the forward and backward computation in a time-multiplexed manner as long as the throughput requirements are met [49].

A closer look at the internal data-path of the turbo and LDPC decoders reveals some similarities as well. Multi-family decoder can be built either reusing some of the LDPC data-path to decode turbo codes [51] or vice versa [52][53]. Figure 30 illustrates how the computational steps for both decoding algorithms can be unified within the same Flexible Functional Unit (FFU) [50]. As in the case of multi-standard Viterbi decoder, some additional multiplexers are required to switch between the codes, which slightly increases the critical path. However, significant reuse options are only valid for high-throughput turbo decoders, which require several parallel SISO units. Otherwise, the area percentage of the data-path compared to the memory is almost negligible, as shown in Figure 29. The memory reuse potential of about 10 % can be achieved if the access bandwidths of the turbo and LDPC parts are comparable [2].

The conclusion from the previous discussion leaves Viterbi-turbo and LDPC-turbo designs as the only options for a multi-family decoder implementation. Some case studies of such designs are known from literature [2][17][49][50][54]. Table XI summarizes the results. As expected, some portions of the data-path and memory can be reused resulting in silicon area savings of 10–42 %, while throughput is not affected in most cases of the combined Viterbi-turbo decoder. If adaptable designs are compared with a single family decoder, the overhead of introducing an additional code family is very low. For instance, adding the LDPC functionality to the existing turbo design comes at only 10–20 % additional silicon area [50].

TABLE XI. COMPARISON OF MULTI-FAMILY FEC DECODERS.

Architecture	Metric	Reference	Results
Viterbi-Turbo Decoders			
VITURBO [17]	silicon area	Viterbi decoder ($K = 3 \ldots 9$)	+5 % logic +25 % memory
[49] (fully parallel)	silicon area	separate turbo and Viterbi decoder	$-14,5\,\%$
	throughput	separate turbo and Viterbi decoder	no change
[49] (time-multiplex)	silicon area	separate turbo and Viterbi decoder	$-28,1\,\%$
	throughput	separate turbo and Viterbi decoder	same for Viterbi 1/2 for turbo
[54]	silicon area	turbo decoder	$+20\,\%$
LDPC-Turbo Decoders			
[2]	silicon area	separate turbo and LDPC decoder	$-10\,\%$
[50]	silicon area	turbo decoder	$+10 \ldots +20\,\%$
		LDPC decoder	$+15 \ldots +20\,\%$
		reused vs. separate data-paths	$-39 \ldots -42\,\%$

It should be noted that reusing the same hardware for different decoders limits the possibilities for parallel decoding of differently encoded data streams. Depending on the throughput requirements, either a time-multiplexed (temporal scheduling) or a multi-data-path (spatial scheduling) solution can be applied to address this problem. Since the application scenarios are known in advance, these constraints can be taken into account before the decision on the final decoder architecture is made.

V. COMPARATIVE ANALYSIS OF ADAPTIVE DECODER ARCHITECTURES

The findings from the last three sections provide enough data for a generalized analysis of all presented architectures. Although it is clear that adaptive decoder implementation comes at an additional cost, the figures vary depending on the code family, the code parameters involved, and the degree of flexibility.

Viterbi decoders can be enabled to decode a few additional codes with almost negligible overhead, as long as the constraint length is kept constant. Otherwise, the increase in the silicon footprint and the drop in throughput can be quite high. Generally speaking, adding several generator polynomials to an existing architecture or changing the code rate has much less impact on the area and performance figures than increasing the constraint length. From an implementation perspective, major changes in the overall decoder architecture are limited to making the data path configurable by adding multiplexers and inserting adaptable address generators instead of simple counters into the trace-back stage, while the memory capacity of the trace-back unit must be dimensioned for the worst case.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Throughput is more severely affected compared to the other discussed decoder families, especially if all code parameters are variable.

Changing the generator polynomials of the code and the underlying finite field for Reed-Solomon decoders has a much greater impact than in the Viterbi case, as it introduces major changes to the GF-multipliers used for syndrome and error location computation. However, changing the degree of the GF generator polynomial (which equals the symbol bit-width) is even more critical. In general, the modifications required to enable the basic RS decoder to process different codes are much more profound. For larger block sizes, the impact on the chip area of the overall decoder is limited, as the FIFO memory becomes the dominating factor. Reduced clock frequency is usually not an issue, since the decoding chain relies on symbols composed of multiple (usually 8) bits. For very small values of m, throughput plays a more important role. However, such RS codes are rarely used in real-world applications.

For multi-family decoders, selecting the proper code combinations is the key factor. Because the goal of such combined implementations is to reuse parts of the data-path and memory, adaptable Viterbi-turbo and LDPC-turbo architectures are the only meaningful alternatives. Changes to the data-path are comparable to the Viterbi decoder case, provided the same decoding algorithms and likelihood metrics are used (which may slightly degrade the BER figures). Otherwise, the potential for sharing silicon area across different decoders is limited to memory, which does not deliver a significant gain. Throughput is only slightly affected – or even completely unaffected – when multi-family decoders are compared to their single-family counterparts. However, the throughput requirements of the decoders should be closely matched to maximize the benefits of a combined implementation. In general, finding the optimal settings for a feasible multi-family decoder architecture is far more difficult than tuning a single-family decoder to process codes with variable parameters.

VI. Conclusion and Future Work

This paper discusses several options for the adaptive FEC decoder design. The findings based on literature study and own research suggest that the overhead of extending a decoder towards other codes of the same family comes at a moderate cost. As expected, this overhead (compared to a decoder implementation for a fixed parameter set) increases with the amount of flexibility and code complexity, but can be justified by the fact that separate designs per code instance are no longer necessary in most cases.

At the same time, decoders covering several code families can take advantage of resource sharing, reducing the overall silicon footprint compared to a dedicated implementation of one particular decoder per family. The potential advantages, however, strongly depend on the use cases and the envisioned set of standards. In particular, combining iterative decoders with significantly different throughput requirements will not result in noticeable area savings.

One aspect not discussed herein is the impact of the adaptive decoder implementation on power consumption. On one hand, reconfigurability comes at the cost of additional area, which can lead to higher energy-per-bit figures. On the other hand, reusing parts of the data-path and memory for different decoding algorithms is expected to yield overall energy savings. Supporting these assumptions with concrete metrics is a promising direction for future research.

References

[1] S. Sawitzki, "Adaptive Architectures for Forward Error Correction", in *CENICS 2025: The Eighteenth International Conference on Advances in Circuits, Electronics and Micro-Electronics*, T. Bostelmann, Ed., Barcelona, Spain: International Academy, Research, and Industry Association, Oct. 2025, pp. 1–7.

[2] J. T. M. H. Dielissen, N. Engin, S. Sawitzki, and C. H. van Berkel, "FEC Decoders for Future Wireless Devices: Scalability Issues and Multi-Standard Capabilities", in *Circuits and Systems for Future Generations of Wireless Communications*, A. Tasić, W. A. Serdijn, L. E. Larson, and G. Setti, Eds., ser. Series on Integrated Circuits and Systems, Berlin: Springer, 2009, pp. 271–297.

[3] T. Vogt and N. Wehn, "A Reconfigurable ASIP for Convolutional and Turbo Decoding in an SDR Environment", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, no. 10, pp. 1309–1320, Oct. 2008.

[4] P. M. Velayuthan, "Towards Optimized Flexible Multi-ASIP Architectures for LDPC/Turbo Decoding", PhD thesis, Université de Bretagne-Sud, Dec. 2012.

[5] M. Scarpellino, A. Singh, E. Boutillon, and G. Masera, "Reconfigurable Architecture for LDPC and Turbo Decoding: A NoC Case Study", in *Proceedings of IEEE 10th International Symposium on Spread Spectrum Techniques and Applications*, Bologna, Italy: IEEE, Aug. 2008, pp. 671–676.

[6] J. T. M. H. Dielissen, N. Engin, S. Sawitzki, and C. H. van Berkel, "Multi-Standard FEC Decoders for Wireless Devices", *IEEE Transactions on Circuits and Systems II*, vol. 55, no. 3, pp. 284–288, May 2008.

[7] S. Sawitzki, "Embedded Reconfigurable Computing: Architekturen, Anwendungen und Entwurfswerkzeuge (Architectures, Applications, Tools)", Habilitation Thesis, Dresden University of Technology, 2024.

[8] P. Elias, "Coding for Noisy Channels", *IRE Convention Record*, vol. 3, no. 4, pp. 37–46, 1955.

[9] T. K. Moon, *Error Correction Coding. Mathematical Methods and Algorithms*. John Wiley & Sons, 2005.

[10] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[11] I. Habib, Ö. Paker, and S. Sawitzki, "Design Space Exploration of Hard-Decision Viterbi Decoding: Algorithm and VLSI Implementation", *IEEE Transactions on VLSI Systems*, vol. 18, no. 5, pp. 794–807, May 2010.

[12] A. P. Hekstra, "An Alternative to Metric Rescaling in Viterbi Decoders", *IEEE Transactions on Communications*, vol. 37, no. 11, pp. 1220–1222, Nov. 1989.

[13] J. Dragaš, "Design Trade-offs in the VLSI Implementation of High-Speed Viterbi Decoders and their Application to MLSE in ISI Cancellation", Master's Thesis, Integrated Systems Laboratory, Swiss Federal Institute of Technology Zurich, Mar. 2011.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[14] P. J. Black and T. H. Meng, "Hybrid Survivor Path Architectures for Viterbi Decoders", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, MN, USA, Apr. 1993, pp. 433–436.

[15] W. Tang, N. Engin, F. A. Steenhof, M. Klaassen, A. Hekstra, and S. Sawitzki, *Multi-Standard Viterbi Processor*, US Patent 8,904,266 B2, NXP B. V., Eindhoven, The Netherlands, Dec. 2014.

[16] K. Chadha and J. R. Cavallaro, "A Reconfigurable Viterbi Decoder Architecture", in *Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems & Computers*, M. B. Matthews, Ed., vol. 1, Pacific Grove, CA, USA: IEEE, Nov. 2001, pp. 66–71.

[17] J. R. Cavallaro and M. Vaya, "VITURBO: A Reconfigurable Architecture for Viterbi and Turbo Decoding", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong: IEEE, Apr. 2003, pp. 497–500.

[18] T. Vogt, N. Wehn, and P. Alves, "A Multi-Standard Channel-Decoder for Base-Station Applications", in *17th Symposium on Integrated Circuits and Systems Design (SBCCI 2004)*, Porto de Galinhas, Brazil: IEEE, Sep. 2004, pp. 192–197.

[19] K. A. Bush, "Orthogonal Arrays of Index Unity", *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 426–434, Sep. 1952.

[20] W. C. Huffman and V. Pless, *Fundamentals of Error-Correcting Codes*. Cambridge, NY, USA: Cambridge University Press, 2003.

[21] I. S. Reed and G. Solomon, "Polynomial Codes over Certain Finite Fields", *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, Jun. 1960.

[22] H.-J. Kang and I.-C. Park, "A High-Speed and Low-Latency Reed-Solomon Decoder based on a Dual-Line Structure", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, May 2002, pp. 3180–3183.

[23] A. Kumar, "High-Throughput Reed-Solomon Decoder for Ultra Wide Band", Master's Thesis, Technische Universiteit Eindhoven and National University of Singapore, Dec. 2004.

[24] H. Lee, "High-Speed VLSI Architecture for Parallel Reed-Solomon Decoder", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 2, pp. 288–294, Apr. 2003.

[25] W. Peterson, "Encoding and Error-Correction Procedures for the Bose-Chaudhuri Codes", *IRE Transactions on Information Theory*, vol. 6, no. 4, pp. 459–470, Sep. 1960.

[26] D. E. Gorenstein and N. Zierler, "A Class of Error Correcting Codes in p^m Symbols", *Journal of the Society of Industrial and Applied Mathematics*, vol. 9, no. 2, pp. 207–214, Jun. 1961.

[27] R. T. Chien, "Cyclic Decoding Procedures for the Bose-Chaudhuri-Hocquenghem Codes", *IEEE Transactions on Information Theory*, vol. 10, no. 4, pp. 357–363, Oct. 1964.

[28] G. D. Forney, "On Decoding BCH Codes", *IEEE Transactions on Information Theory*, vol. 11, no. 4, pp. 549–557, Oct. 1965.

[29] E. R. Berlekamp, *Algebraic Coding Theory*, revised. Aegean Park Press, Jun. 1984.

[30] R. D. Kudryavtsev and M. K. Samarin, "Lagrange Interpolation Formula", in *Encyclopedia of Mathematics*, M. Hazewinkel, Ed., London, UK: Springer, 2001.

[31] C. K. P. Clarke, "Reed-Solomon Error Correction", British Broadcasting Corporation, White paper WPH 031, Jul. 2002.

[32] A. Vourdas, "Harmonic Analysis on a Galois Field and Its Subfields", *Journal of Fourier Analysis and Applications*, vol. 14, no. 1, pp. 102–123, Feb. 2008.

[33] P. L. Montgomery, "Modular Multiplication without Trial Division", *Mathematics of Computation*, vol. 44, no. 170, pp. 519–521, Apr. 1985.

[34] C.-C. Lin, F.-K. Chang, H.-C. Chang, and C.-Y. Lee, "An Universal VLSI Architecture for Bit-Parallel Computation in $\mathrm{GF}(2^m)$", in *The 2004 IEEE Asia-Pacific Conference on Circuits and Systems*, Piscataway, NJ, USA: IEEE, Dec. 2004, pp. 125–128.

[35] H.-C. Chang, C. B. Shung, and C.-Y. Lee, "A Reed-Solomon Product-Code (RS-PC) Decoder Chip for DVD Applications", *IEEE Journal of Solid-State Circuits*, vol. 36, no. 8, pp. 229–238, Feb. 2001.

[36] A. Kumar and S. Sawitzki, "High Throughput and Low Power Reed Solomon Decoder for Ultra Wide Band", in *Intelligent Algorithms in Ambient and Biomedical Computing*, W. Verhaegh, E. Aarts, and J. Korst, Eds., ser. Philips Research Book Series, vol. 7, Dordrecht, The Netherlands: Springer, 2006, pp. 299–316.

[37] A. G. M. Strollo, N. Petra, D. De Caro, and E. Napoli, "An Area-Efficient High-Speed Reed-Solomon Decoder in $0.25\,\mu$m CMOS", in *Proceedings of the IEEE 30th European Solid State Circuits Conference*, Leuven, Belgium: IEEE, Sep. 2004, pp. 479–482.

[38] H.-Y. Hsu and A.-Y. Wu, "VLSI Design of a Reconfigurable Multi-Mode Reed-Solomon Codec for High-Speed Communication Systems", in *IEEE Asia-Pacific Conference on ASIC Proceedings*, Taipei, Taiwan: IEEE, Aug. 2002, pp. 359–362.

[39] J.-C. Huang, C.-M. Wu, M.-D. Shieh, and C.-H. Wu, "An Area-Efficient Versatile Reed-Solomon Decoder for ADSL", in *Proceedings of the 1999 IEEE International Symposium on Circuit and Systems (ISCAS'99)*, vol. 1, Orlando, FL, USA: IEEE, May 1999, pp. 517–520.

[40] F.-K. Chang, C.-C. Lin, H.-C. Chang, and C.-Y. Lee, "Universal Architectures for Reed-Solomon Error-and-Erasure Decoder", in *Asian Solid-State Circuits Conference*, Hsinchu, Taiwan: IEEE, Nov. 2005, pp. 229–232.

[41] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo Codes", in *Proceedings of the IEEE International Conference on Communication*, Geneva, Switzerland: IEEE, May 1993, pp. 1064–1070.

[42] J. T. M. H. Dielissen, "VLSI Design and Implementation of a Turbo Decoder", Master's Thesis, Eindhoven University of Technology, May 2000.

[43] B. Sklar and F. J. Harris, *Digital Communications: Fundamentals and Applications*, third edition. Hoboken, NJ, USA: Pearson Education, Inc., 2021.

[44] J. Hagenauer, E. Offer, and L. Papke, "Iterative Decoding of Binary Block and Convolutional Codes", *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 429–445, Mar. 1996.

[45] R. G. Gallager, "Low-Density Parity-Check Codes", Sc. D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1963.

[46] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate", *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 284–287, Mar. 1974.

[47] P. Robertson, E. Villebrun, and P. Hoeher, "A Comparison of Optimal and Sub-Optimal MAP Decoding Algorithms Operating in the Log Domain", in *IEEE International Conference on Communications*, Seattle, WA, USA: IEEE, Jun. 1995, pp. 1009–1013.

[48] A. J. Viterbi, "An Intuitive Justification and a Simplified Implementation of the MAP Decoder for Convolutional Codes", *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 2, pp. 260–264, Feb. 1998.

[49] G. Krishnaiah, N. Engin, and S. Sawitzki, "Scalable Reconfigurable Channel Decoder Architecture for Future Wireless Handsets", in *Design, Automation and Test in Europe Conference and Exhibition, DATE 2007*, Nice, France: European Design and Automation Association, Apr. 2007, pp. 1563–1568.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[50] Y. Sun and J. R. Cavallaro, "A Flexible LDPC/Turbo Decoder Architecture", *Journal of Signal Processing Systems*, vol. 64, no. 1, pp. 1–16, Jul. 2011.

[51] D. J. C. MacKay, "Good Error-Correcting Codes Based on Very Sparse Matrices", *IEEE Transactions on Information Theory*, vol. 45, no. 2, pp. 399–431, Mar. 1999.

[52] M. M. Mansour and N. R. Shanbhag, "High-Throughput LDPC Decoders", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 6, pp. 976–996, Dec. 2003.

[53] J. Lu and J. M. F. Moura, "Turbo-like Decoding of LDPC Codes", in *IEEE International Magnetics Conference (Intermag 2003)*, Boston, MA, USA: IEEE, Mar. 2003, DT–11.

[54] I. Ahmed and T. Arslan, "A Reconfigurable Viterbi Decoder for a Communication Platform", in *Proceedings of the 2005 International Conference on Field Programmable Logic and Applications (FPL)*, T. Rissa, S. Wilton, and P. Leong, Eds., Tampere, Finland: IEEE, Aug. 2005, pp. 435–440.

2026, © Copyright by authors, Published under agreement with IARIA - www.iaria.org