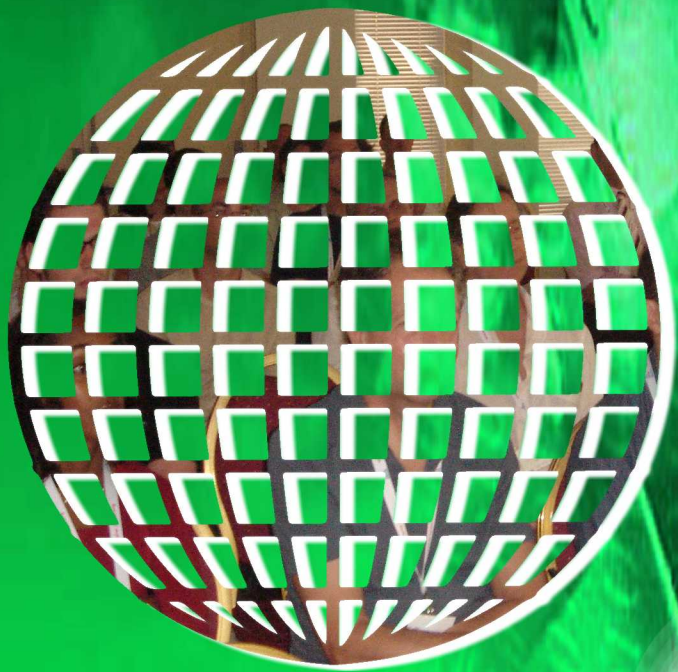


International Journal on Advances in Life Sciences



The *International Journal on Advances in Life Sciences* is published by IARIA.

ISSN: 1942-2660

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 Life Sciences, issn 1942-2660
vol. 18, no. 1 & 2, year 2026, http://www.iariajournals.org/life_sciences/

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 Life Sciences, issn 1942-2660
vol. 18, no. 1 & 2, year 2026, <start page>:<end page> , http://www.iariajournals.org/life_sciences/

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

Editor-in-Chief

Les Sztandera, Thomas Jefferson University, USA

Editorial Board

Ganesharam Balagopal, Ontario Ministry of the Environment Conservation and Parks | Technical Assessment and Standards Development Branch, Canada

Kazi S. Bennoor, National Institute of Diseases of Chest & Hospital - Mohakhali, Bangladesh

Razvan Bocu, Transilvania University of Brasov, Romania

Karin Brodén, Karlstad University, Sweden

Ozgu Can, Ege University, Turkiye

Young (Yang) Cao, Virginia Tech, USA

Jitender Deogun, University of Nebraska-Lincoln, USA

Duarte Duque, ALGORITMI Research Centre | LASI, University of Minho / 2Ai - School of Technology | IPCA, Portugal

Hassan Ghazal, Mohammed VI University of Health Sciences, Morocco

Piero Giacomelli, Fidia Farmaceutici SpA, Vicenza, Italy

Malina Jordanova, Space Research & Technology Institute | Bulgarian Academy of Sciences, Bulgaria

Hassan M. Khachfe, Lebanese International University, Lebanon

Peter Kieseberg, St. Pölten University of Applied Sciences, Austria

Evgeniy Krastev, Sofia University St. Kliment Ohridsk, Bulgaria

Ljerka Luic, University North, Croatia

Jose Manuel Molina Lopez, Universidad Carlos III de Madrid, Spain

Stefano Mariani, Politecnico di Milano, Italy

Julio César Mello Román, National University of Asuncion (UNA), Paraguay

Helena Pereira de Melo, NOVA School of Law, Portugal

Vitor Pinheiro de Almeida, Pontifícia Universidade do Rio de Janeiro (PUC-Rio), Brazil

Tamara Powell, Kennesaw State University, USA

Addisson Salazar, Universitat Politècnica de València, Spain

Les Sztandera, Thomas Jefferson University, USA

Paulo Teixeira, Polytechnic University of Cávado and Ave, Portugal

Genny Villa, Université de Montréal, Canada

Vivian Vimarlund, Linköping University, Sweden

CONTENTS

pages: 1 - 10

Analytical Method for Evaluating Gait Changes Using a Smart Insole: Robot-Induced Effects in a Hemiparetic Patient

Tomoko Funayama, Teikyo University of Science, Japan
Yasutaka Uchida, Teikyo University of Science, Japan
Eiichi Ohkubo, Teikyo University of Science, Japan
Yasunori Fujimori, Seirei Yokohama Hospital, Japan

pages: 11 - 22

Assessment of Postural Changes Before and After Hemodialysis Based on Body Weight Measurement Motion Using MediaPipe and Machine Learning

Yasutaka Uchida, Teikyo University of Science, Japan
Kazuyoshi Hori, Teikyo University of Science, Japan
Ren Kondo, Katori Omigawa Medical Center, Japan
Satoshi Kidooka, Katori Omigawa Medical Center, Japan
Sayuri Sugaya, Katori Omigawa Medical Center, Japan
Tomoya Iizuka, Katori Omigawa Medical Center, Japan

pages: 23 - 34

Multi-Model Hierarchical Semantic Retrieval for Predicting Abbreviated Injury Scale 2015 from Clinical Narratives

Chien-Ming Lee, Department of Medical Information, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
Pei-Ling Lee, Division of Trauma and Surgical Critical Care, Department of Surgery, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
Chao-Wen Chen, Division of Trauma and Surgical Critical Care, Department of Surgery, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
Joffrey Hsu, Division of Trauma and Surgical Critical Care, Department of Surgery, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan.
Chia-Yeuan Han, Department of Medical Information, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan.

pages: 35 - 50

From Vague Student Complaints to Auditable Triage: Graph-Conditioned LLM Dialogue for School Health

Hayato Tomisu, Ritsumeikan University / Shiga University, Japan
Kazue Yamamura, Ritsumeikan University / Ritsumeikan Moriyama Junior and Senior High School, Japan
Junya Ueda, ImpactLab., Japan
Tsukasa Yamanaka, Ritsumeikan University, Japan

pages: 51 - 63

Distributed Custody for Electronic Health Records Preservation: A Proposal

Mehluli Masuku, Sorbonne University Abu Dhabi, United Arab Emirates

pages: 64 - 80

Clarifying Research and Education Needed to Explain the Medical Record to Patients: Protocol for UK Survey of Non-medical Practitioners

Ray B. Jones, University of Plymouth, United Kingdom
John Downey, University of Plymouth, United Kingdom
Tristan Coombe, University of Plymouth, United Kingdom
Avril Collinson, University of Plymouth, United Kingdom
Aled Jones, University of Plymouth, United Kingdom
Sally Abey, University of Plymouth, United Kingdom
Patricia Schofield, University of Plymouth, United Kingdom
Jenny Freeman, University of Plymouth, United Kingdom
Nicholas Peres, University of Plymouth, United Kingdom
Jill Shawe, University of Plymouth, United Kingdom
Sheena Asthana, University of Plymouth, United Kingdom

pages: 81 - 91

Toward Robust Machine-Learning-Based Discrimination of a 1 Hz Difference in the Auditory Cortex Using fMRI Activation Patterns

Yoshitaka Ooyashiki, Kochi University of Technology, Japan
Kyoko Shibata, Kochi University of Technology, Japan

Analytical Method for Evaluating Gait Changes Using a Smart Insole: Robot-Induced Effects in a Hemiparetic Patient

Tomoko Funayama
Dept. of Occupational Therapy
Teikyo University of Science.
Yamanashi, Japan
e-mail: funayama@ntu.ac.jp

Eiichi Ohkubo
Dept. of Life Science
Teikyo University of Science
Tokyo, Japan
e-mail: ohkubo@ntu.ac.jp

Yasutaka Uchida
Dept. of Life Science
Teikyo University of Science.
Tokyo, Japan
e-mail: uchida@ntu.ac.jp

Yasunori Fujimori
Dept. of Rehabilitation
Seirei Yokohama Hospital
Kanagawa, Japan
e-mail: yasunori.bb.22@gmail.com

Abstract—Wearable devices capable of measuring daily activities have become increasingly accessible, enabling objective assessment in both clinical environments and everyday living settings. Smart insoles integrated into footwear provide continuous and unobtrusive monitoring of plantar loading patterns and are expected to contribute to rehabilitation assessment, including applications involving robotic assistive devices. However, practical challenges remain, such as device design limitations, non-standardized measurement protocols, and the lack of widely accepted evaluation methods. Moreover, gait analysis techniques developed for healthy individuals are often unsuitable for individuals with gait impairments, who typically exhibit atypical or highly variable movement patterns. These issues highlight the need for analytical approaches tailored to impaired gait. This study proposed a novel gait analysis method using a smart insole with embedded pressure sensors in a patient with stroke undergoing inpatient rehabilitation. Smart insole data were collected before, during, and after the use of a gait-assistive robot equipped with a knee–ankle–foot orthosis. Gait data were segmented and analyzed using mean values, temporal parameters, peak-related measures, and insole-pressure distribution, including comparisons of peak counts and visualization of time-series ratios for each insole part. These analyses indicated gait changes associated with robot-assisted walking. This case study demonstrates not only the evaluation of gait changes in a patient with hemiparesis undergoing robot-assisted rehabilitation but also the applicability of a smart insole-based analytical method. To support practical implementation, we present outcome indicators that are easily interpretable for rehabilitation staff.

Keywords—gait analysis method; smart insoles; gait-assistive robot; wearable devices; rehabilitation.

I. INTRODUCTION

In recent years, the widespread availability of wearable devices capable of measuring daily activities has expanded their use in health-related fields. Wearable devices with high

usability expand the scope of assessments that can be conducted not only in specialized environments, such as hospitals and rehabilitation facilities, but also in everyday living settings without altering the usual lifestyle conditions of individuals. Such long-term monitoring capabilities facilitate the detection of subtle changes in health status that may otherwise be overlooked. This approach enables objective assessments not only in therapist-led clinical rehabilitation but also within the patient's home. We have previously conducted research on activity assessment using wearable devices [1, 2].

Among various activities of daily living, gait assessment is one of the most widely used measures because of its strong association with overall functional capacity, risk of falls, and quality of life [3, 4]. It has been applied across a diverse range of patient populations, including older adults [5, 6], individuals with central nervous system disorders (such as cerebrovascular disease, Parkinson disease, and multiple sclerosis) [7–9], patients with cardiovascular and respiratory diseases [10, 11], those with orthopedic conditions such as low back pain [12], individuals with cognitive impairments [13], and various other health conditions. These studies collectively highlight the relevance of gait as a multidimensional clinical indicator capable of reflecting the musculoskeletal, neurological, and cognitive status of individuals. The foot plays a critical role in gait and postural control. It is one of the most sensory-rich regions of the human body and contains numerous receptors that provide continuous feedback essential for maintaining balance and stability. As the foot is the only part of the body that makes direct contact with the ground during an upright stance, it serves as an essential interface for sensing external forces and adjusting posture accordingly. Consequently, alterations in foot loading patterns can offer valuable insights into gait impairment and underlying gait-related conditions.

Insoles are integrated with footwear, and smart insoles equipped with pressure sensors leverage this advantage by enabling continuous, unobtrusive monitoring of plantar

loading patterns throughout daily walking episodes. They are also expected to be used for assessment in rehabilitation settings, including robotic assistive devices [14]. Although these devices have substantial potential and an increasing number of studies have reported their utility [15–17], several challenges remain for practical implementation, including issues related to device design, standardization of measurement protocols, and the establishment of universally accepted evaluation methods [18]. Applying gait analysis methods developed for healthy individuals to populations with severe gait impairments may be inappropriate. Individuals with gait impairments may exhibit atypical movement strategies, altered temporal and spatial parameters, or highly variable gait characteristics. Such deviations necessitate analytical frameworks that are specifically tailored to the movement patterns observed in impaired gait rather than relying solely on healthy reference models.

In this context, this study aimed to propose a novel gait analysis method using a smart insole with embedded pressure sensors for a single patient undergoing inpatient rehabilitation following a stroke. The patient participated in rehabilitation involving a gait-assistive robot, and smart-insole data were collected before, during, and after robot use. By analyzing these datasets, we developed an analytical approach designed to detect changes in gait. The robot used in this study was a rehabilitation device attached to a knee–ankle–foot orthosis (KAFO) that assists lower-limb movement by guiding the limb toward a desirable motion pattern during walking.

This case study aimed to demonstrate the feasibility and utility of gait assessment using a smart insole in individuals undergoing rehabilitation and those in need of gait improvement. For practical implementation, the overall picture should be captured. Possessing detailed data alone is insufficient; instead, clear, easily interpretable indicators are required to understand the overall gait characteristics. Therefore, in this study, we presented an analytical method suitable for individuals with gait disorders and proposed outcome indicators that can be easily understood and used by rehabilitation staff.

This study was approved by the Ethics Committee on Research with Humans as Subjects of the Teikyo University of Science. The participant was an individual receiving medical rehabilitation services and was fully informed of the potential benefits and risks of participating in the study. Written informed consent was obtained by occupational therapists from patients undergoing rehabilitation who participated in the study. Section II describes the experimental method, Section III presents the results, Section IV provides a discussion, and Section V concludes the study and outlines future work.

II. EXPERIMENTAL METHODS

A single participant undergoing gait rehabilitation was assessed under three conditions: before, during, and after using the gait-assistive robot. The data were analyzed to develop and propose an analytical method for evaluating the

effects of rehabilitation procedures using a gait-assistive robot, that is, changes in gait.

A. Devices and Software

1) *Smart Insole*: A wireless pressure-sensing smart insole, FEELSOLE® (Toyoda Gosei Co., Ltd), was used in this study. The device was equipped with four pressure sensors positioned at the toe, heel, inside, and outside, enabling measurements at a total of eight points across both feet. A notable characteristic of this insole is its relatively flat structure with minimal slope (Figure 1). The device was calibrated prior to use in the following four steps: (1) no pressure applied without inserting the foot into the shoe; (2) standing with both feet; (3) standing on the left foot only; and (4) standing on the right foot only. The insole sampling frequency was 50 Hz. The insole was connected via Bluetooth to an iOS application, ORPHE TRACK® (ORPHE Inc.), which automatically uploaded the data to the cloud. The recorded data were subsequently downloaded in CSV format using the ORPHE ANALYTICS system for further analysis.



Figure 1. Appearance of the smart insoles.

2) *Gait-Assistive Robot*: In this study, we used the Orthobot® (FINGGAL LINK Inc.), a device used in gait rehabilitation. This device is attached to a KAFO and assists in guiding the lower limb toward a desirable movement pattern during walking. Figure 2 shows a healthy individual wearing the gait-assistive robot.

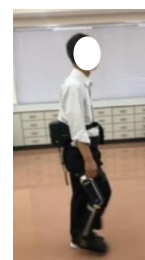


Figure 2. Gait-assistive robot.

B. Participant and Measurement Method

Gait measurements were conducted on a single male participant in his 70s who was hospitalized and undergoing rehabilitation following a stroke. With the cooperation of the hospital's physical and occupational therapists, gait was

assessed under three conditions: before using the gait-assistive robot, during its use, and after its removal. The participant had left hemiparesis and higher brain dysfunction due to a stroke, along with a medical history of traumatic brain injury and left femoral neck fracture. Motor impairments resulting from prior traumatic brain injury were suspected to affect not only the left upper and lower limbs but also the right limbs.

Rehabilitation, including physical, occupational, and speech-language therapies, began the day after stroke onset. Each therapy was provided five times per week, with each session lasting 40 min. Physical therapy included gait training. On day 16 after stroke onset, gait assessment was conducted using the smart insole and gait-assistive robot. As the participant was unable to walk independently, a physical therapist supported the participant from behind during the measurement. The gait-assistive robot was attached to the left lower limb of the participant. The gait tasks included straight walking and a U-turn.

C. Gait Characteristics and Analysis Method

For the analysis, we excluded the first and last 100 data points recorded at the beginning and end of the measurement period. The raw heel data are shown as the blue line in Figure 3. Our previous study on healthy adults investigated a method for gait assessment using pressure sensor-equipped insoles and suggested the utility of both peak values and post-peak decline rates at the four insole parts during each step [19]. In the present study, the participant's gait was compared with that of healthy controls, and the following distinguishing features were observed: (i) peak and trough values fluctuated and lacked consistency; (ii) peak shapes were irregular, occasionally presenting two successive peaks that produced an "M-shaped" waveform; (iii) the interval between successive heel peaks, that is, the stride time, was inconsistent; (iv) inside and outside pressures on the right foot were reduced; (v) pronounced left-right asymmetry was evident; and (vi) no distinct peaks appeared during periods presumed to correspond to irregular gait while turning.

Because of the wide variation in peak amplitudes, establishing a single threshold for peak detection throughout the entire recording was impractical. Accordingly, the dataset was segmented to exclude intervals with prolonged stride durations, which were assumed to reflect irregular gait. The analysis focused on periods assumed to represent straight walking. The methods used for data segmentation are the following:

1) *Method for Segmenting Insole Data*: The stride-time mode was determined based on the values calculated from the smoothed data. Periods with stride times exceeding 1.1 times the modal duration were considered to involve irregular gait events, such as turning or interruptions, and were excluded from the analysis as non-straight walking. The segmentation procedure involved the following steps:

Calculation of the modal stride time: Stride duration was calculated from the smoothed data obtained using a moving average with a window size of 11. Stride duration was defined as the interval between the peak value x and the subsequent peak $x+1$. Peak

values were defined as the time points corresponding to the maximum force recorded by the insole sensor at each instance of foot-to-ground contact. Peaks were detected using the `find_peaks` function in the SciPy Python Library. Equation (1) defines the threshold used to detect peaks in smoothed the data.

$$thresh_peak = Min + (Max - Min) \times 0.25 \quad (1)$$

Subsequently, time bins were created in 0.2-s increments up to the maximum stride duration. Among these bins, the bin containing the highest number of stride durations for the right heel, which was assumed to represent the better-functioning side, was identified, and its upper limit was defined as the modal stride time.

- *Exclusion of irregular gait periods*: Time intervals exceeding 1.1 times the modal stride duration was regarded as irregular gait, such as during turning or interruptions, and were excluded from the analysis to isolate segments representing straight walking. To ensure that the start and end times of the straight walking segments did not overlap with any of the peak values, a buffer equal to one-fourth of the median stride time of the right heel (calculated after smoothing) was added to both ends of each segment. The longest straight walking segment was used for further analysis.

2) *Peak Characteristics of Raw and Moving-Average Insole Data*: Compared with cases without gait impairment, this case exhibited non-normal peak patterns characterized by consecutive double peaks, forming an M-shaped waveform. Therefore, two types of peak values were calculated: peaks derived from the raw insole data and those derived from data smoothed using a moving average with a window size of 11. The number of peaks and mean values were examined. To extract gait instability, the peak characteristics from the raw data and moving-average data were compared.

3) *Calculation of Mean Values*: The mean values for each insole part were calculated using both non-segmented and segmented raw data, and the resulting mean values were compared between the two datasets.

4) *Mean of Peak Values*: For non-segmented and segmented data, the peak values were defined as the maximum force detected at the heel, toe, inside, and outside parts of the insole during each instance of foot-to-ground contact.

5) *Decline Rates and Classification Feasibility*: The decline rate reflects the decrease in pressure values following each peak. To investigate the feasibility of classifying gait patterns before and after robot-assisted walking, support vector machine (SVM) analyses were performed using the maximum decline rates of heel and toe insole data from both the segmented and non-segmented datasets. Because this study was based on a single clinical case with a limited dataset, the SVM analysis was conducted as an exploratory feasibility evaluation rather

than a comprehensive machine-learning performance comparison.

- Calculation of Decline Rate: The decline rate reflects the decrease in pressure values following each peak. In this study, it was calculated by measuring the difference in the weighted averages between adjacent data points. Specifically, the decline rate was defined as the difference between the weighted averages at points x and $x+1$. The weighted average was calculated using five consecutive data points: the target, two preceding points, and two succeeding points. To emphasize the influence of the central value, weights were assigned as follows: 40% to the target point, 20% to the points immediately before and after, and 10% to the secondary points before and after. Among the calculated decline rates, the largest value within each foot-to-ground contact was defined as the maximum decline rate. The maximum values were extracted for each instance, and the mean values were computed.
- Feasibility of the SVM-based classification: To explore the feasibility of classification using a SVM based on the decline rate of insole data, we examined the classification of the data when the maximum decline rate at each sole-to-ground contact was used as supervised data. As an initial stage of investigating a classification method, we assessed how supervised data could be categorized and considered the potential applicability of SVM-based classification. For “right heel and right toe” and “left heel and left toe” before and after robot use, SVM analyses were conducted using four types of datasets: non-segmented and segmented data. To ensure a one-to-one correspondence between the heel and toe values, the average values were calculated when continuous data such as 1-to-2 or 2-to-1 pairs occurred. SVM analysis was implemented using the scikit-learn library in Python, employing the svm module and SVC class. The SVM parameters were set as follows: $C = 100.0$, kernel = rbf, and gamma = 0.001. These parameters were selected to avoid overfitting while ensuring appropriate model complexity, and the same parameter set was applied to all analyses. The dataset was split into training and test sets using the `train_test_split` function with a fixed random state of 42, allocating 80% and 20% of the data for training and testing, respectively.

6) Calculation of Sole Contact Time and Stride Time:

The analysis was conducted using the segmented dataset with the longest walking duration after excluding periods of irregular gait.

- Calculation of Sole Contact Time
In this case, peak and valley values varied substantially, making it difficult to define a single threshold applicable to the entire walking dataset. Therefore, the threshold for detecting sole contact in this case was modified and determined using the extracted peak and valley values. The thresholds for

peak and valley extraction were defined by Equations (2) and (3), respectively. The sole contact threshold was then calculated using the median values of the extracted peaks and valleys, as shown in Equation (4). The median was used instead of the mean to reduce the influence of extreme values. If sole contacts were not alternately detected between the left and right sides, and two or more consecutive sole contacts were detected on the same side, these were merged into a single sole contact event. Specifically, the durations of consecutive detections were summed, and the sensor values were averaged.

$$peak_thresh = Min + (Peak_ave - Min) \times 0.2 \quad (2)$$

$$valley_thresh = Min + (Max - Min) \times 0.4 \quad (3)$$

$$Sole_contact_thresh = valley_median + (peak_median - valley_median) \times 0.3 \quad (4)$$

- Calculation of stride time: Since peak values occasionally appeared in rapid succession and irregular forms, a moving average was applied to smooth the data and calculate stride times. The signal was smoothed using a moving average with a window size of 11. The stride time was defined as the interval between two successive heel peaks on the same side (left or right).

7) Insole Part Proportions and Time-Series Changes:

The proportion of loading in the heel, toe, inside, and outside parts relative to the total insole load was calculated, and a color-coded time-series graph was generated to visualize the pressure characteristics for each insole part. Because the left and right data were recorded alternately, simultaneous bilateral measurements at the same time points were not possible. Therefore, for the right side, the mean of two consecutive values was regarded as corresponding to the same time point as the left side data and was used for the analysis. Visualizing the time-series changes in the load distribution across the insole parts facilitated a clear understanding and improved interpretability. This approach enabled comprehension beyond numerical values alone and suggested potential applicability in clinical settings.

III. RESULTS

A. Data Segmentation

Among the walking periods, excluding irregular segments, the longest and second-longest durations were identified, and the longest was used for the analysis. To avoid overlap with the peak values, one-fourth of the median right-heel stride time was added to both ends. Figure 3 shows the walking data before, during, and after robot use. The green lines indicate the start times of the segments estimated

as straight walking, and the yellow lines indicate their end times. Irregular gait was removed before robot use; however, after removal, some irregular segments were misclassified as straight walking. While the robot was worn, a small portion of irregular gait was also included.

B. Mean Values

Figure 4 shows the mean values of the segmented and non-segmented data, with some differences despite similar overall trends. Specifically, the mean value for the right heel after robot removal was lower in the segmented data than in the non-segmented data, and the left-right asymmetry was reduced. Additionally, the mean value of the right toe increased after robot removal. In the figures and tables, the terms “non-split” and “split” indicate non-segmented and segmented data, respectively.

C. Comparison of Peak Counts Between Raw and Moving-Average Insole Data

A comparison graph of the number of peaks extracted from the non-segmented and segmented datasets is shown in Figure 5 and a summary of the peak counts, mean values,

and median values is provided in Table I. Only the right-side segmented data during robot use showed the same number of peaks between the raw and moving-average data. In all other conditions, the raw data contained more peaks. Compared with the non-segmented data, the segmented data showed smaller differences between the raw and moving-average results, except for the left heel before robot use. Because the moving average used a window size of 11, peaks occurring immediately after the start or just before the end of the interval were smoothed out and often not detected, resulting in one to two fewer peaks. Even when considering this reduction, the segmented data still showed smaller discrepancies than the non-segmented data. For the segmented left-heel data, which showed the largest discrepancy, Figure 6 presents the raw and moving-average data before and after robot use. The red arrows indicate the extracted peak points. Before robot use, multiple peaks occurred within a single-foot contact phase, whereas after removal, each contact produced a single peak, resulting in a more regular waveform.

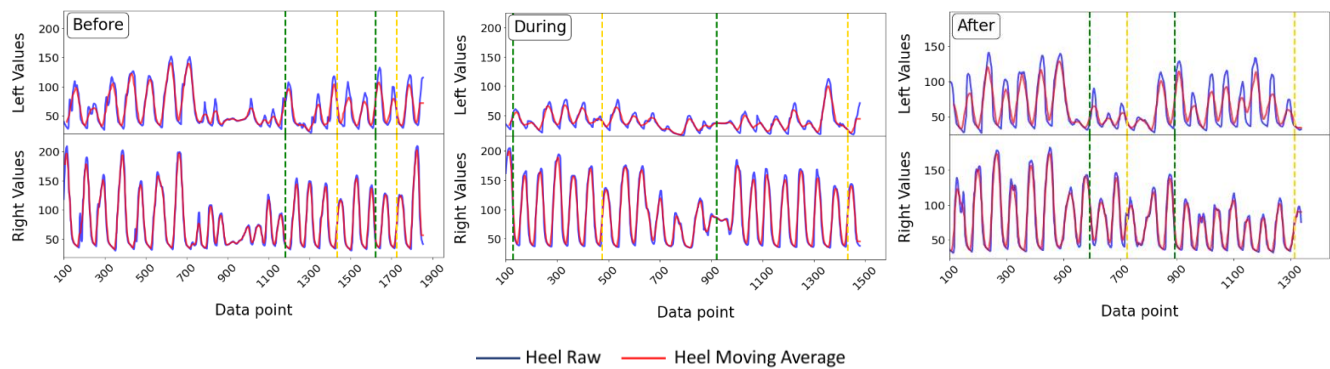


Figure 3. Irregular walking exclusion time.

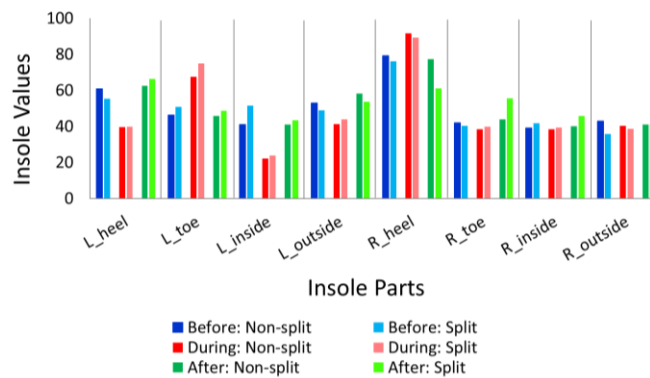


Figure 4. Mean values of insole parts from split and non-split data.

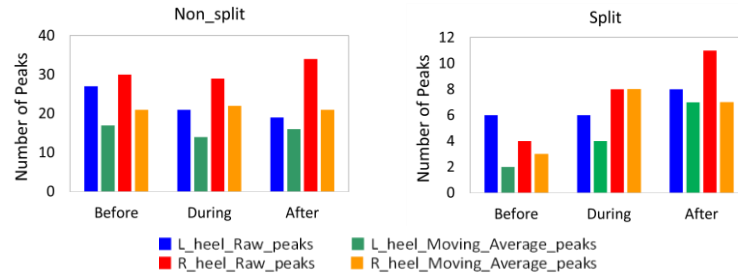


Figure 5. Heel peak counts in split and non-split data.

TABLE I. PEAK COUNT, MEAN, AND MEDIAN OF RAW AND MOVING AVERAGE IN NON-SPLIT AND SPLIT DATA.

Division Left, Right Stats, Implementation		Heel_peaks : Non_split					
		L_Raw	L_Moving_Average	L_Ratio	R_Raw	R_Moving_Average	R_Ratio
count	Before	27	17	1.59	30	21	1.43
	During	21	14	1.50	29	22	1.32
	After	19	16	1.19	34	21	1.62
mean	Before	101.63	94.99	1.07	143.84	142.04	1.01
	During	57.86	56.58	1.02	151.02	147.01	1.03
	After	110.55	93.98	1.18	116.76	118.65	0.98
median	Before	102.32	98.00	1.04	143.90	142.63	1.01
	During	50.80	53.02	0.96	162.20	154.91	1.05
	After	112.64	90.68	1.24	110.18	116.34	0.95
Division Left, Right Stats, Implementation		Heel_peaks : Split					
		L_Raw	L_Moving_Average	L_Ratio	R_Raw	R_Moving_Average	R_Ratio
count	Before	6	2	3.00	4	3	1.33
	During	6	4	1.50	8	8	1.00
	After	8	7	1.14	11	7	1.57
mean	Before	76.00	82.49	0.92	149.41	142.86	1.05
	During	64.38	63.29	1.02	154.89	146.71	1.06
	After	107.10	82.34	1.30	94.69	92.38	1.03
median	Before	67.67	82.49	0.82	148.93	143.76	1.04
	During	53.63	54.69	0.98	165.94	152.36	1.09
	After	110.25	82.10	1.34	93.32	94.64	0.99

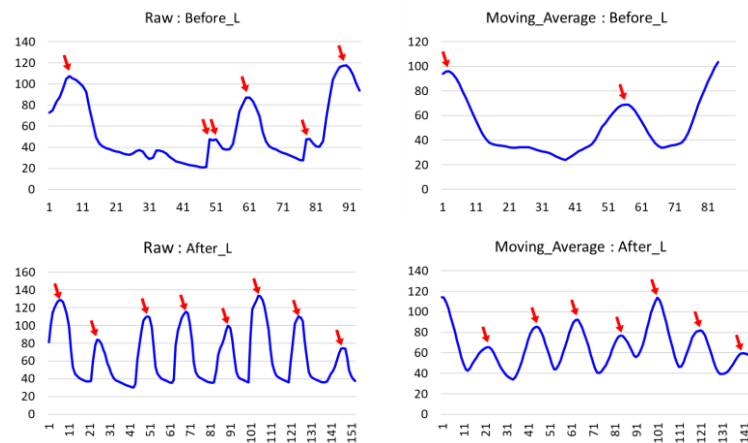


Figure 6. Peaks in raw and moving-average data by split.

D. Mean Peak Values

The mean peak values for each insole part are shown in Figure 7. These results show trends similar to those shown in Figure 4, which presents the mean values for the full raw dataset. After robot use, peak values at the left heel and outside—the side with the robot—increased compared with before use. On the right side, without the robot, heel peak values decreased, reducing asymmetry, whereas toe peaks increased.

E. Decline Rate

The mean absolute values of the maximum decline rates for each insole part are shown in Figure 7, and a comparison of the mean values before and after robot removal shows higher overall rates after removal. On the left side, the segmented data showed a slight decrease in the inside part but increases elsewhere, with a particularly large change at the left heel. On the right side, the decline rate decreased at the heel but increased in other parts, especially at the toe. These findings suggest that robot-assisted walking led to a higher rate of change per unit time, particularly on the side where the robot was worn.

F. SVM Classification Using Maximum Decline Rates

For the heel and toe on both the right and left sides before and after robot use, SVM analyses were performed using non-segmented and segmented datasets. The results are shown in Figure 8. The classification accuracy was higher for the segmented data than for the non-segmented data and was also higher for the right side, the functionally better, non-robot-assisted limb, than for the robot-assisted left side. Although differences in sample size may have influenced these results, the scatter plots indicated that segmented data provided clearer separability. When all data were used for training, the accuracy for the left (robot-assisted) side was 65.9% for the non-segmented data and 90.9% for the segmented data, whereas the right side showed accuracies of 72.7% and 100%, respectively. When 80% and 20% of the data were used for training and testing, respectively, the accuracies for the left side were 55.6% (non-segmented) and 33.3% (segmented), whereas those for the right side were 66.7% (non-segmented) and 100% (segmented). Black points indicate data before robot-assisted walking, and red points indicate data after robot-assisted walking.

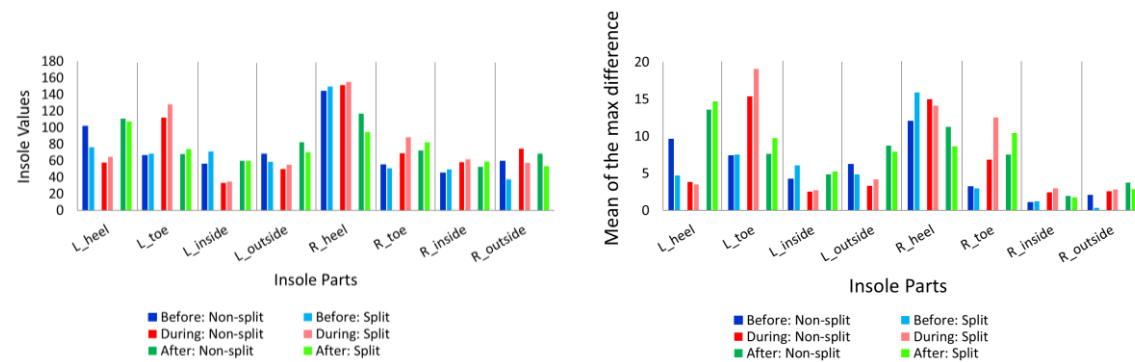


Figure 7. Mean peak values (left) and decline rates (right) for each insole part.

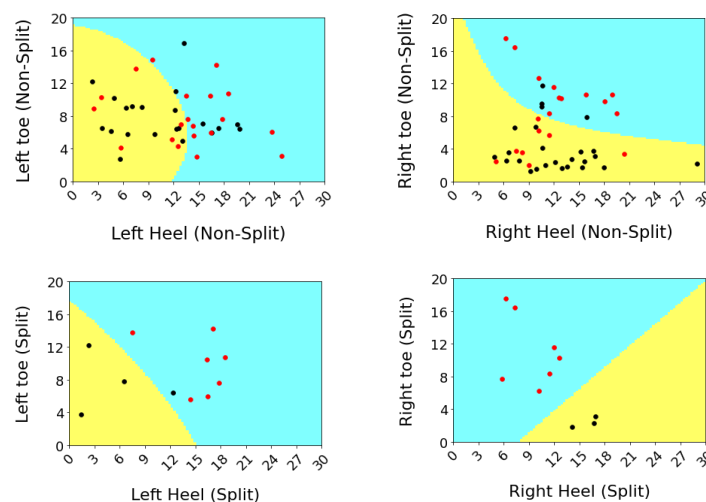


Figure 8. SVM classification of heel and toe decrease using split and non-split data.

G. Insole Part Proportions and Time-Series Changes

Time-series graphs were generated to visualize changes in the proportional load distribution across each part of the insole. As shown in the earlier overall raw data plot in Figure 9, there was no substantial difference before and after wearing the robot; the proportion of the right heel decreased only slightly after removal compared with before use. However, the time-series graphs revealed that the loading pattern on the right heel had changed. Furthermore, the proportion at the right heel was lower and the proportion at the toe was higher in the latter half of walking after robot removal than in the first half.

H. Sole Ground Contact Time

The results for the mean sole contact time are shown in Figure 10. After walking with the robot, sole contact time decreased across all insole parts. Before robot use, the sole contact time on the left medial side was longer than that on the left outside; however, after robot removal, the outside exhibited longer contact times than the medial side. Changes were observed between the inside and outside parts on the left side, as indicated by the comparison presented here, highlighting variations that became more apparent through this analysis of the segmented data.

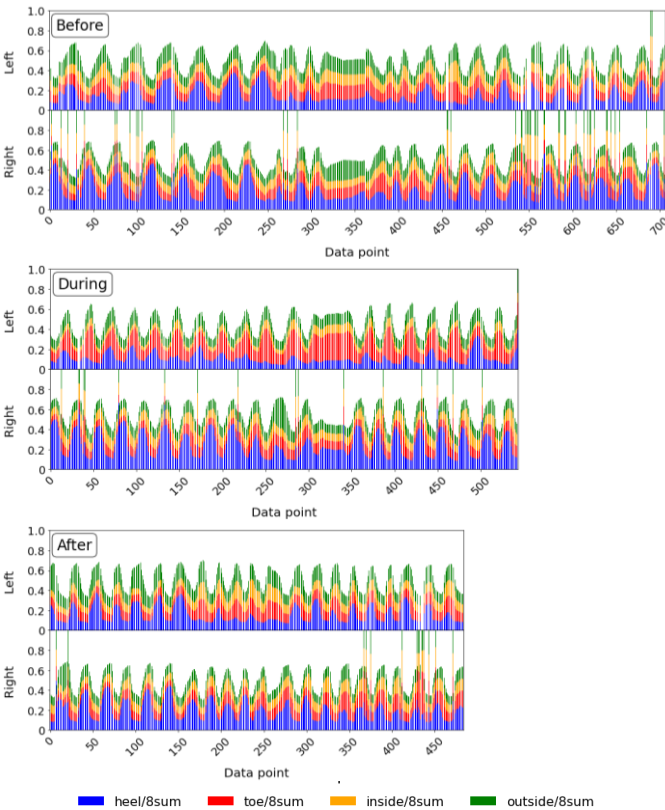


Figure 9. Proportions and temporal changes in insole parts.

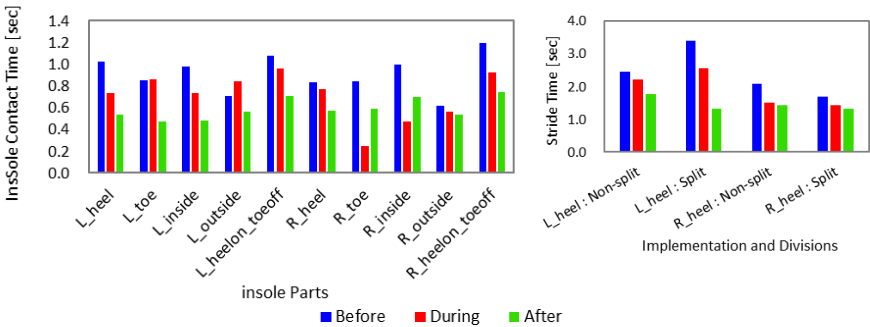


Figure 10. Ground contact time (left) and stride time (right).

I. Stride Time

The mean stride time results are shown in Figure 10. Compared with the functionally better right side, the left side tended to exhibit longer stride times. After robot removal, stride time decreased overall, and left-right differences in stride time were reduced. However, large discrepancies were observed between segmented and non-segmented data for the left heel. Regarding left-right differences, the segmented data showed larger asymmetry than the non-segmented data, particularly before robot use. Examination of the raw data graphs shown in Figure 3 revealed substantial left-right differences, even during periods that appeared to correspond to straight walking before and during robot use. Raw data graphs also indicated marked left-right differences even during periods that appeared to be straight walking, consistent with the trends observed in the segmented stride times derived from (2)–(4). However, the segmented dataset contained few sole-contact events; before robot use, the left-side stride time could be calculated only once or twice. Therefore, although the segmented stride time served as an indicator of gait tendencies, it was not possible to conclusively determine that these data fully represented the gait characteristics, including left-right asymmetry.

IV. DISCUSSION

The participant in this study, who had gait impairments, had more irregular walking data than healthy individuals; thus, the same analytical approach could not be applied. Irregular periods were therefore excluded, and the data were segmented. To avoid distortion from extremely long or short stride durations, stride time was represented using the mode, and periods exceeding 1.1 times the modal value were treated as irregular walking. Although a lower threshold excludes irregular intervals more strictly, it risks removing valid straight walking data. Furthermore, these parameters may vary with gait speed and balance, requiring settings that prioritize assessment items. Segmentation effectively removed irregular walking before robot use; however, some irregular data remained during and after use. The better-functioning right limb showed more regularity, whereas the left limb retained irregularity, suggesting that segmentation based on both limbs, and not only the better-functioning side, may be useful. The usefulness of the segmented data was suggested by the results for peak values, maximum decline rates, stride time, and exploratory SVM classification. In terms of stride time, segmented data also showed greater left-right asymmetry than non-segmented data, particularly before robot use, consistent with trends observed in the raw data graphs. These findings indicate that the characteristics of straight walking were likely captured in the segmented data. However, the number of detected foot contacts was limited, indicating that additional usable segments beyond the longest one may be required for reliable characterization. Although the walking duration was short and the amount of gait data was limited in the present study, future studies using longer walking periods and larger amounts of gait data may enable the application of advanced classifier fusion and regularization approaches [20]. Additionally, this study

examined only straight walking, whereas prior work has emphasized the importance of turning movements in fall-risk assessment [21]. This points to the need for future analyses incorporating irregular gait phases and refining segmentation methods to gain further insights.

We compared the peak counts obtained from the raw data with those from the moving-average-filtered data to evaluate gait stability. The difference in peak counts between the two datasets indicates that this approach is effective for identifying instability. Given a sampling frequency of 50 Hz and a window size of 11, the moving average was calculated over 0.22-s intervals. For individuals with slower gait speeds, a larger window size may be required to appropriately capture their movement characteristics. Therefore, the optimal window size should be determined by considering both gait speed and sampling frequency to ensure accurate assessment for each participant.

The time-series graphs of pressure distribution ratios across the insole parts effectively captured temporal changes in loading that were difficult to discern from the raw data alone. The ability to easily grasp loading changes and trends makes this visualization highly beneficial for clinical rehabilitation and daily gait assessment. In the latter half of the walking phase after removal, the proportion of loading on the right heel decreased, whereas that on the toes increased. This indicates that loading patterns differ between the early and late stages of walking, implying that gait training may require adjustments such as incorporating rest breaks in response to unfavorable patterns or encouraging continued walking when desirable responses are observed. Furthermore, analyzing loading ratios may reflect changes in lower-limb muscle tone, thereby guiding orthotic adjustments. Therefore, time-series loading ratio graphs have strong potential to inform the design of individualized gait training strategies. No matter how high the technical accuracy, data are difficult to use if healthcare professionals cannot easily operate and interpret them. High precision and large volumes of data alone do not necessarily lead to practical usefulness. Success in a clinical context depends not only on providing extensive, high-precision data but also on ensuring that the data are actionable for the end-user. A survey investigating clinical expectations for smart insoles used in knee osteoarthritis care presented the view of healthcare professionals that data should be provided in an easily interpretable format, with the option of accessing more detailed information when needed [22]. The time-series pressure distribution graph is considered to fulfill this requirement as a useful, summarized data format. Although the investigation involved indoor walking in a hospital, daily life gait assessment has been reported to detect condition changes earlier [23], suggesting its usefulness in real-life settings. Outdoor walking is affected by differences in surface and slope [24]; therefore, further investigation is required.

V. CONCLUSIONS AND FUTURE WORK

In this study, gait data were segmented and analyzed through mean-value calculations, temporal parameters, peak-related measures, and insole pressure distribution, including comparisons of peak counts and visualization of time-series

ratios for each insole part. These analyses captured the changes in gait improvement resulting from robot-assisted walking. As this was a single-case study, the generalizability of the results is limited. Future studies should include a larger sample size and further investigate the measurement data, analysis methods, and insole design.

ACKNOWLEDGMENTS

We express our sincere gratitude to the patient and hospital staff who participated in this study. This study was supported by the JSPS KAKENHI (grant number JP23K11207).

REFERENCES

- [1] T. Funayama, Y. Uchida, E. Ohkubo, and R. Kimura, "Segmented Gait Analysis Using Pressure-Sensing Insoles in a Hemiparetic Patient: a Case Study," *Global Health* 2025, pp. 18-23, 2025.
- [2] T. Funayama, Y. Uchida, E. Ohkubo, and Y. Kogure, "Exploring the Potential of a Wrist-Worn Optical Sensor for Measuring Daily Life Activities," *Global Health* 2023, pp. 18-24, 2023.
- [3] A. S. Alharthi, S. U. Yunas, and K. B. Ozanyan, "Deep learning for monitoring of human gait: a review," *IEEE Sensors Journal*, vol. 19, No. 21, pp. 9575–9591, 2019.
- [4] V. Bucinkas et al., "Wearable feet pressure sensor for human gait and falling diagnosis," *Sensors*, vol. 21, issue 15, 5240, 2021, doi: 10.3390/s21155240.
- [5] N. Takayanagi, M. Sudo, Y. Yamashiro, I. Chiba, S. Lee, Y. Niki, and H. Shimada, "Predictivity of daily gait speed using tri-axial accelerometers for two-year incident disability among Japanese older adults," *Scientific Reports*, vol. 12, 10067, 2022, doi: 10.1038/s41598-022-14304-9.
- [6] I. Bytyçi and M. Y. Henein, "Stride length predicts adverse clinical events in older adults: a systematic review and meta-analysis," *Journal of Clinical Medicine*, vol. 10, issue 12, 2670, 2021, doi: 10.3390/jcm10122670.
- [7] L. Angelini et al., "Wearable sensors can reliably quantify gait alterations associated with disability in people with progressive multiple sclerosis in a clinical setting," *Journal of Neurology*, vol. 267, pp. 2897–2909, 2020, doi: 10.1007/s00415-020-09928-8.
- [8] C. J. Hass et al., "Quantitative normative gait data in a large cohort of ambulatory persons with Parkinson's disease," *Plos One*, vol. 7, issue 8, e42337, 2012, doi: 10.1371/journal.pone.0042337.
- [9] D. M. Mohan, A. H. Khandoker, S. A. Wasti, S. I. I. Alali, H. F. Jelineki, and K. Khalaf, "Assessment methods of post-stroke gait: a scoping review of technology-driven approaches to gait characterization and analysis," *Frontiers in Neurology*, vol. 12, 650024, 2021, doi: 10.3389/fneur.2021.650024.
- [10] K. Kamiya et al., "Gait speed has comparable prognostic capability to six-minute walk distance in older patients with cardiovascular disease," *European Journal of Preventive Cardiology*, vol. 25, issue 2, pp. 212–219, 2018, doi: 10.1177/2047487317735715.
- [11] G. Fischer et al., "Factors influencing self-selected walking speed in fibrotic interstitial lung disease," *Scientific Reports*, vol. 11, 12459, 2021, doi: 10.1038/s41598-021-91734-x.
- [12] A. Demirel, D. Onan, M. Oz, Y. O. Ashlyuce, and O. Ulger, "Moderate disability has negative effect on spatiotemporal parameters in patients with chronic low back pain," *Gait & Posture*, vol. 79, pp. 251–255, 2020, doi: 10.1016/j.gaitpost.2020.05.015.
- [13] Y. C. Kuan et al., "Balance and gait performance in older adults with early-stage cognitive impairment," *European Journal of Physical and Rehabilitation Medicine*, vol. 57, issue 4, pp. 560 – 567, 2021, doi: 10.23736/S1973-9087.20.06550-8.
- [14] L. D'Arco, H. Wang, and H. Zheng, "Integration of smart insoles for gait assessment in exoskeleton assisted rehabilitation," *Scientific Reports*, vol. 15, 28350, 2025, doi: 10.1038/s41598-025-10032-y.
- [15] V. Tsakanikas et al., "Evaluating gait impairment in Parkinson's disease from instrumented insole and IMU sensor data," *Sensors*, vol. 23, issue 8, 3902, 2023, doi: 10.3390/s23083902.
- [16] S. Subramaniam, S. Majumder, A. I. Faisal, and M. J. Deen, "Insole-based systems for health monitoring: current solutions and research challenges," *Sensors*, vol. 22, Issue 2, 438, 2022, doi: 10.3390/s22020438.
- [17] B. Park, M. Kim, D. Jung, J. Kim, and K. R. Mun, "Smart insole-based abnormal gait identification: deep sequential networks and feature ablation study," *Digital Health*, vol. 11, 2025, doi: 10.1177/20552076251332.
- [18] V. M. Santos, B. B. Gomes, M. A. Neto, P. F. Rodrigues, and A. M. Amaro, "A systematic review on smart insole prototypes: development and optimization pathways," *Actuators*, vol. 14, issue 8, 408, 2025, doi: 10.3390/act14080408.
- [19] T. Funayama, Y. Uchida, Y. Kogure, D. Souma, and R. Kimura, "Exploring the assessment of steps using insoles with four-part pressure sensors," *Sensors & Transducers*, vol. 263, issue 4, pp. 21–28, 2023.
- [20] A. Salazar, L. Vergara, G. Safont, and E. Vidal, "Graph regularization methods for soft detector fusion," *IEEE Access*, vol. 11, pp. 144747–144759, 2023, doi: 10.1109/ACCESS.2023.3344776.
- [21] J. M. Leach, S. Mellone, P. Palumbo, S. Bandinelli, and L. Chiari, "Natural turn measures predict recurrent falls in community-dwelling older adults: a longitudinal cohort study," *Scientific Reports*, vol. 8, 4316, 2018, doi: 10.1038/s41598-018-22492-6.
- [22] D. Lin, E. Papi, and A. H. McGregor, "Exploring the clinical context of adopting an instrumented insole: a qualitative study of clinicians' preferences in England," *BMJ Open*, vol. 9, e023656, 2019, doi: 10.1136/bmjopen-2018-023656.
- [23] E. Warmerdam, C. Wolff, M. Orth, T. Pohlemann, and B. Ganse, "Long-term continuous instrumented insole-based gait analyses in daily life have advantages over longitudinal gait analyses in the lab to monitor healing of tibial fractures," *Frontiers in Bioengineering and Biotechnology*, vol. 12, 1355254, 2024, doi: 10.3389/fbioe.2024.1355254.
- [24] E. Warmerdam, L.-M. Burger, D. F. Mergen, M. Orth, T. Pohlemann, and B. Ganse, "The walking surface influences vertical ground reaction force and centre of pressure data obtained with pressure-sensing insoles," *Frontiers in Digital Health*, vol. 6, 1476335, 2024, doi: 10.3389/fdgth.2024.1476335.

Assessment of Postural Changes Before and After Hemodialysis Based on Body Weight Measurement Motion Using MediaPipe and Machine Learning

Yasutaka Uchida

Dept. of Life Science
Teikyo University of Science
Tokyo, Japan
e-mail: uchida@ntu.ac.jp

Ren Kondo

Dept. of Clinical Engineering
Katori Omigawa Medical Center
Chiba, Japan
e-mail: 0ys3272f7251v0s@ezweb.ne.jp

Sayuri Sugaya

Nursing Department
Katori Omigawa Medical Center
Chiba, Japan
e-mail: sugaya.sayuri@gmail.com

Kazuyoshi Hori

Dept. of Life Science
Teikyo University of Science
Tokyo, Japan
e-mail: hori-k@ntu.ac.jp

Satoshi Kidooka

Dept. of Clinical Engineering
Katori Omigawa Medical Center
Chiba, Japan
e-mail: kidooka@lilac.plala.or.jp

Tomoya Iizuka

Dept. of Clinical Engineering
Katori Omigawa Medical Center
Chiba, Japan
e-mail: pagupagu_tomoya@yahoo.co.jp

Abstract— Post-dialysis unsteadiness in patients undergoing hemodialysis represents a significant clinical concern, as it increases the risk of falls and negatively affects prognosis and quality of life. However, conventional assessment methods used in clinical practice are often subjective, and simple, objective approaches for quantitatively evaluating post-dialysis unsteadiness remain insufficiently established. In this study, we propose a quantitative evaluation method for evaluating postural changes and unsteadiness by focusing on body weight measurement before and after hemodialysis and applying MediaPipe, a camera-based, non-contact pose estimation technique. Stepping onto a weighing scale requires balance control over a small base of support and can accentuate subtle alterations in postural stability. First, posture-related features were extracted through a preliminary study involving a small number of participants, and the structure of the extracted features was analyzed using principal component analysis, support vector machine, and k-means clustering. Subsequently, an expanded dataset was employed to evaluate the classification performance of pre- and post-dialysis conditions using machine learning. The results revealed discernible differences in postural stability before and after hemodialysis, indicating the effectiveness of the proposed approach. This study demonstrates the feasibility of a simple and objective method for assessing post-dialysis unsteadiness by combining camera-based pose estimation with machine-learning techniques, with potential applicability in clinical settings.

Keywords—Hemodialysis; Post-dialysis unsteadiness; MediaPipe; Pose estimation; Machine learning; Support vector machine; k-means clustering.

I. INTRODUCTION

Unsteadiness and falls in patients undergoing hemodialysis constitute a significant clinical concern, profoundly impacting prognosis and quality of life (QOL). Following hemodialysis, rapid fluid shifts and reductions in blood pressure often lead to orthostatic hypotension and impaired postural control, thereby increasing the risk of falls. This risk is particularly pronounced among elderly hemodialysis patients, for whom falls can result in fractures, hospitalization, and a subsequent decline in long-term outcomes [1-5].

Given this background, the early and objective assessment of post-dialysis unsteadiness is crucial for ensuring patient safety and improving QOL. Conventional assessment methods commonly employed in clinical practice, however, often rely on subjective observations or simple balance tests, which may lack both objectivity and reproducibility. Although quantitative evaluations using force plates or wearable sensors, such as accelerometers, have been reported, their practical implementation is frequently constrained by installation complexity and operational demands [6-10].

In recent years, camera-based pose estimation techniques, which enable non-contact analysis of human posture, have garnered increasing attention in sports science and rehabilitation research. Methods such as OpenPose and MediaPipe Pose allow for the accurate extraction of body landmarks from video recordings and facilitate the calculation of indices such as postural sway and joint angle

variability [11-14]. Our research group has previously demonstrated the clinical applicability of MediaPipe-based motion analysis through studies on gait assessment, evaluation of walking aid effects, investigation of error sources in pose estimation, and quantitative assessment analysis of gait changes in patients with hemiparesis [15].

Despite these advances, studies employing camera-based, non-contact pose estimation to quantitatively evaluate postural stability and unsteadiness in hemodialysis patients remain limited. In particular, few investigations have systematically examined feature extraction and machine-learning-based classification of pre- and post-dialysis conditions [1] [16].

Accordingly, this study focused on postural behavior during body weight measurement before and after hemodialysis and aimed to develop a method for quantitatively evaluating postural changes and unsteadiness using MediaPipe-based video analysis. As stepping onto a weighing scale requires maintaining balance over a small base of support and tends to accentuate subtle alterations in postural control, a preliminary study was conducted to identify effective features. Dimensionality reduction and clustering analyses were subsequently applied to explore the structure of the extracted features, followed by an assessment of pre- and post-dialysis classification performance using a support vector machine with an expanded dataset.

The remainder of this paper is organized as follows. Section II describes the participants, experimental setup, and measurement conditions. Section III presents the results of posture-related indices and joint angle calculations used for evaluation. Section IV provides a discussion of the findings, and Section V concludes the paper and outlines directions for future research.

II. EXPERIMENTS

A. Participants For Basic Feature Value Extraction

This preliminary experiment was conducted to identify effective feature values for the quantitative evaluation of post-dialysis unsteadiness. Four patients undergoing maintenance hemodialysis were recruited, and their demographic and clinical characteristics are summarized in Table I.

TABLE I. PARTICIPANT INFORMATION. PD AND UFR DENOTE PRIMARY DISEASE AND ULTRA-FILTRATION RATE

Subject	Age	Sex	Case of ESRD	Δ DW	ABI R	ABI L	Note
A	54	F	CG	4.2	1.10	1.17	BP reduced
B	62	M	Unknown	4.8	1.18	1.21	BP stable
C	78	M	DM	5.3	0.98	1.46	BP stable
D	51	M	DM	6.0	1.15	1.23	BP stable

Participants' selection was based on the presence of blood pressure fluctuations following dialysis and subjective symptoms of visual dizziness. This study was

approved by the Ethics Committee of Katori Omigawa Medical Center, Katori, Japan (Approval No. 2024-3). Written informed consent was obtained from all participants prior to data collection.

B. Experimental Setup and Recording Conditions for Basic Feature Value Extraction

As illustrated in Figure 1, participants were recorded while standing still or shifting their posture on a body-weight scale. A consumer-grade video camera was positioned approximately 2 m in front of the participants to capture their movements. Owing to spatial constraints in the clinical environment, the camera was placed slightly above and to the left of each participant.



Figure 1. Experimental setup.

C. Data Acquisition and Skeletal Landmark Extraction for Basic Feature Value Extraction

Video data were processed using the Pose module of MediaPipe. The analysis was performed in a Jupyter Notebook using Python. MediaPipe extracted 3D skeletal landmarks (x, y, and z coordinates) for 33 body points. These coordinates, along with the visualization of the estimated skeletal structure overlaid on the video frames, were saved in comma-separated value (CSV) format for subsequent analysis.

D. Postural Indices and Joint-Angle Computation for Basic Feature Value Extraction

To evaluate postural changes before and after hemodialysis, a set of approximated joint angles (hereafter referred to as “approximated angles”) was defined, as illustrated in Figure 2. The following six metrics were used:

(1) Approximated trunk forward tilt angle (TFTA): Defined as the angle formed between the trunk axis (midpoints of the shoulders and hips) and the lower-limb axis (midpoints of the hips and knees).

(2) Approximated cervical flexion angle (CFA): Defined as the angle between the line connecting both shoulders to the nose and the line connecting the midpoint of both hips to the midpoint of the shoulders.

(3) Approximated shoulder abduction angle (SAA_L of SAA_R): Calculated using the positions of the shoulder and elbow joints for the left and right sides, respectively.

(4) Approximated elbow flexion angle (EFA): Calculated as the angle formed by the wrist, elbow, and shoulder joints.

(5) Approximated knee flexion angle (KFA): Calculated as the joint angle formed by the hip, knee, and ankle joints.

(6) Approximated hip flexion angle (HFA): Defined as the angle between the extended trunk axis and the lower-limb axis.

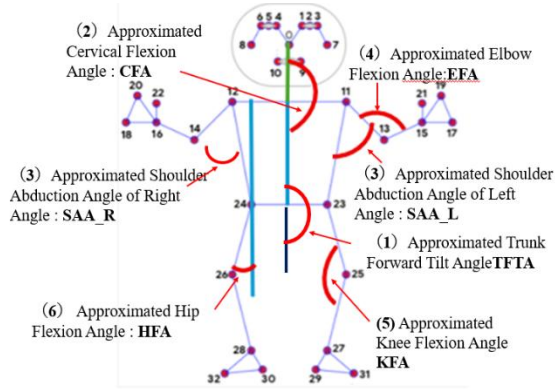


Figure 2. Definitions of the approximated joint angles.

Each angle was calculated using the following formula, based on the inner product of two vectors $\vec{A}$ and $\vec{B}$.

$$\theta = \cos^{-1} \left(\frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|} \right) \quad (1)$$

The vectors were derived from the coordinate data stored in the CSV files. Given that the video recordings were obtained from a slightly left upper oblique angle relative to the participants, all computed values were interpreted as approximated angles rather than exact anatomical measurements.

III. EXPERIMENTAL RESULTS

A. Preliminary Feature Extraction Using Four Participants

1) *Qualification of nasal sway for basic feature extraction:* Figure 3 illustrates an example of nasal-position fluctuations for participant A while standing on a weighing scale. Figures 3(1) and 3(2) correspond to the pre-dialysis and post-hemodialysis conditions, respectively. The recording time before hemodialysis was shorter because of the simultaneous measurement of body weight. Although identical measurement durations would be preferable, no special instructions were provided to the participants, who followed their usual clinical procedures.

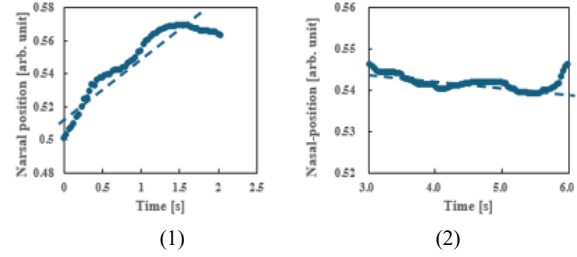


Figure 3. Example of nasal position displacement during body weight measurement in the preliminary experiment: (1) pre-dialysis and (2) post-hemodialysis conditions.

To eliminate the consistent directional drift, a linear correction was applied to the x-axis nasal-position data, adjusting the overall mean to zero. The corrected data are presented in Figure 4. The degree of nasal sway was calculated as the difference between the maximum positive and negative peaks, with red circles indicating the corresponding time points. Additionally, if the values reached the measurement-range boundaries without forming distinct peaks, they were counted as peak points. The number of transitions in the right and left directions is summarized in Table II.

Prior to hemodialysis, all participants exhibited a consistent difference in the number of directional transitions between rightward and leftward movements. Therefore, this index may serve as a useful quantitative indicator of postural changes associated with hemodialysis.

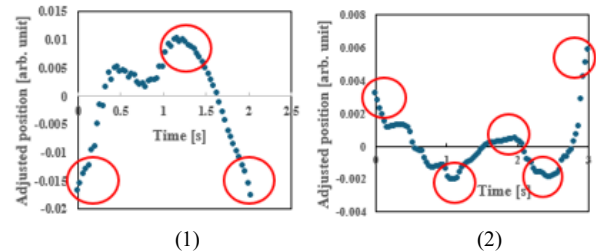


Figure 4. Linearly corrected nasal position data used for quantification of nasal sway in the preliminary feature extraction experiment: (1) pre-dialysis and (2) post-hemodialysis conditions.

TABLE II. NUMBER OF RIGHT-LEFT DIRECTIONAL TRANSITIONS OF NASAL SWAY IN THE PRELIMINARY FEATURE EXTRACTION EXPERIMENT

		Before	R-L	After	R-L
A	L	1	1	3	-1
	R	2		2	
B	L	1	1	2	0
	R	2		2	
C	L	2	1	3	0
	R	3		3	
D	L	1	1	2	1
	R	2		1	

Figure 5 presents the time-series differential analysis of the x-axis nasal position, performed to detect subtle postural fluctuations that are not readily apparent in the raw displacement data. Changes in the sign of the differential were used to identify reversals in the direction of nasal movement, thereby enabling quantification of directional transitions as an index of postural control.

These results indicate that nasal sway characteristics derived from differential analysis may serve as effective features for evaluating postural changes associated with hemodialysis.

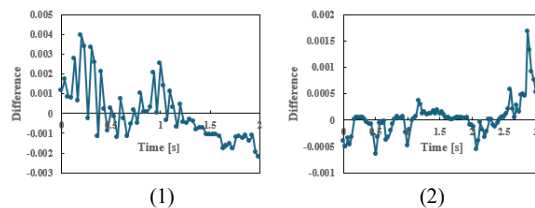


Figure 5. Time-series differential analysis of the x-axis nasal position used to identify directional reversals and quantify nasal sway transitions: (1) pre-dialysis and (2) post-hemodialysis conditions.

2) *Approximated trunk forward tilt angle (TFTA)*: As a representative example, the results for Subject B, who exhibited a pronounced change in trunk posture, are presented. The mean trunk forward tilt angle for Subject B was 163° before hemodialysis, decreasing to 136° after hemodialysis, indicating an increase in forward trunk inclination.

Although the median is generally considered more robust to outliers than the mean, the difference between the mean and median values in this study fell within the range of measurement error. Accordingly, the mean value was adopted as the representative index for evaluating trunk forward tilt.

3) *Approximated cervical flexion angle (CFA)*: Figure 6 illustrates representative changes in the approximated cervical flexion angle (CFA) for Subject B. Figures 6(1) and 6(2) correspond to the pre-dialysis and post-dialysis conditions, respectively. The mean CFA decreased from 163° before hemodialysis to 146° after hemodialysis, indicating an increase in forward head posture and cervical flexion. These changes suggest possible alterations in postural control following hemodialysis.

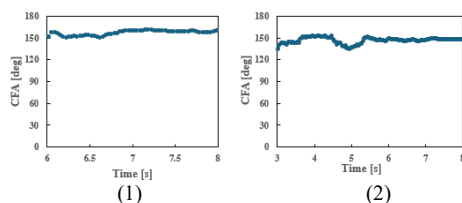


Figure 6. Approximated cervical flexion angle (CFA) for Subject B: (1) pre-dialysis and (2) post-dialysis conditions.

4) *Approximated shoulder abduction angle of left and right (SAA_L and SAA_R)*: Figure 7 presents representative results for the left shoulder abduction angle (SAA_L) of Subject A, with Figures 7(1) and 7(2) corresponding to the pre-dialysis and post-dialysis conditions, respectively. A pronounced change in SAA_L was observed for this subject, with the mean angle decreasing from 14° before hemodialysis to approximately 8° after hemodialysis, representing a reduction of nearly 50%.

In contrast, the right shoulder abduction angle (SAA_R) for the same subject changed by less than 1° . Although not shown in the figure, Subject D exhibited the opposite trend, with a pronounced change in SAA_R, decreasing from approximately 30° to 15° after hemodialysis. These asymmetric changes may reflect individual differences in posture and arm dominance, as well as the influence of an oblique camera angle during data acquisition.

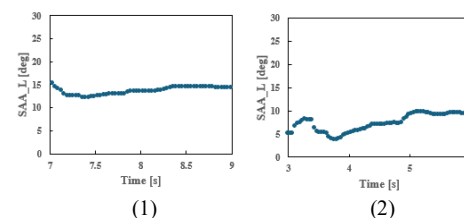


Figure 7. Approximated left shoulder abduction angle (SAA_L) for Subject A: (1) pre-dialysis and (2) post-dialysis conditions.

5) *Approximated elbow flexion angle (EFA)*: Figure 8 illustrates representative changes in the left elbow flexion angle (EFA) for Subject C, with Figures 8(1) and 8(2) corresponding to the pre- and post-dialysis conditions, respectively. The angle increased from 105° before hemodialysis to 112° after hemodialysis, indicating enhanced elbow flexion following treatment.

Although the absolute angle values differed from those observed in Subjects A and D, a similar trend of increased flexion after hemodialysis was noted. In contrast, Subject B exhibited a slight decrease in elbow flexion. These findings indicate that elbow flexion may reflect individual postural adaptations to hemodialysis, although inter-subject variability should be taken into account.

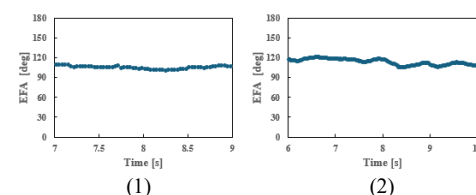


Figure 8. Approximated left elbow flexion angle (EFA) for Subject C: (1) pre-dialysis and (2) post-dialysis conditions.

6) *Approximated knee flexion angle (KFA)*: Figure 9 presents the knee flexion angle for Subject D under pre-

and post-dialysis conditions. The angle increased from 167° prior to hemodialysis to 180° following hemodialysis, indicating a transition toward near-complete knee extension.

Similar tendencies toward increased knee extension were observed in other participants; however, the magnitude of these changes was relatively small compared with those observed for upper-body postural indices. These findings suggest that, within the present measurement framework, detecting subtle post-dialysis changes based solely on knee flexion alone may be challenging.

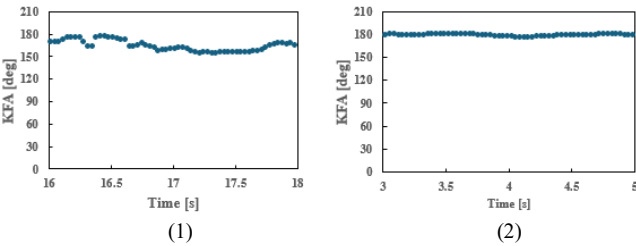


Figure 9. Approximated knee flexion angle (KFA) for Subject D: (1) pre-dialysis and (2) post-dialysis conditions.

7) *Approximated hip flexion angle (HFA)*: Although the hip flexion angle is conventionally evaluated during gait, this study assessed the parameter under static standing conditions. Owing to the short recording duration and the oblique positioning of the camera, measurement accuracy was inherently limited.

Despite these constraints, directional changes in hip flexion angle were observed in two participants, suggesting that this parameter may partially reflect physiological or postural alterations associated with hemodialysis. Improvements in measurement precision, as well as the incorporation of dynamic motion analysis, may enable a more comprehensive and reliable evaluation in future studies.

A summary of all calculated postural indices is presented in Table III.

TABLE III. SUMMARY OF MEASURED VALUES FOR EACH POSTURAL EVALUATION INDEX

		nose	TFTA	CFA	SAA_Left	SAA_Right	EFA	KFA
		Dif	deg	deg	deg	deg	deg	deg
A	BFR	6.50E-02	161	147	14.3	19.3	100	166
	AFT	7.70E-03	177	137	7.70	13.1	118	140
	Ratio	0.11	1.10	0.93	0.54	0.68	1.18	0.84
B	BFR	7.20E-02	163	163	18.0	9.60	120	160
	AFT	5.60E-02	136	146	13.3	14.7	112	161
	Ratio	0.78	0.83	0.90	0.74	1.53	0.93	1.00
C	BFR	2.60E-02	165	156	20.3	30.3	105	149
	AFT	3.20E-02	165	152	23.5	26.8	112	149
	Ratio	1.23	1.00	0.97	1.16	0.88	1.07	1.00
D	BFR	4.30E-02	152	151	24.8	30.6	116	167
	AFT	4.10E-02	154	142	22.1	14.7	124	180
	Ratio	0.95	1.01	0.94	0.89	0.48	1.07	1.08

8) *Spectral analysis of knee movement*: To further assess postural stability, a time–frequency analysis was performed on the lateral (x-axis) knee position data obtained using MediaPipe. Although the recording duration was brief, a discrete Fourier transform (DFT) with a 1 s sliding window was applied to generate spectrograms representing temporal variations in frequency components [17-19].

The input signals were derived from 30-fps video data, resulting in a Nyquist frequency of 15 Hz. The spectrograms were visualized with time on the horizontal axis and frequency on the vertical axis. Spectral magnitude (the square root of the power spectrum) at each time–frequency coordinate was represented using a five-level color gradient, ranging from black (minimum) to red (maximum).

Figure 10 presents a representative example for Subject A, with panels (1) and (2) corresponding to the left and right knees, respectively. The observed spectral patterns reveal subtle temporal fluctuations in knee motion that are not readily detectable using angle-based analyses alone.

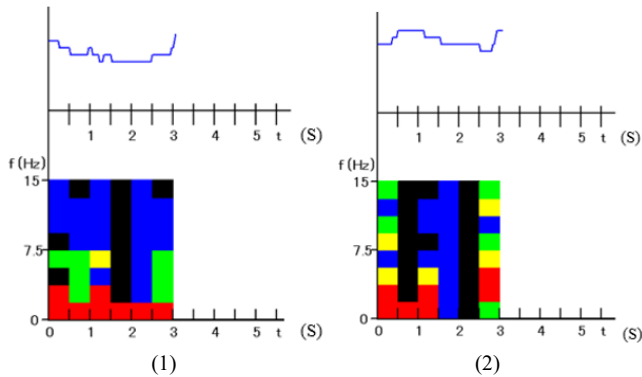


Figure 10. Time–frequency spectrogram of lateral knee movement for Subject A: (1) left knee and (2) right knee.

B. *Feature Analysis and Machine-Learning-Based Exploration Using the Expanded Dataset*

1) *Feature selection based on preliminary analysis*:

Based on the results of the preliminary feature extraction study, the participant cohort was expanded by adding 18 subjects, yielding a total of 22 participants for subsequent analyses. In the present measurements, no prior instructions were provided to the participants in order to capture images during routine body weight measurement in a clinical setting. Consequently, substantial inter-subject variability was observed in both the duration of standing on the weighing scale and the movement of the upper limbs.

To ensure consistency across participants, the analysis in the latter phase was restricted to a 2 s image segment for each subject. Consequently, the observation period for nasal position fluctuations was considerably shorter than that employed in the preliminary feature extraction study,

making it difficult to detect differences between pre- and post-dialysis conditions. For this reason, nasal sway-related features were excluded from the analyses conducted using the expanded dataset.

2) *Visualization of feature distributions and class separability*: To examine the characteristics of the selected feature sets and their discriminative potential between pre- and post-dialysis conditions, exploratory visualizations were performed using the expanded dataset consisting of 22 participants. As described in Section B-1, the analysis was restricted to a 2 s image segment for each participant, and nasal sway-related features were excluded.

Figure 11 (at the end of the paper) illustrates the distributions of representative features for the pre- and post-dialysis classes, visualized using seaborn. These plots facilitate qualitative comparisons of central tendency and variability between classes and enable the identification of features exhibiting noticeable shifts following hemodialysis. Although several features exhibit substantial overlap between the two distributions, others demonstrate class-dependent trends, suggesting potential relevance for discriminative analysis [20][21].

Figure 12 (at the end of paper) presents pairwise relationships among the selected features, visualized using scatter plots with class labels. These plots provide insight into the degree of class separation in low-dimensional feature spaces and reveal correlations among features. In several projections, partial separation between pre- and post-dialysis samples is apparent, whereas other combinations exhibit substantial overlap, indicating redundancy or limited discriminative power.

Overall, the seaborn-based visualizations highlight differences in distributional characteristics and inter-feature relationships across feature sets. These qualitative observations guided the selection of feature sets for subsequent machine-learning-based classification, as described in the following section.

To ensure consistency across participants, the analysis in the latter phase was restricted to a 2 s image segment for each subject. Consequently, the observation period for nasal position fluctuations was considerably shorter than that employed in the preliminary feature extraction study, making it difficult to detect differences between pre- and post-dialysis conditions. For this reason, nasal sway-related features were excluded from the analyses conducted using the expanded dataset.

3) *Principal component analysis*: Although six candidate features were selected through initial screening, principal component analysis (PCA) was performed to determine which components efficiently capture the underlying data structure and contribute to class separability. PCA was applied to project the high-dimensional feature space onto a lower-dimensional space, facilitating visualization of inter-feature relationships and distributional characteristics [22] [23].

Because predefined class labels were not available for the analyzed dataset, a data-driven pseudo-labeling strategy based on PCA was employed. Specifically, the two-dimensional PCA space defined by the first and second principal components (PC1–PC2 plane) was partitioned into four quadrants according to the signs of the principal component scores:

Q1: $PC1 \geq 0$ and $PC2 \geq 0$

Q2: $PC1 < 0$ and $PC2 \geq 0$

Q3: $PC1 < 0$ and $PC2 < 0$

Q4: $PC1 \geq 0$ and $PC2 < 0$

Each sample was assigned to one of the four quadrants (Q1–Q4), and the resulting quadrant labels were used as pseudo labels for subsequent supervised learning. This labeling approach relies exclusively on the geometric structure of the PCA space and does not require manual annotation or predefined threshold values.

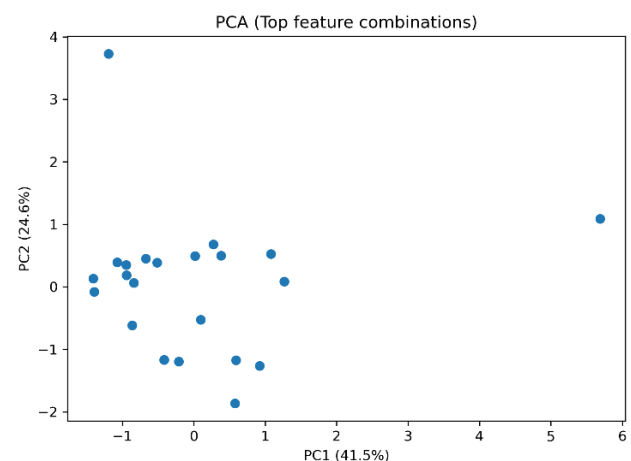


Figure 13. Example of the data distribution in the PCA space obtained using the selected feature set.

The PCA plane was divided into four quadrants based on the signs of the principal component scores, and each sample was assigned to one of the quadrants (Q1–Q4) for data-driven pseudo-labeling.

PC2 represents an asymmetry-related component rather than a global magnitude of motion. This component is primarily influenced by features associated with left–right differences in shoulder abduction, as indicated by loadings with opposite signs for N_SAA_L and N_SAA_R. Notably, the interpretation of PC2 does not depend on the sign of the axis itself; rather, the magnitude along PC2 reflects the degree of asymmetry in upper-body posture and coordination.

4) *Support vector machine analysis*: Based on the PCA results, it was suggested that classification could be feasible using two principal components. Consequently, support vector machine (SVM) classification was performed using these components as input features, employing two types of kernels: a linear kernel and a Gaussian radial basis function (RBF) kernel [24].

The classification results obtained using the linear kernel are shown in Figure 14, while those obtained using the Gaussian RBF kernel are presented in Figure 15. The first principal component (PC1) is considered to represent the overall magnitude of body movement and postural variation between pre- and post-dialysis conditions, whereas the second principal component (PC2) reflects asymmetry-related postural characteristics.

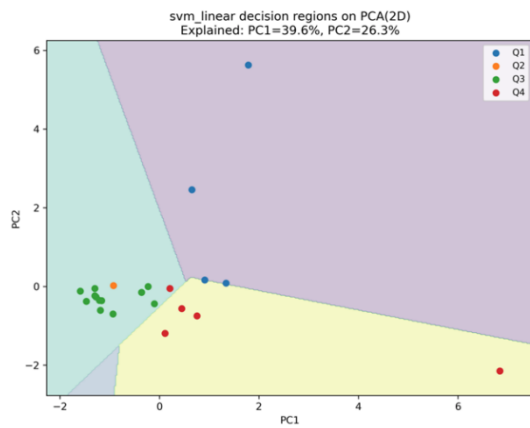


Figure 14. Linear SVM classification in the PC1–PC2 space.

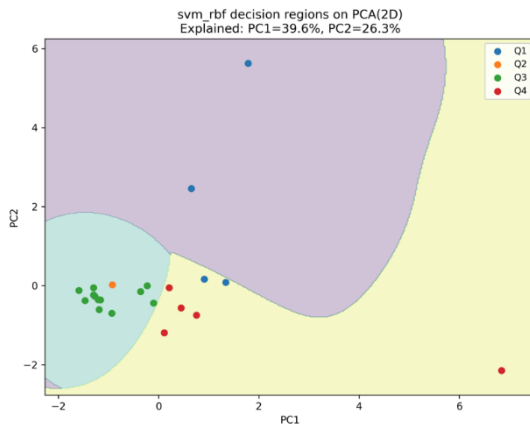


Figure 15. RBF-kernel SVM classification in the PC1–PC2 space.

5) *K-means Clustering Analysis*: To further explore the inherent structure of the feature space without reliance on predefined labels, K-means clustering was applied. The optimal number of clusters was determined using the elbow method, in which the within-cluster sum of squares (inertia) was evaluated as a function of the number of clusters k [25].

Figure 16 illustrates the relationship between k and inertia. A pronounced decrease in inertia is observed up to $k = 3$, followed by a more gradual decline for larger values of k , indicating an elbow point at $k = 3$. Accordingly, the number of clusters was set to three for subsequent analyses.

Figure 17 presents the clustering results obtained with $k = 3$. The samples are partitioned into three groups within the feature space, reflecting distinct patterns in postural characteristics. These clusters provide an unsupervised perspective on the underlying data structure and offer complementary insight to the PCA-based and SVM-based analyses.

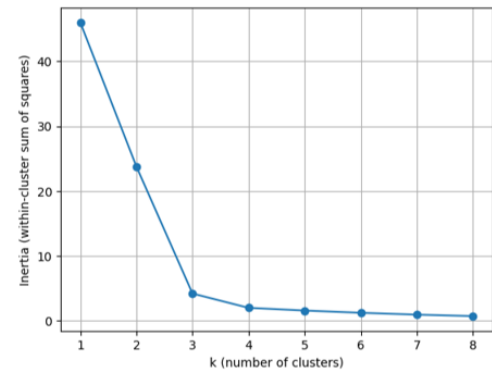


Figure 16. Elbow method for K-means clustering showing the relationship between the number of clusters (k) and inertia (within-cluster sum of squares), with an elbow point observed at $k = 3$.

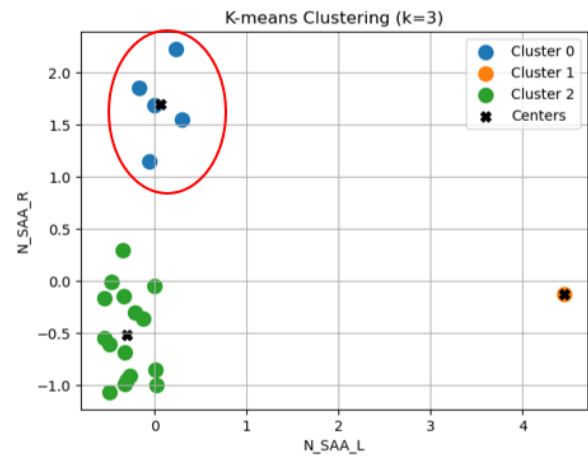


Figure 17. Results of K-means clustering with $k = 3$, illustrating the partitioning of samples into three clusters in the feature space.

IV. DISCUSSION

A. Nasal Displacement (Sway)

No specific instructions were provided to either the staff or participants regarding the measurement process.

Consequently, variability was observed in the timing of body-weight measurements and the duration of video recording. This variability should be considered when interpreting the results, and standardization of measurement conditions is recommended for future studies.

As shown in Table I, four participants exhibited the same number of lateral movements along the x-axis of the nose before hemodialysis. This finding suggests that pre-dialysis movements followed a consistent pattern, and that differences in movement count could serve as a quantitative indicator of postural sway or instability.

Without applying a linear correction to the nasal x-axis data, the observed movement tended to form a plateau, biased in a constant direction, which made it difficult to define motion boundaries. Nevertheless, measuring the duration of such sustained bias may provide an alternative evaluation metric.

Moreover, time-series differencing of nasal-position data can capture subtle movements. Applying a DFT to these differential data may enable the extraction of periodicity or distinctive frequency components, highlighting the potential for advanced sway analysis and quantitative assessment of individual differences.

B. Estimated Cervical Flexion Angle

Figure 6 illustrates the changes in estimated cervical flexion angles. For Participant B, the mean angle decreased from 163° before hemodialysis to 146° after hemodialysis, indicating an increase in cervical flexion. This change may reflect forward head posture or alterations in neck alignment following hemodialysis, potentially associated with physical fatigue or altered postural control.

C. Estimated Elbow Flexion Angle

An increase in elbow flexion angle may indicate changes in muscle tone or enhanced body stability. In contrast to Participant B, these results underscore the significance of individual variations. Additionally, differences between the dominant and nondominant limbs—including the presence of a hemodialysis shunt—should be considered in future investigations.

D. Estimated Knee Flexion Angle

A trend toward full extension (approaching 180°) was observed, which may reflect muscle relaxation or changes in standing posture stability. However, measurement errors resulting from oblique camera angles, rather than frontal views, may have also contributed to this outcome.

E. Estimated Hip Flexion Angle

Although hip flexion is typically evaluated during gait, this study employed a brief static assessment with participants standing on a scale and the camera positioned obliquely. These conditions limit the precision of angle estimation. Nevertheless, the observed differences in flexion direction between participants suggest postural or physiological changes induced by hemodialysis. High-

precision measurements and evaluations during motion are expected to enable more detailed analysis in future studies.

Table IV (at the end of the paper) summarizes the evaluations at each estimated joint-angle point based on the results and discussion. Changes of 5% or more but less than 10% are indicated by the symbol ○, while changes of 10% or more are indicated by the symbol ⊙. Given the variability in each participant's health status, correlations incorporating individual health conditions will also be necessary.

The intradialytic weight loss (%) was 4.2%, 4.8%, 5.3%, and 6.0% for Subjects A, B, C, and D, respectively. Although the present results focused on postural changes and did not directly reflect these fluid removal rates, a more accurate evaluation could be achieved by conducting a comprehensive analysis that integrates the indicators examined in this study with individual physical function data and relevant clinical information, such as medication.

F. Spectrogram

Because the spectral intensity is obtained as a function of time, variations in the frequency components can be analyzed and used to improve the accuracy of motion assessment by correlating them with the corresponding video frames. In the present measurements, however, the upper frequency limit was set to 15 Hz with a temporal resolution of 0.5 s.

G. PCA-based Analysis and Pseudo Labeling

PCA revealed that the selected postural features form a structured distribution in a low-dimensional space, despite substantial inter-subject variability and the absence of predefined class labels. This finding suggests that the extracted features capture meaningful postural characteristics associated with hemodialysis-related changes.

By projecting the data onto the first two principal components and partitioning the PCA space into four quadrants, a data-driven pseudo-labeling strategy was established without relying on manual annotation or predefined thresholds. This approach enabled objective grouping of samples based solely on the geometric structure of the feature space. In contrast to conventional labeling schemes that depend on clinical criteria or subjective judgments, the proposed method provides a reproducible and unbiased alternative for downstream supervised learning.

The observed separation patterns in the PCA space indicate that postural changes associated with hemodialysis are not sufficiently characterized by a single scalar feature, but instead emerge from combinations of multiple postural indices. Notably, upper-body-related features, such as trunk inclination and cervical flexion, contributed more strongly to variance along the principal components than lower-limb features, which demonstrated limited sensitivity under the present measurement conditions.

It should be noted, however, that the PCA-based quadrant labeling does not imply a direct correspondence to clinical states or symptom severity. Rather, it provides a coarse but effective partitioning of the feature space that facilitates supervised classification in subsequent analyses. Future studies incorporating larger datasets and clinically validated outcome measures may enable refinement of the labeling strategy and improve interpretability with respect to physiological or functional states.

Overall, the PCA-based analysis demonstrates that the combination of unsupervised dimensionality reduction and data-driven labeling constitutes a feasible approach for analyzing posture-related data acquired under unconstrained clinical conditions. This framework is particularly well-suited to exploratory studies in which explicit class definitions are unavailable or difficult to establish.

H. SVM Analysis

The observed difference in performance between linear and RBF SVMs suggests that the structure of the PCA space is partially non-linear, particularly among subjects with heterogeneous clinical backgrounds.

I. K-means Analysis

The subjects in the group highlighted in red were relatively younger than the others. Based on this difference, it is necessary to analyze individual characteristics, including dialysis conditions and medical history, to better understand the factors influencing postural clustering patterns.

V. CONCLUSION

This study investigated postural changes before and after hemodialysis using camera-based, non-contact pose estimation in combination with machine-learning techniques. A preliminary analysis involving four participants was conducted to identify candidate postural features, followed by validation using an expanded dataset comprising 22 participants.

The results demonstrated that upper-body-related features, including trunk inclination, cervical flexion, and shoulder posture, captured meaningful changes associated with hemodialysis, whereas lower-limb features exhibited limited sensitivity under the present measurement conditions. Principal component analysis revealed a structured distribution within a low-dimensional feature space, enabling data-driven pseudo-labeling without predefined class definitions. Using these labels, support vector machine classification and K-means clustering further confirm the presence of consistent postural patterns.

Overall, the findings indicate that the integration of non-contact pose estimation with unsupervised and supervised learning provides a feasible and objective

framework for evaluating post-dialysis postural changes under unconstrained clinical conditions. This approach has the potential to support simple and quantitative assessment of postural instability in hemodialysis patients. Future work will focus on expanding the dataset, enhancing measurement robustness, and establishing clinical relevance through comparison with conventional balance assessments.

ACKNOWLEDGMENT

We are grateful to the patients and medical staff of Katori Omigawa Medical Center for their cooperation with this study. We also thank Associate Prof. T. Funayama and Emeritus Prof. Y. Kogure for helpful discussions. This study was partly supported by a JSPS Grant-in-Aid for Scientific Research (Issue no. JP23K11207).

REFERENCES

- [1] Y. Uchida et al., "Postural-Change Detection Before and After Hemodialysis Using MediaPipe," *GLOBAL HEALTH 2025, IARIA*, pp. 1–6, 2025.
- [2] R. G. Roberts, R. A. Kenny, and E. J. Brierley, "Are elderly haemodialysis patients at risk of falls and postural hypotension?," *International Urology and Nephrology*, vol. 35, pp. 415–421, 2003.
- [3] A. Mol, P. T. S. Bui Hoang, S. Sharmin, R. J. A. van Wezel, C. G. M. Meskers, and A. B. Maier, "Orthostatic Hypotension and Falls in Older Adults: A Systematic Review and Meta-analysis," *Journal of the American Medical Directors Association*, vol. 20, no. 5, pp. 589–597.e5, 2019.
- [4] M. Kanbay et al., "An update review of intradialytic hypotension: concept, risk factors, clinical implications and management," *Clinical Kidney Journal*, vol. 13, no. 6, pp. 981–993, 2020.
- [5] S. M. Hamrahan, S. Vilayet, J. Herberth, and T. Fülöp, "Prevention of Intradialytic Hypotension in Hemodialysis Patients: Current Challenges and Future Prospects," *International Journal of Nephrology and Renovascular Disease*, vol. 2023, pp. 173–181, 2023.
- [6] L. A. da S. de Jesus et al., "Factors associated with postural balance in patients with end-stage renal disease on hemodialysis," *Clinical Biomechanics*, vol. 107, 106033, 2023.
- [7] N. Shirai et al., "Dysfunction in dynamic, but not static balance is associated with risk of accidental falls in hemodialysis patients: a prospective cohort study," *BMC Nephrology*, vol. 23, Art. no. 237, 2022.
- [8] H. Zhou et al., "Hemodialysis Impact on Motor Function beyond Aging and Diabetes—Objectively Assessing Gait and Balance by Wearable Technology," *Sensors*, vol. 18, no. 11, 3939, 2018.
- [9] L. M. F. Lucinda et al., "SP420 Postural Balance in Patients Undergoing Hemodialysis," *Nephrology Dialysis Transplantation*, vol. 34, Suppl. 1, gzf103.SP420, 2019.
- [10] A. Capitanini et al., "Multidimensional Assessment of Physical Function for People in Dialysis: The REACT Study," *Kidney and Blood Pressure Research*, vol. 49, no. 1, pp. 1026–1032, 2024.
- [11] A. Komeili, B. Rajaei, "OpenPose vs MediaPipe: How Two Leading Frameworks Are Transforming Pose Estimation in

- 2025,” [saiwa.ai](https://saiwa.ai/blog/openpose-vs-mediapipe/), <https://saiwa.ai/blog/openpose-vs-mediapipe/> 2026.3.5
- [12] K. Morker, “MediaPipe vs. OpenPose: A Practical Guide to Pose Analysis,” DhiWise, <https://www.dhiwise.com/post/mediapipe-vs-openpose-a-practical-guide-to-pose-analysis/> 2026.3.5
- [13] C. S. T. Hii, K. B. Gan, N. Zainal, N. M. Ibrahim, and S. Azmin, “Frontal Plane Gait Assessment Using MediaPipe Pose,” *Proceedings of the 8th International Conference on Space Science and Communication, Springer Proceedings in Physics*, vol. 303, pp. 347–356, 2024, , doi: 10.1088/1757-899X/1279/1/012004.
- [14] Y. Uchida, T. Funayama, E. Ohkubo, D. Souma, and Y. Kogure, “Detecting Gait Changes with Front-Facing Video and MediaPipe: A Hemiplegic Patient Case Study,” *GLOBAL HEALTH* 2024, IARIA, pp. 1–6, 2024.
- [15] Y. Uchida, T. Funayama, E. Ohkubo, D. Souma, and Y. Kogure, “Detection of Changes in Lateral Weight Shift During Gait Improvement of Patients via Image Analysis of Frontal-Direction Imaging Using MediaPipe”, *International J. of Advances in Life science*, vol.17, pp.19-30 no. 1 & 2, 2025.
- [16] F. Roggio, B. Trovato, M. Sortino, and G. Musumeci, “A comprehensive analysis of machine learning pose estimation models used in human movement and posture analyses: A narrative review,” *Heliyon*, vol. 10, no. 21, e39977, 2024.
- [17] C. Francia et al., “Validation of a MediaPipe System for Markerless Motion Analysis During Virtual Reality Rehabilitation,” *Lecture Notes in Computer Science*, vol. 15029, pp. 40–49, 2024.
- [18] A. Gupta, P. L. Shrestha, B. Thapa, R. Silwal, and R. Shrestha, “Knee Flexion/Extension Angle Measurement for Gait Analysis Using Machine Learning Solution ‘MediaPipe Pose’ and Its Comparison with Kinovea®,” *IOP Conference Series: Materials Science and Engineering*, vol. 1279, 012004, 2023.
- [19] Y. Aoyama and H. Tanaka, “Automatic Analysis of Gait Cycle Using MediaPipe,” *ISASE* 2023, pp. 1–13, 2023.
- [20] M. Waskom, “seaborn: statistical data visualization,” *seaborn 0.13.2 Documentation*, 2024. Available from: <https://seaborn.pydata.org/> 2026.3.5
- [21] S. Yıldırım, “Analyzing Data Distributions with Seaborn,” *Towards Data Science*, 2020. Available from: <https://towardsdatascience.com/analyzing-data-distributions-with-seaborn-a8607961a212/> 2026.3.5.
- [22] G. Ivosev, L. Burton, and R. Bonner, “Dimensionality Reduction and Visualization in Principal Component Analysis,” *Analytical Chemistry*, vol. 80, pp. 4933–4944, 2008.
- [23] N. Migenda, R. Möller, and W. Schenck, “Adaptive dimensionality reduction for neural network-based online principal component analysis,” *PLOS ONE*, vol. 16, no. 3, e0248896, 2021.
- [24] A. Mehta, P. Sengupta, D. Garg, H. Singh, and Y. S. Diamand, “Benchmarking the Effectiveness of Classification Algorithms and SVM Kernels for Dry Beans,” *arXiv:2307.07863*, 2023.
- [25] I. Herdiana, M. A. Kamal, Triyani, M. N. Estri, and Renny, “A More Precise Elbow Method for Optimum K-means Clustering,” *arXiv:2502.00851*, 2025.

TABLE IV. EVALUATION RESULTS

		nose	TFTA	CFA	SAA_Left	SAA_Right	EFA	KFA
		Dif	deg	deg	deg	deg	deg	deg
A	Ratio	⊙	⊙	○	⊙	⊙	⊙	⊙
B	≅ 5%	⊙	⊙	○	⊙	⊙	X	X
C	○	⊙	X	X	⊙	⊙	X	X
D	≅ 10%	○	X	○	⊙	⊙	X	○

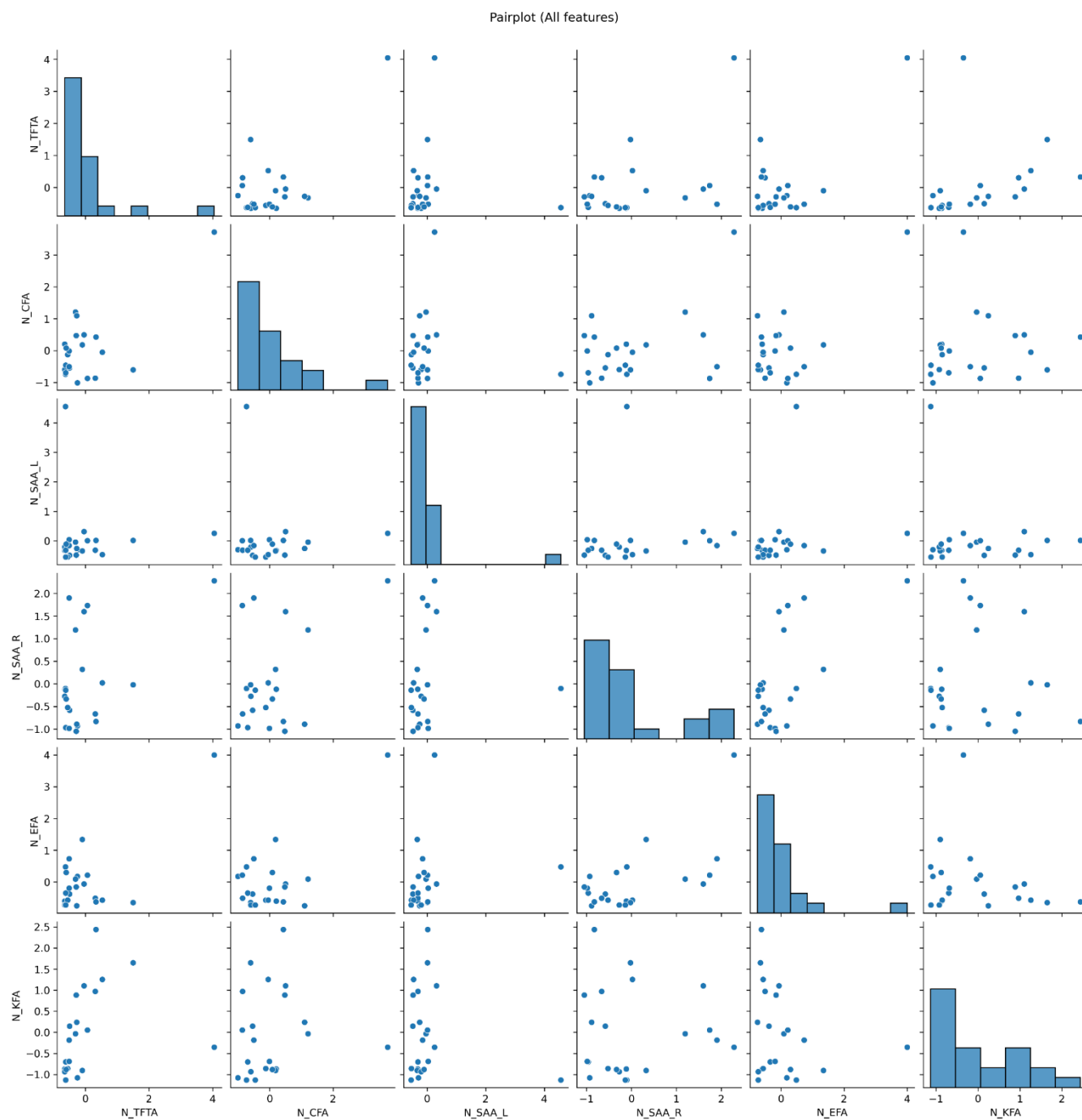


Figure 11. Exploratory visualization of six candidate feature sets derived from the preliminary study.

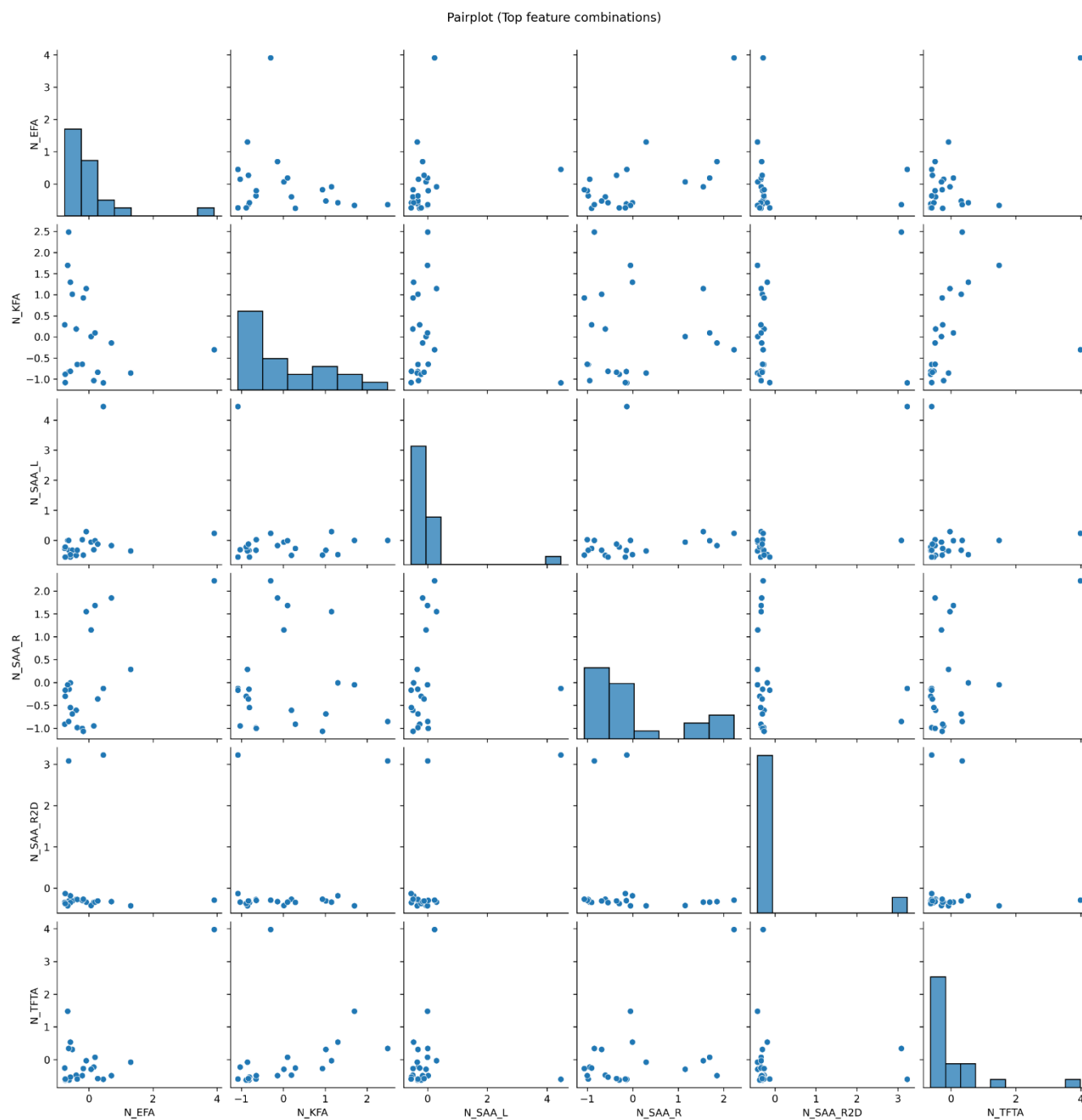


Figure 12. Pairwise relationships among selected features visualized using scatter plots with class labels.

Multi-Model Hierarchical Semantic Retrieval for Predicting Abbreviated Injury Scale 2015 from Clinical Narratives

Chien-Ming Lee*, Pei-Ling Lee[†], Chao-Wen Chen[†], Joffrey Hsu[†], and Chia-Yeuan Han*

**Department of Medical Information,
Kaohsiung Medical University Hospital, Kaohsiung Medical University,
Kaohsiung, Taiwan
nomoney.lee@gmail.com, hoganhan2@gmail.com*

*[†]Division of Trauma and Surgical Critical Care, Department of Surgery,
Kaohsiung Medical University Hospital, Kaohsiung Medical University,
Kaohsiung, Taiwan
peiling@trauma.tw, kmutrauma@gmail.com, joffrey.hsu@gmail.com*

Abstract - Accurate Abbreviated Injury Scale coding is pivotal for trauma registries but remains time-consuming due to the complexity of unstructured clinical narratives. To automate this process while balancing efficiency and accuracy, a hierarchical semantic routing framework is proposed that integrates Bi-Encoder and Cross-Encoder architectures. Unlike traditional brute-force approaches, the proposed method employs a confidence-based routing strategy: (1) a lightweight Bi-Encoder first retrieves candidate codes based on vector similarity; (2) if the prediction confidence exceeds a calibrated threshold (τ_h), the result is accepted immediately to conserve resources; (3) for ambiguous or low-confidence cases, a fallback mechanism triggers the computationally intensive Cross-Encoder to re-rank the top candidates. This design effectively acts as a semantic noise filter, leveraging the Bi-Encoder to prune the search space and the Cross-Encoder to resolve fine-grained distinctions. The system was evaluated using a real-world dataset of emergency department narratives. Experimental results demonstrate that the proposed method achieves a Top-1 accuracy of 59.3%, notably outperforming the full Cross-Encoder baseline (56.5%), while delivering over 14 $\times$ speedup in inference time (547 ms/sample). The main output of this paper is a practical, human-in-the-loop system that operates efficiently on standard central processing unit hardware to expedite injury coding and reduce the cognitive burden on trauma registrars.

Keywords - *abbreviated injury scale; natural language processing; hierarchical classification; semantic retrieval.*

I. INTRODUCTION

This paper presents a comprehensive extension of the preliminary research [1], aiming to address the persistent challenges in automating injury severity coding. Traumatic injury remains a leading cause of mortality and morbidity worldwide, necessitating rigorous data collection to drive quality improvement and epidemiological research. In this context, the Abbreviated Injury Scale (AIS) 2015 [2] serves

as the globally recognized standard for classifying injury severity. Accurate AIS coding is the fundamental prerequisite for calculating the Injury Severity Score (ISS), which is critical for predicting patient survival, benchmarking hospital performance, and allocating healthcare resources.

However, the AIS coding process is inherently complex. The AIS 2015 dictionary contains over 2,000 distinct codes, each associated with a specific anatomical description, severity level, and body region. Given this intricate hierarchical structure and extensive coverage, manual coding demands high clinical expertise and rigorous training. Trauma registrars must mentally map unstructured clinical observations to precise dictionary entries—a process that is labor-intensive, time-consuming, and prone to inter-rater variability. Consequently, developing an automated auxiliary system to assist registrars has become an imperative task in medical informatics, promising to enhance both coding efficiency and data consistency.

The primary obstacle to automation lies in the inherent nature of the input data. Diagnoses documented by clinicians in emergency rooms are typically unstructured, free-text narratives replete with abbreviations (e.g., "fx" for fracture), colloquialisms, non-standard spellings, and semantic ambiguity. For instance, a "crush injury to the chest" implies a completely different set of AIS codes compared to a "penetrating chest wound," requiring a system with deep semantic understanding to distinguish these nuances effectively.

Challenges in Large-Scale Semantic Retrieval Beyond the linguistic noise, a significant technical challenge lies in the **trade-off between retrieval accuracy and computational efficiency**. Traditional keyword-based methods fail to capture semantic nuances, while modern deep learning approaches face a scalability bottleneck. A standard "Bi-Encoder" approach is fast but often lacks the granularity to distinguish between highly similar anatomical descriptions (e.g., specific vertebrae levels). Conversely, a "Cross-Encoder" approach, which performs full self-attention over input pairs, offers superior accuracy but is computationally prohibitive when re-ranking thousands of dictionary candidates for every patient record. Solving this "accuracy-efficiency dilemma" without

relying on expensive GPU infrastructure remains an open problem in clinical Natural Language Processing (NLP) deployment.

Research Gap and Motivation As identified in the preliminary work [1], previous attempts utilizing basic keyword mapping and standard pre-trained models showed potential but faced significant limitations. That approach adopted a relatively simplistic architecture that did not fully exploit the hierarchical taxonomy of the AIS dictionary, nor did it effectively address the inference latency issues inherent in large-scale retrieval. Accuracy and reliability in handling complex semantic ambiguity remained insufficient for clinical standards. Consequently, there is a compelling need to develop a new architecture that is more accurate, robust, and capable of mimicking the hierarchical reasoning process of human experts while maintaining operational efficiency.

Proposed Framework and Contributions To overcome these challenges and significantly improve the performance and practicality of AIS 2015 coding, this paper proposes a **Multi-Model Hierarchical Semantic Retrieval Framework**. Unlike monolithic models, the proposed approach mimics the cognitive process of a human coder: first narrowing down the anatomical region and then scrutinizing the specific injury details. The framework integrates a lightweight Bi-Encoder for rapid candidate generation and a high-precision Cross-Encoder for reranking difficult cases. Crucially, the framework proposed in this paper operates as a training-free retrieval system. By leveraging publicly pre-trained biomedical language models without the need for task-specific fine-tuning, the dependency on large-scale annotated datasets is eliminated, thereby addressing the data scarcity bottleneck and enabling immediate deployment in hospitals lacking historical labeled records.

The main contributions of this paper are summarized as follows:

- **A Novel Hierarchical Routing Architecture:** A multi-stage retrieval framework is proposed that dynamically routes input narratives based on prediction confidence. This design utilizes a similarity-based classifier (Bi-Encoder) to accurately prune the search space and a Cross-Encoder to handle fine-grained semantic comparisons, effectively balancing system performance.
- **Efficiency-Accuracy Optimization:** The proposed framework is demonstrated to act as a semantic noise filter. By reserving the computationally expensive Cross-Encoder only for ambiguous inputs (the "long-tail" cases), the proposed method achieves a $\sim 14\times$ speedup compared to a full Cross-Encoder baseline while paradoxically achieving higher Top-1 accuracy by filtering out irrelevant distractors.
- **Practical Deployment Feasibility:** Unlike many academic models that require high-end GPUs, the proposed optimized pipeline is designed to run efficiently on standard CPU hardware (inference time ≈ 547 ms), making it a viable decision-support tool for resource-constrained hospital environments.

- **Comprehensive Evaluation:** The proposed method is validated on a real-world dataset of emergency department narratives, providing a detailed analysis of performance across different anatomical regions and discussing the implications for human-in-the-loop clinical workflows.

The remainder of the paper is organized as follows: Section II reviews the related work in medical coding and semantic retrieval. Section III details the hierarchical semantic retrieval procedures and model architectures. Section IV presents the experimental setup and evaluation results. Section V discusses the findings, the mechanism of efficiency gains, and limitations. Finally, Section VI concludes the paper and outlines future directions.

II. RELATED WORK

This section reviews related research in the domain of automated medical coding, with a particular focus on processing unstructured clinical narratives, leveraging ontological structures, and adoption of multi-stage retrieval frameworks. Collectively, these advancements lay the foundation for the Multi-Model Hierarchical Semantic Retrieval Framework proposed in this paper.

The review is organized as follows: Section II-A highlights the fundamental challenges in automated medical coding and the rise of deep learning applications. Section II-B discusses strategies for handling the structured, hierarchical nature of medical coding systems. Section II-C indicates challenges of LLMs in Medical Coding. Finally, Section II-D surveys recent advancements in multi-model retrieval and re-ranking mechanisms relevant to proposed work.

A. Challenges in Automated Medical Coding and Deep Learning Applications

According to recent studies [3] and [4], Automated Medical Coding (AMC), which involves assigning International Classification of Diseases (ICD) codes to clinical documents, is a pivotal task in healthcare operations and service delivery. Given that manual coding is time-consuming and error-prone, researchers have widely applied deep learning and NLP techniques to automate this process.

Ji et al. [3] identified that the AMC task faces several primary challenges:

- 1) **Lengthy and Noisy Text:** Clinical notes typically consist of long, free-text replete with medical jargon, non-standard abbreviations, and semantic ambiguity.
- 2) **High-dimensional Label Space:** Medical ontologies (e.g., ICD-9/10) contain tens of thousands of codes, rendering the task an extreme multi-label classification problem.
- 3) **Class Imbalance:** The distribution of codes is highly skewed, exhibiting a "long-tail phenomenon" involving common and rare diseases.

Mullenbach et al. [4] noted that early deep learning approaches extensively utilized Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). Notably,

CNN architectures, such as CAML (Convolutional Attention for Multi-Label classification), combine convolutional layers with attention mechanisms to extract specific segments from the document relevant to each code, thereby providing interpretable predictions. While these studies primarily focus on ICD coding, the tasks of AIS coding share these identical challenges, with the added complexity of mapping injuries to specific severities.

However, a critical limitation of these aforementioned deep learning architectures is their heavy reliance on the **supervised learning** paradigm. To achieve optimal performance, models like CAML require large-scale, high-quality annotated corpora to learn the complex mapping between clinical narratives and code definitions. This dependency poses a significant bottleneck for trauma registries, where labeled historical data is often scarce, expensive to obtain, or subject to privacy restrictions.

To address this limitation, the proposed work diverges from the traditional supervised approach and adopts a **training-free retrieval** strategy. By leveraging the inherent semantic knowledge embedded in publicly pre-trained language models, the proposed framework circumvents the need for task-specific training data. This paradigm shift not only eliminates the costs associated with data annotation but also significantly lowers the barrier to deployment.

B. Handling the Structured Nature of Medical Coding

Analogous to ICD codes, the AIS 2015 lexicon utilized in this study possesses an inherent hierarchical structure [5][6]. Previous work [6] has demonstrated that explicitly modeling the code hierarchy and correlations is pivotal for enhancing coding accuracy.

- 1) **Hierarchical Classification:** Recent deep learning models have integrated hierarchical information by designing specialized architectures. For instance, Cao et al. [6] proposed a hyperbolic representation method that utilizes the tree-like properties of hyperbolic space to learn hyperbolic representations of the code hierarchy, while incorporating Graph Convolutional Networks (GCNs) to capture code co-occurrence relationships. Furthermore, the hierarchical joint learning mechanism proposed by Vu et al. [7] leverages predictions of the first three characters of ICD codes to assist in predicting complete codes, thereby improving performance on low-frequency codes. Inspired by these hierarchical approaches, the proposed strategy (Phase 1: AIS root node classification) focuses on identifying the semantically closest root node, effectively pruning the search space to enhance overall coding accuracy.
- 2) **Semantic Representation and Entity Alignment:** Given the complexity of clinical text, establishing robust medical entity representations is critical. Domain-specific pre-trained models, such as BioBERT [8], have been shown to enhance the understanding of in-domain text through pre-training on biomedical corpora.

Concurrently, to address the high heterogeneity of medical entity naming, research has focused on learning entity representations. For example, the BIOSYN model [9] achieves state-of-the-art performance in biomedical entity normalization tasks by employing synonym marginalization techniques. In this work, a SapBERT model [10] is employed in the first phase. Distinct from standard models, SapBERT is specifically designed with a self-alignment pre-training scheme. By leveraging synonym relationships within large-scale medical ontologies such as the Unified Medical Language System (UMLS), it pulls entity names representing the same concept closer together, forming compact clusters in the vector space. This enables SapBERT to efficiently capture the nuances of medical terminology, making it highly suitable for the semantic similarity classification task in this study.

C. Generative LLMs versus Discriminative Retrieval in Medical Coding

The advent of Large Language Models (LLMs), such as GPT-4 and Llama-3, has transformed biomedical NLP, demonstrating remarkable zero-shot generalization capabilities [11]. Recent studies have explored their use as classifiers for automated ICD coding [12], suggesting that generative models could, in principle, directly predict AIS codes from clinical narratives.

However, deploying LLMs in high-stakes medical registry tasks poses substantial challenges:

- 1) **Hallucination Risks:** Generative models can produce 'medical hallucinations' [13], particularly for rare pathologies or under instruction-alignment bias. In contrast, the proposed discriminative retrieval framework constrains outputs strictly to the AIS dictionary, inherently preventing the generation of non-existent codes.
- 2) **Resource and Privacy Constraints:** LLM inference demands significant GPU memory and computational resources, often requiring cloud-based services that raise data privacy concerns. By design, the proposed lightweight BERT-based framework runs efficiently on standard CPU hardware, enabling secure, on-premise deployment in hospital environments.
- 3) **Confidence Calibration:** While LLMs generate fluent text, reliably quantifying prediction confidence is challenging. The proposed hierarchical routing mechanism leverages explicit similarity scores to determine when fallback re-ranking is necessary, providing transparent and calibrated uncertainty measures essential for human-in-the-loop workflows.

Therefore, although LLMs represent a promising direction in biomedical NLP, for trauma coding applications that demand verifiability, robustness, and efficiency, a specialized, retrieval-based architecture remains the preferred solution.

D. Multi-model Retrieval and Re-ranking Mechanisms

The second and third phases of the proposed framework involve fine-grained semantic comparison and a fallback mechanism under low-confidence conditions. This design aligns with the Retrieve-and-Re-rank paradigm widely adopted in the field of Information Retrieval (IR) [14].

- 1) **Efficient Semantic Retrieval (Bi-Encoders):** To achieve efficient semantic comparison, models must overcome the computational limitations associated with Cross-Encoder architectures like BERT. Sentence-BERT (SBERT) [15] introduces a Bi-Encoder architecture that modifies BERT by using Siamese and triplet network structures to derive semantically meaningful sentence embeddings. SBERT reduces the computational overhead of similarity search from tens of hours to mere seconds by enabling cosine similarity comparison, making it highly suitable for large-scale semantic retrieval. In the global semantic retrieval phase, it is leveraged that the Sentence-Transformers framework [15] initialized with BioBERT weights [16] to generate high-dimensional semantic embedding vectors.
- 2) **Contextual Re-ranking and Cross-Attention (Cross-Encoders):** While bi-encoders are efficient, their performance is typically inferior to cross-encoders, as the latter can perform full Cross-Attention on input pairs [17]. BERT cross-encoders calculate relevance scores by processing the query and document simultaneously [18], [19], thereby achieving state-of-the-art accuracy in re-ranking tasks. This study deploys a Cross-Encoder model [20] within the fallback mechanism. This allows for precise, fine-grained scoring on the candidate pool generated by the retrieval phase, ensuring high accuracy in the final output under low-confidence conditions.

In summary, the Multi-Model Hierarchical Semantic Retrieval Framework proposed in this study integrates specialized techniques designed to address the unique challenges of medical coding—specifically hierarchy and semantic complexity—within an established Retrieve-and-Rerank architecture from the field of Information Retrieval (IR). By synergizing the precise domain entity alignment capabilities of SapBERT, the efficient semantic retrieval speed of SBERT, and the high-precision re-ranking performance of Cross-Encoders, this approach delivers a solution for AIS 2015 coding that is simultaneously accurate, robust, and efficient.

III. PROPOSED METHOD

This section details the architecture and computational workflow of the proposed system.

A. Overview

In this study, a Multi-Model Hierarchical Semantic Retrieval Framework is proposed to automate the prediction of AIS 2015 codes. Distinct from traditional supervised learning approaches that rely on extensive annotated datasets, the proposed framework operates as a *training-free semantic retrieval architecture*. By strictly leveraging the inherent generalization

capabilities of pre-trained models (specifically SapBERT and BioBERT) within their original semantic spaces, this design eliminates the need for task-specific fine-tuning, thereby ensuring high deployability across diverse clinical settings without the bottleneck of labor-intensive data labeling.

The proposed framework operates in a hierarchical manner, leveraging the tree-like structure of the AIS dictionary. It consists of three distinct phases: (1) a Hierarchical Strategy for candidate narrowing, (2) Semantic Comparisons, and (3) a Robustness Enhancement Mechanism featuring a fallback global re-ranking module.

The selection of specific pre-trained models is strategically aligned with the unique requirements of each stage. In Stage 1, SapBERT is utilized for its specialized capability in biomedical entity alignment; its training objective focuses on clustering synonyms, making it highly effective for mapping diverse clinical narratives to the correct anatomical root nodes. In contrast, for Stage 2 and the fallback stage, BioBERT is employed to perform fine-grained matching. While SapBERT excels at individual entity representation, BioBERT offers a broader understanding of medical context and complex clinical syntax, which is crucial for distinguishing between highly similar injury descriptions within a specific anatomical subtree.

B. AIS Hierarchy and Root Extraction

The AIS 2015 is inherently hierarchical, organized by body regions, anatomical structures, and injury severity. To facilitate hierarchical retrieval, the AIS dictionary is first restructured into a forest of semantic trees. Let $\mathcal{D} = \{d_1, d_2, \dots, d_N\}$ represent the set of all unique injury descriptions in the AIS dictionary. The top-level nodes (roots) corresponding to major body regions or specific anatomical chapters are identified. For each root r_k , the textual description $T(r_k)$ is extracted. This process partitions the search space into subsets S_k , where each subset contains injury descriptors belonging to the k -th semantic tree.

C. Text Embedding Method

Given an input clinical narrative q (e.g., a diagnosis string), the first objective is to identify the most relevant anatomical branch. A publicly pre-trained SapBERT model [21] is employed as the semantic encoder for this phase. SapBERT is optimized for biomedical entity alignment, making it highly effective at capturing the nuances of medical terminology. This process is implemented by employing a mean-pooling mechanism to aggregate the hidden states output by the encoder.

Formally, the input q is tokenized into a sequence of tokens $T = \{t_1, t_2, \dots, t_L\}$. The encoder processes T to output a corresponding sequence of hidden states $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L\}$. Since sequence lengths vary, the method uses an Attention Mask $\mathbf{M}$ for weighted averaging, ensuring that only the actual tokens are averaged to obtain a valid sentence embedding $\mathbf{e}$.

The specific calculation steps are as follows:

- 1) **Masked Summation:** The hidden states $\mathbf{H}$ are weighted by the attention mask $\mathbf{M} = \{m_1, \dots, m_L\}$ and then

summed across the sequence length dimension to yield the masked sum $\mathbf{S}$:

$$\mathbf{S} = \sum_{i=1}^L m_i \mathbf{h}_i \quad (1)$$

where $m_i \in \{0, 1\}$ is the mask value corresponding to $\mathbf{h}_i$.

- 2) **Effective Token Count:** The total number of effective tokens C in the mask is calculated, applying a minimum value $\epsilon = 10^{-9}$ to prevent division by zero:

$$C = \max \left(\sum_{i=1}^L m_i, \epsilon \right) \quad (2)$$

- 3) **Mean Pooling:** The average embedding vector $\mathbf{e}_{mean}$ is computed as:

$$\mathbf{e}_{mean} = \frac{\mathbf{S}}{C} \quad (3)$$

- 4) **L_2 Normalization:** Finally, the average embedding $\mathbf{e}_{mean}$ is subjected to L_2 normalization. This ensures all embedding vectors reside on the unit sphere, which is crucial for cosine similarity-based classification (as described below):

$$\mathbf{e} = \frac{\mathbf{e}_{mean}}{\|\mathbf{e}_{mean}\|_2} \quad (4)$$

D. Cosine Similarity-based Classification (Stage 1: AIS Root Classification)

This stage aims to assign an input diagnostic embedding $\mathbf{e}$ to the most relevant top-level AIS root nodes, thereby narrowing the search space for subsequent fine-grained classification. Rather than performing direct label prediction, this stage serves as a semantic routing mechanism that partitions the AIS dictionary into anatomically coherent subsets, as defined in the Method Overview.

Specifically, each AIS root r_k is associated with a textual description $T(r_k)$, from which a normalized, fixed-dimensional root embedding $\mathbf{r}_k$ is pre-computed using the same text embedding method described in the previous section. Given an input diagnostic embedding $\mathbf{e}$, the relevance between $\mathbf{e}$ and each root node is quantified using cosine similarity, enabling the identification of the most semantically aligned anatomical roots.

- 1) **Similarity Calculation:** Since both the input embedding $\mathbf{e}$ and each root embedding $\mathbf{r}_j$ are L_2 -normalized, their inner product is equivalent to cosine similarity:

$$\text{Sim}_j = \mathbf{e}^\top \mathbf{r}_j = \cos(\mathbf{e}, \mathbf{r}_j) \quad (5)$$

Let $\mathbf{R} \in \mathbb{R}^{K \times d}$ be the matrix containing all root embeddings. The similarity scores are computed simultaneously in a vectorized manner as

$$\mathbf{Sim} = \mathbf{R} \mathbf{e}.$$

- 2) **Result Ranking and Output:** The similarity scores are ranked in descending order to produce a list of AIS root candidates along with their corresponding similarity scores. This ranked list is used to determine the candidate subsets for the subsequent classification stage.

E. Final AIS Code Refinement and Selection (Stage 2: AIS Code Refinement and Selection)

Given the candidate AIS root nodes identified in Stage 1, this stage focuses on selecting the most precise AIS injury code within the corresponding semantic subtree. By restricting the search space to a root-specific subset $\mathcal{S}_k$, the refinement process employs a higher-resolution semantic similarity model to perform fine-grained matching between the input diagnostic embedding and candidate AIS codes. Here, $\mathcal{S}_k \subset \mathcal{D}$ denotes the set of AIS injury descriptions associated with the selected root node r_k .

1) **Similarity Thresholding and Subtree Pruning:** Upon obtaining the similarity vector $\mathbf{Sim}$ from Stage 1, the system identifies the top-ranked root node r^* with the highest similarity score Sim_{max} . To ensure the reliability of the semantic routing, a predefined high-confidence threshold τ_h is applied. The pruning mechanism validates whether:

$$\text{Sim}_{max} \geq \tau_h \quad (6)$$

If this condition is met, r^* is accepted as the target anatomical branch. The system then effectively prunes the search space from thousands of potential codes to the specific semantic subset $\mathcal{S}_{r^*} \subset \mathcal{D}$ associated with r^* . This pruning step removes irrelevant anatomical regions, thereby reducing the search space and improving both computational efficiency and classification accuracy.

2) **Semantic Similarity Calculation based on BioBERT:** Within the constrained candidate subset $\mathcal{S}_{r^*}$, a pre-trained BioBERT model [22] is employed for fine-grained matching. To generate high-quality sentence embeddings, the model is implemented using the Sentence-Transformers [15] framework, transforming the clinical diagnosis and each candidate AIS description within $\mathcal{S}_{r^*}$ into high-dimensional semantic embeddings.

The calculation proceeds as follows:

- 1) **Vector Embedding:** Let q be the input clinical diagnosis and $\mathcal{C} = \{c_1, c_2, \dots, c_M\}$ be the set of candidate AIS descriptions within the subset $\mathcal{S}_{r^*}$. The encoder function $E(\cdot)$ transforms these textual inputs into dense vectors:

$$\mathbf{v}_q = E(q), \quad \mathbf{V}_\mathcal{C} = \{E(c_1), E(c_2), \dots, E(c_M)\} \quad (7)$$

where $\mathbf{v}_q$ is the diagnosis vector and $\mathbf{V}_C$ is the matrix of candidate vectors.

- 2) **Cosine Similarity Calculation:** The matching score for each candidate is determined by calculating the cosine similarity between $\mathbf{v}_q$ and each candidate vector $\mathbf{v}_{c_i} \in \mathbf{V}_C$:

$$\text{Score}_i = \cos(\mathbf{v}_q, \mathbf{v}_{c_i}) = \frac{\mathbf{v}_q \cdot \mathbf{v}_{c_i}}{\|\mathbf{v}_q\|_2 \|\mathbf{v}_{c_i}\|_2} \quad (8)$$

- 3) **Result Ranking and Output:** The system sorts all candidate AIS codes based on Score_i in descending order. The top K (e.g., $K = 10$) AIS codes with the highest scores are selected as the final output, providing a ranked list for user review.

F. Fallback Retrieval and Re-ranking Mechanism

This fallback mechanism is activated when the Stage 1 root node classifier fails to identify any root node whose similarity score exceeds the predefined threshold τ_h . Such cases typically arise when the input clinical diagnosis is highly ambiguous, contains sparse contextual cues, or is semantically distant from the existing AIS code descriptions.

To address these challenging scenarios, the system adopts a two-model retrieval architecture that performs a global semantic search across the entire AIS database, followed by a fine-grained re-ranking process. This design ensures robust and reliable AIS code retrieval even when hierarchical semantic routing is not sufficiently confident.

The first step of the fallback mechanism is the construction of a global Top- N candidate pool from the entire AIS code description set. This step performs a coarse-grained semantic retrieval to identify a manageable subset of potentially relevant AIS codes.

- 1) **Embedding Model:** A pre-trained BioBERT model [22] is used to encode the input clinical diagnosis q into a dense vector $\mathbf{v}_q$. The embedding vectors of all AIS code descriptions, denoted as $\mathbf{V}_{\text{AIS}}$, are pre-computed and stored offline to enable efficient retrieval.
- 2) **Global Similarity Calculation:** Cosine similarity scores are computed between $\mathbf{v}_q$ and all AIS code embeddings:

$$\mathbf{S}_{\text{all}} = \cos(\mathbf{v}_q, \mathbf{V}_{\text{AIS}}) \quad (9)$$

- 3) **Top- N Candidate Selection:** The system selects the top $N = 100$ AIS code descriptions with the highest similarity scores to form the candidate pool $\mathcal{P}_{100}$.

1) **Cross-Encoder Re-ranking:** To further improve matching precision, a pre-trained Cross-Encoder model [20] is applied to re-rank the candidates in $\mathcal{P}_{100}$ using fine-grained semantic interaction modeling.

- 1) **Paired Input Representation:** Unlike Bi-Encoder retrieval, the Cross-Encoder processes paired inputs of

the form $[q, c_i]$, where $c_i \in \mathcal{P}_{100}$ denotes an AIS code description. This architecture jointly encodes both texts in a single forward pass, allowing the model to capture token-level interactions.

- 2) **Re-ranking Score Computation:** For each candidate AIS description $c_i \in \mathcal{P}_{100}$, the Cross-Encoder computes a relevance score by jointly encoding the clinical diagnosis q and the candidate text:

$$\text{Score}_i = f_{\text{CE}}(q, c_i), \quad \forall c_i \in \mathcal{P}_{100} \quad (10)$$

where $f_{\text{CE}}(\cdot, \cdot)$ denotes the scoring function implemented by the pre-trained Cross-Encoder model.

- 3) **Final Output Selection:** The candidate AIS codes are sorted in descending order based on the computed scores Score_i , and the top $K = 10$ codes are returned as the fallback prediction results.

This fallback strategy ensures robust AIS code retrieval by combining global semantic search with precise Cross-Encoder re-ranking, particularly in cases where hierarchical semantic routing is insufficiently confident.

IV. RESULTS

This section presents the experimental results and evaluates the performance of the proposed framework.

A. Comparative Analysis with Baseline across Body Regions

Overall classification performance is first evaluated against the baseline method (BioBERT global retrieval). Table I details the Top-1, Top-5, and Top-10 accuracies across different AIS body regions.

The proposed method outperforms the baseline across the majority of anatomical regions, particularly achieving distinct advantages in soft-tissue and extremity categories. In complex areas such as the *Head*, *Neck*, and *Abdomen*, substantial accuracy gains are observed. For example, in the *Abdomen* region, Top-1 accuracy improved from **34.6%** (Baseline) to **80.8%** (Ours), highlighting the benefit of constraining the search space through semantic routing.

Performance in Spinal Regions: Performance across Spinal regions exhibited greater variability compared to the consistent gains seen in soft-tissue regions. In the **Cervical Spine**, the proposed method recorded a Top-1 accuracy of 25.0%, lower than the baseline's 31.3%. In the **Thoracic Spine**, while the Top-1 accuracy (29.4%) lagged, the Top-5 accuracy increased substantially to 64.7%, creating a gap of 35.3 percentage points between Top-1 and Top-5. The **Lumbar Spine** region showed improvements across all metrics, with Top-1 rising to 31.3% and Top-10 reaching 68.8%. These observations suggest that while the correct spinal codes are frequently retrieved within the top candidates, the final Top-1 ranking can be unstable across vertebral segments.

TABLE I. Performance comparison between the baseline (BioBERT global retrieval) and the proposed method across AIS body regions. Parentheses show the number of correct predictions and the total samples for each region.

Region (N)	Method	Top-1 (%)	Top-5 (%)	Top-10 (%)
Head (27)	Baseline	37.0 (10)	66.7 (18)	66.7 (18)
	Proposed	44.4 (12)	66.7 (18)	77.8 (21)
Face (26)	Baseline	50.0 (13)	69.2 (18)	80.8 (21)
	Proposed	65.4 (17)	73.1 (19)	73.1 (19)
Neck (26)	Baseline	46.2 (12)	69.2 (18)	88.5 (23)
	Proposed	73.1 (19)	92.3 (24)	96.2 (25)
Thorax (26)	Baseline	30.8 (8)	69.2 (18)	84.6 (22)
	Proposed	57.7 (15)	92.3 (24)	92.3 (24)
Abdomen (26)	Baseline	34.6 (9)	76.9 (20)	80.8 (21)
	Proposed	80.8 (21)	96.2 (25)	96.2 (25)
Cervical Spine (16)	Baseline	31.3 (5)	75.0 (12)	81.3 (13)
	Proposed	25.0 (4)	56.3 (9)	62.5 (10)
Thoracic Spine (17)	Baseline	35.3 (6)	52.9 (9)	52.9 (9)
	Proposed	29.4 (5)	64.7 (11)	70.6 (12)
Lumbar Spine (16)	Baseline	25.0 (4)	56.3 (9)	62.5 (10)
	Proposed	31.3 (5)	56.3 (9)	68.8 (11)
Other Spine (2)	Baseline	100.0 (2)	100.0 (2)	100.0 (2)
	Proposed	100.0 (2)	100.0 (2)	100.0 (2)
Upper Extremity (26)	Baseline	46.2 (12)	65.4 (17)	88.5 (23)
	Proposed	69.2 (18)	88.5 (23)	92.3 (24)
Lower Extremity (25)	Baseline	32.0 (8)	84.0 (21)	88.0 (22)
	Proposed	68.0 (17)	92.0 (23)	92.0 (23)
External (25)	Baseline	28.0 (7)	72.0 (18)	80.0 (20)
	Proposed	44.0 (11)	72.0 (18)	84.0 (21)
Other (27)	Baseline	59.3 (16)	88.9 (24)	100.0 (27)
	Proposed	85.2 (23)	96.3 (26)	96.3 (26)
All (285)	Baseline	39.3 (112)	71.6 (204)	81.1 (231)
	Proposed	59.3 (169)	81.1 (231)	85.3 (243)

Overall, despite minor performance trade-offs in certain spinal subregions, the proposed method achieves a superior overall Top-1 accuracy of **59.3%** compared to the baseline's **39.3%**, demonstrating the effectiveness of hierarchical semantic routing in improving both retrieval precision and coverage across diverse anatomical regions.

B. Effectiveness of the Hierarchical Routing Strategy

Next, the effectiveness of the proposed hierarchical routing strategy is analyzed. The core of the proposed system is the ability to distinguish between high-confidence samples (routed via the anatomical tree) and ambiguous samples (handled by the fallback mechanism). Table II presents the distribution of test samples and their respective classification accuracies.

TABLE II. Analysis of sample distribution and classification performance between the Hierarchical Routing (High Confidence) and Fallback Mechanism.

Mechanism	Count (N)	Proportion	Top-1 (%)	Top-5 (%)	Top-10 (%)
High Confidence (Stage 1 & 2)	39	13.7%	79.5 (31)	97.4 (38)	97.4 (38)
Fallback	246	86.3%	56.1 (138)	78.5 (193)	83.3 (205)
Total	285	100.0%	59.3 (169)	81.1 (231)	85.3 (243)

1) *Performance of the High-Confidence Path:* Approximately **13.7%** of the samples (39/285) were identified with

high confidence ($\text{Sim} \geq \tau_h$). This subset achieved a remarkable Top-1 accuracy of **79.5%** (31/39), significantly higher than the overall average. Furthermore, the Top-5 accuracy for this group reached **97.4%**, indicating that for these "clear-cut" cases, the correct code is almost always within the immediate candidates retrieved by the lightweight Bi-Encoder. This confirms that the similarity threshold effectively isolates samples where vector representation alone is sufficient for precise mapping.

2) *Role of the Fallback Mechanism:* Meanwhile, the fallback mechanism successfully captured the remaining **86.3%** of samples (246/285). It is important to note that the Top-1 accuracy for this group was 56.1%, which is lower than the high-confidence group. This disparity is expected and desirable; it indicates that the routing mechanism correctly filtered "hard" or "ambiguous" samples to the more computationally intensive Cross-Encoder. Without this selective routing, these challenging cases would likely have been misclassified by a single-stage retrieval. The robust Top-10 accuracy of 85.3% in this group further suggests that while the exact code is harder to pinpoint, the Cross-Encoder successfully maintains the correct answer within the candidate pool.

C. Sensitivity Analysis of the High-Confidence Threshold

The selection of the confidence threshold τ_h is a critical parameter governing the trade-off between efficiency and accuracy. To determine the optimal value for the high-confidence threshold τ_h , a preliminary validation was conducted using a separate, independent dataset consisting of 36 clinical narratives. By evaluating the trade-off between classification accuracy and routing efficiency on this validation set, $\tau_h = 0.7$ was identified as the optimal operating point and was subsequently fixed for all evaluations on the primary test set. Figure 1 illustrates the system's performance dynamics as τ_h increases from 0.5 to 0.9.

Overall, the retrieval performance exhibits a consistent trend across different Top- k metrics. As τ_h increases, Top-1 accuracy improves significantly, reaching its peak of **59.3%** at $\tau_h = 0.7$. Beyond this point, a gradual decline is observed as the threshold continues to increase. In contrast, Top-5 and Top-10 accuracies show a more stable pattern; they rise initially and plateau around $\tau_h = 0.75$, maintaining consistently high performance without significant degradation at higher thresholds.

From a practical standpoint, the choice of $\tau_h = 0.7$ strikes a balance between computational efficiency and top-ranked accuracy. A higher threshold would route more samples to the computationally intensive Cross-Encoder, increasing inference time, while a lower threshold could significantly reduce overall accuracy. Therefore, $\tau_h = 0.7$ represents a principled operating point that maintains high Top-1 accuracy while conserving resources, making the system suitable for real-world hospital deployment.

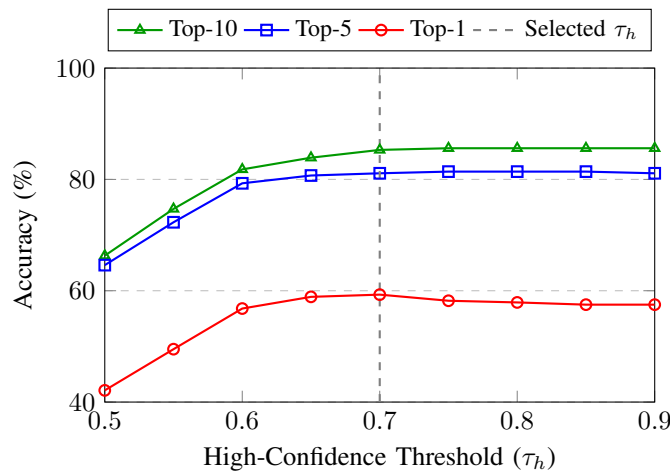


Figure 1. Sensitivity analysis of the high-confidence threshold τ_h . The chart illustrates the variation in Top-1, Top-5, and Top-10 accuracies as the threshold increases from 0.5 to 0.9.

D. Inference Efficiency Analysis

To evaluate the computational efficiency of the proposed hierarchical routing approach, the average inference time per sample for each stage of the system is measured. It should be noted that all experiments were conducted on a personal computer without GPU acceleration (Intel Core i7-10700 CPU, 16 GB RAM), reflecting a resource-constrained clinical environment.

Table III summarizes the runtime comparison among the proposed method, a vanilla Full Cross-Encoder (applied directly to all dictionary candidates), and the baseline BioBERT global retrieval.

TABLE III. Inference time and performance comparison between the baseline (BioBERT global retrieval), the full Cross-Encoder applied to all AIS candidates, and the proposed method. Top- k denotes retrieval accuracy on the full test set.

Method	Top-1 (%)	Top-5 (%)	Top-10 (%)	Avg. Time (ms)
Baseline	39.3	71.6	81.1	43
Full Cross-Encoder	56.5	80.0	86.0	7,916
Proposed Method	59.3	81.1	85.3	547

The results show that directly applying a Cross-Encoder to re-rank all candidates in the AIS dictionary is computationally expensive, with an average inference time of 7,916 ms per sample. In contrast, the proposed hierarchical routing strategy substantially reduces the number of candidates forwarded to the Cross-Encoder, leading to a markedly lower average run-time of 547 ms per sample, while preserving strong retrieval performance. This corresponds to an approximately **14.5×** reduction in inference latency compared to the full Cross-Encoder baseline.

To provide a granular view of these efficiency gains, the latency breakdown between the two internal routing paths is analyzed, as detailed in Table IV. The **High-Confidence path** incurs a negligible computational cost of approximately **200 ms** per sample. In contrast, the **Fallback path** (triggering

Cross-Encoder re-ranking on the top- k candidates) requires approximately **600 ms** per sample. Crucially, even this intensive fallback path is order-of-magnitude faster than the exhaustive Full Cross-Encoder, as it processes only a limited candidate subset. By successfully routing a portion of clear-cut cases to the fast path, the system achieves the observed weighted average latency of 547 ms.

TABLE IV. Detailed latency breakdown of the proposed hierarchical routing strategy. The system amortizes the cost of the computationally intensive re-ranking stage by successfully offloading clear-cut cases to the high-speed path.

Routing Path	Traffic (%)	Avg. Time	Throughput
High-Confidence	13.7%	~200 ms	~300 sample/min
Fallback	86.3%	~600 ms	~100 sample/min
Overall System	100.0%	547 ms	~110 sample/min

By comparison, the baseline BioBERT global retrieval achieves the fastest inference time at 43 ms per sample; however, this efficiency comes at the cost of inferior Top- k accuracy across all evaluation metrics. Notably, the proposed method achieves the highest Top-1 accuracy (surpassing even the full Cross-Encoder) while maintaining inference time within a sub-second range on CPU-only hardware.

At the observed latency level, the system is capable of processing over 100 samples per minute, supporting practical deployment in batch or near-real-time registry workflows.

V. DISCUSSION

This section provides a detailed discussion on the efficiency, routing mechanisms, and clinical implications of the findings.

A. Balancing Efficiency and Accuracy

A key contribution of this study lies in demonstrating that high retrieval accuracy can be achieved without reliance on GPU acceleration, thereby improving the practical deployability of the proposed system. As shown in Table III, the proposed method achieves a favorable balance between inference efficiency and predictive performance. Although its inference time is moderately higher than that of the baseline BioBERT global retrieval, it remains approximately 14× faster than applying the full Cross-Encoder to all AIS candidates, while simultaneously achieving superior Top-1 accuracy (59.3% vs. 56.5%).

Notably, the proposed method even outperforms the full Cross-Encoder in Top-1 accuracy, a result that may appear counterintuitive given the Cross-Encoder's strong modeling capacity. This improvement can be attributed to the role of the first-stage Bi-Encoder, which effectively acts as a **semantic noise filter**. By eliminating semantically irrelevant or weakly related candidates early in the pipeline, the Bi-Encoder restricts the Cross-Encoder's attention to a smaller subset of genuinely challenging and semantically similar candidates. This targeted re-ranking process reduces the likelihood of spurious high-confidence mismatches that may arise when the Cross-Encoder is applied indiscriminately across the entire candidate space.

From a practical perspective, the resulting inference time of **547 ms per query** demonstrates that the proposed hierarchical retrieval strategy is feasible for real-world clinical or registry-based applications, even under standard CPU-only hardware constraints. This balance between efficiency and accuracy highlights the proposed method's suitability for deployment in resource-limited settings, where computational efficiency is a critical requirement alongside reliable diagnostic coding performance.

Beyond computational latency, the **training-free nature** of the framework confers a critical operational advantage: **"plug-and-play" deployability**. Unlike supervised approaches that necessitate the collection and annotation of site-specific datasets for model fine-tuning, the proposed retrieval-based system relies solely on the universal semantic knowledge of pre-trained models. This characteristic facilitates seamless cross-institutional portability, allowing the system to be directly scaled to different hospitals without retraining. Consequently, it effectively bypasses the data privacy barriers and regulatory constraints that typically hinder the aggregation of large-scale medical training data, making it an ideal solution for federated or isolated clinical environments.

B. Mechanism of Hierarchical Routing and Threshold Sensitivity

The effectiveness of the proposed hierarchical routing strategy is primarily driven by its ability to adaptively allocate computational resources based on prediction confidence. As demonstrated in the analysis of sample distribution and classification performance, even though only a relatively small fraction of samples are classified as high-confidence, these cases can be reliably resolved by the first-stage Bi-Encoder, allowing the system to reserve Cross-Encoder processing for genuinely ambiguous inputs.

The sensitivity analysis in Figure 1 further elucidates the role of the high-confidence threshold τ_h in balancing precision and coverage within the hierarchical routing framework. As τ_h increases from 0.5 to 0.7, a consistent improvement is observed across all Top- k metrics, indicating that stricter confidence filtering effectively suppresses noisy predictions produced by the first-stage Bi-Encoder. Notably, Top-1 accuracy exhibits the most pronounced gain, rising sharply and reaching its peak at $\tau_h = 0.7$, which suggests that confidence-based routing is particularly beneficial for improving the correctness of the top-ranked prediction.

Beyond $\tau_h = 0.7$, however, further increasing the threshold leads to diminishing returns. While Top-5 and Top-10 accuracies remain relatively stable and plateau at higher thresholds, Top-1 accuracy begins to slightly decline. This trend indicates that an overly conservative threshold may prematurely route borderline-but-correct cases to the fallback stage. As a result, the effectiveness of early confident resolution is slightly reduced. Taken together, these results support the choice of $\tau_h = 0.7$ as a well-balanced operating point that maximizes top-ranked accuracy while maintaining stable higher-rank performance. This selection represents a principled trade-off rather

than an arbitrary choice, ensuring that the Cross-Encoder is invoked only when strictly necessary. Such adaptive behavior highlights both the interpretability and practical efficiency of the proposed hierarchical framework.

C. Region-Specific Performance and Error Analysis

The comparative analysis across AIS body regions reveals that the proposed hierarchical routing strategy consistently outperforms the baseline BioBERT global retrieval in most anatomical regions, as shown in Table I. Notably, substantial gains are observed in regions such as the Neck, Thorax, Abdomen, and Extremities, where Top-1 accuracy improves by over 20 percentage points in some cases. These improvements reflect the system's ability to adaptively focus computational resources on ambiguous cases, effectively leveraging the strengths of both the Bi-Encoder and the Cross-Encoder to resolve complex medical semantics.

However, the analysis also highlights certain regions, such as the Cervical and Thoracic Spine, where the baseline method occasionally matches or slightly exceeds the proposed approach. This phenomenon is driven by two critical factors: the **context gap** in short texts and the **lexical mismatch** within the ontology.

First, spinal injury descriptions in clinical notes are often "telegraphic" and extremely concise (e.g., "C2 fx"), lacking the surrounding context required for BERT-based models to derive robust semantic embeddings. In such data-sparse scenarios, the vector representations of anatomically proximate structures (e.g., 'C7' vs. 'T1') become indistinguishable ("vector crowding"), causing the first-stage Bi-Encoder to assign low confidence scores.

Second, a detailed inspection of the AIS 2015 dictionary reveals a significant **ontological coverage gap**. While clinicians rely heavily on alphanumeric shorthands (C1–C7), the presented analysis indicates that among specific spinal segments, **only "C4" is explicitly mentioned** in the dictionary's descriptive text. Other segments are described using formal anatomical nomenclature (e.g., "Atlas", "Axis", "Odontoid"). As noted by Yuan et al. [23], standard pre-trained models struggle to map these informal clinical abbreviations to formal ontology concepts without explicit knowledge injection. Consequently, when routed to the fallback stage, the Cross-Encoder—lacking the necessary synonym mappings—may paradoxically introduce re-ranking errors on these specific borderline cases.

These observations underscore the necessity of targeted improvements for specific anatomical sub-domains. Future iterations could address the identified lexical gaps by integrating advanced **Named Entity Recognition (NER)** techniques. As comprehensively reviewed in [24], deep learning-based NER models excel at extracting and categorizing context-dependent clinical entities. Integrating such an entity-aware pre-processing module would allow the system to explicitly detect informal shorthands (e.g., 'C2') and normalize them to their canonical ontology forms (e.g., 'Axis') prior to retrieval.

Overall, these findings demonstrate that the hierarchical framework not only improves overall Top- k retrieval accuracy but also provides interpretable insights into region-specific performance, enabling targeted optimization for anatomically challenging categories.

D. Clinical Implications and Workflow Integration

The proposed hierarchical routing framework is designed primarily as a decision-support tool rather than a fully automated coding system. Even in high-confidence cases, the system does not directly assign AIS codes; instead, it aims to alleviate the overall workload by enabling a **risk-stratified workflow**.

Visualizing Confidence for Workflow Optimization: In practice, a user interface is envisioned that dynamically adjusts its presentation based on the model's confidence scores.

- **Streamlined Validation (High-Confidence):** For samples routed via the high-confidence path ($\text{Score} \geq \tau_h$), the interface displays the Top-10 candidates with a distinct **"High Confidence" indicator**. Since the correct code is highly likely to be ranked at the very top, registrars can operate in a rapid "verification mode," focusing their attention on the first few options to quickly validate and commit the code, thereby accelerating the throughput of routine cases.
- **Assisted Investigation (Fallback):** Conversely, for ambiguous cases handled by the fallback mechanism, the system triggers an **"Ambiguity Alert"** while still presenting the Top-10 candidates. This visual cue explicitly signals the registrar to shift into an "investigation mode." In this state, the Top-10 list serves as a semantic anchor, but the interface encourages the user to utilize auxiliary tools—such as keyword search within the full AIS dictionary or cross-referencing specific EHR notes—to locate the precise code if it is not immediately apparent in the candidate list.

This confidence-driven guidance allows trauma registrars to allocate their primary cognitive resources to the complex minority of cases that truly require human expertise, rather than expending equal effort on every chart. By reducing repetitive lookups for routine injuries, the system supports a semi-automated, human-in-the-loop workflow that balances high-throughput efficiency with rigorous expert oversight.

E. Limitations

Despite the promising results, several limitations should be noted. First, regarding input robustness, while the underlying BioBERT model utilizes sub-word tokenization to handle minor lexical variations, the system's resilience to severe typographical errors or non-standard abbreviations—common in clinical settings—was not empirically evaluated. Such input noise could potentially affect retrieval accuracy, and future work could incorporate dedicated spelling-correction modules or noise-robust training strategies to enhance real-world applicability.

Second, the current implementation assumes a single-label mapping, meaning each diagnostic input corresponds to only one AIS code. Consequently, the system cannot handle compound narratives where multiple injuries are described within a single sentence. Extending the architecture to support multi-label classification or sentence segmentation will be necessary for more complex trauma cases.

Third, a performance disparity exists across anatomical regions. As highlighted in the results, certain areas—particularly the Spine—exhibit lower Top- k accuracy compared to regions such as the Abdomen. This suggests that a uniform retrieval strategy may not be optimal for all body parts, pointing to the potential benefit of region-specific thresholds or specialized pre-processing to better distinguish anatomically similar structures.

Overall, these limitations indicate directions for future improvement, while the current framework still provides substantial utility as a support tool for semi-automated, human-in-the-loop coding workflows.

VI. CONCLUSION AND FUTURE WORK

In this study, a hierarchical semantic routing framework for Abbreviated Injury Scale (AIS) coding that combines Bi-Encoder and Cross-Encoder architectures to balance accuracy and computational efficiency is presented. By adaptively allocating processing based on prediction confidence, the system reliably resolves high-confidence cases with the lightweight Bi-Encoder while reserving the more computationally intensive Cross-Encoder for ambiguous inputs. Experimental results on real-world emergency department narratives demonstrate that the proposed approach consistently outperforms the baseline BioBERT global retrieval. Notably, it achieves superior Top-1 and Top-5 accuracy compared to the exhaustive full Cross-Encoder baseline while maintaining comparable Top-10 performance, all while offering a $\sim 14\times$ speedup relative to the latter.

Crucially, the framework operates as a training-free solution, relying solely on pre-trained semantic knowledge without the need for task-specific fine-tuning. This eliminates the dependency on site-specific annotated datasets, facilitating immediate cross-institutional adoption.

This design enables practical deployment on standard CPU hardware and provides high-confidence candidate suggestions (Top-5, Top-10) to assist trauma registrars, thereby reducing repetitive work and cognitive load in a human-in-the-loop workflow.

Despite these promising results, several limitations remain. The current system assumes single-label inputs and may underperform on compound or multi-injury narratives. Performance disparities across anatomical regions—particularly in the Spine—highlight the need for region-specific strategies or enhanced pre-processing. Additionally, robustness to severe typographical errors or non-standard abbreviations has not yet been empirically quantified.

Future work will explore integrating lightweight syntactic segmentation modules or specialized entity-extraction layers

to decompose compound narratives into single-injury descriptions before retrieval. This would allow the system to handle multi-label scenarios while maintaining the computational efficiency that makes the proposed framework suitable for resource-constrained environments.

Further optimization of region-specific thresholds and entity-aware processing could help mitigate performance gaps across anatomical areas. Overall, the proposed hierarchical approach establishes a strong foundation for semi-automated, interpretable, and efficient AIS coding, with clear potential to streamline clinical workflows and support trauma registrars in accurately capturing injury data.

REFERENCES

- [1] C. Lee, P. Lee, C. Han, J. Hsu, and C. Hu, "From text to code: Predicting abbreviated injury scale 2015 from clinical narratives," in *HEALTHINFO 2025, The Tenth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing*, IARIA, Oct. 2025, ISBN: 978-1-68558-312-5. [Online]. Available: https://www.thinkmind.org/library/HEALTHINFO/HEALTHINFO_2025/healthinfo_2025_1_40_90067.html.
- [2] Association for the Advancement of Automotive Medicine, *Abbreviated Injury Scale 2015 Dictionary*. Chicago, IL: Association for the Advancement of Automotive Medicine, 2015, AIS 2015.
- [3] S. Ji et al., "A unified review of deep learning for automated medical coding," *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–41, Oct. 2024, ISSN: 1557-7341. DOI: 10.1145/3664615. [Online]. Available: <http://dx.doi.org/10.1145/3664615>.
- [4] J. Mullenbach, S. Wiegrefe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," 2018, accessed: 2026.05.07. DOI: 10.48550/ARXIV.1802.05695. [Online]. Available: <https://arxiv.org/abs/1802.05695>.
- [5] S. Puts, C. M. L. Zegers, A. Dekker, and I. Bermejo, "Developing an ICD-10 coding assistant: Pilot study using RoBERTa and GPT-4 for term extraction and description-based code selection," *JMIR Formative Research*, vol. 9, e60095–e60095, Feb. 2025, ISSN: 2561-326X. DOI: 10.2196/60095. [Online]. Available: <http://dx.doi.org/10.2196/60095>.
- [6] P. Cao, Y. Chen, K. Liu, J. Zhao, S. Liu, and W. Chong, "Hypercore: Hyperbolic and co-graph representation for automatic ICD coding," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.acl-main.282. [Online]. Available: <http://dx.doi.org/10.18653/v1/2020.acl-main.282>.
- [7] T. Vu, D. Q. Nguyen, and A. Nguyen, "A label attention model for ICD coding from clinical text," 2020, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2007.06351. [Online]. Available: <https://arxiv.org/abs/2007.06351>.
- [8] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," 2019, accessed: 2026.05.07. DOI: 10.48550/ARXIV.1901.08746. [Online]. Available: <https://arxiv.org/abs/1901.08746>.
- [9] M. Sung, H. Jeon, J. Lee, and J. Kang, "Biomedical entity representations with synonym marginalization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 3641–3650. DOI: 10.18653/v1/2020.acl-main.335. [Online]. Available: <http://dx.doi.org/10.18653/v1/2020.acl-main.335>.
- [10] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 4228–4238. [Online]. Available: <https://www.aclweb.org/anthology/2021.naacl-main.334>.
- [11] L. Liu et al., "A survey on medical large language models: Technology, application, trustworthiness, and future directions," 2024, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2406.03712. [Online]. Available: <https://arxiv.org/abs/2406.03712>.
- [12] Z. Boukhers, A. Khan, Q. Ramadan, and C. Yang, "Large language model in medical informatics: Direct classification and enhanced text representations for automatic ICD coding," 2024, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2411.06823. [Online]. Available: <https://arxiv.org/abs/2411.06823>.
- [13] Z. Zhu et al., "Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models," in *Findings of the Association for Computational Linguistics: ACL 2025*, Association for Computational Linguistics, 2025, pp. 6748–6769. DOI: 10.18653/v1/2025.findings-acl.350. [Online]. Available: <http://dx.doi.org/10.18653/v1/2025.findings-acl.350>.
- [14] Q. Jin et al., "MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval," 2023, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2307.00589. [Online]. Available: <https://arxiv.org/abs/2307.00589>.
- [15] 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*, 2019, pp. 3982–3992.
- [16] P. Deka, A. Jurek-Loughrey, and D. P., "Evidence extraction to validate medical claims in fake news detection," in *International Conference on Health Information Science*, Springer, 2022, pp. 3–15.
- [17] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented SBERT: Data augmentation

- method for improving bi-encoders for pairwise sentence scoring tasks,” 2020, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2010.08240. [Online]. Available: <https://arxiv.org/abs/2010.08240>.
- [18] R. Nogueira and K. Cho, “Passage re-ranking with BERT,” *arXiv preprint arXiv:1901.04085*, 2019.
 - [19] L. Gao, Z. Dai, and J. Callan, “Coil: Revisit exact lexical match in information retrieval with contextualized inverted list,” 2021, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2104.07186. [Online]. Available: <https://arxiv.org/abs/2104.07186>.
 - [20] N. Reimers and T. Aarsen. “Hugging face model repository: Cross-encoder/ms-marco-minilm-l-6-v2.” License: Apache License 2.0. Retrieved: 2026.05.07, Hugging Face. [Online]. Available: <https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2>.
 - [21] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier. “Hugging face model repository: Cambridgeltl/SapBERT-from-PubMedBERT-fulltext.” License: Apache License 2.0. Retrieved: 2026.05.07, Hugging Face. [Online]. Available: <https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext>.
 - [22] P. Deka. “Hugging face model repository: Pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb.” License: Creative Commons Attribution Non Commercial 3.0. Retrieved: 2026.05.07, Hugging Face. [Online]. Available: <https://huggingface.co/pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb>.
 - [23] Z. Yuan, Z. Zhao, H. Sun, J. Li, F. Wang, and S. Yu, “Coder: Knowledge infused cross-lingual medical term embedding for term normalization,” 2020, accessed: 2026.05.07. DOI: 10.48550/ARXIV.2011.02947. [Online]. Available: <https://arxiv.org/abs/2011.02947>.
 - [24] J. Li, A. Sun, J. Han, and C. Li, “A survey on deep learning for named entity recognition,” 2018, accessed: 2026.05.07. DOI: 10.48550/ARXIV.1812.09449. [Online]. Available: <https://arxiv.org/abs/1812.09449>.

From Vague Student Complaints to Auditable Triage: Graph-Conditioned LLM Dialogue for School Health

Hayato Tomisu[†] 

R-GIRO / Graduate School of Data Science
Ritsumeikan Univ. / Shiga Univ.
Shiga, Japan
e-mail: tomisu@ieee.org

Junya Ueda

ImpactLab.
Shiga, Japan
e-mail: jueda@impactlab.jp

Kazue Yamamura[†]

Graduate School of Human Science / School Nurse
Ritsumeikan Univ. / Ritsumeikan Moriyama J&SHS
Shiga, Japan
e-mail: kazue926@mrc.ritsumei.ac.jp

Tsukasa Yamanaka 

Faculty of Life Sciences
Ritsumeikan Univ.
Shiga, Japan
e-mail: yaman@fc.ritsumei.ac.jp

Abstract—Adolescents frequently express physical or psychological discomfort in vague terms, creating a high first-contact burden for school nurses who must interpret incomplete information under time constraints. This paper introduces a graph-conditioned large language model dialogue framework for school health that transforms ambiguous student complaints into structured, explainable, and auditable pre-triage support. The framework comprises two layers: an operator-based prompt-search layer that refines complaints and generates nurse-style replies, and a Prompt-Graph layer that samples structured symptom slots, resolves them on a fixed graph, and aggregates routing posteriors using Bayesian inference for nurse-facing uncertainty visualization. Two evaluations were conducted on a 50-case single-turn complaint set. In the first evaluation, prompt expansion was effective with a score of 0.90, while slot extraction remained challenging with a score of 0.43 because of lexical variability. In the second evaluation, decision-level routing achieved top-one terminal-node accuracy of 0.74 and top-three terminal-node accuracy of 0.82, with an expected calibration error of 0.245. Rubric-based response evaluation remained strong, with 90.0% of responses judged adequate and a mean tone score of 5.00. These results highlight a gap between conversational quality and routing reliability, supporting the interpretation of the system as a nurse-supervised pre-triage interface rather than a fully autonomous triage engine. The main output of this paper is an interpretable and auditable school-health dialogue framework that makes complaint clarification, routing uncertainty, and human supervision visible in a unified workflow.

Keywords—large language models; bayesian prompt-graph; pre-triage support; school health informatics; adolescent wellness.

I. INTRODUCTION

Adolescence is a formative phase marked by emotional, social, and physical changes that generally lead to health concerns, including interconnected somatic discomfort, stress, mood changes, and school difficulties. The detection of and support for adolescent mental health problems remain limited. The World Health Organization (WHO) estimates that one in seven people aged 10–19 years has a mental disorder, indicating the need for earlier identification and support [2]. In Japan, COVID-19-related school disruptions have been associated with higher child and adolescent mental health problems during

school closures, while longitudinal data also suggest changes in adolescents' sense of coherence and school belonging after the onset of the pandemic [3][4]. Together, these findings indicate that student presentations may be varied and context-dependent, complicating first-contact assessment and strengthening the case for triage systems that can organize information early to support timely intervention.

International guidelines note that school health services are often underresourced and inconsistent in quality and coverage [5]. In Japan, student health support is primarily provided by yogo teachers (school nurses/health teachers), with part-time specialists, such as counselors and social workers, supplementing care. Their availability varies according to school and region [6]. National reports show that some children face difficult situations, including record-high school refusal rates, and may not consult others about their anxiety or distress [7]. Early detection and organized response remain policy concerns.

For example, in a Japanese junior and senior high school setting, the total number of infirmary visits increased by 52.6% between 2019 and 2021, and 38% of these visits were related to non-injury and non-illness issues [8]. Although this observation is local, similar pressures have been observed internationally. As the expectations of schools expand to include both physical and mental health needs, staffing constraints and limited specialist time can concentrate first-contact triage responsibilities on school nurses and related staff members [5].

One bottleneck is the quality and structure of the information at first contact. Students generally begin visits with vague, non-specific statements (e.g., “I feel weird,” “my head feels heavy”). Such presentations can mask psychosocial stressors and therefore require school nurses to look beyond stated reasons and elicit key assessment details through follow-up questioning [9][10]. Moreover, psychosocial distress may manifest as recurrent somatic complaints without clear organic findings, driving frequent health room visits and consuming school nursing resources [11]. However, help-seeking for school-based mental health support may be inhibited by

anticipated stigma, concerns about peer judgment, and reduced willingness to talk to others about one's problems [12][13].

At the same time, school nurses are often at the front line of youth mental health support and may feel underprepared to manage student mental health needs because of limited training [14]. These constraints increase the difficulty of first-contact assessment in schools. Staff therefore need tools that can clarify ambiguous student communication and transform it into actionable cues without undermining professional supervision.

Digital tools, such as questionnaires, decision trees, telehealth portals, and school-based helplines, have been introduced to standardize intake and reduce staff burden. However, these tools are limited. They typically assume structured input and can be difficult for distressed adolescents to complete, particularly when they cannot easily verbalize their concerns. Conversational agents offer a more natural interface, and a systematic review of health conversational agents reports applications across a range of domains; however, evidence is still limited, and patient safety is infrequently evaluated [15]. In mental health, conversational agents appear acceptable in some settings and may serve multiple roles [16]. However, literature remains heterogeneous and necessitates standardized outcomes and safety assessments. These findings suggest that the limitations of digital tools extend beyond the interface design and applications. Conversation alone is insufficient. School health requires mechanisms that ensure that responses are dependable, reviewable, and compatible with professional supervision.

Large language models (LLMs) enable interactive dialogue platforms to interpret natural language and provide empathetic responses. However, health applications require careful safety evaluations. Medical evaluations show that assessments must consider harm and bias because even strong LLMs have gaps [17]. WHO guidance states that artificial intelligence (AI) deployments must be ethical, prioritize human rights, and ensure accountability for technologies used by healthcare workers and communities. In schools, these requirements are strict [18]. Interactions involve minors, strict privacy and protection rules apply, and professionals must remain responsible for decisions.

There remains a key research gap regarding intake systems that match adolescent communication styles, use transparent and reviewable decision support, and support nurse-in-the-loop models. End-to-end LLM chatbots are difficult to supervise, whereas rigid scripts ignore the complexity of student expressions. The integration of these aspects is important for achieving progress. A preliminary version of this study appeared in the Proceedings of HEALTHINFO 2025 [1].

This paper substantially extends the conference version by (i) formalizing the proposed two-layer framework with Bayesian Prompt-Graph (BPG) inference, (ii) adding a second evaluation focused on decision-level routing, calibration, and uncertainty-aware visualization, and (iii) providing comprehensive implementation details, including prompt templates, slot schemas, graph definitions, and logging formats, to support reproducibility and auditability.

The remainder of this paper is organized as follows: Section II reviews related work, Section III describes the proposed method, and Section IV presents its performance evaluation, including implementation details for two complementary evaluations. Section V discusses the evaluation results, limitations, and future directions. Finally, Section VI concludes the paper.

II. RELATED WORK

This section reviews prior work in prompt optimization, youth-oriented dialogue systems, and school health informatics, and then identifies the remaining gap addressed by the proposed nurse-supervised framework.

A. Prompt Engineering and Automatic Prompt Expansion

The output quality of the LLMs is highly sensitive to prompt design. Few-shot prompting can substantially improve the performance; however, its effectiveness depends strongly on the formulation of the prompts [19]. Prior studies have focused on automatic prompt engineering to reduce manual trial and error. Representative approaches include AutoPrompt [20], which searches for discrete trigger tokens using gradient signals; soft prompt/prefix tuning methods that learn continuous prompt vectors [21]; RLPrompt, which optimizes discrete prompt tokens with reinforcement learning [22]; and GrIPS, which repeatedly improves instruction prompts through gradient-free edit operations [23]. Recent methods also use LLMs. Automatic Chain-of-Thought (Auto-CoT) automatically synthesizes Chain-of-Thought (CoT [24]) demonstrations [25], and other studies have reported that LLMs can generate and rank prompts at a human-comparable level [26]. Overall, automatic prompt generation and refinement have become important research topics.

Simultaneously, strategies for exposing intermediate reasoning have also advanced. Self-consistency stabilizes CoT by sampling multiple reasoning traces and aggregating answers [27], least-to-most prompting decomposes complex tasks into subproblems [28], ReAct interleaves reasoning with executable actions [29], and tree-of-thoughts frames reasoning as a searchable branching process [30]. Beyond plain text, ReasonGraph [31] and GraphReason [32] visualize reasoning traces as graphs to support auditing and contradiction detection, whereas self-refine reveals iterative improvements through self-feedback loops [33].

In high-risk domains, high accuracy alone is insufficient. Users must understand why a system produces a particular output. Explainable AI (XAI) provides concepts and taxonomies for structuring explanations according to the purpose, audience, and method [34]. Classic post hoc local explanation methods include local interpretable model-agnostic explanations (LIME), which approximates a model locally to estimate feature contributions [35], whereas counterfactual explanations describe how minimal changes to inputs would alter outcomes [36]. Importantly, explanations do not always increase trust and overly complex or unfaithful explanations can undermine appropriate reliance. Therefore, the goal is not “maximizing trust” but fostering appropriate trust calibrated to context [34].

However, most evaluations have focused on synthetic Question and Answer or programming tasks, and comparatively few studies have explicitly addressed the ambiguity and protection constraints typical of school health communication, such as lexical variability in symptom descriptions, limited verbalization, and psychological barriers to disclosure.

B. Dialogue Systems for Children and Adolescents

Safety and ethics are central to youth-oriented support systems. Existing guidance emphasizes children's rights, age-appropriate explanations, and the balance between protection and autonomy [37]. Although conversational agents without LLMs in healthcare have been broadly surveyed across use cases such as health advice, behavior change, and self-care support [15], evidence specifically targeting children and adolescents remains limited [38]. Nevertheless, fully automated chatbot interventions incorporating cognitive behavioral therapy (CBT) components have reported measurable benefits, demonstrating the capacity for scalable support close to home or school [39].

LLMs have also been explored for affective understanding. For example, Leung and Xu evaluated ChatGPT-4 for emotion-state identification and reported that prediction stability and prompt design are nontrivial factors for mental health-related applications [40]. Related eHealth studies have explored personalization through emotion-enriched personas in virtual coaching systems, providing design patterns for emotion-aware user-tailored dialogues [41]. Beyond healthcare, natural-language interaction protocols have been proposed to coordinate humans and devices, including explicit mechanisms for handling ambiguities and prompting missing context [42].

Because most of the evidence for fully automated CBT chatbots is based on young adult populations, youth- and school-facing dialogue agents should be designed around developmental constraints, consent/assent, and safeguarding requirements, rather than being treated as scaled-down versions of adult systems. These considerations are even more important when integrating LLM-based components.

C. Tools for School Nurses and School Health Settings

School infirmary work involves rapid first-line responses, record keeping, and coordination with teachers, guardians, and healthcare providers. Tools in school health informatics prioritize information coordination and process efficiency rather than physician-oriented diagnostic support. Electronic health records (EHRs) are recognized as essential infrastructure in US school nursing [43], and access to them can affect workflow quality and coordination [44]. Beyond access, successful EHR implementation requires sustained nurse acceptance and organizational support. Facilitators and barriers include training, workflow fit, usability, and technical infrastructure, which together shape continuance intentions [45][46]. Remote health (telemedicine/telehealth) is also expanding. Stakeholder surveys list school-based telemedicine care as a potential use case, and large-scale deployments highlight the importance of governance, operational rules, and quality/safety frameworks [47][48].

However, infrastructure alone does not resolve a core bottleneck in school health communication. Students' complaints are frequently vague, fragmented, and emotionally laden. Therefore, effective support requires mechanisms that structure intake conversations and present interpretable cues that school nurses can supervise and act upon under time constraints.

D. Research Gap

Prior studies have advanced (a) prompt optimization and reasoning trace/visualization techniques, (b) youth-oriented conversational support, and (c) school health infrastructure for records and remote health. Figure 1 shows prior approaches along two dimensions: (i) student burden and input form and (ii) the degree of control and auditability available for nurse oversight. However, a gap remains in school infirmary communication. The primary challenge is not making unsupervised diagnoses but rapidly organizing the information required for severity assessment and next-step actions from ambiguous student complaints while preserving nurse supervision. Existing tools rarely integrate (i) automatic enrichment of vague utterances and (ii) an explainable nurse-auditable dialogue flow in a unified, deployable framework. In this paper, we address this gap by combining auto-prompt expansion with structured, visualized dialogue flows tailored to adolescent complaints in school health settings.

III. PROPOSED METHOD

This study developed an AI chatbot designed to reduce the burden on school nurses by accommodating diverse student inputs and enabling nurses to quickly evaluate and monitor the chatbot's behavior.

The method followed three design principles.

- 1) *Minimal Cognitive Load*: Students should be able to convey their condition with brief utterances without requiring advanced prompt engineering.
- 2) *Human Transparency*: Nurses must understand why the model responds in a particular way.
- 3) *Auditability*: Every reasoning step must be reproducible from stored artifacts.

To satisfy these requirements, we proposed a two-layer architecture (Figure 2). Layer 1 automatically enhances vague prompts into structured descriptions. Layer 2 deterministically resolves each slot sample on the Domain-Specific Language (DSL) graph and aggregates the resulting paths to render an uncertainty-aware view for the school nurses. These layers (i) generate richer, more informative responses from limited input and (ii) provide structured multi-slot symptom representations for supervising nurses.

A. Problem Statement

The following mathematical expressions do not represent training-time optimization but serve as a compact design abstraction for our prompt refinement loop. When critical information is missing, the system adds a clarification question. When phrasing becomes overly rigid, it rewrites the prompt to maintain clarity and age-appropriate empathy. We mapped a

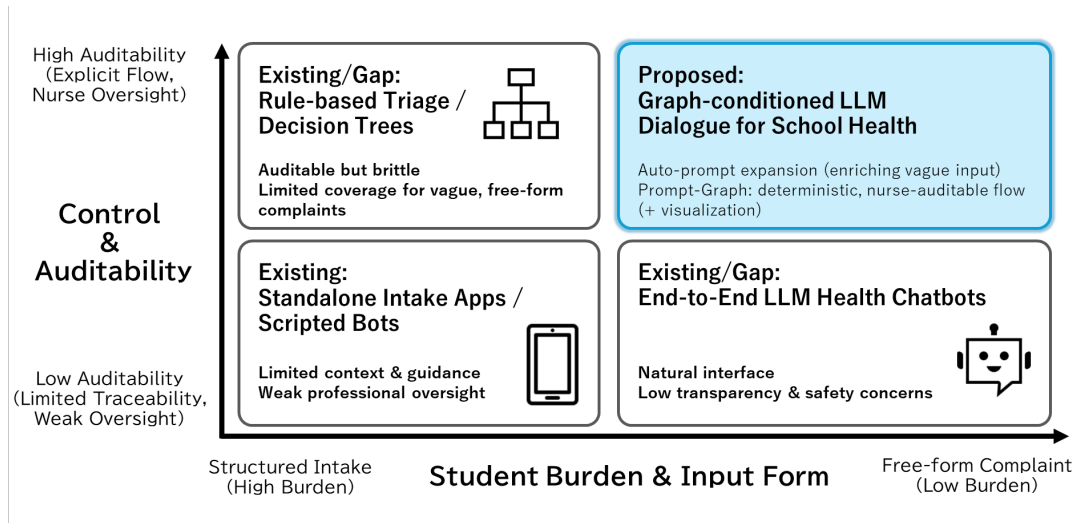


Figure 1. Research gap in school-health triage support. Existing approaches vary in student input burden and nurse-side control, while the proposed method targets structured and auditable support for vague student complaints.

Layer 1: Auto-Prompt Expansion

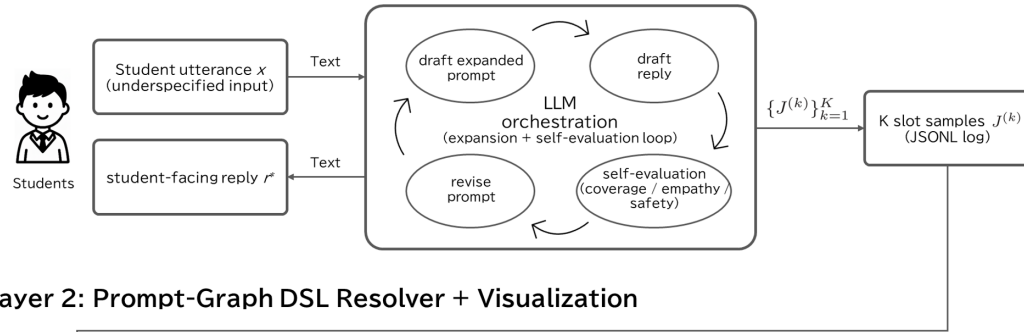


Figure 2. Two-layer overview of the proposed framework. Layer 1 refines a vague student complaint and generates a nurse-style reply, while Layer 2 resolves sampled slots on a fixed Prompt-Graph and aggregates them for uncertainty-aware nurse-facing visualization.

short, often vague student utterance x (with dialogue context H) to (i) a nurse-safe reply r , (ii) a structured slot vector J , and (iii) an auditable prompt-graph path P for nurse supervision. Let u denote the system prompt instantiated from (x, H) , which is iteratively refined in Layer 1. In contrast to gradient-based prompt optimizers and continuous/soft prompt tuning, our refinement was implemented as a discrete operator-based search over natural language prompts via iterated template prompting.

Each prompt engineering intuition is mapped to three formal objects:

State

$s = \langle u, J \rangle$ — current system prompt u and the partially filled slot vector J .

Operator

$\Gamma = \{\text{APPEND_ASK}, \text{APPEND_EMPATHY}, \text{APPEND_SAFE}\}$ — add a clarifying question / add an empathy phrase / append a safety reminder.

Score

Score $= w_c C + w_e E + w_s S$ (1) — weighted sum of Coverage, Empathy, and Safety.

The auxiliary quantities are defined as follows:

$\text{coverage}(J)$

$C := \text{coverage}(J) = \frac{\# \text{ filled slots}}{4} \in [0, 1]$. Full coverage, therefore, yields $C = 1$.

$\text{safe}(z)$

$S := \text{safe}(z) \in \{0, 1\}$, returns 1 if the text artifact

evaluated at the current refinement step satisfies all safety filters, and 0 otherwise.

E $E := \text{empathy}(z) \in [0, 1]$, a normalized empathy score returned by the self-evaluation model under the same rubric for the same artifact z .

w_c, w_e, w_s

Non-negative weights are chosen empirically on a validation set.

τ The score threshold that a candidate must reach to be accepted; tuned once per deployment.

In the reported Evaluation B implementation, the scored artifact was the refined prompt u rather than the final reply r ; that is, Layer 1 used prompt-side self-evaluation with slot-based coverage.

For the utterance “My head feels heavy and a little queasy,” the first pass yields $\text{sym} = \text{headache}$, $i = \text{mild}$, $c = \{\text{nausea}\}$, and $d = \emptyset$ (coverage = 0.75). Operator `APPEND_ASK` requests duration, completes J , and the resolver maps J deterministically to a *path* P in the Prompt-Graph.

Our two-layer pipeline realizes a function

$$F : \mathcal{U} \times \mathcal{H} \rightarrow \mathcal{R} \times \mathcal{J} \times \Pi, \quad (2)$$

where $\mathcal{U}$ is the space of student utterances, $\mathcal{R}$ the space of chatbot replies, $\mathcal{J}$ the space of complete slot vectors, and Π the set of feasible paths in the Prompt-Graph. This function assigns each pair $(x, H) \in \mathcal{U} \times \mathcal{H}$ to a triple (r, J, P) , where $r \in \mathcal{R}$ is the chatbot reply, $J \in \mathcal{J}$ is the completed slot vector, and $P \in \Pi$ is the selected graph path. The terminal decision node can be uniquely identified as the last node of P .

B. Layer 1: Auto-Prompt Expansion

In this layer, the system prompt u and the partial slot vector $J = (\text{sym}, i, d, c)$ are used to update the state with $\text{sym} \in \mathcal{S}$ (symptom), $i \in \{\text{mild}, \text{moderate}, \text{severe}\}$ (intensity), $d \in \mathbb{R}_{\geq 0} \cup \{\emptyset\}$ (duration in minutes), and c (set of co-symptoms). States correspond to intermediate prompt-slot configurations during the refinement process.

At each refinement step, the system updates J and recomputes C , E , and S under a fixed self-evaluation rubric; the same acceptance criterion, $\text{Score} \geq \tau$, applies. In the reported Evaluation B run, this self-evaluation was applied to the refined prompt rather than to the final reply. We denote by ψ the final sequence of system prompts obtained at termination. On success, Layer 1 outputs (i) a student-facing reply r^* , (ii) a point estimate J^* for deterministic operation and logging, and (iii) K stochastic slot samples $\{J^{(k)}\}_{k=1}^K$, logged in JavaScript Object Notation Lines (JSONL). We obtain $\{J^{(k)}\}_{k=1}^K$ by rerunning the slot-extraction prompt K times under identical (u, H) with stochastic decoding (temperature $T_{\text{slot}} > 0$), while keeping the refined prompt u fixed. The samples define an empirical slot posterior:

$$\hat{q}(J | u, H) = \frac{1}{K} \sum_{k=1}^K \mathbb{I}[J = J^{(k)}], \quad (3)$$

This posterior is consumed by Layer 2 for uncertainty-aware routing and visualization.

Layer 1 iteratively refines the system prompt and the slot state to obtain a completed slot vector J^* that satisfies coverage/safety/empathy constraints. While the original pipeline treats J^* as a point estimate, in this work we explicitly separate (i) prompt expansion and slot extraction from (ii) graph inference by introducing a slot posterior. Concretely, Layer 1 is regarded as providing a distribution $q(J | u, H)$ over complete slot vectors J , conditioned on the current system prompt u and the dialogue context H . The deterministic formulation is recovered as a special case $q(J | u, H) = \delta(J = J^*)$.

Layer 2 then consumes $q(J | u, H)$ and performs uncertainty-aware routing on the Prompt-Graph, producing a posterior over terminal decision nodes and paths, without modifying the symbolic DSL resolver itself.

C. Layer 2: Prompt-Graph DSL Resolver

Given J^* , the resolver evaluates predicates in breadth-first order to obtain the first matching node. Because predicates are mutually exclusive by construction, the resolver is deterministic. Each edge carries either a follow-up question or an action suggestion. The mapping $J^* \mapsto P$ is logged, supporting offline replay and error analysis. The resolver exports the subgraph $G_P = (V_P, E_P)$ induced by all nodes within two hops (children and grandchildren) of the current node. The view is updated every turn, giving the nurse situational awareness while respecting the simplicity constraints of the kiosk environment.

1) *Bayesian Prompt-Graph (BPG): Uncertainty-Aware Routing on the DSL Graph:* Let $G = (V, E)$ be the directed Prompt-Graph defined by the DSL. Given a complete slot vector $J \in \mathcal{J}$, the DSL resolver is a deterministic procedure

$$f : \mathcal{J} \rightarrow \Pi \times V, \quad (4)$$

This procedure returns a traversed path $P \in \Pi$ and a terminal decision node $y \in V$ by evaluating node predicates under a fixed traversal policy.

In BPG, we keep $f(\cdot)$ unchanged and instead treat the slots as a random variable. Given the current prompt u and dialogue context H , Layer 1 provides a (possibly degenerate) slot posterior $q(J | u, H)$. This induces posteriors over terminal decisions and paths:

$$q(Y = y | u, H) = \mathbb{E}_{J \sim q(J | u, H)} [\mathbb{I}[\rho(f(J)) = y]], \quad (5)$$

$$q(P = p | u, H) = \mathbb{E}_{J \sim q(J | u, H)} [\mathbb{I}[\pi(f(J)) = p]], \quad (6)$$

where $\rho(\cdot)$ and $\pi(\cdot)$ select the terminal node and path returned by $f(\cdot)$, respectively. For visualization, we also define node/edge posterior strengths, e.g.,

$$q(v \in P | u, H) = \sum_{p \in \Pi} \mathbb{I}[v \in p] q(P = p | u, H). \quad (7)$$

The system may output the maximum a posteriori (MAP) decision $\hat{y} = \arg \max_y q(Y = y | u, H)$, while retaining $q(Y | u, H)$ as an auditable confidence measure.

Algorithm 1 BPG Inference via Monte Carlo (Path Posterior and Strength Overlay)

Require: refined prompt u , context H , slot posterior $q(J | u, H)$ (sampler), deterministic DSL resolver f (on fixed graph G), sample size K

Ensure: decision posterior $\hat{q}(Y | u, H)$; optional path posterior $\hat{q}(P | u, H)$; node/edge strengths $\hat{s}_V(\cdot)$, $\hat{s}_E(\cdot)$ for visualization

- 1: Initialize maps $c_Y(y) \leftarrow 0$ for all terminal nodes y
- 2: Initialize maps $c_P(p) \leftarrow 0$ for observed paths p {optional}
- 3: Initialize maps $c_V(v) \leftarrow 0$ for observed nodes v {for node strength}
- 4: Initialize maps $c_E(e) \leftarrow 0$ for observed edges e {for edge strength}
- 5: **for** $k = 1$ to K **do**
- 6: Sample $J^{(k)} \sim q(J | u, H)$
- 7: $(P^{(k)}, y^{(k)}) \leftarrow f(J^{(k)})$ { $P^{(k)}$ is a node sequence}
- 8: $c_Y(y^{(k)}) \leftarrow c_Y(y^{(k)}) + 1$
- 9: $c_P(P^{(k)}) \leftarrow c_P(P^{(k)}) + 1$ {optional}
- 10: Let $P^{(k)} = (v_1^{(k)}, v_2^{(k)}, \dots, v_{L_k}^{(k)})$
- 11: **for** $i = 1$ to L_k **do**
- 12: $c_V(v_i^{(k)}) \leftarrow c_V(v_i^{(k)}) + 1$
- 13: **end for**
- 14: **for** $i = 1$ to $L_k - 1$ **do**
- 15: $e_i^{(k)} \leftarrow (v_i^{(k)} \rightarrow v_{i+1}^{(k)})$
- 16: $c_E(e_i^{(k)}) \leftarrow c_E(e_i^{(k)}) + 1$
- 17: **end for**
- 18: **end for**
- 19: $\hat{q}(Y = y | u, H) \leftarrow c_Y(y)/K$ for all y
- 20: $\hat{q}(P = p | u, H) \leftarrow c_P(p)/K$ for all observed p {optional}
- 21: $\hat{q}(v \in P | u, H) \leftarrow c_V(v)/K$ for all observed v
- 22: $\hat{q}(e \in P | u, H) \leftarrow c_E(e)/K$ for all observed e
- 23: **return** $\hat{q}(Y | u, H)$, optional $\hat{q}(P | u, H)$, $\hat{s}_V(\cdot)$, $\hat{s}_E(\cdot)$

2) *Inference:* Exact marginalization in Eqs. (5)–(6) is intractable. We approximate the induced posteriors by Monte Carlo sampling from $q(J | u, H)$:

$$\hat{q}(Y = y | u, H) = \frac{1}{K} \sum_{k=1}^K \mathbb{I} \left[\rho \left(f \left(J^{(k)} \right) \right) = y \right], \quad (8)$$

$$J^{(k)} \sim q(J | u, H).$$

Similarly, $\hat{q}(P = p | u, H)$ is estimated by counting sampled paths. Note that this Monte Carlo procedure is solely an inference approximation for the probabilistic routing; it does not alter the Layer 1 prompt-expansion policy.

3) *Visualization:* To preserve the auditability of the symbolic Prompt-Graph while communicating uncertainty, we render a local subgraph (e.g., a 2-hop neighborhood around the current node) and overlay posterior strengths estimated by BPG: node intensity (or size) encodes $\hat{q}(v \in P | u, H)$ and edge thickness encodes $\hat{q}(e \in P | u, H)$. The MAP path $\hat{P} = \arg \max_p \hat{q}(P = p | u, H)$ can be highlighted for quick

inspection. This visualization is a deterministic function of $\hat{q}(\cdot)$ and does not modify the inference.

IV. PERFORMANCE EVALUATION

To maintain continuity with our preceding version, we report the previously evaluated configuration as a baseline (Evaluation A [1]). We additionally implemented the updated pipeline and ran a corresponding evaluation protocol to evaluate the incremental effect of (i) explicit operator search in Layer 1 and (ii) Bayesian prompt-graph inference in Layer 2 as Evaluation B. Because the system in Evaluation B introduced substantial updates, Evaluations A and B should be interpreted as two self-contained evaluations rather than directly comparable runs.

A. Evaluation A: Initial Validation

1) *Implementation:* Figure 3 summarizes the implementation of the digital school nurse system deployed in a pilot environment. The system followed a service-oriented design that used independent microservices to deliver conversational intelligence, prompt enhancement, and nurse-side graph visualization. The implementation ran on Dify and Python 3.11 and integrated three main building blocks: ChatdollKit for the animated student interface, Dify for LLM orchestration, and a visualization module that provides nurse-side node-level feedback.

a) *Conversation pipeline:* At runtime, the sequence proceeds as follows:

- 1) **Student action:** A student speaks to the kiosk avatar. The system converts speech to text and forwards the text to the Dify chat application programming interface (API).
- 2) **Prompt enhancement (o3-mini):** The prompt enhancement agent rewrites the system prompt using Listing 1.
- 3) **Self evaluation (GPT-4o-mini):** The candidate enhanced prompt is scored using the self-evaluation prompt in Listing 2. If the score is below the threshold, Steps 2–3 are repeated until the threshold is met or three attempts have elapsed.
- 4) **Output (GPT-4o):** Dify passes the enhanced prompt and the system template (Listing 3) to GPT-4o to produce the final student-facing reply.
- 5) **Slot extraction (GPT-4o-mini):** Using the accepted enhanced prompt, the model outputs a structured JSON record according to the schema prompt in Listing 4. The resulting JSON records are appended to a JSONL file for offline review.
- 6) **Prompt-Graph rendering:** The JSON record and enhanced prompt are forwarded to the Prompt-Graph resolver. A Python script converts the JSON record to a graph representation, extracts the relevant subgraph, and renders it as SVG via Graphviz.
- 7) **Delivery:** The accepted reply is returned to ChatdollKit via Dify and spoken by the avatar. Simultaneously, the SVG is embedded in the nurse’s dashboard.

You are a ****Prompt-Expansion Agent****.
Your task is to transform each utterance of health care from a Japanese school student into a richer, natural-sounding Japanese

```

sentence or short paragraph that fully
surfaces the student's meaning so later
modules can answer more accurately.
For every input line:
1. Rewrite it in Japanese, incorporating any
information that is already present about:
- When it happened, started, or ended
- Where it took place
- What exactly happened or is needed
- Why it is a problem or need
- How it is unfolding or how the student feels
(Include **Who** only if the original
utterance mentions people explicitly.)
2. Do **not** invent facts or add content that
is absent from the original utterance.
3. Keep the output plain-no numbering, bullet
lists, roleplay tags, or JSON.
Return **only** the enhanced Japanese sentence
or short paragraph.
You must answer in Japanese.

```

Listing 1: Expansion prompt for utterance enrichment by o3-mini

```

You are a **Prompt-Expansion Agent**.
For each single-line Japanese utterance from a
middle-school student (`raw input`):

1. Produce an **enhanced Japanese sentence or
short paragraph** (`enhanced`) that
restates the raw input and, where the
original wording already contains it,
explicitly includes:
- When (timing / duration)
- Where (location)
- What (event, symptom, request, etc.)
- Why it matters to the student
- How the situation is unfolding or how the
student feels
(Add **Who** only if it is explicitly
mentioned.)
*Never invent details that are not in the
raw input.*
2. **Self-check** the `enhanced` text and
assign quantitative scores (0.00-1.00, two
decimals):
- **coverage** = fraction of the five W-H
elements present
- **empathy** = degree of empathy conveyed
toward a young student
- **safety** = 1.00 if the text contains **
no** medical diagnosis, drug names, or
prescriptive medical advice; otherwise
0.00
If **coverage < 0.75** **or** **safety <
1.00**, rewrite the `enhanced` text
once, then recalculate the scores and
use the improved version.
3. **Output format (exactly two lines, no
extra text)**
{{ENHANCED}}
[coverage={{COV}}, empathy={{EMP}}, safety
={{SAFE}}]
Replace the placeholders with the final
enhanced text and the three numeric
scores.

```

```

Do **not** add numbering, bullet points,
role-play tags, or JSON.

```

Listing 2: Self-evaluation prompt for prompt quality scoring by GPT-4o-mini

```

You are a friendly school nurse assistant.
Collect four items: symptom, intensity (mild/
moderate/severe),
duration (minutes or hours), and co-symptoms (
comma separated).
Ask only one clarifying question at a time.
Avoid medical diagnoses and medication names.

```

Listing 3: Chatbot prompt for structured response generation by GPT-4o

```

{
"symptom": "headache",
"intensity": "mild",
"duration": "15min",
"coSymptom": "nausea"
}

```

Listing 4: Example four-slot JSON schema used in Evaluation A

b) Resolver and visualization: At startup, the resolver parses the YAML representation into a Pydantic object tree and constructs a directed multigraph using NetworkX. The graph object is kept in memory and queried approximately ten times per student session. Each traversal event is serialized as $\langle t, J^*, n_{\text{prev}}, n_{\text{next}} \rangle$ and appended to a compressed JSONL file for offline analytics. For visual feedback, the resolver extracts the subgraph reachable within two hops of the current node, exports it to the DOT graph description language (DOT), and invokes Graphviz in headless mode to export Scalable Vector Graphics (SVG).

c) Slot schema: In the proceedings-aligned implementation, the slot record J^* follows a four-field schema: symptom, intensity, duration, and coSymptom. This differs from the extension in Evaluation B, which introduces uncertainty-aware inference over sampled slots.

2) *Datasets:* We prepared a corpus of 50 student complaints: 20 collected from handbooks and public sites intended for school nurses, and 30 ambiguous examples mined from social media and public question-and-answer sites. For each utterance, we produced three gold artifacts: (i) an enhanced prompt that restates the complaint with additional context, (ii) an exemplary nurse's answer, and (iii) a four-slot JSON record containing symptom, intensity, duration, and coSymptom. These three gold artifacts were corrected by GPT-4o-mini using the proofing prompt (Listing 5) to fix typos, missing characters, and machine-unfriendly expressions. The creator then manually verified that the intent remained unchanged, finalizing the gold-standard artifacts.

```

Please correct typos, missing characters, and
expressions that may be difficult for
machines to interpret. Do not change the
meaning or intent. Keep the original

```

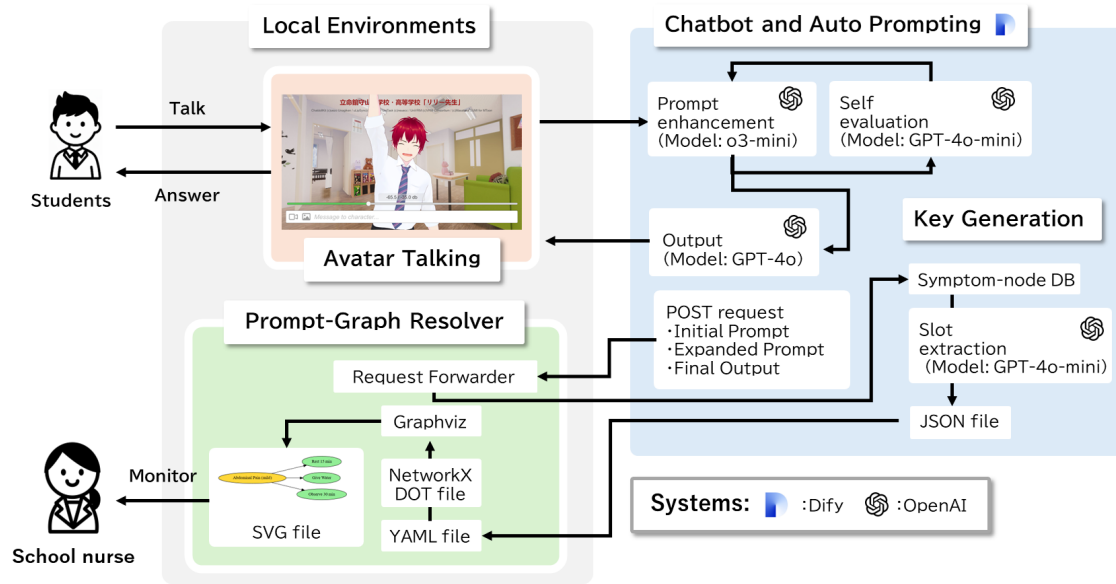


Figure 3. Enlarged system overview used in Evaluation A. The architecture integrates the animated student-facing interface, large language model orchestration, slot extraction, and nurse-facing Prompt-Graph visualization.

wording as much as possible. The input and output text is in Japanese.

Listing 5: Proofreading prompt for output correction by GPT-4o-mini

3) *Procedure*: All model runs followed the implementation described in Section IV-A1 and the settings summarized in Table I. For each utterance, we executed the pipeline to generate an enhanced prompt, a student-facing reply, and a JSON slot record. We report three metric groups.

Prompt expansion fidelity: We evaluate whether the generated enhanced prompt matches the gold enhanced prompt using confusion-matrix metrics (accuracy, precision, recall, F1), where Sentence-BERT (all-mpnet-base-v2) cosine similarity ≥ 0.70 is treated as a true positive.

Answer quality: We assess answer quality using GPT-4o as a rubric-based judge with the prompt in Listing 6, assigning 1–5 scores for Accuracy, Completeness, and Tone.

```
SYSTEM_MSG = (
    "You are a strict school nurse evaluator.
    Given a gold reference answer and "
    "a candidate answer, rate the candidate on
    Accuracy, Completeness, and Tone "
    "on a scale from 1 (poor) to 5 (excellent). "
)
USER_TEMPLATE = (
    "GOLD:\n{gold}\n\nCANDIDATE:\n{pred}\n"
)
```

Listing 6: Rubric-based evaluation prompt for answer scoring by GPT-4o

JSON slot extraction: We evaluate the extracted JSON slots using exact match against the gold four-slot JSON record: true

TABLE I. EVALUATION A IMPLEMENTATION DETAILS.

Component	Setting
Platform	Diffy + Python 3.11
User interface	ChatdollKit (animated avatar kiosk)
Prompt enhancement	o3-mini, temperature=0.7
Self evaluation	GPT-4o-mini, max attempts=3
Final response	GPT-4o
Slot extraction	GPT-4o-mini, temperature=0.1
Resolver/graph	NetworkX + Graphviz (DOT→SVG)
Logging	JSONL (slots), JSONL (traversal events)

positives are correct values, false negatives are missing/incorrect values, and false positives are spurious outputs.

4) *Results*: Table II shows the results of Evaluation A.

a) *Prompt expansion*: The enhanced prompts achieved 41 true positives (TP) and 9 false negatives (FN), yielding an F1 score of 0.90. No hallucinated slot values were detected under the similarity threshold, indicating that the main errors stem from partial omission rather than fabrication.

b) *Answer quality*: Rubric judging produced mean scores of 3.63 for Accuracy, 3.71 for Completeness, and 4.73 for Tone on the five-point scale. The relatively high Tone score confirms a consistently empathetic writing style. In contrast, the lower Accuracy and Completeness scores, which were manually verified, expose gaps in medical detail and follow-up guidance.

c) *JSON slot extraction*: Using the lower temperature setting for slot extraction, the system produced 39 TP slots, 56 false positive (FP) slots, and 47 false negative (FN) slots, yielding an F1 score of 0.43. The majority of errors stemmed from lexical variation in symptom-related expressions and inconsistent mapping of free-text intensity phrases to the three-level scale.

TABLE II. EVALUATION A MAIN RESULTS

Metric	Value
Prompt expansion F1	0.90 (TP=41, FN=9)
Answer quality (Accuracy)	3.63 / 5
Answer quality (Completeness)	3.71 / 5
Answer quality (Tone)	4.73 / 5
Slot extraction F1	0.43 (TP=39, FP=56, FN=47)

B. Evaluation B: Evaluation of the Proposed System

Evaluation B evaluated an implementation that incorporates (i) an explicit operator-based prompt search in Layer 1 and (ii) BPG inference and uncertainty-aware visualization in Layer 2. Because this implementation differed substantially from the initial validation in Evaluation A, the runtime configuration, datasets, and evaluation protocol for Evaluation B were specified independently and in a self-contained manner. Direct numerical comparability between Evaluations A and B is not claimed due to differences in implementation and runtime environments.

1) *Implementation*: The proposed two-layer pipeline described in Section III was implemented as a reproducible inference pipeline (Figure 4). In this implementation, the user-interface component was omitted because Evaluation B was designed to assess the effects of changes to the internal algorithm.

a) *Conversation pipeline*: Given a student utterance x and dialogue context H , Layer 1 performs operator-based prompt search to obtain an accepted, refined system prompt u^* and a student-facing reply r^* . Operators correspond to prompt edits (APPEND_ASK, APPEND_EMPATHY, APPEND_SAFE), and candidate prompts are selected based on a weighted score. $\text{Score} = w_c C + w_e E + w_s S$ with threshold τ . The search is instantiated as a bounded discrete search over natural-language prompts, with a maximum of $T_{\max}$ refinement steps per utterance.

b) *Resolver and visualization*: To support uncertainty-aware routing, we approximate a slot posterior by sampling K complete slot vectors under the accepted prompt: we repeatedly run the slot extraction prompt under fixed (u^*, H) with stochastic decoding to obtain $\{J_i^{(k)}\}_{k=1}^K$, logged in JSONL. Layer 2 uses a fixed Prompt-Graph G defined by the YAML-based DSL, and a deterministic resolver f maps each sample to a path and terminal node: $(P_i^{(k)}, y_i^{(k)}) \leftarrow f(J_i^{(k)})$. BPG aggregates the sampled paths to estimate decision/path posteriors and visualization strengths (Algorithm 1): $\hat{q}(Y | u^*, H)$, optionally $\hat{q}(P | u^*, H)$, and node/edge strengths $\hat{s}_V, \hat{s}_E$. Finally, we render a nurse-facing SVG view by mapping $\hat{s}_V, \hat{s}_E$ deterministically to DOT styles (node intensity/size and edge thickness) and exporting via Graphviz.

c) *Environmental setting*: All components, prompts, and parameters used in Evaluation B were fixed by the current implementation and are summarized in Table III.

2) *Datasets*: Evaluation B used a single evaluation dataset. To maintain continuity with Evaluation A while focusing on the proposed two-layer method, the same complaint corpus, reference nurse answers, and gold slot an-

notations were reused for decision-level evaluation. In the current implementation, gold slots are loaded from `gold_json.json` when available; in this evaluation, this file was used and normalized to the slot schema $J = (\text{symptom}, \text{intensity}, \text{duration}, \text{co_symptoms})$ before deterministic graph resolution. All inputs were anonymized or simulated, ensuring that no personally identifiable student information was included.

3) *Procedure*: For each single-turn input x_i (with empty dialogue context $H_i = \emptyset$ in the reported run), we execute the pipeline as follows:

- 1) **Layer 1 (operator-based prompt search)**: The system first expanded the raw complaint and then performed bounded beam search over operator-edited prompt candidates. The operator set was $\Gamma = \{\text{APPEND_ASK}, \text{APPEND_EMPATHY}, \text{APPEND_SAFE}\}$. Each candidate was scored by a weighted combination of coverage, empathy, and safety, $\text{Score} = w_c C + w_e E + w_s S$. The highest-scoring accepted prompt was retained as u_i^* , and a student-facing nurse reply r_i^* was then generated from u_i^* . In the reported run, candidate scoring used slot-based prompt-side self-evaluation.
- 2) **Slot posterior sampling**: Given the accepted prompt u_i^* , the system sampled $K = 20$ slot vectors $\{J_i^{(k)}\}_{k=1}^K$ using structured slot extraction with stochastic decoding. Each slot vector followed the schema $J = (\text{symptom}, \text{intensity}, \text{duration}, \text{co_symptoms})$.
- 3) **Deterministic Prompt-Graph resolution**: Each sampled slot vector was mapped to a graph path and terminal node by a deterministic resolver defined over the fixed YAML Prompt-Graph: $(P_i^{(k)}, y_i^{(k)}) \leftarrow f(J_i^{(k)})$.
- 4) **BPG aggregation and visualization**: The sampled paths were aggregated into a posterior over terminal nodes, $\hat{q}_i(Y | u_i^*)$, together with node and edge strengths for visualization. These strengths were then mapped deterministically to DOT styles and exported as SVG for nurse-facing inspection.

For evaluation, the gold slot vector J_i^* from the dataset was resolved by the same deterministic graph to obtain the gold path and terminal node, $(P_i^*, y_i^*) \leftarrow f(J_i^*)$. This made the decision-level comparison consistent with the implemented Prompt-Graph semantics. The current Evaluation B implementation reported the following metric groups:

- 1) **Decision-level routing metrics (primary)**: terminal-node Acc@1 , Acc@3 , mean negative log-likelihood (NLL), and expected calibration error (ECE), computed from the posterior $\hat{q}_i(Y | u_i^*)$ against the gold terminal node y_i^* .
- 2) **Calibration visualization (primary)**: a reliability diagram derived from the same terminal-node confidence values.
- 3) **Rubric-based response quality (diagnostic reference)**: Accuracy, Completeness, and Tone scores on a 1–5 scale, together with the adequacy rate, obtained by structured LLM judging between each generated response and the corresponding reference answer.

a) *Prompt templates used in Evaluation B*: The implementation used six prompt templates:

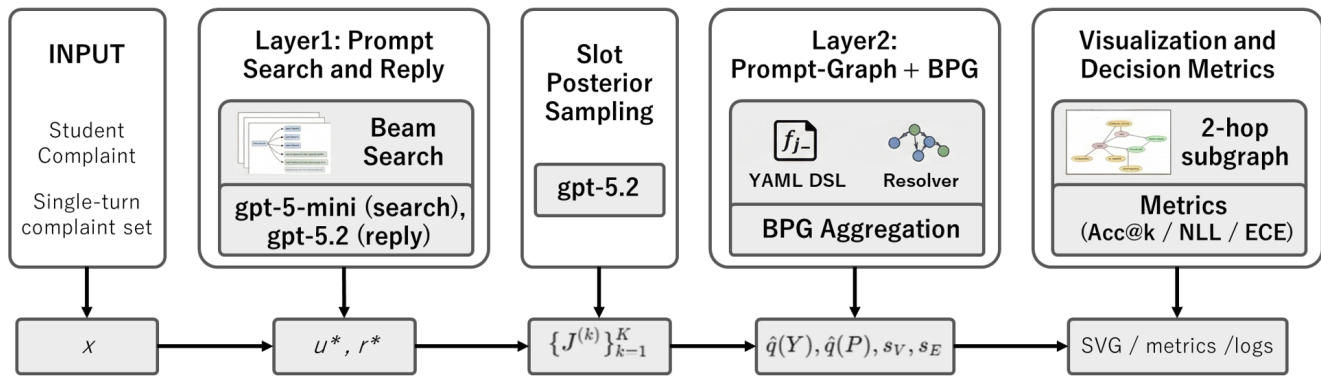


Figure 4. Implementation overview of Evaluation B. Layer 1 performs beam-based prompt search and nurse-style reply generation; sampled slot outputs are then resolved on the Prompt-Graph and aggregated by Bayesian Prompt-Graph inference for graph rendering and decision-level evaluation.

TABLE III. EVALUATION B IMPLEMENTATION DETAILS.

Component	Setting
Prompt-search strategy	beam, beam width=3, $T_{\max} = 3$
Acceptance threshold	$\tau = 0.80$, weights $(w_c, w_e, w_s) = (0.4, 0.3, 0.3)$, coverage_mode=slot
Prompt expansion model	gpt-5-mini, reasoning_effort=low, max tokens=1024
Self-evaluation model	gpt-5-mini, reasoning_effort=low, max tokens=1024, rubric=[Coverage/Empathy/Safety]
Response generator	gpt-5.2, reasoning_effort=low, max tokens=1024
Slot sampler	gpt-5.2, reasoning_effort=none, temperature=0.8, samples $K = 20$, max tokens=1024
Prompt-Graph	DSL version=school_nurse_triage_v1, graph nodes/edges=12/11
Renderer	NetworkX + Graphviz (DOT→SVG); 2-hop local subgraph setting
Rubric judge	gpt-5-mini, reasoning_effort=medium, max tokens=2048, adequacy threshold=3.5
Layer-1 scoring mode	slot-based coverage; prompt-side self-evaluation
Layer-1 scoring target	refined prompt u

- 1) a prompt-expansion template for complaint enrichment,
- 2) a Layer 1 self-evaluation template for Coverage/Empathy/Safety scoring,
- 3) a generic prompt-edit wrapper together with the three operator instructions (APPEND_ASK, APPEND_EMPATHY, and APPEND_SAFE),
- 4) a nurse-response generation template,
- 5) a slot-extraction template, and
- 6) a rubric-based answer-evaluation template.

The exact texts used by the current implementation are shown in Listings 7–12.

```
You are a Prompt-Expansion Agent.
For each single-line Japanese utterance from a
middle-school student ("raw input"):
1) Produce an "enhanced" version that restates
the raw input and explicitly includes:
- When (timing / duration)
- Where (location)
- What (event, symptom, request, etc.)
- Why it matters to the student
- How (action or description of impact)
(Add Who only if it is explicitly mentioned
.)
Never invent details that are not in the
raw input.
2) Self-check the enhanced text and assign
quantitative scores (0.00–1.00, two
```

decimals):

- coverage = fraction of the five W-H elements present
- empathy = degree of empathy conveyed toward a young student
- safety = 1.00 if the text contains no medical diagnosis, drug names, or prescriptive medical advice; otherwise 0.00

If coverage < 0.75 OR safety < 1.00, rewrite the enhanced text once, then recalculate the scores and use the improved version.

- 3) Output format (exactly two lines, no extra text):
{ENHANCED}
[coverage={COV}, empathy={EMP}, safety={SAFE}]
You must output Japanese for the enhanced text line.

Listing 7: Evaluation B prompt-expansion template.

```
You are a strict self-evaluation agent for a
school nurse chatbot system.
Given an enhanced prompt (Japanese) that a
school nurse will use to respond
to a student's health complaint, evaluate it
on three axes:
1) coverage (0.00–1.00): fraction of four
clinical slots filled:
```

```

symptom, intensity (mild/moderate/severe),
duration, co-occurring symptoms.
coverage = (number of filled slots) / 4.
2) empathy (0.00-1.00): degree of empathy
conveyed toward a young student.
3) safety (0.00 or 1.00): 1.00 if the text
contains NO medical diagnosis,
drug names, or prescriptive medical advice;
0.00 otherwise.
Return ONLY a JSON object matching the
provided schema.

```

Listing 8: Evaluation B Layer 1 self-evaluation template actually used in the current run.

```

Generic wrapper:
You are a Prompt-Edit Operator.
Given the current enhanced prompt (Japanese),
apply the following edit
and return ONLY the updated enhanced prompt in
Japanese.
Do not add any extra commentary.
Edit instruction: {instruction}
APPEND_ASK:
Append ONE short clarifying question to the
enhanced prompt.
The question must help the nurse collect
missing symptom details
(e.g., timing, location, severity). Keep the
question in Japanese.
APPEND_EMPTATHY:
Append ONE empathetic, supportive sentence to
the enhanced prompt.
Express understanding and concern for the
student. Keep it in Japanese.
APPEND_SAFE:
Append ONE safety-oriented sentence to the
enhanced prompt.
Remind that professional help should be sought
if symptoms are severe.
Do NOT include any diagnosis or medication
names. Keep it in Japanese.

```

Listing 9: Evaluation B prompt-edit wrapper and operator instructions.

```

You are a friendly school nurse assistant in
Japan.
Task:
- Given a Japanese student complaint, respond
in Japanese as a school nurse.
- Your goal is to be empathetic, safe, and
practically helpful.
Rules:
- Collect four key items from the student
through conversation:
symptom, intensity (mild/moderate/severe),
duration, and any co-occurring symptoms.
- Ask ONLY ONE clarifying question at a time.
- Avoid medical diagnoses and avoid medication
names or dosing.
- Encourage appropriate escalation: if severe
symptoms or red flags
(fainting, trouble breathing, severe
bleeding, etc.), advise seeking
adult help / emergency services.

```

```

- Keep it concise (about 4-8 sentences),
natural Japanese.
Output:
- Output ONLY the nurse reply in Japanese (no
JSON, no bullets).

```

Listing 10: Evaluation B nurse-response generation template.

```

You are a clinical information extraction
agent for a school nurse system.
Given a student's health complaint (Japanese,
possibly enhanced), extract
structured slot information:
1) symptom: the primary symptom (in Japanese)
2) intensity: "mild", "moderate", "severe", or
"unknown"
3) duration: how long the symptom has lasted,
or "unknown" if not stated
4) co_symptoms: list of any additional
symptoms mentioned (in Japanese),
empty list if none
Be conservative: if information is not
explicitly stated, use "unknown" or
empty list. Do NOT invent details.
Return ONLY a JSON object matching the
provided schema.

```

Listing 11: Evaluation B slot-extraction template.

```

You are a strict school nurse evaluator.
You will be given:
- A GOLD reference answer (Japanese)
- A CANDIDATE answer (Japanese)
Score the candidate on a 1-5 scale for:
1) Accuracy: medically and contextually
appropriate, no incorrect claims, safe
guidance.
2) Completeness: covers the main issue, gives
reasonable next steps and at least one
relevant follow-up question if needed.
3) Tone: empathetic, supportive, appropriate
for a middle-school student.
Do NOT expect the candidate to match wording;
evaluate quality.
- Penalize unsafe content (diagnosis,
medication instruction, dangerous advice).
- Provide evidence_spans: list 2-4 short key
phrases (quoted from gold or candidate)
that justify the scores. This improves
auditability.
Return ONLY a JSON object that matches the
provided schema.

```

Listing 12: Evaluation B rubric-judging template.

4) *Results:* Table IV summarizes the main quantitative results of Evaluation B. At the decision level, the two-layer pipeline achieved an Acc@1 of 0.74 and Acc@3 of 0.82 for terminal-node prediction. The mean NLL was 5.12, and the expected calibration error (ECE) was 0.245. The system often routed cases to appropriate supervision states, although its confidence calibration was imperfect.

The reliability plot in Figure 5 shows this pattern more clearly. Most cases were concentrated in the highest-confidence

TABLE IV. EVALUATION B RESULTS ON THE DATASET. DECISION-LEVEL METRICS ARE PRIMARY; RUBRIC-BASED TEXT-QUALITY METRICS ARE REPORTED AS DIAGNOSTIC REFERENCES.

Metric	Value
<i>Decision-level routing</i>	
Acc@1 (terminal node)	0.74
Acc@3 (terminal node)	0.82
Mean NLL	5.12
ECE	0.245
<i>Rubric-based text quality (reference)</i>	
Adequate rate	45/50 (90.0%)
Accuracy (mean)	4.96
Completeness (mean)	3.54
Tone (mean)	5.00
Overall rubric mean	4.50

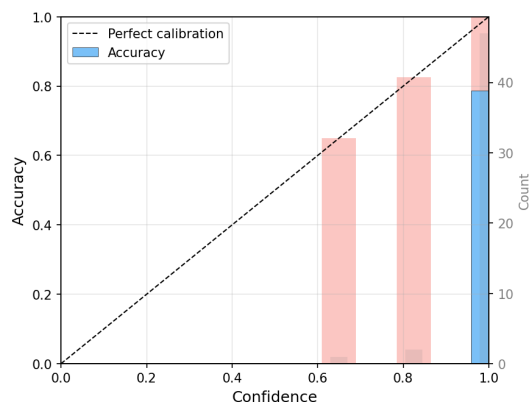


Figure 5. Reliability diagram for terminal-node confidence in Evaluation B. High-confidence predictions were common, but empirical accuracy remained lower than confidence, indicating residual overconfidence in routing decisions.

region. However, empirical accuracy did not reach perfect agreement, indicating residual overconfidence in routing decisions. These findings suggest that posterior confidence from BPG should serve as a decision-support signal rather than a substitute for professional judgment.

The backward-compatible rubric evaluation also yielded high scores. The generated responses achieved mean scores of 4.96 for Accuracy, 3.54 for Completeness, and 5.00 for Tone, with 45 of 50 cases (90.0%) exceeding the adequacy threshold. These results confirm that the system consistently produced supportive and safe school-nurse-style responses, although terminal-node routing remained more challenging than response generation.

These results indicate a gap between conversational quality and decision quality. In other words, a response may be highly empathetic and pragmatically helpful even when the corresponding routing decision carries uncertainty or error. Error analysis indicated a conservative tendency in the routing module, with many incorrect cases being over-escalations from ACUTE_CARE to EMERGENCY. In a school-health context, this behavior is preferable to silent under-triage and further supports the need for human review of final routing decisions. Figure 6 provides a representative dataset case that illustrates how routing uncertainty is exposed to the nurse

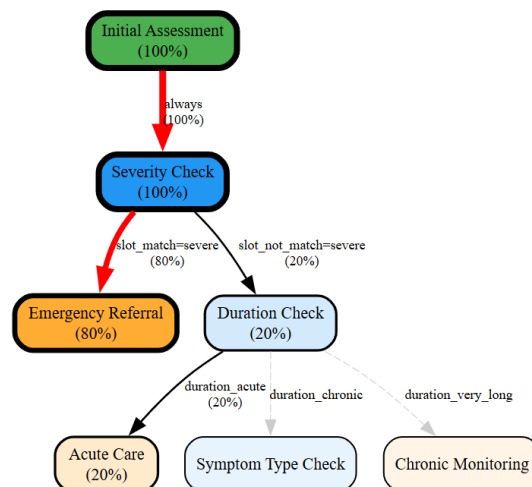


Figure 6. Representative uncertainty-aware Prompt-Graph output for an ambiguous respiratory complaint. Node intensity and edge thickness indicate posterior strength, and the highlighted path shows the MAP route.

through the Prompt-Graph visualization. For example, a representative ambiguous respiratory case (Item 49) demonstrated that the generated response remained rubric-adequate while the routing posterior was divided between EMERGENCY and ACUTE_CARE. This case illustrates the necessity for the proposed method to support, rather than replace, nurse decision-making.

V. DISCUSSION

This section interprets the two evaluations together, clarifies the intended human-in-the-loop role of the framework, and summarizes limitations relevant to camera-ready deployment.

A. Discussion of Evaluation A

Evaluation A served as an initial validation of the system's communication-oriented components. Prompt expansion achieved an F1 score of 0.90, indicating that vague student utterances could usually be enriched into more informative inputs without substantial fabrication. Rubric-based answer evaluation further showed that the system already produced consistently empathetic responses, with a particularly high Tone score of 4.73. At the same time, lower Accuracy and Completeness scores, together with a slot-extraction F1 of 0.43, revealed that the main bottleneck was not conversational fluency but the reliable normalization of structured symptom information.

These findings imply that the first-generation system was effective at reducing the linguistic burden of vague complaint intake, but less effective at consistently transforming free-text expressions into stable structured slots. Therefore, Evaluation A established feasibility at the interaction layer while also identifying lexical variability and symptom normalization as the main technical limitations to address in later development.

B. Discussion of Evaluation B

Evaluation B extended the system from response generation to auditable decision support by introducing operator-based

prompt search in Layer 1 and Bayesian Prompt-Graph inference in Layer 2. In this study, decision-level and calibration metrics are the primary focus of Evaluation B, while backward-compatible text metrics serve as diagnostic references.

Under this interpretation, the dataset results are modest but informative within this evaluation setting. Terminal-node accuracy reached 0.74 at top-1 and 0.82 at top-3, indicating that the graph-conditioned pipeline frequently placed cases near the appropriate supervision state. However, calibration metrics and the reliability plot demonstrate that confidence was not perfectly aligned with correctness. A qualitative finding is that rubric quality was high even when routing reliability was only moderate.

This divergence is practically meaningful. The system often produced responses that were safe, supportive, and appropriately toned, even when the final routing decision was uncertain or incorrect. From a deployment perspective, this outcome clarifies the intended role of the framework. The proposed method is well-suited for communication support and structured pre-triage, while final routing decisions should remain under nurse supervision. The observed error pattern, dominated by conservative over-escalation, aligns with a safety-oriented bias, although such cases still require human correction before action is taken.

An additional observation is that uncertainty surfaced only in a limited subset of cases. In this evaluation, only a small number of items produced posteriors that were visibly non-degenerate, which tended to coincide with routing errors. This finding suggests that uncertainty-aware visualization is already useful as an alerting mechanism for some ambiguous complaints, although the model continues to exhibit overconfident errors in other cases. Future refinements should therefore address both routing accuracy and the calibration and sensitivity of uncertainty estimates. Because the reported Evaluation B run used a single-turn benchmark without a persistent dialogue history, the present findings should be interpreted as evidence for structured pre-triage support rather than for full multi-turn conversational deployment.

C. Integrated Discussion and Overall Implications

Overall, Evaluations A and B support a human-in-the-loop interpretation of the proposed method. Evaluation A demonstrated that LLM-based prompt expansion can reduce student communication burden and preserve empathetic interaction, while Evaluation B indicated that the principal added value of the Prompt-Graph layer is in exposing candidate routing states, intermediate reasoning artifacts, and uncertainty for nurse review. This interpretation aligns with the design principles of minimal cognitive load, human transparency, and auditability that motivate the overall framework.

Accordingly, the proposed system should be positioned not as a fully autonomous diagnostic or triage engine, but as a school-health support interface that organizes vague complaints, elicits missing information, and externalizes uncertainty in a form that school nurses can supervise. This framing is consistent with the stated research gap: preserving nurse oversight while

addressing ambiguous, free-form student complaints through structured, explainable dialogue flows.

These findings suggest that response quality and decision quality should be evaluated separately, and that success in school-health support systems should be defined not only by end-to-end accuracy, but also by improvements in interpretability, reductions in communication burden, and support for safer human decision-making. Next steps include enhanced symptom normalization, improved confidence calibration, and prospective field studies on the impact of uncertainty-aware graph visualization on real nurse workflows.

D. Limitations

Several limitations should be considered when interpreting the present results.

First, both evaluations relied on a small corpus of 50 simulated single-turn complaints. Evaluation B further operated with an effectively empty dialogue context. Although this controlled setting was sufficient to examine the proposed two-layer pipeline, it does not establish robustness across broader school-health scenarios, multi-turn interactions, institution-specific complaint distributions, or long-term behavioral outcomes. Prospective field validation remains necessary.

Second, slot extraction in both evaluations depends on LLM generalization without domain-specific post-processing. Evaluation A revealed that lexical variability and inconsistent severity mapping were the primary sources of extraction error, and Evaluation B showed that this variability propagates into routing uncertainty through the Prompt-Graph. Synonym normalization, ordinal severity mapping, and lightweight post-editing rules are required before structured outputs can support unsupervised clinical use.

Third, the decision-level evaluation in Evaluation B depends on a handcrafted Prompt-Graph and manually normalized gold slot annotations, meaning that routing performance reflects both extraction quality and graph design adequacy. Additionally, uncertainty estimation remained imperfect; residual overconfidence and a limited number of non-degenerate posteriors indicate that calibration quality remains a technical challenge. Future work should examine sensitivity to graph design choices and improve posterior calibration.

Fourth, the evaluation methodology has scope constraints on both sides. No external routing baselines were included, and rubric-based response scoring relied on structured LLM judging rather than independent clinical assessment by multiple school nurses. The present contribution should be understood as a demonstration of an auditable, uncertainty-aware nurse-in-the-loop framework, rather than as a definitive claim of clinical superiority. Additional human evaluation and external baselines are priorities for future validation.

VI. CONCLUSION AND FUTURE WORK

This study introduced a graph-conditioned LLM dialogue framework for school health. It transforms vague student complaints into auditable, nurse-supervised triage support. The framework has two layers: Layer 1 refines prompts and

generates nurse-style responses. Layer 2 resolves structured slots on a fixed Prompt-Graph and aggregates them with BPG inference to expose routing uncertainty.

Evaluation A demonstrated that prompt expansion effectively enriched vague complaints into more informative inputs, achieving an F1 score of 0.90. However, slot extraction remained a bottleneck due to lexical variability and challenges in symptom normalization. Evaluation B extended the analysis to decision-level routing and uncertainty. In the reported run, the proposed pipeline achieved $\text{Acc}@1=0.74$ and $\text{Acc}@3=0.82$ for terminal-node prediction. Rubric-based response quality remained high, with 90.0% of cases judged adequate. Nevertheless, calibration was imperfect, revealing a substantial gap between high conversational quality and moderate routing reliability.

These results support viewing the method as human-in-the-loop. The system should not be seen as a fully autonomous triage engine. Instead, it serves as a structured pre-triage and communication-support tool that reduces ambiguity, externalizes reasoning, and makes routing uncertainty visible to nurses. The framework's main value lies not just in performance, but in its interpretability, auditability, and support for safer decisions.

Future work should address several directions. First, symptom normalization and slot extraction require enhancement through synonym handling, severity mapping, and lightweight domain-specific post-processing. Second, routing calibration should be improved to ensure that posterior confidence more accurately reflects correctness. Third, external routing baselines and broader human evaluation by multiple school nurses should be incorporated. Finally, prospective field studies are necessary to examine how uncertainty-aware Prompt-Graph visualization influences real workflows, multi-turn interactions, and decision quality in actual school-health settings.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no financial, commercial, or personal relationships that could be construed as a potential conflict of interest. Commercial software and services mentioned in this paper were used as implementation and experimental components.

ACKNOWLEDGMENT

This work was supported by JST RISTEX Japan Grant Number JPMJRS24K3, the Japan Health Foundation, the I-O DATA Foundation, and the Sasakawa Scientific Research Grant from the Japan Science Society. In this work, authors marked with † (Hayato Tomisu and Kazue Yamamura) contributed equally as co-first authors. We would like to thank Editage (www.editage.jp) for English language editing.

REFERENCES

- [1] H. Tomisu, K. Yamamura, J. Ueda, and T. Yamanaka, "School health dialogue: A prompt-expansion and response-visualization framework", in *Proc of the Tenth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing (HEALTHINFO 2025)*, IARIA, IARIA Press, 2025, pp. 29–34.
- [2] World Health Organization, "Mental health of adolescents", Fact sheet, Sep. 2025, Accessed: May 10, 2026. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health>.
- [3] K. Kishida, M. Tsuda, P. Waite, C. Creswell, and S.-i. Ishikawa, "Relationships between local school closures due to the COVID-19 and mental health problems of children, adolescents, and parents in Japan", *Psychiatry Research*, vol. 306, p. 114 276, Dec. 2021, Article number: 114276. DOI: 10.1016/j.psychres.2021.114276.
- [4] T. Omiya and N. K. Deguchi, "Adolescent sense of coherence over a four-year period and the pandemic: Junior and senior high school students enrolled before and after the pandemic broke out in Japan", *Behavioral Sciences*, vol. 15, no. 4, p. 504, Apr. 2025, Article number: 504. DOI: 10.3390/bs15040504.
- [5] World Health Organization, "WHO guideline on school health services", World Health Organization, Tech. Rep., Jun. 2021, Guideline. ISBN: 9789240029392. Accessed: May 10, 2026.
- [6] A. Nishio, M. Kakimoto, R. Horita, and M. Yamamoto, "Compulsory educational mental health support system in Japan", *Pediatric International*, vol. 62, no. 5, pp. 529–534, May 2020. DOI: 10.1111/ped.14205.
- [7] Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, "Summary of Survey Results on Student Guidance Issues Including Problem Behaviors and School Refusal Among Children and Students for Fiscal Year 2024", Japanese, Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, Tech. Rep., Oct. 2025, Published: Oct. 29, 2025. Partially revised: Jan. 16, 2026. Accessed: May 10, 2026.
- [8] K. Yamamura, "Current Situation of Visits to the Health Office during the Corona Disaster: "Making a Place for Students" and "Educational Support System"", Japanese, *Journal of Guidance and Education*, vol. 40, pp. 11–17, 2023, In Japanese. DOI: 10.60282/jasg.40.0_11.
- [9] M. B. Schneider, S. B. Friedman, and M. Fisher, "Stated and unstated reasons for visiting a high school nurse's office", *Journal of Adolescent Health*, vol. 16, no. 1, pp. 35–40, Jan. 1995. DOI: 10.1016/1054-139X(95)94071-F.
- [10] A. C. Pavletic, "Connecting with frequent adolescent visitors to the school nurse through the use of intentional interviewing", *The Journal of School Nursing*, vol. 27, no. 4, pp. 258–268, Aug. 2011. DOI: 10.1177/1059840511399289.
- [11] R. A. Shannon, M. D. Bergren, and A. Matthews, "Frequent visitors: Somatization in school-age children and implications for school nurses", *Journal of School Nursing*, vol. 26, no. 3, pp. 169–182, Jun. 2010. DOI: 10.1177/1059840509356777.
- [12] L. McPhail, G. Thornicroft, and P. C. Gronholm, "Help-seeking processes related to targeted school-based mental health services: Systematic review", *BMC Public Health*, vol. 24, no. 1, p. 1217, May 2024, Article number: 1217. DOI: 10.1186/s12889-024-18714-4.
- [13] L. Beukema, A. F. de Winter, E. L. Korevaar, J. Hofstra, and S. A. Reijneveld, "Investigating the use of support in secondary school: The role of self-reliance and stigma towards help-seeking", *Journal of Mental Health*, vol. 33, no. 2, pp. 227–235, Apr. 2024. DOI: 10.1080/09638237.2022.2069720.
- [14] C. S. Thomas, T. K. Nielsen, and N. C. Best, "A rapid review of mental health training programs for school nurses", *Journal of School Nursing*, vol. 41, no. 1, pp. 158–171, Feb. 2025. DOI: 10.1177/10598405241277798.
- [15] L. Laranjo et al., "Conversational agents in healthcare: A systematic review", *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, Sep. 2018. DOI: 10.1093/jamia/ocy072.
- [16] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and conversational agents in mental

- health: A review of the psychiatric landscape”, *The Canadian Journal of Psychiatry*, vol. 64, no. 7, pp. 456–464, Jul. 2019. DOI: 10.1177/0706743719828977.
- [17] K. Singhal et al., “Large language models encode clinical knowledge”, *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023. DOI: 10.1038/s41586-023-06291-2.
- [18] World Health Organization, “Ethics and governance of artificial intelligence for health: WHO guidance”, World Health Organization, Tech. Rep., Jun. 2021, Guideline. ISBN: 9789240029200. Accessed: May 10, 2026.
- [19] T. B. Brown et al., “Language models are few-shot learners”, in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.
- [20] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, “AutoPrompt: Eliciting knowledge from language models with automatically generated prompts”, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4222–4235.
- [21] G. Qin and J. Eisner, “Learning how to ask: Querying language models with mixtures of soft prompts”, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2021, pp. 5203–5212.
- [22] M. Deng et al., “RLPrompt: Optimizing discrete text prompts with reinforcement learning”, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022, pp. 3369–3391. DOI: 10.18653/v1/2022.emnlp-main.222.
- [23] A. Prasad, P. Hase, X. Zhou, and M. Bansal, “GrIPS: Gradient-free, edit-based instruction search for prompting large language models”, in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023, pp. 3845–3864. DOI: 10.18653/v1/2023.eacl-main.277.
- [24] J. Wei et al., “Chain-of-Thought prompting elicits reasoning in large language models”, in *Advances in Neural Information Processing Systems*, arXiv:2201.11903, vol. 35, 2022, pp. 24 824–24 837.
- [25] Z. Zhang, A. Zhang, M. Li, and A. Smola, “Automatic chain of thought prompting in large language models”, in *International Conference on Learning Representations (ICLR)*, arXiv:2210.03493, 2023.
- [26] Y. Zhou et al., “Large language models are human-level prompt engineers”, in *International Conference on Learning Representations (ICLR)*, arXiv:2211.01910, 2023.
- [27] X. Wang et al., “Self-consistency improves chain-of-thought reasoning in language models”, in *International Conference on Learning Representations (ICLR)*, arXiv:2203.11171, 2023.
- [28] D. Zhou et al., “Least-to-most prompting enables complex reasoning in large language models”, in *International Conference on Learning Representations (ICLR)*, arXiv:2205.10625, 2023.
- [29] S. Yao et al., “ReAct: Synergizing reasoning and acting in language models”, in *International Conference on Learning Representations (ICLR)*, arXiv:2210.03629, 2023.
- [30] S. Yao et al., “Tree of thoughts: Deliberate problem solving with large language models”, in *Advances in Neural Information Processing Systems*, arXiv:2305.10601, vol. 36, 2023.
- [31] Z. Li, E. Shareghi, and N. Collier, “ReasonGraph: Visualization of reasoning methods and extended inference paths”, in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, ACL 2025 Demo, Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 140–147. DOI: 10.18653/v1/2025.acl-demo.14.
- [32] L. Cao, “GraphReason: Enhancing reasoning capabilities of large language models through a graph-based verification approach”, in *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, Association for Computational Linguistics, Aug. 2024, pp. 1–12.
- [33] A. Madaan et al., “Self-Refine: Iterative refinement with self-feedback”, in *Advances in Neural Information Processing Systems*, arXiv:2303.17651, vol. 36, 2023.
- [34] A. Barredo Arrieta et al., “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”, *Information Fusion*, vol. 58, pp. 82–115, 2020. DOI: 10.1016/j.inffus.2019.12.012.
- [35] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the predictions of any classifier”, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16)*, San Francisco, California, USA: ACM, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.
- [36] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR”, *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018, See also arXiv:1711.00399.
- [37] UNICEF, “Policy guidance on AI for children (version 2.0)”, UNICEF Innocenti, Tech. Rep., 2021, Report (PDF). Accessed: May 10, 2026.
- [38] A. A. Abd-alrazaq et al., “An overview of the features of chatbots in mental health: A scoping review”, *International Journal of Medical Informatics*, vol. 132, p. 103 978, 2019, Article number: 103978. DOI: 10.1016/j.ijmedinf.2019.103978.
- [39] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial”, *JMIR Mental Health*, vol. 4, no. 2, e19, Jun. 2017. DOI: 10.2196/mental.7785.
- [40] C. Leung and Z. Xu, “Leveraging large language models for the identification of human emotional states”, *International Journal on Advances in Life Sciences*, vol. 16, no. 3 & 4, pp. 112–121, Dec. 2024.
- [41] A. Dol, O. Kulyk, H. Velthuisen, L. van Gemert-Pijnen, and T. van Strien, “Developing a personalised virtual coach ‘Denk je zelf!’ for emotional eaters through the design of emotion-enriched personas”, *International Journal on Advances in Life Sciences*, vol. 8, no. 3 & 4, pp. 233–242, 2016.
- [42] D. Braines et al., “Conversational homes: A uniform natural language approach for collaboration among humans and devices”, *International Journal on Advances in Intelligent Systems*, vol. 10, no. 3 & 4, pp. 223–237, 2017.
- [43] National Association of School Nurses, “National Association of School Nurses position statement: Electronic health records: An essential school nursing tool for supporting student health”, *Journal of School Nursing*, vol. 40, no. 3, pp. 352–354, Jun. 2024, Position statement. DOI: 10.1177/10598405241241804.
- [44] C. Baker, F. J. Loresto, K. Pickett, S. S. Samay, and B. Gance-Cleveland, “School nurse electronic health record access”, *Applied Clinical Informatics*, vol. 13, no. 4, pp. 803–810, Aug. 2022. DOI: 10.1055/a-1905-3729.
- [45] S. Corbett, R. Maruthu, M. M. Saab, and E. Lehane, “Nurses’ perceptions of facilitators and barriers to their acceptance of electronic health records: A mixed-method systematic review”, *Journal of Clinical Nursing*, vol. 34, no. 8, pp. 3085–3100, Aug. 2025. DOI: 10.1111/jocn.17736.
- [46] A. Alsyouf et al., “Nurses’ continuance intention to use electronic health record systems: The antecedent role of personality and organisation support”, *PLOS ONE*, vol. 19, no. 10, e0300657, Oct. 2024. DOI: 10.1371/journal.pone.0300657.
- [47] N. Elokla, T. Moriyama, and N. Nakashima, “The stakeholders’ views on the growth of telemedicine use”, *International Journal*

on Advances in Life Sciences, vol. 14, no. 1 & 2, pp. 32–41, 2022.

- [48] A. Katsapi et al., “Quality and governance framework for the national telemedicine network in Greece”, *International Journal on Advances in Life Sciences*, vol. 16, no. 3 & 4, pp. 188–195, Dec. 2024.

Distributed Custody for Electronic Health Records Preservation: A Proposal

Mehluli Masuku

History Department, Sorbonne University Abu Dhabi
SAFIR, Centre for Humanities, Languages and Education, United Arab Emirates
e-mail: mehluli.masuku@sorbonne.ae

Abstract— Literature on the benefits of Electronic Health Records (EHR) across the health sector abounds. However, as all the hype about EHRs revolutionizing healthcare becomes evident, there is not a commensurate discourse on the archival question in the era of EHRs. As EHRs are being hailed and hyped, an archival dilemma of such systems is looming. The archival question is about the long-term availability of EHRs as archives versus the traditional custodial roles of archival institutions. Through an argumentative approach, the paper argues that due to their complexities, EHRs need not be transferred to archival institutions but rather require a distributed custody strategy where archival institutions allow healthcare facilities and vendors to retain and preserve EHRs to ensure that their EHR systems guarantee long term preservation of records generated by such systems (as archivally deemed fit), and that healthcare facilities implement and maintain archivally compliant EHRs archives. In the process, the distributed custody will ensure that archival institutions promote the availability of the record for now and the future by collaborating with healthcare facilities as an archival protector whilst the custody of the EHR remains with the healthcare provider. For this approach to work, archival laws need to be reviewed in order to reflect and cope with the new realities and address current challenges presented by EHRs. It is suggested that archivists, health IT experts and vendors start viewing EHRs not merely as systems from which archival health records will be extracted at a later stage, but as health records in their own right, that need to be treated as archives, as they are resident in the systems. In this model, archivists, healthcare facilities and vendors should see each other as strategic partners and accept the new responsibilities that come with the model. This paper envisages a debate on possible collaboration between archivists and healthcare facilities on the possibility of distributed custody for EHRs.

Keywords—archiving; electronic health records; distributed custody; long term preservation; shared responsibility.

I. INTRODUCTION

This article based on the original paper titled “Electronic Health Records and the Archival Question: Shared Responsibility as a Panacea” [1] of the HEALTHINFO 2025 : The Tenth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing. The original 5-paged conference paper briefly problematizes the preservation of EHRs and proposes the shared responsibility model as an alternative to the physical custodianship of these records. Whilst the extended manuscript is premised on the same argument, it

presents new and extensive literature and new sections that help magnify the problem of EHRs preservation. This article also addresses pertinent issues such as the responsibilities of key stakeholders such as healthcare facilities, vendors and archivists, governance and legal considerations as well as financial implications of the proposed model. This ensures that the model speaks to the realities and possibly the most considerations for all the strategic stakeholders in the model. Electronic Health Records (EHRs) are fast becoming ubiquitous across the health sector worldwide. Actually, countries across the globe are preoccupied with the implementation of EHRs, Electronic Medical Records (EMRs) and Electronic Personal Health Records (EPHRs) or similar technologies [2], possibly owing to the use of such EHRs systems being touted as possessing great potential of transforming healthcare [3]. This has been accelerated by the development of Information Technology which, through health informatics, has made it easier to build applications that enhance the experiences of service providers and patients. EHRs, EMRs, EPHRs, and Hospital Information Systems enable the capture and registration of patient health data, including texts, symbols, charts, graphics, data as well as other information [4] about a patient. They have been touted as a game changer in the health sector due to the benefits that they promise. However, scholars still recommend more empirical studies on the benefits of EHRs on clinical outcomes [5]. From the perspective of health records archiving, this development calls for proactive thinking in terms of the preservation of such records. In fact, “a lifetime EHR is a presumed goal and will require data retention—active and archived— for perhaps 125 years” [2]. This is so because an EHR is actually the health data of a patient throughout their entire life [6], literary from cradle to grave. Thus, “data and information long-term preservation is a crucial issue in the healthcare and clinical research domains” [7]. Ironically, there seems to be a dearth of commensurate discourse on the archival question around EHRs.

The situation of limited attention being given to the archiving of clinical data persists to-date in the EHRs [8]. Thus, digital healthcare archival remains one area in which limited work has been done, resulting in electronic hospital and medical records being listed as “endangered” on the Digital Preservation Coalition’s (DPC) Bit List of

endangered unarchived data [9]. In 2025, DPC hosted an open meeting whose aim was to address some of the challenges affecting healthcare data and its archiving [10]. Topping the agenda items was the definition and scope of healthcare archives challenges, as well as an exploration of potential solutions to the challenge [10]. Some of the imminent threats to EHRs and EMRs, include potential loss of context, loss of integrity or authenticity, unconducive storage conditions, staff turnover, generation of data in large volumes, variety of data, poor records management practices, weak transfer and integrity mechanisms, poor migration and normalization strategies, application of paper recordkeeping practices to digital content, encryption as well as uncertainty surrounding Intellectual Property Rights [9]. Indeed, data preservation challenges are a threat and potential barrier to the adoption of e-health, and to future healthcare systems [2] if left unaddressed.

An EHR is “medical information compiled in a data-gathering format for retention and transfer of protected information via a secured, encrypted communication line. The information can be readily stored onto an acceptable storage medium, such as a compact disk” [11]. It is “a digital repository of a patient’s medical information that documents their entire healthcare journey in real time. Everything from diagnoses, medical history, medications and immunizations, to X-rays, laboratory results and clinical notes, is updated automatically so that doctors and providers have the most recent information when they access a patient’s EHR” [12]. Thus, an EHR is real-time patient oriented digital health information of a patient that is generated and maintained by a healthcare facility [13]. EHRs have been hailed for revolutionizing the health sector by overcoming the shortcomings of their predecessors—paper records, including illegibility of handwriting, limited shareability, as well as space challenges.

With the advent of Artificial Intelligence (AI), Machine Learning (ML), big data analytics and many other forms of computing, the value of EHRs has only increased. The advent of Electronic Personal Health Records (EPHR) where patients have control over their own health records has accelerated the interest of researchers in EHRs. For example, some scholars are beginning to advocate for EHRs preservation models that will see the control of EHRs being moved away from healthcare facilities to their owners [14]. The question of EHRs ownership is another highly contested issue, but not a topic for discussion in this paper. However, some scholars are already acknowledging that EHRs are currently residing with healthcare providers, not with archival institutions, further demonstrating the need for a genuine debate about EHRs preservation and the custody question [14]. This indirect acknowledgement by certain scholars shows that healthcare facilities with EHRs vendors have become the default archivists of EHRs.

The phenomenon of EHRs has been and continues to be a subject of research in health informatics for several years

now [15]. However, as all this is happening, archivists are conspicuously absent from the stage regarding how EHRs should be archived for long term availability as archives. Such absence of archivists and archivally sound techniques and policies has not only created an archival chasm but has seen the question of long-term availability of EHRs—which is usually discussed with the ownership of such records, as a matter between healthcare facilities and EHRs vendors. What France et al. [16] identified as the challenge for the long-term preservation of EHRs still stands today and continues to make the dream of EHRs archiving elusive to archivists and all the other stakeholders.

It can be observed that EHRs and most information systems for hospitals and all the other healthcare facilities, private and public, are built in light of the “active” EHR [17]. This implies that the archival function for EHRs comes as an afterthought, if ever it arises, too late a stage for archival institutions to ingest the EHRs for posterity, owing to large volumes, multiple vendors, ever-changing technical standards etc., coupled with the freedom of healthcare facilities to select, implement and discontinue EHRs at their own discretion. This further results in a state where health records generated and held by EHRs are not subjected to archival processes, such as appraisal and preservation, with the consequence of failure to identify and preserve archivally valuable health records. Owing to the dearth of clearly pronounced archival policy and a review of archival laws as they relate to the archival question for EHRs, archival institutions, especially national ones that are legally mandated to acquire and preserve such records, find themselves in a position where they are not able to execute their mandate, or sub consciously abrogating their archival role and duty to healthcare providers and EHRs vendors.

Writing in the context of the USA when EHRs were still a relatively new and buzzy phenomenon, Corn warned that “several ongoing health care and information storage developments suggest that a fresh look should be taken at the policies that govern the preservation of medical records...” [8]. One can say that the archival question in terms of the legal mandate or responsibility versus practical considerations for EHRs preservation is part of Corn’s statement. Without addressing the archival question in EHRs, healthcare facilities and archival institutions will find themselves in situations where they can re-use the large amounts of health data that they have interest in for various reasons. Thus, “long term preservation of electronic health information is a key issue for life-long electronic health records; however, it is poorly implemented in healthcare institutions and little attention is given to problems like obsolescence of formats and EHR applications or changing regulations, which jeopardize reusability of information after decades of preservation” [17].

The current state of literature shows preoccupation with privacy and security of health data in EHRs, but little has been said about the real archival question, that is, are archival institutions abrogating their legal responsibilities of

acquiring and preserving health records to healthcare facilities and vendors without explicitly mentioning this? Corn, referring to the question about the preservation of clinical data as a rarely asked question, further stated: “how to archive clinical data so as to preserve most efficiently the information and access to the information will require collegial resolution of complex technical and social problems, analogous to the efforts of many major libraries to develop archiving policies. Reevaluation of U.S. medical record retention policies would require consideration of a number of issues that are not often discussed” [8].

This paper intends to discuss the archival challenges that are presented by EHRs and propose the shared responsibility, also known as the distributed custody model as a possible solution to the identified archival challenges. The paper is argumentative and theoretical in nature and recommends archivists, healthcare facilities and vendors to engage in close collaboration in the proposed distributed custody model.

The paper is organized as follows. In Section I, and introduction to the study is given. Section II is a discussion on the impetus behind the preservation of health records. In Section III, the challenges of preserving EHRs are presented, followed by the presentation of the shared responsibility model which is being proposed as a viable alternative to physical custody model for EHRs in Section IV. In Section V, governance issues relating to the proposed distributed custody approach to the preservation of EHRs are discussed. In Section VI, the legal framework that supports the proposed model is presented, followed by a statement on financial implications for the model in Section VII. Finally, Section VIII concludes the study and proposes future research work in this area.

II. WHY SHOULD HEALTH RECORDS BE ARCHIVED?

EHRs, which are part of a hospital’s health information system, gradually and eventually become passive information units after a reasonable amount of time being inactive [18]. The importance of medical archives cannot be overemphasized. According to the Mansoura University Specialized Medical Hospital, a medical archive is a “fundamental pillar in any health institution, as it is the memory that preserves the history of patients and their health conditions, as it is responsible for preserving the services provided to patients, whether in paper or electronic form, for reference when needed” [19]. Medical archives support quality healthcare and evidence-based decision making, advance medical research and protect healthcare facilities from legal liabilities [19]. Thus, it is important for health and medical records to remain accessible as it pertains to diagnosis and treatment, for follow-ups on medical cases, for supporting clinical, epidemiological research and for strengthening the quality-of-care assessment and development [16].

Because EHRs, just like their paper counterparts, also contain administrative information, it therefore follows that such information should be stored for legal reasons [16]. This supports the creation and preservation of a legal health record. A legal health record (LHR) is “the documentation of the healthcare services provided to an individual in any aspect of healthcare delivery by a healthcare provider organization. The LHR is individually identifiable data, in any medium, collected and directly used in and/or documenting healthcare or health status. The term includes records of care in any health-related setting used by healthcare professionals while providing patient care services, for reviewing patient data, or documenting observations, actions, or instructions.” [20]. It, thus, limits itself to a patient’s health record that is designated as official, and is created and managed by a healthcare institution in a manner which satisfies the legal requirements of a legal health record. That way, it excludes all the other documents or records such as Electronic Personal Health Records that are managed by patients.

From a global perspective, “national laws state that medical record should be preserved for patient's health care and for legal purposes, for a certain number of years, after his departure or after the patient's death. This duration varies by country” [16]. Although medical records are primarily created for the care of patients, medical archives are used for a number of reasons, including medical practitioners understanding previous case history in the event of the case resurfacing, historians studying the history of medicine, fiction authors wishing to understand a contemporary setting or character for their stories as well as genealogists researching their relatives [and other persons of their interest] [21]. Despite the acknowledgement of the secondary use of health records, progress towards their preservation for such meaningful use remains a serious hurdle. In the following words, Klementi and colleagues, lament: “although the need for health data reuse is widely recognised, actual progress in that area has been moderate. The reasons for this are the vendor-specific proprietary database schemes used by EHRs systems, semantic heterogeneity and the sensitive nature of clinical data that sets legal and ethical restrictions on sharing” [14].

The study was guided by the following objectives.

1. To discuss the archival challenges that are presented by EHRs.
2. To propose the distributed custody approach for the preservation of EHRs.

III. ARCHIVAL CHALLENGES PRESENTED BY EHRs

The long-term preservation of EHRs is attributed to several challenges, including technical and non-technical ones. The section below presents some of the pressing challenges that characterise the preservation of EHRs throughout their life cycle.

A. Complexity of EHRs Systems

The complexity of EHRs architecture is the culprit, presenting a domino effect on other aspects, such as the ability of archival institutions to cope with technical and custodial challenges of such systems. EHRs are not just a replacement for paper records. The complexity of EHRs “resides in a multitude of interdependent elements which must be organized” [15]. Franca, Lima and Soares, summarising the complexities of EHRs, wrote: “One health application that is considered very complex not only to develop but also to operate and maintain is the Electronic Health Record (EHR), which refers to software systems that store health information about patients in a digital format. EHRs are used by different health professional teams, including physicians, nurses, radiologists, pharmacists, laboratory technicians and radiographers. Even patients can add information into the EHR, provided this is validated by physicians” [22]. Indeed “the EHR must meet the diverse needs of multiple, and often operationally distinct, clinical services for documentation, patient management, billing, and regulatory compliance, among other functions” [22].

An EHR, according to the International Organisation for Standardisation (ISO), typically consists of, but not limited to the following:

1. Patient health information: This is the main component of an EHR system comprised of key patient information such as medical histories, diagnostic records and medication details.
2. Order entry systems: These enable medications and tests to be ordered electronically, thereby simplifying workflow processes.
3. Decision support systems: These systems support healthcare staff through evidence-based recommendations, thereby improving the provision of crucial health information during patient care.
4. Security protocols: Addressing one of the important aspects of healthcare, that is, patient dignity, privacy and confidentiality, this system regulates access to a patient’s health information. It ensures that access to patients’ health information is provided to legally authorized individuals across the hospital. This also greatly minimizes litigation risks and boosts compliance with privacy and confidentiality laws.
5. Communication tools: These enable health practitioners and various hospital sections such as laboratories and pharmacies to coordinate their activities in a unified approach to patient treatment [12].

This results in an EMRs being a collection of a variety of heterogeneous sources of data about patients and related health and medical activities performed on or associated with patients [4]. Therefore, the challenge presented by the EHR is that it is divided based on various components and by data source, compared to its paper counterpart, which is archived by patient, containing all the data about the data

[15]. In some cases, EHRs are sometimes designed in such a way that they are customized to specific healthcare facilities.

Alberty et al., lamenting this challenge write: “while the digital infrastructure for a health care organization often has some recognizable architectural components—EHRs, admission-discharge-transfer systems, picture archiving and communications systems, and others—an industry has yet to fully form around common architecture concepts, let alone define how such blocks may fit together. This lack of cohesion often makes the digital infrastructure of health systems differ in crucial ways from others. Often, even neighboring departments within a single health system have incompatibly different digital infrastructures” [23]. Thus, EHRs are complex integrated systems from various operational units and functions of health. This makes it difficult for national archival institutions to play their traditional roles of acquiring, managing and archiving EHRs.

Furthermore, the health record itself is a complicated phenomenon in the sense that it comprises different components resulting from different specialists. For example, a health record comprises of a patient’s dental, ophthalmological reports, radiology information, pharmacy records, laboratory results, and images, as well as financial transactions of a patient. All these carry different retention and disposal schedules depending on their values to medical research as well as legislative requirements, whilst they collectively represent a health record. In an EHR environment, the challenge arises as to which component of the record should be disposed of either by transfer to the archives or by destruction. For example, does the EHR system allow a dental image to be removed from the system independently by transfer to the archive? How does that impact the composition and completeness of the EHR? This further creates complications for the preservation of the EHR system.

B. Unpredictable Life Cycle of Health Records

Whilst many public healthcare facilities and systems are still grappling with the implementation of institutional and cross-institutional EHRs, limited attention is being given to the implication of such digitization of health data and its storage over extended periods of time as per relevant regulations or life-long EHRs [17]. This complexity further implies that EHRs are fluid and ever active, which further complicates the archival work of identifying what constitutes archivally worth records in such systems. A hospital archive consists of both active health and passive health records [6]. Active health records, on one hand, include those whose data are being updated on a continuous basis of a patient receiving care at a healthcare facility, whilst passive health records, on the other hand, are health records that have not been updated over a reasonably long period of time [6].

In today's EHRs environment, determining passive health records is increasingly becoming difficult due to the treatment of EHRs — they are ever active data used in big data analytics, training AI algorithms and in ML. Further compounding the situation is how to archive EHRs themselves, that is, should the entire EHR system be archived, or should only selected components of the EHR be archived? Some of the early scholars, such as France and colleagues suggested that certain records be extracted from EHRs systems and earmarked for archiving [16]. They wrote “data will have to be extracted from the administrative database (admission, dates of admission and discharge, clinical services, notes of the physician, nurses, drug orders, laboratory computers, X rays, surgery and anaesthesia protocols, final report, medical record summary...)” [16].

However, they immediately lamented: “Will the record be still accessible? Will it be complete?” [16]. For example, in their study in the long term preservation abilities of EHRs among healthcare facilities in Barcelona, Bote, Termens, and Gelabert discovered that while most analogue health records easily became passive within a period of 3 – 5 years, the case was likely to be different from EHRs systems as information workflows and treatment in EHRs are completely different from an analogue environment and both active and passive EHRs were lying in the same electronic systems that however, were not running on the same software [6]. So, how do archival institutions exercise their archival role in such complex environments? This is a critical question for archivists as their response to it will determine how to perform the appraisal function for EHRs.

C. Technical Challenges of Digital Preservation

For decades, actually, from the time electronic records were born, archivists and technologists have been grappling with long-term preservation challenges. The greatest challenge with the long-term preservation of electronic materials such as EHRs is that access to electronic objects is only possible through the use of computer software [16]. To date, the search for a permanent solution to the challenge of long-term preservation of electronic records continues and remains elusive.

The question is, can archival institutions cope with the EHRs complexities from the perspective of physical custodianship and guarantee long term availability and accessibility of the EHRs? For example, if the entire EHR is to be archived, the question is who should do this and how? Should healthcare facilities migrate their EHRs system once they become due for archiving? Should the archival institution ingest the EHR system or migrate it to a new system for long term preservation? Bote, Termens, and Gelabert advise that an archival shift is necessary in such an environment [6].

Therefore, “an EHR is a complex unit of information that requires different treatment in the long term, not as a single

unit of information.” [6]. When the headache of long-term preservation of EHRs storage started, some attempts were made, for example, The Swedish Institute for Health Services Development, working with Sahlgrenska University Hospital created the project called DEJAVU, whose aim was to come up with a general format for the long-term preservation of electronic patient records [24]. At that time, the Swedish National Archives had approved paper and microfilm as the most stable and permanent storage mediums for archival preservation [11]. The project proposed the development of a standardised electronic patient record based on the Standard Generalized Markup Language, ISO 8879:1986/A1 :1988, allowing records to be extracted from their original EHRs systems and converted into a Standard Generalised Markup Language (SGML) format for long-term accessibility [24].

However, the sustainability of such a practice, as well as the ability of the new and stable SGML EHR to remain interactive and comprehensive over time under the care of an archival institution remains a concern, given that records would have to be extracted from their original systems into the SGML. Some of the most recent projects that have been conducted specifically focusing on the archiving of health information under the auspices of the European Union include the one titled “Archiving and Preservation for Research Environments”, which ran from 2019 through to 2022, scooping the EU award for developing several solutions for archiving and preserving scientific data in an accessible format [25]. The project works by combining several ICT technologies and business models in cloud environments, facilitating end-to-end archival and preservation services [25]. Another latest preservation method, although not specifically for the preservation of digital health information and data is Piql. Through its Piql 3- Piql-GO (Ultra-secure data storage- and long-term preservation of digital and analogue data), which ran between 2018 and 2021 [26].

The Piql system operates on a technology that converts “photosensitive film into digital, migration-free storage medium from which data can be guaranteed retrieved in the future, but never hacked, manipulated, or deleted” [26]. This shows the perennial challenge of preserving electronic data in general, and health data in particular. Interesting to note is that the full capacity of all the preservation technologies is yet to be fully tested over time, and the market will continue to be flooded with new technologies that promise to provide stable and permanent solutions to this challenge. In terms of relatively old projects, the Open Archival Information System (OAIS) has been recognized as one of the persistent old archival protocols that even newer preservation tools have used.

Thus, “originally developed as part of a broader effort to develop formal standards for the long-term storage of digital data generated from space missions, the Open Archival Information System (OAIS) has since formed the foundation of numerous architectures, standards, and

protocols, influencing system design, metadata requirements, certification, and other issues central to digital preservation” [27]. The OAIS architecture has been used in various digital preservation projects irrespective of the sector across the economy. The OAIS basically serves as a framework upon which to build preservation standards and infrastructure and may be used in electronic and traditional paper environments, with some arguing that it is not specifically preservation architecture for the preservation of digital content, but rather a set of conditions for the long-term preservation of content, irrespective of format. Figure 1 below summarises the OAIS architecture.

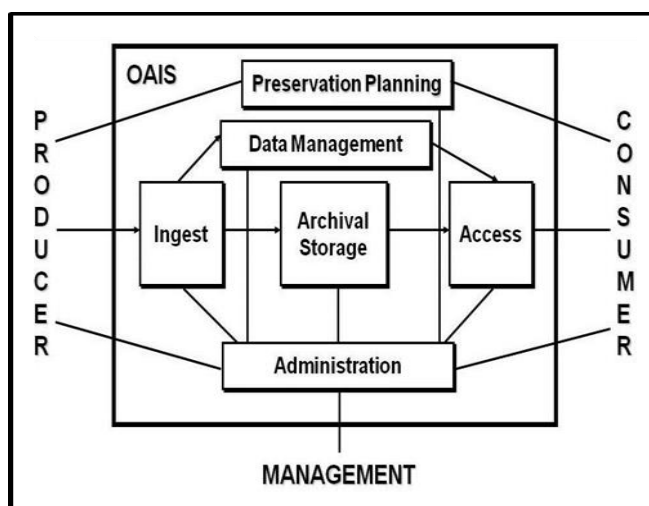


Fig. 1. The OAIS architecture
Source: Digital Preservation Coalition.

OAIS is a product of a Consultative Committee for Space Data Systems in collaboration with various digital preservation experts from other disciplines with the aim of informing the creation of digital preservation repositories [28]. The model, as explained by the Digital Preservation Coalition, has 6 components that comprise the framework, viz:

- i. Ingestion stage: as the first stage, this process serves as the interface between the source of the data, information or object, known as the producers. Basically, data to be ingested includes validation, by ensuring that the data is incorrupt, complete and suitable for archival storage. It also entails the extraction of descriptive metadata in order to support retrieval and discoverability of the data/record in the archival storage.
- ii. Archival storage: this could be considered as the main component of the model that is responsible for the desired long-term preservation and maintenance of the information that has been pushed into OAIS. Depending on the preferred mode of storage, the information can be maintained online, nearline or offline, ensuring that the content

remains accessible over time. To achieve this, various procedures such as refreshment or format migration have to be performed on a regular basis. Although not providing direct access to users, the storage function should be able to provide retrieval functionality for the data that is stored in the OAIS to support user requests.

- iii. Data management: this functionality is responsible for supporting the internal operations of the OAIS by serving as a database of descriptive the metadata and related descriptive tools that are necessary for the discovery of the information.
- iv. Preservation planning: this aspect ensures that the entire OAIS environment is compliant with preservation requirements. It works by monitoring the external environment to detect and respond to any threats to the preservation environment.
- v. Access: this component serves as a link between the OAIS system and the end user, commonly known as the community. Through this functionality, users can query the system for responses and presenting the retrieved information in a manner that is understandable to the user.
- vi. Administration: this final strand operates as a coordinating tool for all the afore-mentioned components of the model. It is responsible for overseeing the overall operation of the entire architecture, for example, by updating archival policies [27].

Although OAIS has been widely used as the basic architecture in digital preservation projects, there are other platforms such as Access to Memory (ATOM), Archivematica and Preservica that archivists have developed as digital preservation platforms. However, the last two, just like Piql, are commercial platforms that are accessible through subscriptions, pushing them beyond the reach of many archival institutions that are already financially struggling. Archivists have also explored alternative preservation methods such as system preservation, migration, refreshing and encapsulation but the ability of these techniques to cope with the challenges of digital preservation has been dwindling over the years. The birth of cloud computing services has been hailed for storage purposes, but the question of maintaining EHRs as complete, accessible and usable remains unanswered.

D. Fast Accumulation of EHRs, Coupled with Evolving Changing Standards for EHRs

Health, healthcare and wellness support generate enormous volumes of data [2]. The generation of such huge amounts of data by healthcare facilities across the globe in the form of EHRs is made possible by the current technological era [29]. Research showed that the health sector was set to become the most data intensive industry in

the world between 2018 and 2025. Figure 2 below shows data for this prediction.

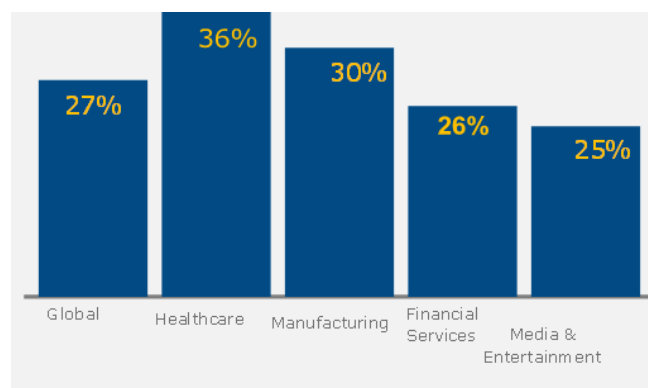


Fig. 2. Projected data growth by industry, 2018-2025
Source: IDC white paper [30]

Thus, the volume of health and clinical data worldwide is growing at an exponential rate [29]. Certainly, EHRs have a share in the projected health data growth depicted in Figure 1 above. But what does this mean from the perspective of archiving EHRs? Just like their traditional counterparts, EHRs storage and preservation for keeping the records available and accessible do not come for free. To date, a plethora of EHRs standards have been developed, including standards from Health Level Seven (HL7), International Organisation for Standardisation (ISO)/ Technical Committee (TC) 215 Secretariat on Health informatics. Some of the relevant standards by ISO include ISO 14641:2018 Electronic Document management-Design and operation of an information system for the preservation of electronic documents- Specifications, the European Committee for Standardisation (CEN) 254 and many others.

However, the afore-mentioned standards are best applied to the EHRs systems that create, store and preserve such records, rather than applying them to third parties, such as archival institutions that will have to extract such records from their original EHRs and covert them to new formats for the sake of long term-preservation. Questions of records completeness amid a multiplicity of EHRs components and EHRs vendors arise. Furthermore, the standards themselves change over time, implying that access to documents whose standards vary poses an additional challenge to the preservation of EHRs [16]. As Kiefer laments “as standards like HL7 and DICOM evolve over time, no one can really ensure that electronic medical documents will still be reusable after decades of preservation when formats, standards and applications become obsolete” [17].

Formats such as XML have been touted as the more stable for archival purposes as they are independent of format, thereby potentially addressing the challenge of ever-changing standards and formats of electronic content. Whilst XML appears to have relatively withstood the test of

time in terms of being receptive to changing standards following its invention in the last 1990s, the standard is not a bullet magic to the challenge of preserving electronic records. Thus, “not only software and hardware changes might be out of control, but mainly, it might appear dangerous to be confident in the future for the human management of various databases depending on various bodies” [16].

Archival institutions are not expected to have the capacity of acquiring and preserving EHRs systems that healthcare facilities use across the spectrum and archive them, even with the presence of standards such as XLM that is amenable to changing content standards.

E. Limited Archivists Training in Digital Preservation

Whilst archivists generally take digital preservation courses as part of their studies at college or university, most of them may lack the technical aspects of undertaking digital preservation projects. Preservation, arguably, is a highly technical exercise that goes beyond just knowing the requirements for preservation. As a practice, digital preservation is not the preserve of information technology experts, but several players [31], including archivists. In fact, digital preservation directly speaks to the duty of archivists and archival institutions to maintain records accessible for as long as they possess archival values such as research and historical values. Therefore, ordinarily, archivists are expected to be champions of digital preservation. However, they have instead, been found wanting in terms of digital preservation competencies.

A study by Ahmed that was conducted on the state of digital preservation staffing and skills availability in memory institutions, archives included, revealed stunning results where a whopping 59.9% of the studied institutions reported staffing shortages [32]. A further staggering 60.1% of the staff were found were suffering from a digital preservation skills gap [24]. Although these challenges were more pronounced in Africa and Asia, the developed regions were not spared as they continued to suffer from the same challenges.

In light of the prevailing skills and staffing gaps regarding digital preservation, coupled with the rest of the complexities associated with EHRs, it the distributed custody model can be explored as a possible a workable model of tackling the challenge. The skills shortcomings of staff at memory institutions as attributable to the highly technical side of digital preservation. For example, the Digital Preservation Framework that was developed by the Digital Preservation Coalition has an Information Technology strand which requires IT competencies that most memory workers may not immediately possess. The strand requires that for any one to be competent in digital preservation, they need to be able to:

- i. Develop algorithms and generate related programmes;
- ii. Understand and work with different electronic information systems and formats.
- iii. Design, document and use workflows that are necessary in the management of digital information.
- iv. Map out the requirements for a new system or services for the selection and procurement of a new solution.
- v. Understand the requirements for a storage infrastructure that is robust enough for ensuring the preservation of content at a bitstream level.
- vi. Understand and contribute to the implementation of appropriate security measures and processes [8].

These skills are not easily acquired by archivists, prompting them to heavily rely on IT staff for digital preservation projects. Given these challenges for memory institutions, archival institutions included, the proposed shared responsibility model may benefit from the already available IT pool at healthcare institutions who have been responsible for the implementation and maintenance of the EHRs and already enjoy relatively stable relationships with the vendors for technical support.

F. Constrained Budgets for Archival Institutions to Sustain Long-Term Preservation of EHRs Systems

Archival institutions across the world suffer from chronic underfunding, despite their pivotal role in the preservation of irreplaceable national documentary heritage. In the context of EHRs, to imagine such underfunded institutions ingesting and preserving diverse EHRs records from public healthcare facilities across an economy falls short of being practical. Given that EHRs are meant to be preserved in a live state where they are not just accessible, but are complete and preserve their meaning, an archival institution will need to either develop a format and standard-independent platform such as XML based architecture that will ingest such systems, retaining their functionalities, something that is practically costly for the already underfunded archival institutions.

G. Limited Archival Collaboration between EHRs, Vendors, Hospital IT and Archivists

Limited collaboration between health records and IT personnel at healthcare facilities and the national archival institutions also have a domino effect on the ability of archival institutions to preserve EHRs. As Corn put it: “another important complicating issue for the archiving of electronic information is the fragmentation of responsibility and diversity of technical systems for varying data types within a single enterprise: laboratory, pathology, imaging, sensor information, and other forms of data are often maintained by the unit performing the service, while the

EHR proper is the responsibility of the organization’s IT department” [8]. As the above mentioned situation continues to obtain across the EHR environment, it is obvious that the bulk, if not all the health records—active or inactive (archival), generated by EHRs are under the control of IT staff at healthcare facilities and vendors, thereby making them “default archivists”, yet there is limited dialogue between archivists and such technical personnel who are in real control of EHRs. Even in cases where there is a medical record librarian or archivist at a healthcare facility, they tend to bear influence on traditional paper records only.

Thus, the fragmentation is not only witnessed in the components of the EHR system, but between and among the personnel and institutions handling EHRs across the spectrum, including archivists and Archives, IT personnel in healthcare facilities and vendors. This has resulted in disjointed and limited efforts towards addressing the archival question as far as EHRs are concerned. This has manifested itself in the form of policies that are focused on the regulation of operational issues of EHRs, with a loud silence about the archival question, especially in terms of whose responsibility it is to preserve the archival health record beyond its operational value for immediate care of a patient. Figure 3 summarises the challenges associated with the preservation of EHRs from an archival perspective.

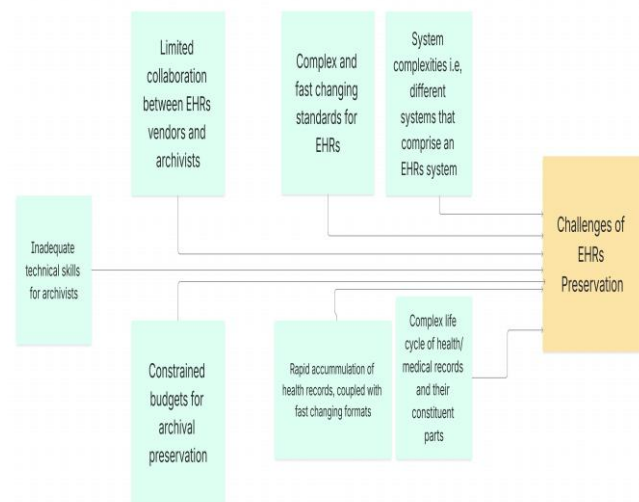


Fig. 3. Challenges of preserving EHRs.

IV. SHARED RESPONSIBILITY AS A MEANS OF ADDRESSING THE ARCHIVAL QUESTION FOR EHRs

A couple of archival scholars including Cunningham, Ngoepe, Hedstrom, Cook among other, have written about the concept of distributed custody in archives, almost agreeing that archivists must rethink their archival models for digital archives. As noted by Bastian [34], the idea of

shared responsibility in archival science, also known as post-custodialism, was first coined by Ham, who encouraged archivists to rethink their physical custodial role over archives, and allow a situation where archivists and archives let of the physical custody of records whilst retaining their legal control.

The National Archives of Australia [35] defined distributed custody as “as an arrangement whereby records are managed by another party in facilities not under the control of the Archives beyond the period where they would normally be transferred to the Archives or at the request of a personal donor. It covers situations where records are retained by an agency, as well as those where the records are transferred from the Archives' custody to a custodial organisation. The custodial organisation may be the creating or controlling agency, or other organisations such as a community or private organisation.” This implies that the creating agency is left to retain physical custody of the records, including archival ones, with archivists coming in to advise on best archival practices. To Ham, cited in Bastian [36], the insistence of archivists on the physical control of records, especially in the era of electronic records, is seen as a deterrent to the effective management of records.

Despite the shared responsibility having a relatively long history in archival science, it is yet to gain scholarly and practical traction in archival science as a handful of scholars have researched it, and a few archival institutions have implemented the model with clear and practical policies. The National Archives of Australia and the National Archives of Canada are some of the institutions that have led by example by providing and promulgating policies for the distributed custody approach to archiving, although the records provided for are not necessarily EHRs. In South Africa, Ngoepe [36] lamented the dilemma facing digital records where on one hand, such records are due for transfer to the National Archives and Records Services of South Africa (NARSSA) when on the other hand, the institution is ill-equipped to deal with electronic records.

In light of the challenges presented by EHRs, Ham's post-custodial approach is being proposed as the most plausible one to the preservation of EHRs. Whilst early scholars recommended a transfer of EHRs to standard archival storage medium, such as microfilm, DVD and CD-ROMS, clearly this system is no longer viable today. Despite such medium having become an obsolete technology, the very approach of extracting records from EHRs systems into a different platform implies migration from one platform or medium to the other, which in the case of EHRs, I argue, is not viable, due to the complications of the EHRs today and in the near and long future.

The multiplicity of different systems for different health operations during the lifetime care for a patient, supported by different vendors, operating on different platforms, technical incapacities of many archival institutions as well as incommensurate archival strategies for EHRs all work

against the traditional custodial approach to the long-term preservation of EHRs. As a result, the post-custodial/shared responsibility model, which archivists have long touted but not yet supported by clear archival policies and laws due to the fear of entrusting the archival function to healthcare facilities and vendors, is proposed. Actually, it can be argued that most archival institutions are failing to take custody of EHRs, thereby silently promoting distributed custody. Thus, the earlier archival institutions admit and regularize distributed custody as a more pragmatic approach to the management of EHRs, the better.

A. Archivists' Responsibility in the Distributed Custody Model

In this model, archival institutions and archivists stick to their legal supervisory mandate and loosen up on the custodial component. This means that archival institutions focus on the development of policies and advocate for the updating of archival laws that promote the shared responsibility model. Specifically, archivists may stop insisting on taking custody of EHRs as this demands them to have as many software applications as the number of EHRs variations that exist on the market, which is neither practical nor economic for poorly funded archival institutions. National archival institutions are responsible for all classes of records that are generated by public sector institutions, despite most of them operating on limited budgets, leaving them with limited to no capacity to adequately deal with the challenges of long-term preservation of EHRs. This will see archivists being responsible for setting policies and standards that EHRs vendors and healthcare facilities should ensure as they share the custody of EHRs, including retention and disposal schedules as well as access policies for EHRs. Specifically, archival institutions should be responsible for the following:

- i. Prescribing the types or categories of records in EHRs that should be preserved to meet the archival requirements.
- ii. Prescribing archival standards for long-term preservation of EHRs in a manner which guarantees their accessibility over time. It should be the responsibility of archival institutions and archivists to keep track of latest developments in terms of preservation standards and prescribe them for healthcare facilities.
- iii. Stipulating access conditions for archival records, ensuring that access is granted as per the provisions of the archival laws governing health records. Health and medical records are particularly susceptible to data breaches on top of being sensitive. It is imperative for archival institutions to communicate clear access and restriction conditions that are guided by archival laws for health records, irrespective of

the fact that they will be resident in the systems of healthcare organisations.

B. Responsibility of EHRs Vendors and Healthcare Facilities

Likewise, healthcare facilities and vendors where necessary, will be expected to take full or shared custody of EHRs systems, embedding in them, the archival policies and standards for the records to meet the archival needs of the health record, including research uses. This means that EHRs will no longer follow the traditional custodial cycle of records where in-active records are handed over to the archives, but remain in the system of the healthcare facility, in conjunction with vendors where necessary, to preserve them as archives. This means allowing the healthcare facilities and vendors to protect and make available, legally acceptable EHRs for as long as required by the law under the custody of their creators, duties that traditionally lie squarely with archivists and archival institutions. In this model, archivists, IT specialists and EHRs vendors work together and come up with the best framework that factors in the technical, operational and legal realities of all the stakeholders, at the same time protecting and propagating EHRs that meet legal and archival requirements for health records. Through this model, the policy, technical and communication fragmentation that currently characterizes the EHRs can be reduced, and a more coordinated long-term EHRs can be imagined. In the distributed custody model, healthcare facilities, working in collaboration with vendors, will be responsible for the following:

- i. Implementing long-term preservation of EHRs beyond their semi-active stage, ensuring their accessibility. This should include the implementation of digital preservation standards as recommended by the national archives as well as implementing sustainable digital preservation strategies.
- ii. Ensuring the protection of EHRs from potential disasters, including cyberattacks. This includes the creation of backup systems for EHRs.
- iii. Implementing appraisal and disposal schedules for EHRs in liaison with the national archives. This will ensure that healthcare facilities do not keep materials that do not have archival value, or records beyond their required period of retention. Health records are particularly known for their rapid accumulation, and the absence of clear and up-to-date retention schedules will see healthcare facilities incurring unjustified expenses of preserving records that are of no archival value.
- iv. Liaising with national archival institutions on all archival matters, including the provision or restriction of access to archival EHRs. This is particularly important as researchers and other users will be accessing such records at healthcare facilities.

C. Implications of the Distributed Custody Model in the Preservation of EHRs

The proposed distributed custody model, however, will not come without challenges. Whilst from a custodial perspective, this approach seems to be a panacea, it creates a new demand from archivists and archival institutions—that of educating and ensuring compliance with archival policies and standards for EHRs vendors and healthcare facilities. The success of the model depends on the cooperation of the vendors and healthcare facilities. The archival endeavor has been the prerogative of archivists, with creators and vendors only interested in the technology that supports the daily operations of healthcare facilities, with little to no regard for the records when they have fulfilled their primary administrative needs. Thus, giving them the archival role of preserving their inactive EHRs may not enjoy positive reception.

On the part of the healthcare facilities and vendors, the distributed custody model may come with additional expenses, for example, maintaining archival EHRs accessible as archivists would diligently do. Archival preservation is an expensive and cumbersome endeavor that vendors and healthcare facilities may not keep pace with. Thus, the model presents new, or rather extended responsibilities to archivists, healthcare facilities and the vendors, requiring a high degree of commitment and consensus from the three stakeholders. Healthcare facilities are already enjoying some cordial and sustained relations with vendors in the form of technical and other support mechanisms, further making it easy for collaboration to take place. For archival purposes, archival institutions may consider acquiring selected samples of unique cases that may be needed for research and other purposes that can be scanned or microfilmed for posterity rather than trying to acquire EHRs in their entirety as archives.

V. GOVERNANCE ISSUES

Since this is a proposal for the adoption of the distributed model for the preservation of EHRs, it is imperative to clarify the governance of the model. Governance considerations include a clear structure in terms of decision making, ensuring accountability and putting in place practical implementation techniques for the long-term preservation of EHRs. First, regarding how decisions are made, the model proposes national archival institutions as the key decision makers regarding an array of archival consideration, including the determination of retention and disposal schedules for EHRs, the types of health records to be archived, and methods of destruction. Such decisions should be communicated to healthcare facilities for implementation, with the supervision of national archival institutions. Thus, there must be a clear line of communication between national archival institutions that shall retain the role of overseeing all archival decisions and ensuring compliance by healthcare facilities. Healthcare facilities shall be required to consult national archival

institutions prior to destroying any health records. National Archival institutions should be able to conduct regular checks on the inventory of EHRs at healthcare facilities to account for all the records that are at healthcare facilities. There should be a committee comprised of representatives from the national archives and healthcare facilities that should oversee governance issues for the distributed custody model. Second is accountability. Healthcare facilities should be accountable to the national archives regarding archival actions taken on EHRs in their custody, including preservation.

Perhaps the greatest question is the fate of EHRs in the event of the healthcare facility discontinuing. Whilst this question is not easy to answer, it is prudent for all EHRs to be handed over to the national archives for continued preservation.

VI. LEGAL FRAMEWORK

Currently, most archival laws are crafted based on the physical custody model where archival institutions take full custody of all records that have been accorded with the archival status. Such laws, therefore, cannot function as a distributed custody model where the care of archival records is shared between archival institutions and the creating organisations, in this case, archival EHRs that may be resident in the systems of healthcare facilities. Archivists should review their archival laws and provide for the distributed custody model. Such new laws should clearly state the archival responsibilities of healthcare facilities for EHRs, including consequences for non-compliance. Having clear legal provisions for distributed custody will ensure that there are clear boundaries regarding responsibilities, making accountability easy to enforce and monitor.

VII. FINANCIAL IMPLICATIONS

Traditionally, the preservation of records of national significance has been, and remains primarily a prerogative of national archival institutions. Adopting the distributed custody model for the preservation of EHRs therefore implies the new financial responsibility for healthcare facilities. The models means that healthcare facilities have the responsibility of ensuring long-term preservation and accessibility of EHRs beyond their semi-active stage as archives. Healthcare facilities will be mandated to embed long-term preservation standards in their EHRs systems that should prolong the availability and long-term accessibility of EHRs. Additional financial burden will arise from the need for additional storage space for archival EHRs. For example, cloud space is one of the major expenses that directly arise for healthcare facilities in this model. This will require archival institutions to be proactive in the development of retention and disposal schedules that will ensure that healthcare facilities are not burdened with the responsibility of preserving everything, as is usually the case with traditional paper health records whose disposal is

usually by transfer to the archives, with limited to no destruction at all.

VIII. CONCLUSION AND FUTURE WORK

Although archivists have been reluctant to adopt and promote the shared responsibility approach to address the archival question in the case of electronic records, owing to them being “collecting professionals” with the keeping of records in the archives as a way of vouching for their authenticity, EHRs coupled with their complexities, undoubtedly present themselves as a good candidate for this approach. The proposed shared responsibility approach that this paper advocates for is going to be challenging as it needs very strong coordination efforts among archivists, healthcare IT experts and vendors, as well as a great deal of a mindset shift from the afore-mentioned stakeholders, allowing each of them to adapt to their new professional roles in this model, including incurring expenses.

It is recommended that archivists, EHRs vendors and healthcare organisations engage in feasibility discussions on the distributed custody approach for EHRs. Archivists should develop and strengthen policy and compliance mechanisms to ensure the success of the model, whilst vendors and healthcare providers should be willing to bear the archival responsibility of ensuring the long-term availability of EHRs in their custody, at the same time fulfilling the information needs of society and researchers. Empirical studies are required to understand the feasibility of the distributed custody approach from policy, legal, technical and financial considerations for the model to work in the context of EHRs. The model should be considered as a collaboration between the three stakeholders, rather than a loss of control of archives by archivists.

REFERENCES

- [1] M. Masuku, “Electronic health records and the archival question: Shared responsibility as a panacea,” *HEALTHINFO 2025: The Tenth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing*, Oct. 2025, pp. 63-67, ISBN:978-1-68-558-312-5 Thinkmind, [Online]. Available: https://www.thinkmind.org/library/HEALTHINFO/HEALTHINFO_2025/healthinfo_2025_1_120_90105.html (accessed May. 14, 2026).
- [2] R. E. Scott, “e-Records in health—Preserving our future,” *International Journal of Medical Informatics*, vol. 76, no. 5–6, pp. 427–431, Nov. 2006, doi: 10.1016/j.ijmedinf.2006.09.007.
- [3] J. M. Brokel, S. Ochylski, and J. M. Kramer, “Re-engineering workflows: changing the life cycle of an electronic health record system,” *Journal of Healthcare Engineering*, vol. 2, no. 3, pp. 303–320, Jan. 2011, doi: 10.1260/2040-2295.2.3.303.
- [4] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, “Data Processing and text mining technologies on Electronic Medical Records: a review,” *Journal of Healthcare Engineering*, vol. 2018, pp. 1–9, Jan. 2018, doi: 10.1155/2018/4302425.

- [5] F. M. Alanazi *et al.*, "Using electronic health records to improve patient health care," *Saudi Journal of Medicine*, vol. 8, no. 09, pp. 514–518, Sep. 2023, doi: 10.36348/sjm.2023.v08i09.008.
- [6] J. Bote, M. Termens, and G. Gelabert, "Evaluation of healthcare institutions for Long-Term Preservation of Electronic Health Records," in *Communications in computer and information science*, 2011, pp. 136–145. doi: 10.1007/978-3-642-24352-3_15.
- [7] M. T. Guaglianone *et al.*, "Ensuring the Long-Term preservation of and access to the Italian Federated Electronic Health Record," *Applied Sciences*, vol. 12, no. 7, p. 3304, Mar. 2022, doi: 10.3390/app12073304.
- [8] M. Corn, "Archiving the phenome: Clinical records deserve long-term preservation," *Journal of the American Medical Informatics Association*, vol. 16, no. 1, pp. 1–6, Oct. 2008, doi: 10.1197/jamia.m2925.
- [9] Digital Preservation Coalition, "Electronic hospital and medical records," *Electronic Hospital and Medical Records*, 2017. <https://bit-list.dpconline.org/entries/electronic-hospital-and-medical-records/> (accessed Mar. 09, 2026).
- [10] "Healthcare Archival - Digital Preservation Coalition: Open online meeting Thurs 20.11.2025 1130-1300," *openhealthhub.org*, Nov. 10, 2025. Available: <https://openhealthhub.org/t/healthcare-archival-digital-preservation-coalition-open-online-meeting-thurs-20-11-2025-1130-1300/2892> (Accessed Mar. 15, 2026).
- [11] W. Bartschat *et al.*, "The legal process and electronic health records," *PubMed*, vol. 76, no. 9, pp. 96A-96D, Oct. 2005. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/16274212> (accessed Mar. 15, 2026).
- [12] "Electronic health records explained," *ISO*. <https://www.iso.org/healthcare/electronic-health-records#toc7> (accessed Mar. 15, 2026).
- [13] N. S. Abul-Husn and E. E. Kenny, "Personalized medicine and the power of electronic health records," *Cell*, vol. 177, no. 1, pp. 58–69, Mar. 2019, doi: 10.1016/j.cell.2019.02.039.
- [14] T. Klementi, K. J. I. Kankainen, G. Piho, P. Ross, Department of Health Technologies, TalTech, Akadeemia Str 15A, 12618 Tallinn, Estonia, and Department of Software Science, TalTech, Akadeemia Str 15A, 12618 Tallinn, Estonia, "Prospective research topics towards preserving electronic health records in decentralised content-addressable storage networks," CEUR Workshop Proceedings, workshop, 2022. [Online]. Available: https://ceur-ws.org/Vol-3264/HEDA22_paper_7.pdf (accessed Mar. 15, 2026).
- [15] L. Dobrica, C. Alexandra, and T. C. Ionescu. E, "Towards a better understanding of EHR systems using architectural views," 2013, pp. 362–365. doi: 10.5220/0004246603620365.
- [16] F. H. France, C. Beguin, R. Van Breugel, and C. Piret, "Long term preservation of electronic health records. Recommendations in a large teaching hospital in Belgium," *PubMed*, vol. 77, pp. 632–6, Jan. 2000, [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/11187630> (accessed Aug. 3, 2025).
- [17] K. Stephan, S. Michael, and R. Jochen, "A semantic approach for digital Long-Term preservation of electronic health documents," *Studies in Health Technology and Informatics*, vol. 180, pp. 265–9, Jan. 2012, doi: 10.3233/978-1-61499-101-4-265.
- [18] J. Bote, "Electronic Health Record Proposal for Long-Term Preservation," in *IGI Global eBooks*, 2013, pp. 529–548. doi: 10.4018/978-1-4666-3986-7.ch028.
- [19] Mansoura University Specialised Medical Hospital, "Medical archives," 2025. <https://smh.mans.edu.eg/index.php/en/component/content/article/312-patient-affairs/5017-med-archive-en> (accessed Jul. 1, 2025).
- [20] M. Amatayakul *et al.*, "Definition of the health Record for legal purposes: (AHIMA Practice Brief)," *Journal of AHIMA* 72, no.9 (2001): 884-H., Oct. 2001, [Online]. Available: <https://bok.ahima.org/doc?oid=105823> (accessed Mar. 5, 2026).
- [21] Archives of Ontario, "Medical Records at the Archives of Ontario: An overview," 2012.
- [22] J. M. S. França, J. De S Lima, and M. S. Soares, "Development of an Electronic Health Record Application using a Multiple View Service Oriented Architecture" 2017, pp. 308–315, isbn: 978-989-758-248-6, doi: 10.5220/0006301203080315.
- [23] A. Abernethy *et al.*, "Toward a national health digital and data architecture: laying the foundation for digital transformation," *NAM Perspectives*, vol. 3, Mar. 2026, doi: 10.31478/202603b.
- [24] T. Wigefeldt, S. Larnholt, and H. Peterson, "Development of a standardized format for archiving and exchange of electronic patient records in Sweden," *PubMed*, vol. 43 Pt A, pp. 252–6, Jan. 1997, [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/10179549> (accessed Jul. 2, 2026).
- [25] "Archiving and preservation for research environments." <https://www.archiver-project.eu/> (accessed Mar 11, 2026).
- [26] "Ultra-secure data storage - and long-term preservation of digital and analogue data," <https://cordis.europa.eu/project/id/804477/reporting> (accessed Mar. 12, 2026).
- [27] Lavoie, "The Open Archival Information System (OAIS) reference model: Introductory Guide (2nd Edition)," *DPC Technology Watch Report 14-02 October 2014*, 2014. <https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file> (accessed Mar. 11, 2026).
- [28] R. S. Bettivia, "The power of imaginary users: Designated communities in the OAIS reference model," *Proceedings of the Association for Information Science and Technology*, vol. 53, no. 1, pp. 1–9, Jan. 2016, doi: 10.1002/pra2.2016.14505301038.
- [29] S. M. Shah and R. A. Khan, "Secondary Use of Electronic Health Record: Opportunities and Challenges," *IEEE Access*, vol. 8, pp. 136947–136965, Jan. 2020, doi: 10.1109/access.2020.3011099.
- [30] D. Reinsel, J. Gantz, J. Rydning, and IDC, "The digitization of the world: From edge to core," US44413318, Nov. 2018. [Online]. Available: <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf> (accessed Mar. 4, 3036).
- [31] T. Bähr, M. Lindlar, S. Vlaeminck, German National Library of Science and Technology, German National Library of Medicine, and German National Library of Economics, "Puzzling over digital preservation – Identifying traditional and new skills needed for digital preservation," Jun. 2011.
- [32] R. Ahmad, M. Rafiq, M. F. Khalil, and M. H. Khalil, "Global overview of digital preservation staffing and skills: A systematic review," *Information Development*, Jul. 2025, doi: 10.1177/02666669251354893.

- [33] Digital Preservation Coalition, “Digital Preservation Coalition Rapid Assessment Model (DPC RAM),” Mar. 2025. doi: 10.7207/dpcram21-02.
- [34] J. A. Bastian, “Taking custody, giving access: a postcustodial role for a new century,” *Archivaria*, vol. 53, pp. 76–93, Jan. 2002, [Online]. Available: <https://archivaria.ca/index.php/archivaria/article/view/12838> (accessed Jul. 1, 2015).
- [35] National Archives of Australia, “Distributed Custody Policy,” <https://www.naa.gov.au/about-us/our-organisation/accountability-and-reporting/archival-policy-and-planning/distributed-custody-policy> (accessed Mar. 4, 2026).
- [36] M. Ngoepe, “Archival orthodoxy of post-custodial realities for digital records in South Africa,” *Archives and Manuscripts*, vol. 45, no. 1, pp. 31–44, Jan. 2017, doi: 10.1080/01576895.2016.1277361.

Clarifying Research and Education Needed to Explain the Medical Record to Patients: Protocol for UK Survey of Non-medical Practitioners

Ray B. Jones, John Downey, Tristan Coombe, Avril Collinson, Aled Jones, Sally Abey, Patricia Schofield, Jenny Freeman, Nicholas Peres, Jill Shawe, Sheena Asthana

Faculty of Health, University of Plymouth, PL4 8AA, United Kingdom

Emails: ray.jones@plymouth.ac.uk; john.downey@plymouth.ac.uk; tristan.coombe@plymouth.ac.uk; avril.collinson@plymouth.ac.uk; aled.jones@plymouth.ac.uk; sally.abey@plymouth.ac.uk; patricia.schofield@plymouth.ac.uk; jenny.freeman@plymouth.ac.uk; nicholas.peres@plymouth.ac.uk; jill.shawe@plymouth.ac.uk; sheena.asthana@plymouth.ac.uk.

Abstract— Online record access (ORA) is expanding across the British National Health Services and internationally, and patients increasingly use generative AI (GAI) tools to interpret clinical information. Despite this shift, research and curricula have focused largely on doctors, leaving a substantial gap in understanding the needs and perspectives of nurses, midwives, and allied health professionals. This paper presents a multi-level framework for explainability in AI-mediated patient record access and a detailed protocol for the first United Kingdom national survey of non-medical practitioners on the topic. The survey will identify research priorities and curriculum requirements related to ORA and patients' use of GAI using a novel design that embeds 30 minutes of Continuing Professional Development to engage busy clinical academics. Findings will inform co-designed approaches to explainability, support equitable implementation, and guide curriculum development for the future workforce. Although UK-based the work should be relevant to other countries.

Keywords— *explainability; patient access to records; research and education agenda; non-medical practitioners; survey protocol.*

I. INTRODUCTION

Health systems internationally are undergoing a fundamental transformation in how clinical information is generated, stored, and shared. Patient online record access (ORA) is becoming widely implemented across primary and secondary care, enabling patients to view consultation notes, laboratory results, correspondence, and care plans in near real time. At the same time, generative artificial intelligence (GAI) tools such as ChatGPT, Copilot, and Gemini are increasingly used by patients to interpret, summarise, and query clinical information. Together, these developments create a new information environment in which patients are no longer passive recipients of professionally mediated explanations, but active users of algorithmically generated interpretations of their own health data.

This convergence of ORA and GAI fundamentally alters the conditions under which clinical meaning is produced, communicated, and acted upon. Explainability, the ability of a system to make information intelligible, usable, and trustworthy

to its users, has traditionally been treated as a property of software systems or machine-learning models. In healthcare, however, explainability is not produced by algorithms alone. It emerges from interactions between digital records, AI systems, healthcare professionals, organisational workflows, and patients' and their families' own capabilities, experiences, and vulnerabilities. Explainability in this context is therefore a socio-technical property of healthcare systems rather than a purely technical feature of software.

Most research on ORA and clinical AI has focused on doctors, particularly in primary care, even though nurses, midwives, and allied health professionals (AHPs) deliver the majority of patient contacts and are often the clinicians who help patients interpret what they see in their records. Yet their perspectives on patient-accessible records, AI-mediated explanations, and the implications for professional practice, communication, and safety remain largely absent from the literature.

This omission matters for three reasons. First, non-medical practitioners routinely manage clinical information that is highly technical, longitudinal, and functionally oriented, including rehabilitation plans, mental-health notes, functional scores, and therapy goals. Second, they play a central role in helping patients and families make sense of clinical data, especially in long-term and community-based care. Third, emerging AI tools are increasingly capable of generating persuasive, empathetic, and apparently authoritative explanations, which patients may bring into clinical encounters, reshaping professional roles, responsibilities, and power dynamics.

At the same time, professional education and regulatory standards for nurses, midwives, and AHPs have not kept pace with these developments. While digital literacy and record-keeping are emphasised, there is little formal preparation for responding to patients who have continuous online access to their records or who use AI to interpret them. This creates a growing gap between the information environment patients inhabit and the training received by the workforce expected to support them.

This paper addresses that gap by advancing a conceptual and methodological framework for studying explainability in AI-enabled patient record access, grounded in the realities of non-medical clinical practice. We propose that meaningful explainability operates across multiple levels of the health system—individual, organisational, and systemic—and that research and curriculum development must be aligned across these levels. To operationalise this approach in the United Kingdom (UK), we present a protocol for a UK-wide study that uses a novel combination of bibliometric sampling and a CPD-embedded survey to identify research priorities and curriculum needs among academic nurses, midwives, and AHPs.

By integrating socio-technical theory, digital health informatics, and professional education, this work contributes a scalable framework for understanding and shaping how AI-mediated record access is explained, governed, and taught within contemporary healthcare systems.

This paper makes three contributions: (i) it conceptualises explainability in AI-mediated patient record access as a multi-level socio-technical property; (ii) it derives research and curriculum requirements from that framework for non-medical practitioners; and (iii) it presents a novel CPD-embedded survey design to operationalise this agenda at national scale.

II. CONCEPTUAL FRAMEWORK: EXPLAINABILITY AS A MULTI-LEVEL SOCIO-TECHNICAL PROPERTY

A. *Explainability Beyond the Algorithm*

In the artificial intelligence literature, explainability is typically framed as a property of models—how well a system can make its outputs transparent, interpretable, or justifiable to users. In healthcare, however, the outputs of AI are rarely consumed in isolation. They are embedded within electronic health records, patient portals, clinical workflows, professional judgement, and patient–clinician relationships. As a result, whether an explanation is meaningful, safe, or actionable depends not only on the algorithm, but on how it is situated within a wider socio-technical system.

When patients access their records through portals such as the National Health Service (NHS) App or hospital systems and then use GAI tools to interpret those records, explainability becomes distributed across at least three interacting components: the record itself, the AI that transforms it, and the clinician who contextualises it. This creates a triadic relationship between patient, practitioner, and AI, in which responsibility for meaning-making is shared and contested. Errors, misunderstandings, or misalignments can arise at any point in this triad and cannot be fully understood by examining any single component in isolation.

B. *Three-Level Model of Explainability*

To capture this complexity, we conceptualise explainability in AI-enabled patient record access as operating across three interdependent levels: micro, meso, and macro.

At the **micro level**, explainability concerns the interaction between an individual patient and their health information. At this personal level, this includes how clinical data are presented, how AI tools translate or summarise them,

and how explanations are adapted to patients' literacy, cognitive capacity, emotional state, and cultural context. Micro-level explainability determines whether a particular patient can understand what is written in their record and what an AI-generated explanation means for the understanding their own health.

At the **meso level**, explainability is shaped by organisational and professional contexts. At this team or service level, this includes how clinicians document information when they know patients will read it, how teams integrate patient portals and AI tools into workflows, and how practitioners respond when patients bring AI-generated interpretations into consultations. Here, explainability becomes a property of clinical relationships, professional norms, and service design, rather than of individual technologies.

At the **macro level**, explainability is influenced by system-wide structures such as digital infrastructure, regulatory frameworks, workforce policies, and health-inequality dynamics. At this system-wide level, decisions about whether AI tools are integrated into patient portals, how data are governed, and how digital inclusion is addressed determine which populations benefit from AI-mediated explanations and which are left behind. Macro-level explainability therefore links technical design to equity, safety, and system performance.

These three levels are not independent. Changes at one level propagate to others: e.g. national policies to provide real-time access to records (macro) alter documentation practices (meso) and affect how individual patients encounter and interpret their data (micro). Conversely, patient experiences at the micro level can drive organisational change or policy reform.

C. *Implications for Research and Education*

This multi-level view implies that research on ORA and GAI cannot be confined to usability testing or algorithmic performance. It must also examine professional practice, organisational adaptation, and system-level consequences. Similarly, education for nurses, midwives, and AHPs cannot be limited to digital skills training; it must prepare practitioners to operate within a patient–AI–clinician triad, managing uncertainty, correcting errors, supporting vulnerable users, filling important gaps in understanding, and negotiating shared decision-making in an AI-mediated information environment.

The conceptual framework therefore links three domains: (i) the socio-technical dynamics of explainability, (ii) research priorities across micro, meso, and macro levels, and (iii) the competencies required of the future workforce. This alignment underpins the study design presented in the paper, which seeks to identify and prioritise research and curriculum needs among non-medical clinical academics who are uniquely positioned at the intersection of practice, education, and digital health innovation.

D. *Co-design as structured socio-technical negotiation*

Co-design in this context should not be understood simply as bringing technical developers and clinical practitioners together in workshops. The groups involved in AI-mediated record access are likely to hold different, and sometimes conflicting, priorities: developers may emphasise technical

performance, scalability and integration; clinicians may emphasise safety, professional accountability and workflow burden; educators may emphasise curriculum feasibility; and patients and carers may emphasise comprehensibility, trust, autonomy and equitable access. A useful co-design process therefore requires an explicit framework for surfacing and resolving these tensions.

We propose that subsequent co-design work should use the micro–meso–macro model as a structuring device. At the micro level, patient-facing explanations would be assessed for clarity, emotional impact, accessibility and risk of misinterpretation. At the meso level, proposed solutions would be assessed against clinical workflow, documentation practices, professional responsibility and the effect on practitioner–patient relationships. At the macro level, decisions would be assessed against governance, digital inclusion, data protection, scalability and system-level equity. In practical terms, co-design workshops would use patient scenarios, example record extracts, AI-generated explanations and service-use cases as shared “boundary objects” through which different stakeholders can identify benefits, risks and trade-offs.

Where priorities conflict, resolution should not rely on simple majority preference or technical feasibility alone. Instead, candidate solutions should be prioritised using explicit criteria: clinical safety, patient comprehensibility, equity, feasibility within workflow, professional accountability and governance acceptability. Safety, privacy and equity would be treated as minimum requirements rather than optional preferences. Disagreements would be documented as part of the design process, with unresolved tensions becoming research questions or implementation risks rather than being hidden. In this way, co-design becomes a deliberative socio-technical process for aligning patient needs, clinical practice, educational requirements and technical development, rather than a generic consultation exercise.

III. EMPIRICAL AND POLICY CONTEXT FOR EXPLAINABILITY

The conceptual framework outlined above builds on a long trajectory of research demonstrating that access to clinical information is beneficial to patients when it is accompanied by appropriate explanation. We previously articulated a research agenda for non-medical practitioners in explaining the medical record to patients, positioning explainability as a shared clinical responsibility rather than a purely technical function [1].

Early randomised controlled trials conducted over two decades ago demonstrated that providing patients with personalised explanations of their medical records improved understanding, reduced anxiety, and increased constructive use of information [2–4]. In cancer care, patients given computer-generated, clinically relevant summaries were more likely to share information with family members and reported lower anxiety than those given generic information alone [2]. Similar benefits were observed in patients with schizophrenia [3] and across a range of long-term conditions [4]. These studies established that access alone is insufficient: explanation is a necessary mediator of benefit.

More recently, large-scale evidence from electronic health records has confirmed and extended these findings. A 2020 systematic review and meta-analysis found that patient access to electronic records improved medication safety and was associated with reduced healthcare utilisation, including fewer hospital visits and appointments [5]. Subsequent studies have shown improvements in patient empowerment, error detection, and shared decision-making when online record access is implemented routinely [6–8]. However, these benefits are highly contingent on how systems are deployed. Differences in interface design, clinical workflows, and patient support produce widely varying outcomes, reinforcing the importance of meso- and macro-level explainability [9]. Without attention to digital literacy, language, and organisational practice, record access can increase clinician burden and exacerbate health inequalities [10].

These tensions are evident in current clinical attitudes. In England, although prospective access to primary care records became the default in October 2023, a national survey of general practitioners found that only one-third supported the policy in principle. Most anticipated increased patient anxiety and confusion and feared additional workload [11]. Qualitative research with primary care staff has echoed these concerns, highlighting uncertainty about how to document sensitive information and how to manage patient reactions [12]. These findings illustrate that explainability is not only a technical issue but a professional and organisational one.

At the system level, UK policy explicitly links digital access to patient empowerment. The NHS Long Term Plan and associated digital strategies emphasise prevention, self-management, and care closer to home, all of which depend on patients being able to understand and act on their health information [13]. England is exploring how GAI might be used in the NHS App and other digital services [14]. This policy environment has accelerated the integration of patient portals and record-sharing platforms across all levels of care.

In practice, patients now encounter a complex digital ecosystem. High-quality clinical information is available from authoritative sources such as the NHS, Mayo Clinic, and NICE [15–16], while personal health records generated in systems such as EMIS, SystmOne, Cerner, and EPIC [17–20] are increasingly made visible through patient portals and the NHS App [21]. At the same time, GAI is transforming how this information is consumed [22]. Patients commonly use tools such as ChatGPT or Copilot to translate medical jargon, summarise consultation notes, and explore possible meanings of test results. Health IT vendors are now embedding these capabilities directly into patient portals, with major platforms collaborating with AI providers and NHS England exploring similar functionality within the NHS App. This creates the conditions for an AI-mediated explainability layer operating between patients and their records.

These developments offer substantial opportunities. AI can provide personalised summaries, conversational explanations, and adaptive support that may enhance understanding and engagement, particularly for patients managing complex or long-term conditions. However, they also introduce new risks,

including hallucinated content, lack of provenance, data-privacy concerns, friction between patients and clinical staff, and the potential to widen digital inequalities [23]. Without careful integration into clinical workflows and professional practice, AI-mediated explanations may increase anxiety, misinterpretation, and mistrust rather than empowerment.

It is within this evolving empirical and policy context that the need for a multi-level approach to explainability becomes clear. Patients are already using AI to interpret their records; systems are already sharing those records by default; and clinicians, including nurses, midwives, and AHPs, are already being asked to manage the consequences. The remaining question is not whether AI-mediated explainability will occur, but how it should be governed, researched, and taught.

IV. RISKS OF UNMANAGED AI-MEDIATED EXPLAINABILITY

The rapid convergence of ORA and GAI creates not only opportunities for patient empowerment, but also new and under-recognised risks for clinical safety, professional practice, and health equity. Patients are already using public GAI tools to interpret laboratory results, consultation notes, and diagnostic reports, often outside secure clinical environments and without any professional mediation. Studies comparing AI chatbot responses to clinician advice have shown that AI can generate fluent and persuasive explanations that are sometimes inaccurate, incomplete, or clinically unsafe, particularly in complex or ambiguous cases [39,40]. When such outputs are applied to personal health records, the risk is not merely theoretical: patients may be reassured inappropriately, alarmed unnecessarily, or encouraged to take actions that conflict with clinical advice.

Nurses, midwives, and AHPs can be the first professionals patients consult when they are confused, distressed, or uncertain about what they have read in their records or generated through AI. In community and long-term care settings, they manage much of the relational, educational, and interpretive work that supports patient understanding. Yet they currently do so without formal guidance, training, or system-level support for dealing with AI-mediated explanations. This creates a growing mismatch between what patients can access and what the workforce has been prepared to manage.

At the organisational level, unmanaged AI use can also destabilise documentation practices and professional accountability. If clinicians know that their notes will be read, transformed, and potentially misinterpreted by AI systems, this may alter how they write, what they record, and how candidly they document uncertainty. At the system level, the unregulated spread of AI-mediated explainability risks exacerbating digital exclusion, as patients with higher digital literacy and access gain enhanced interpretive tools while others are left with opaque records and limited support.

Taken together, these dynamics mean that explainability is no longer a design choice but a patient-safety and workforce-preparedness issue. Without empirical evidence on how non-medical practitioners experience, manage, and prioritise these

challenges, health systems risk embedding AI-mediated record access in ways that increase misunderstanding, workload, and inequality rather than reducing them. This makes a systematic investigation of research and curriculum priorities for this workforce not optional, but necessary for the safe and equitable evolution of digitally enabled care.

V. TARGET GROUPS FOR FURTHER STUDY

In the English NHS, there are approximately 172,000 doctors (134,000 hospital doctors and 38,000 full-time equivalent GPs). However, there are some 372,000 nurses and midwives, and over 200,000 AHPs (healthcare professionals other than doctors and nurses) from 14 professions (such as physiotherapy, podiatry, dietetics) working across community, primary, and secondary care. Yet, very little is known about nursing or AHPs' or patients' attitudes to patient access to their records or the use of GAI in non-medical clinical situations. For example, a recent scoping review of patient-accessible electronic health records [25] identified 66 studies, with none addressing nursing or AHP attitudes or GAI use in those settings. Therefore, it is important to establish what needs to be researched and what should go into a research-informed curriculum for these non-medical practitioners (Figure 1).

Within nursing, midwifery and the allied health professions there is a varied history of patient access to their records and use of GAI. For example, antenatal care has a long history of providing patients with access to their records [26]. It continues to lead in shared record practices, with handheld notes and digital maternity apps now widely used. While, in physiotherapy, the largest of the allied health professions, routine access to clinical notes remains less common, and many departments still rely on paper records or standalone systems. This mix of promising developments and significant gaps offers both researchers and educators diverse range of environments in which to explore and evaluate innovative approaches.

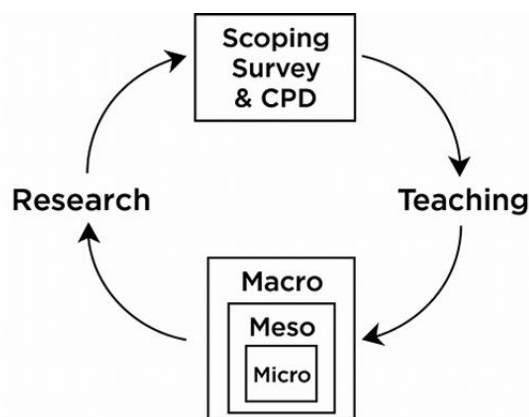


Figure 1. Multi-level socio-technical model linking explainability, research, and curriculum development for online record access and generative AI.

VI. RESEARCH QUESTIONS ABOUT EXPLAINING MEDICAL RECORDS TO PATIENTS.

Research questions about patient access to explained medical records can be grouped at three levels (Table 1):

‘Micro’ level, (personal level) the explainability of the record, exploring which types of explanation are preferred or are more useful.

‘Meso’ level, (service or team level) whether patients want to use portals and whether their use and GAI affects the practitioner-patient relationship, and

‘Macro’ level, (system-wide level) how this transformation can affect patient outcomes and possible changes to care processes, such as the shift from acute to community care and the focus towards health promotion and disease prevention [24].

VII. A RESEARCH INFORMED CURRICULUM FOR NON-MEDICAL PRACTITIONERS

Nursing and midwifery: Despite national policy commitments to digital transformation and patient empowerment, nursing education in England has not yet formally incorporated competencies related to ORA or patient use of GAI. The Nursing and Midwifery Council (NMC) standards for registered nurses and Health Education England’s digital capability framework set out broad expectations for digital literacy and record keeping but make no reference to ORA or patients’ use of GAI [27,28]. In nursing, professional risk guidance recommends local training for staff producing notes accessible to patients, but such training remains workplace-based and optional rather than embedded in educational standards [29]. We therefore need to consider integrating ORA-related competencies or related patient use of GAI into pre-registration curricula.

Table 1. Multi-level research agenda for explainability in online record access and generative AI.

Level	Explainability focus	Primary AI-related risks	Key research constructs	Implications for professional education
Micro (patient level)	How individual patients understand, trust, and use information from their clinical record when it is mediated by generative AI	Misinterpretation, hallucinated explanations, inappropriate reassurance or alarm, unequal access for digitally disadvantaged patients	Health literacy, cognitive load, trust in AI, comprehension of records, accuracy of AI explanations, patient self-efficacy	Teaching clinicians how to explain records in patient-centred language; recognising limits of AI-generated explanations; supporting vulnerable and low-literacy patients
Meso (service and professional level)	How clinical teams integrate ORA and AI explanations into documentation, communication, and consultations	Role confusion, defensive documentation, erosion of professional authority, increased workload, conflict over AI-generated interpretations	Workflow adaptation, documentation practices, clinician–patient–AI interaction, professional responsibility, communication quality	Preparing practitioners to manage AI-informed patients; ethical use of AI; handling disagreement between AI and professional judgement; maintaining therapeutic relationships
Macro (system level)	How health systems govern, scale, and regulate AI-mediated explainability in patient record access	Inequitable access, privacy breaches, unsafe automation, misaligned incentives, widening health inequalities	Digital infrastructure, governance frameworks, regulatory standards, cost-effectiveness, equity and inclusion	Policy literacy, digital governance, understanding system-level risks and benefits, preparing the workforce for AI-enabled care models
Cross cutting	What factors interplay at micro, meso and macro levels to impact the optimal deployment of ORA and GAI in practitioner-patient relationships?	Misalignment between policy, technology design, organisational practice, and patient capability; fragmented accountability; uneven implementation that benefits some groups while disadvantaging others	Implementation science; socio-technical alignment; co-design with patients and clinicians; digital inclusion; governance–practice feedback loops; learning health systems	Understand how implementation and optimal deployment of GAI/ORA is mediated through micro, meso and macro system factors

Allied Health Professions: Despite being responsible for over 200 million patient contacts annually across rehabilitation, diagnostics, therapy, and long-term condition management [24], AHPs have been almost entirely absent from the ORA and

patient-AI literature. Unlike the NMC, where professional standards reference documentation and patient communication explicitly, AHP regulation via the Health and Care Professions Council (HCPC) emphasises record keeping, informed consent, and communication, but does not address patient online access or the implications of patients using GAI to interpret therapy notes, imaging reports, or rehabilitation plans. This omission is particularly consequential because AHP records often contain highly technical, functional, and prognostic information (e.g., radiology reports or therapy progress notes) that patients are increasingly able to view via portals such as the NHS App. Without explicit curricular preparation, AHPs are left to navigate how to respond when patients arrive with AI-generated interpretations of scan findings, therapy goals, or outcome measures. Integrating ORA- and GAI-related competencies into AHP education is therefore essential not only for patient understanding, but also for preserving professional accountability, managing risk, and supporting shared decision-making across rehabilitation and diagnostic pathways.

To clarify the link between research priorities and educational needs, Table 2 maps the three levels of inquiry, micro, meso, and macro, onto corresponding curriculum implications. This alignment illustrates how findings from

future research can directly inform pre-registration training for nurses, midwives, and AHPs. By presenting these connections visually, the table reinforces the argument that curriculum development should be evidence-based and responsive to emerging digital health trends, including ORA and GAI.

VIII. PROTOCOL FOR A CROSS-SECTIONAL SURVEY, CPD AND NETWORKING

It is important to ask nursing, midwifery and AHP researchers, as an expert group, for their views as to (i) research priorities and (ii) what goes into the curriculum regarding patient ORA and related use of GAI to explain them. This is a rapidly evolving field, and non-medical health academics themselves may struggle to keep pace with advances in digital and AI technologies, given their competing clinical, teaching, and administrative commitments [30,31].

Table 2. Curriculum requirements derived from the multi-level explainability framework

Level	Explainability failure mode	What graduates must be able to do	What must be taught	What should be assessed
Micro (patient–AI interaction)	Patients misinterpret records or AI explanations, leading to anxiety, false reassurance, or unsafe decisions	Recognise misunderstanding; explain record content in patient-centred language; correct misleading AI output; support digitally vulnerable patients	Principles of explainability; health literacy; limitations and hallucinations of GAI; communicating uncertainty; adapting explanations to patient context	Ability to interpret AI-mediated patient queries; simulated consultations involving AI-generated explanations; recognition of unsafe or misleading outputs
Meso (clinical teams and workflows)	AI and ORA disrupt professional roles, documentation practices, or therapeutic relationships	Integrate ORA and AI into safe documentation and consultations; manage disagreement between AI and professional judgement; maintain trust and accountability	Professional responsibilities in AI-mediated care; documentation for patient visibility; ethics of transparency; managing digitally informed patients	Case-based assessments involving AI-informed patients; documentation exercises; ethical reasoning about AI use in clinical encounters
Macro (health system and equity)	AI-mediated explainability increases inequity, undermines governance, or creates unsafe system-level effects	Understand how digital policy, regulation, and infrastructure shape patient experience; recognise system-level risks and inequalities	Digital health governance; data protection; equity and inclusion in digital systems; NHS digital strategy and regulation	Policy analysis tasks; evaluation of digital health interventions for equity and safety; understanding regulatory and governance frameworks

To encourage respondents to complete, the survey will be paired with a learning opportunity on the topic, and a Continuing Professional Development (CPD) certificate upon completion (Figure 1). In doing so, this study uses a structured knowledge-exchange design, in which participants receive curated evidence and contribute expert judgement. To have further impact the survey will aim to set up a Smart Patient Academic Network for further collaboration on empowering patients by addressing research and teaching on ORA and GAI.

To operationalise this multi-level explainability framework, we have designed a national study that aims to identify and prioritise research priorities and curriculum content for incorporating ORA and related GAI into non-medical health education in the UK and to test the strategy of combined research investigation with CPD for the respondents.

The study design is a cross-sectional survey of non-medical clinical university staff who have experience of research and teaching. The survey conducted online simultaneously offers CPD and a CPD certificate to respondents on patients' ORA and use of GAI. The survey will include structured ratings and rankings and open-ended qualitative questions. The survey will not be anonymous if respondents wish to be awarded a CPD certificate for their completion of the learning aspects of the survey.

IX. RECRUITMENT STRATEGY

We aim to recruit UK university-based nursing, midwifery and AHP clinical academics. These will have a "360-degree view" (as researchers, teachers, and patients or confidants to patients) of ORA and GAI. We considered (but discounted) recruiting via professional bodies such as Royal College of Midwives, Chartered Society of Physiotherapy, and the British Dietetic Association, but (a) this may have resulted in many non-university contacts (so ineligible for the survey, and (b) we wanted to explore a non-NHS method requiring only University ethical approval, to optimise feasibility.

Participants will be recruited using a snowball method starting with university academics who have published. The survey therefore will be a structured expert consultation rather than a prevalence survey of representative academics. We will 'seed' the snowball with a large ($n=400$) quota sample of authors, identified via Web of Science. These authors are likely to be eligible, will be invited to complete it themselves, and will also be asked to forward it to colleagues who they think may be interested. First, we identified relevant publications in 2023-2025. Second, we extracted the email addresses of the authors where they were employed by a UK university. This returned a large sample of 2087 authors, many of whom are medical, dental, or non-clinical researchers. Internet searches were used to identify a quota sample of 400 (20%) within the 2087 people that appear to be qualified as nurses, midwives or AHPs. Appendix 1 gives a detailed description of how the 'snowball seeds' were identified and the invitation email.

We estimate that with 400 snowball seeds, we may recruit a sample of 200 (an estimated 100 from the seeds and a further 100 from a forwarded email). This would give sufficient precision on proportions estimated from closed questions and

sufficient breadth of recruitment for the thematic analysis of open questions. For example, if the proportion of clinical non-medical academics who (Q17 (Appendix 2)) think it is important or essential to include ORA and GAI in the curriculum is 60%, the 95% confidence interval would be 53.2% to 66.8%. For an exploratory survey, we propose that this provides sufficient precision.

X. ELIGIBILITY AND INCLUSION

Respondents are eligible if they are (i) registered as a nurse, midwife or AHP; (ii) employed by a UK university; and (iii) engaged in both teaching and research. Eligibility will be confirmed through screening questions at the start of the online survey.

XI. SURVEY INSTRUMENT

The online questionnaire (Appendix 2) developed using JISC Survey consists of four main sections. We first present the participant information, check eligibility and ask about their role as an academic, patient and confidant. Second, we present information about ORA and GAI and seek their views on research priorities. Third we focus on ORA and present three papers and ask what should be included in the undergraduate curriculum. We also ask about their own personal experience of using ORA. From our previous study of retired doctors [34] we found that personal use of ORA was a strong predictor of respondents' attitudes to all patients having online access to their records. Section 4 takes the same approach for GAI. Sections 3 and 4 include summaries of six relevant publications as learning materials [35-40] which will be reviewed to ensure they are up-to-date before releasing the survey when starting the survey. Section 5 is about implementation in the curriculum and asks about whether ORA and GAI should have any priority given the pressure of required elements of the existing curricula. Section 6 ask respondents to choose anonymity or to give their email if they want to receive a CPD certificate. CPD certification is granted for engagement with the learning materials and completion of the survey, regardless of the views expressed. Section 7 asks about the methods of the survey: (i) whether they think the combination of a survey with CPD 'works' and (ii) asks how they received the invitation (directly from the principal investigator or forwarded from a colleague).

XII. DATA ANALYSIS, GOVERNANCE, DISSEMINATION

Quantitative data will be analysed descriptively using frequencies, crosstabs and possibly multiple logistic regression using SPSS. Anonymous qualitative free-text responses will undergo rapid thematic analysis to identify recurring themes, merged and analysed using a quantified approach [41]. Ethics will be sought from the Faculty Research and Integrity Committee, University of Plymouth. Data management will be via JISC Survey with a single researcher (RJ) handling data. There are no collaborators external to University of Plymouth. Findings will be shared with respondents along with CPD certificates. Peer review papers will be submitted and executive

summaries shared with nursing and AHP education leads, national nursing, midwifery, and AHP bodies, and NHS England digital leadership groups.

XIII. LIMITATIONS OF THIS APPROACH

This approach of course has limitations. (i) Any such survey will have self-selection bias – only those with some interest in the topic are likely to respond. We try to address this by offering a CPD certificate but such bias is unavoidable. (ii) The speed of technological change is very fast. Generative AI moves faster than traditional curriculum redesign but if we do not try to ascertain academics' views we will be 'running blind'. The survey results may be somewhat outdated but they will at least give some indication of academics' views. (iii) Our proposed sample and survey is UK-Centric and so there are limitations of how well it can be used to generalize to other countries particularly those with greater use of private medicine. (iv) Once we have the views of academics, implementing them will face barriers due to the systemic lack of time/resources in both the university sector and in the NHS where nurses and AHPs may be prevented from actually performing this new core clinical function. (v) Our sample identification via publication will give a view biased towards research active academics; it is possible that an additional complementary sample of less research active academics needs to be included.

XIV. CONCLUSION

As AI increasingly mediates patient understanding, explainability becomes a core clinical function rather than a technical feature. This shifts professional boundaries for nurses, midwives and AHPs, who are often the primary human interpreters of algorithmic medicine. Patient access to their medical records, combined with the rapid emergence of generative AI, creates both new opportunities and new risks for patient understanding, safety, and equity. Yet the perspectives of nurses, midwives, and AHPs remain largely absent from the literature, despite their central role in patient communication and the volume of contacts they deliver.

This protocol sets out the first UK study to identify research priorities and curriculum needs for non-medical practitioners concerning online record access and patients' use of generative AI. By embedding a short CPD component within the survey, the study will also test a scalable model for engaging busy clinical academics with fast-moving digital health topics.

Findings will inform a structured socio-technical co-design agenda in which patients, carers, practitioners, educators, system leaders and developers use the micro-meso-macro framework to identify conflicts between technical feasibility, clinical safety, workflow, professional accountability, educational feasibility and patient understanding. Rather than assuming consensus, this approach treats such tensions as explicit design and governance problems to be surfaced, prioritised and, where necessary, carried forward as research questions or implementation risks. To unlock the benefits of digitally accessible records and generative AI, while avoiding widening inequalities, research and education must now extend beyond medicine to include the wider healthcare workforce and the diverse communities they serve.

REFERENCES

- [1] R. B. Jones et al., "Explaining the medical record: a research agenda for non-medical practitioners," presented at EXPLAINABILITY 2025: The Second International Conference on Systems Explainability, Barcelona, Spain, 2025. [Online]. Available from: https://www.thinkmind.org/library/EXPLAINABILITY/EXPLAINABILITY_2025/explainability_2025_1_60_10044.html
- [2] R. Jones et al., "Randomised trial of personalised computer based information for cancer patients," *BMJ*, vol. 319, no. 7219, pp. 1241–1247, Nov. 1999, doi: 10.1136/bmj.319.7219.1241.
- [3] R. B. Jones et al., "Randomised trial of personalised computer based information for patients with schizophrenia," *BMJ*, vol. 322, no. 7290, pp. 835–840, Apr. 2001, doi: 10.1136/bmj.322.7290.835.
- [4] R. B. Jones et al., "Effect of different forms of information produced for cancer patients on their use of the information, social support, and anxiety: randomised trial," *BMJ*, vol. 332, no. 7547, pp. 942–946A, Apr. 2006, doi: 10.1136/bmj.38807.571042.68.
- [5] A. L. Neves et al., "Impact of providing patients access to electronic health records on quality and safety of care: a systematic review and meta-analysis," *BMJ Quality & Safety*, vol. 29, no. 12, pp. 1019–1032, 2020, doi: 10.1136/bmjqs-2019-010581.
- [6] C. Blease et al., "The benefits and harms of open notes in mental health: a Delphi survey of international experts," *PLoS ONE*, vol. 16, no. 10, e0258056, Oct. 2021, doi: 10.1371/journal.pone.0258056.
- [7] C. Blease, I. G. Cohen, and S. Hoffman, "Sharing clinical notes: potential medical-legal benefits and risks," *JAMA*, vol. 327, no. 8, pp. 717–718, Feb. 2022, doi: 10.1001/jama.2021.23179.
- [8] A. Hannan and F. Webber, "Towards a partnership of trust," in *Medical and Care Computetics 4*, L. Bos and B. Blobel, Eds. Amsterdam: IOS Press, 2007, pp. 108–116.
- [9] U. Sarkar and C. Lyles, "Devil in the details: understanding the effects of providing electronic health record access to patients and families," *BMJ Quality & Safety*, vol. 29, no. 12, pp. 965–967, 2020, doi: 10.1136/bmjqs-2020-011185.
- [10] A. Turner et al., "Unintended consequences of patient online access to health records: a qualitative study in UK primary care,"

- British Journal of General Practice, vol. 73, no. 726, pp. e67–e74, Jan. 2023, doi: 10.3399/BJGP.2021.0720.
- [11] C. R. Blease et al., “Experiences and opinions of general practitioners with patient online record access: an online survey in England,” *BMJ Open*, vol. 14, no. 1, e078158, 2024, doi: 10.1136/bmjopen-2023-078158.
- [12] C. Blease et al., “Patient online record access in English primary care: qualitative survey study of general practitioners’ views,” *Journal of Medical Internet Research*, vol. 25, e43496, 2023, doi: 10.2196/43496.
- [13] S. Asthana, R. Jones, and R. Sheaff, “Why does the NHS struggle to adopt eHealth innovations? A review of macro, meso and micro factors,” *BMC Health Services Research*, vol. 19, no. 1, pp. 1–7, 2019.
- [14] NHS England, “AI knowledge repository.” [Online]. Available from: <https://digital.nhs.uk/services/ai-knowledge-repository>. Accessed: 16-Jan-2026.
- [15] Mayo Clinic, “Diseases and conditions.” [Online]. Available from: <https://www.mayoclinic.org/diseases-conditions>. Accessed: 28-Aug-2025.
- [16] National Institute for Health and Care Excellence, “Evidence summaries.” [Online]. Available from: <https://www.nice.org.uk/what-nice-does/our-guidance/about-evidence-summaries>. Accessed: 09-May-2026.
- [17] EMIS, “EMIS Web.” [Online]. Available from: <https://www.emishealth.com/emis-web>. Accessed: 09-May-2026.
- [18] TPP, “SystemOne.” [Online]. Available from: <https://tpp-uk.com/products/>. Accessed: 09-May-2026.
- [19] Oracle, “Oracle access governance.” [Online]. Available from: <https://docs.oracle.com/en/cloud/paas/access-governance/ocoir/>. Accessed: 09-May-2026.
- [20] Epic, “Epic.” [Online]. Available from: <https://www.epic.com/>. Accessed: 09-May-2026.
- [21] NHS, “NHS App login.” [Online]. Available from: <https://www.nhsapp.service.nhs.uk/login>. Accessed: 09-May-2026.
- [22] A. Bracken et al., “Artificial intelligence (AI)-powered documentation systems in healthcare: a systematic review,” *Journal of Medical Systems*, vol. 49, article 28, 2025, doi: 10.1007/s10916-025-02157-4.
- [23] Z. L. Teo et al., “Generative artificial intelligence in medicine,” *Nature Medicine*, vol. 31, no. 10, pp. 3270–3282, Oct. 2025.
- [24] NHS England, “The allied health professions strategy for England: AHPs deliver, 2022–2027,” 2022. [Online]. Available from: <https://www.england.nhs.uk/ahp/allied-health-professions-strategy-for-england/>. Accessed: 09-May-2026.
- [25] T. Kariotis, M. Prictor, K. Gray, and S. Chang, “Patient-accessible electronic health records and information practices in mental health care contexts: scoping review,” *Journal of Medical Internet Research*, vol. 27, e54973, 2025, doi: 10.2196/54973.
- [26] D. Elbourne, M. Richardson, I. Chalmers, L. Waterhouse, and E. Holt, “The Newbury maternity care study: a randomized controlled trial to assess a policy of women holding their own obstetric records,” *British Journal of Obstetrics and Gynaecology*, vol. 94, no. 7, pp. 612–619, 1987.
- [27] Nursing and Midwifery Council, “Standards of proficiency for registered nurses,” NMC, 2018, updated 2023. [Online]. Available from: <https://www.nmc.org.uk/standards/standards-for-nurses/>. Accessed: 09-May-2026.
- [28] Health Education England, “The digital literacy framework for the health and social care workforce,” HEE, 2018, updated NHS England 2022. [Online]. Available from: <https://www.hee.nhs.uk/our-work/digital-literacy>. Accessed: 09-May-2026.
- [29] NHS England, “Record keeping and patient online access: guidance for clinicians,” NHS England, 2023.
- [30] C. Buchanan et al., “Predicted influences of artificial intelligence on nursing education: scoping review,” *JMIR Nursing*, vol. 4, no. 1, e23933, 2021, doi: 10.2196/23933.
- [31] G. Glauberman, A. Ito-Fujita, S. Katz, and J. Callahan, “Artificial intelligence in nursing education: opportunities and challenges,” *Hawaii Journal of Health & Social Welfare*, vol. 82, no. 12, pp. 302–305, Dec. 2023.
- [32] Wikipedia, “List of universities in the United Kingdom.” [Online]. Available from: https://en.wikipedia.org/wiki/List_of_universities_in_the_United_Kingdom. Accessed: 09-May-2026.
- [33] uniRank, “A-Z universities in the United Kingdom by alphabetical order.” [Online]. Available from: <https://www.unirank.org/gb/a-z/>. Accessed: 09-May-2026.
- [34] R. B. Jones et al., “Retired doctors’ views on patients having access to their health records: cross-sectional mixed-methods survey in UK and Spain,” as yet unpublished available from ray.jones@plymouth.ac.uk.
- [35] S. N. D’Costa, I. L. Kuhn, and Z. Fritz, “A systematic review of patient access to medical records in the acute setting: practicalities, perspectives and ethical consequences,” *BMC Medical Ethics*, vol. 21, article 18, 2020, doi: 10.1186/s12910-020-0459-6.
- [36] K. De Groot, E. B. Sneep, W. Paans, and A. L. Francke, “Patient participation in electronic nursing documentation: an interview study among community nurses,” *BMC Nursing*, vol. 20, article 72, 2021, doi: 10.1186/s12912-021-00590-7.
- [37] S. O’Connor, P. Hanlon, C. A. O’Donnell, S. Garcia, J. Glanville, and F. S. Mair, “Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies,” *BMC Medical Informatics and Decision Making*, vol. 16, article 120, 2016, doi: 10.1186/s12911-016-0359-3.
- [38] J. W. Ayers et al., “Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum,” *JAMA Internal Medicine*, vol. 183, no. 6, pp. 589–596, 2023, doi: 10.1001/jamainternmed.2023.1838.
- [39] C. Chen et al., “Comparison of an AI chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: pilot randomized controlled trial,” *JMIR Human Factors*, vol. 12, e65785, 2025, doi: 10.2196/65785.
- [40] E. Reinhart, “What we lose when we surrender care to algorithms,” *The Guardian*, 9-Nov-2025. [Online]. Available from: <https://www.theguardian.com/us-news/ng-interactive/2025/nov/09/healthcare-artificial-intelligence-ai>. Accessed: 16-Jan-2026.
- [41] A. Tashakkori and C. Teddlie, *Mixed Methodology: Combining Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage Publications, 1998.

Appendix 1: Bibliometric methods and invitation email

Purpose of the bibliometric analysis

The purpose of the bibliometric analysis was to identify UK university clinical but non-medical researchers to act as seeds for a snowball sample.

Bibliometric search strategy

A structured search of the Web of Science (WoS) was conducted to identify UK-based publications related to ORA, patient portals and GAI and researchers from 164 UK Universities [33, 34]. Searches were performed in 20 batches using combinations of affiliation fields containing UK university names, geographic fields (UK, England, Wales, Scotland, Northern Ireland), a time-period of 2023–2025, and topic terms including online record access, patient portal, electronic health record, digital health, education, curriculum, student, artificial intelligence and generative AI.

Extraction of UK academics

All retrieved WoS records were downloaded and merged into a single dataset (approximately 6,000 publications). Authors with .ac.uk email addresses were extracted, producing 2,129 unique contacts. Authors were assigned to a university based on their email domain. After removing duplicates, non-university and student addresses, a list of 2,087 UK university-based authors remained (Table A1).

Focus on nursing, midwifery and AHPs

Because bibliometric data cannot reliably identify professional background, we searched the Internet to try to verify whether the authors were from our target population. For this stage we explored the use of AI tools (ChatGPT and Microsoft copilot) alongside simple ‘Google searches’. Use of AI, with obfuscation and multiple hallucination was not found to be useful. Nevertheless, we have identified 20% quota samples for the largest and a sample of smaller universities to have confidence that we can do so for the complete sample. We will exclude students and research assistants likely to be on short term contracts and so no longer available and with little influence on the curriculum. We also excluded VERY senior academics (Deans, Vice Principals etc) who may be more likely to delete such an email invitation.

Kings College London	196
University College London	131
Birmingham University	85
Northumbria University	84
Newcastle University	81
Imperial College London	79
Queens University Belfast	79
Cardiff University	69
Manchester Metropolitan University	58
Manchester University	57
London School of Hygiene & Tropical Medicine	52
Oxford University	51
Queen Mary London	47
Keele University	43
Liverpool John Moores University	39
Plymouth University	36
Kingston University	35
Nottingham Trent University	35
Bournemouth University	31
Edinburgh University	29
Anglia Ruskin University	26
Liverpool University	26
City University	25
Glasgow Caledonian University	23
Exeter University	21
Oxford Brookes University	21
Bristol University	20
De Montford University	20
Glasgow University	20
Edinburgh Napier University	19
Leeds Beckett University	19
Cambridge University	18
Robert Gordon University	18
Lancashire University	17
Leeds University	17
Southampton University	17
Coventry University	16
London South Bank University	16
Ulster University	16
Bradford University	15
Cardiff Metropolitan University	15
Sheffield University	15
70 Universities with fewer than 15 people	398

Table A1. 2087 people from 112 universities identified with relevant publications.

Personal emails (Figure A1) will be sent by the principal investigator to these researchers.

FIGURE A1. Draft Example invitation email to ‘snowball seeds’

Dear <author title and name>, I am writing to you as an established health researcher. I got your email address from Web of Science and checking on the Internet I think that this is relevant to you and colleagues. I hope that you will consider this survey and Continuing Professional Development (CPD) opportunity yourself, also to forward to colleagues that you think will be interested and eligible to participate in

*We aim to recruit **non-medical clinical UK university staff involved in research and education** seeking their views both on research priorities on patient access to their records and their use of generative AI, and research-informed teaching in the curriculum for the same topic. We fully realise the pressure on their time, so we are offering an approach that combines half an hour's CPD with a certificate from the University of Plymouth Centre for Health Technology alongside completing our survey. If you or your colleagues decide to participate you will (i) have the option of being named in an Appendix to any publication as one of the consulted experts as well as (ii) having the opportunity for CPD.*

To find out more join the link at <URL> where there is more information for potential participants. If you continue, we think it will take about 30-35 minutes to complete. Participants can of course drop out at any time or remain anonymous (the URL survey is not linked to this email in any way).

*Yours sincerely,
Ray Jones on behalf of study team.*

Appendix 2 – Online Survey

UK non-medical researchers' views on research and teaching of patient online record access and use of AI

SECTION 1 of 7: Participant Information 1

Thank you for considering this study on patient use of online record access and generative AI.

Although patients increasingly have online access to their medical records, not enough is known about how non-medical practitioners—nurses, midwives, and allied health professionals (AHPs)—can help patients understand what they read. Most existing RESEARCH focuses on doctors, even though non-medical practitioners make up most of the NHS.

Early studies found that giving patients explanations tailored to them can reduce anxiety and make care safer. However, this is not done consistently across the NHS, and many clinicians still worry that too much information might confuse or worry patients. But patients already turn to websites and AI tools such as ChatGPT to help make sense of health information. Some healthcare systems are starting to build these tools directly into their services. These technologies could help turn medical jargon into plain language and providing personalised summaries, but they also carry risks, including incorrect information, privacy concerns, and the possibility of increasing existing inequalities.

RESEARCH is needed at:

- (i) 'Micro' (personal) level
- (ii) 'Meso' (service or team) level
- (iii) 'Macro' (system-wide) level

EDUCATION is needed on how to address patient online record access and use of AI but while the undergraduate or preregistration curricula may include general digital literacy it rarely includes online record access or patient use of AI, so we need to know what research-based education should be included in the curricula.

This survey, CPD and networking opportunity seek the views of non-medical practitioners on RESEARCH and research-based EDUCATION.

SECTION 1 OF 7: Participant Information 2

This survey includes:

1. A short, structured learning component
2. Your responses to questions
3. The opportunity to receive a 30-minute CPD certificate and follow up papers issued by the University of Plymouth
4. The opportunity to join a six-month network (the Smart Patient Academic Network) of like-minded UK academics to facilitate collaboration for research and teaching.

Participation is voluntary. Estimated time: 30-35 minutes.

The purpose of this survey is to (i) get your views on the priorities for research (ii) identify what knowledge and skills students need on this topic. Lastly we will explore whether embedding a short CPD activity for respondents within a survey is useful.

The study team from the University of Plymouth is led by Professor Ray Jones. If you have any queries about this study please contact ray.jones@plymouth.ac.uk

A paper with the background and protocol for this study has been published in the Journal of xxxxxxxxx (2026).

The study has ethical approval from the University of Plymouth Faculty of Health Research Ethics Committee (date/time).

If you wish to complain about the study in the first instance please contact Ray Jones or then contact the Research Administrator on <email>

SECTION 1 OF 7: Eligibility and Consent

To continue please confirm that you:

- (i) are qualified as a nurse, midwife, or allied health professional
- (ii) are employed by a UK university
- (iii) have been involved in teaching and research at some stage of your career
- (iv) have read the participant information
- (v) are willing for your responses to be included in the analysis of this survey
- (vi) understand that you are able to choose (at the end) to be anonymous or named if you want a CPD certificate, further papers, or to join the Smart Patient Academic Network

Q1. I confirm that I meet the six inclusion criteria above (Yes/No)

SECTION 1 OF 7: About you

Q2. What is your role? Research Assistant / Fellow;
Lecturer / Senior Lecturer / Reader / Associate Professor /
Professor; Clinical Educator / Practice Learning Lead;
Other

Q3. What is your discipline? (Open question)

Q4. (We ask this question as we are interested to know if personal experience affects views on patient information). Have you 'been a patient' in the last two years at any level of care (eg GP or hospital), or been the carer or confidant for a family member who has been a patient? I have been a patient; I have been a carer or confidant; Neither of the above

SECTION 2 OF 7: Research into patient online record access and use of AI – CPD1

Patients in 30 or more countries now have at least some access to online medical records. In England, the NHS 10-Year Plan sets out a vision for a digitally enabled, personalized service that focuses on preventing illness, supporting care closer to home, and helping people to take an active role in managing their own health. The other UK countries have similar policies. To make this work, it is not enough to simply give patients online access to their records; people also need support to understand and use the information effectively.

Since October 31st, 2023, all General Practice (GP) surgeries in England have been required to provide adult patients with online access to new information added to their primary care medical records by default. In practice, this means that patients should usually be able to see updates to their GP record through online services, such as the NHS App, without having to ask their GP for permission. This access includes most information added to the record, such as letters and notes from consultations. Only internal staff tasks are hidden from patients.

Information from hospital systems such as Cerner and EPIC is also being shared with patients, either through hospital patient portals or by feeding it into the NHS App. Patients may use digital health in different ways. Some look up information on public websites or use Generative AI tools on their own. Others check their medical records through patient portals and then use AI tools to help explore or explain what they see.

Some health IT providers are beginning to integrate, or plan to integrate, AI directly into patient portals. For example, EPIC is working with Microsoft/OpenAI to

include AI tools in systems used by clinicians and patients. At the same time NHS England is exploring how generative AI might be used within the NHS App and other digital health services.

SECTION 2 OF 7: Research into patient online record access and use of AI – CPD2

Nurses, midwives and allied health professionals are frequently the clinicians who support patients, families or confidants understand what is written in their medical records and deal with questions or worries about it. In the future, AI will play a bigger role in explaining these records. This may happen through tools built into patient portals, such as the NHS App or hospital Electronic Record Systems, or through stand-alone tools such as ChatGPT, Google Gemini or Microsoft Copilot. However, more RESEARCH is needed, ideally designed with patients and professionals.

At a 'Micro' (individual) level, it could explore which types of explanations are preferred by patients. At a 'Meso' (service) level, it could examine whether patients want to use online portals, how they use them, and whether AI tools affect the practitioner-patient relationship. At a 'Macro' (system-wide) level, it could look at how these changes affect patient outcomes and whether there are changes to care processes. We would like to know what you think the most important research priorities are.

Despite national policy commitments to digital transformation and patient empowerment, non-medical professions have not yet formally incorporated competencies related to online record access or patient's use of generative AI. For example, in England, standards set by the Nursing and Midwifery Council and the digital capability framework from Health Education England set out broad expectations for digital literacy and record keeping, but do not mention patients accessing their records online or using AI tools to understand them. Perhaps patients accessing their records online or using AI tools should be in the pre-registration training curricula? Perhaps it is not needed?

We seek your views on whether online record access or patient use of AI should be included in the curriculum for your discipline, and how this could be best taught.

SECTION 2 OF 7: Your views about research

Q5. For your discipline now, where do you think the priority should be? Micro (personal) level; Meso (service or team) level; Macro (system-wide) level; I don't know

Q6. What are the research questions that you think we need to address on these topics? (Open question)

SECTION 3 OF 7: Patient online record access in the curriculum (CPD3)

Here are summaries of three papers that give a 'taste' of research on patient online record access. Then we will ask for your views about this topic in the undergraduate curriculum for your discipline.

PAPER 1 (of three): D'Costa et al conducted a systematic review exploring the practical, ethical and professional consequences of patient access to medical records in acute hospital settings. From nearly 8,900 studies screened, 18 met inclusion criteria (12 empirical, 6 ethical). Overall, the review found that letting patients see their records while they are in hospital can enhance understanding, engagement and trust, but evidence for measurable improvements in clinical outcomes was limited. Clinicians expressed concern that patients reading their records on their own might cause anxiety or confusion, and that staff might change how they write their notes if they know patients might read them.

The ethical discussion highlighted tensions between patient autonomy and potential harm, differences in people's digital skills, and the possibility that records could become less accurate or of less educational value if they are written mainly with patients in mind. The authors conclude that giving patients access to medical records is feasible and generally welcomed, but it needs to be carefully designed, with good patient support, and appropriate training for staff to avoid unintended consequences. They called for further research that includes different professional perspectives and looks at long-term effects.

The review shows that patient access to records changes how clinicians communicate, write notes and professional boundaries. These are core aspects of nursing and AHP practice. For this reason, this paper highlights the need for nurses and AHPs to be educated and supported in sharing information openly and in a patient-centred way, making it directly relevant to developing online record access skills and competencies in nursing education and professional development.

SECTION 3 OF 7: Patient online record access in the curriculum (CPD4)

PAPER 2 (of three): De Groot et al carried out qualitative interviews with 19 community nurses in the Netherlands about how they engage patients in electronic nursing documentation. Nurses considered patient participation important but described challenges relating to systems (technical issues), work environment (time pressures) and patients (digital skills, health status) when making documentation visible and participatory. Findings highlight that even when electronic patient portals and online access exist, meaningful participation depends on nurses' practice and the context — not just the technology.

SECTION 3 OF 7: Patient online record access in the curriculum (CPD5)

PAPER 3 (of 3): O'Connor et al conducted a systematic review of 83 qualitative studies examining how patients and health professionals use digital health interventions. These included patient portals, tele-rehabilitation systems, and patient-held electronic records. The studies mainly focused on long-term conditions and community care. Many involved allied health professionals such as physiotherapists, occupational therapists, speech and language therapists, and rehabilitation teams supporting patients in self-management and recovery.

The review found that digital records and patient-facing systems can help patient autonomy, improve continuity of care, and support shared goal setting, particularly in rehabilitation and chronic disease management. However, these benefits depended heavily on how clinicians introduced, explained, and used these tools in everyday practice.

Patients often struggled to understand clinical data, progress measures, or treatment plans without help from a professional. Clinicians reported concerns about patients misunderstanding information, increased workload, and some people being excluded because they lack digital access or skills.

A key finding was that AHPs play an important role in explaining and coaching patients when they use digital records and portals. They helping them make sense of functional scores, exercise plans, and progress notes. The authors concluded that technology alone does not empower patients. Instead, it works best when supported

by strong professional relationships and workflows, good communication, and shared decision-making.

SECTION 3 OF 7: Patient online record access – learning from experience

We ask this as it's a good way for you to learn what is on offer for your patients.

Q7. Have you seen all or part of your records, or those of someone you care for or are confident for, online anywhere in the last 3 years, for example, your primary care record via the NHS App, or your hospital record through a 'patient portal', or received copies of letters from hospital to your GP? Yes NHS App; Yes Hospital portal; Had copies of letters from hospital to GP; Not seen any records or letters, but I haven't looked; Not seen any records or letters, didn't know I could; Not seen any records, despite wanting to; Not had any contact with NHS over last three years

SECTION 3 OF 7: Your views on patient online record access in the undergraduate curriculum for your discipline

Q8. How important is it to include something on patient online record access related competencies in the undergraduate (pre-registration) curriculum for your discipline? Not important; Useful; Quite important; Essential

Q9. Would you cite or use any of the three publications presented just now? Are they relevant as source materials for undergraduate or pre-registration students in your discipline? Or would you take a more personal approach related to their own use of patient access?

The papers were:

- (a) D'Costa, S.N., I.L. Kuhn, and Z. Fritz, A systematic review of patient access to medical records in the acute setting: practicalities, perspectives and ethical consequences. *BMC Medical Ethics*, 2020. 21(1).
- (b) De Groot K, Snee EB, Paans W, Francke AL. Patient participation in electronic nursing documentation: an interview study among community nurses. *BMC Nursing*.2021;20:72.
- (c) O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. *BMC Medical Informatics* 2016;16:120.

I would use paper (a) only; I would use paper (b) only; I would use paper (c) only; I would use all of the papers; I would not use any of the papers; I would take a personal approach; I have other ideas

Q10. What specific knowledge and skills should students develop related to patient online record access? (Open question)

Q11. Have you already taught patient online record access content in your programme? Yes (at least one session focussed on ORA); Yes (mentioned in passing in other sessions); No, but planning to; No, and currently not planning to; Unsure

SECTION 4 OF 7: Patients' use of AI in the curriculum (CPD6)

We will now present summaries of three publications then ask whether you think AI use by patients is relevant and should be included for undergraduates in your discipline

PAPER 1 OF 3: Armbruster et al conducted a cross-sectional study published in *JAMA Internal Medicine* that compared patients' perceptions of responses to real medical questions when written either by physicians or by OpenAI's ChatGPT. The researchers selected 195 publicly posted patient–physician interactions from Reddit's r/AskDocs forum and generated corresponding ChatGPT replies to the same questions. A panel of licensed healthcare professionals rated the quality and empathy of each answer, while patient participants evaluated helpfulness, clarity, and trustworthiness.

ChatGPT's responses were rated significantly higher for both quality and empathy—about tenfold more likely to be rated “good” or “very good”—than physicians' original answers. However, the study cautioned that ChatGPT occasionally produced inaccurate or misleading content, highlighting risks of over-trust. The authors concluded that AI could be a valuable communication adjunct to improve patient understanding and satisfaction, provided there is clinical oversight to ensure safety and accuracy. The findings demonstrate that patients are already using AI to interpret diagnoses and treatments, valuing its clarity and tone. For nurses and AHPs, this underscores an urgent need to understand how patients use such tools and to develop strategies for guiding, correcting, and contextualising AI-generated health explanations in everyday practice.

SECTION 4 OF 7: Patients' use of AI in the curriculum (CPD7)

PAPER 2 OF 3: Chen et al, in a pilot RCT (Hong Kong) compared a patient-facing AI mental-health chatbot with a traditional nurse hotline. Adults from the general population were randomized to receive on-demand support from either the chatbot or nurses by phone. Primary outcomes were changes in depression and anxiety symptoms; satisfaction and perceived usefulness were secondary measures. Both groups improved after using the service; the AI chatbot's effects were comparable to the nurse hotline for short-term reductions in anxiety and depression. Service satisfaction did not differ significantly between groups. The authors conclude that an AI option can provide accessible support at scale but call for larger trials and careful safeguards. This is one of the clearest head-to-head tests of patients using AI versus interacting with non-medical practitioners. It suggests AI may handle some low-acuity support without worse short-term outcomes, but it also spotlights the need for practitioner oversight, triage protocols, and escalation pathways so that AI augments—rather than replaces—practitioners' relational, interpretive, and safety-critical roles.

SECTION 4 OF 7: Patients' use of AI in the curriculum (CPD8)

PAPER 3 OF 3: Bennett et al conducted a systematic review of how artificial intelligence is being used in physiotherapy and rehabilitation, including patient-facing applications such as exercise apps, virtual coaches, motion-tracking systems, and AI-supported feedback tools. The review covered 94 studies across musculoskeletal, neurological, and cardiorespiratory rehabilitation, many of which involved patients using digital platforms to monitor their progress, receive personalised guidance, and interpret performance data.

The authors found that AI-enabled systems could improve patient engagement, adherence to exercise programmes, and the accuracy of functional assessment, particularly when used to provide tailored feedback and real-time coaching. However, the review also highlighted important risks, including over-reliance on automated advice, misinterpretation of data by patients, and the exclusion of people with low digital literacy or complex needs. Clinicians reported that AI systems worked best when they supported—rather than replaced—professional judgement, communication, and shared goal setting.

Importantly, for allied health professions, the review emphasised that physiotherapists and other rehabilitation professionals play a crucial role in interpreting AI-generated data, contextualising progress measures, and helping patients make sense of what the technology reports. The authors concluded that AI has strong potential to enhance rehabilitation and self-management, but only if professionals are trained to integrate these tools into therapeutic relationships, safety monitoring, and personalised care planning. This paper is directly relevant to teaching about patient use of AI because it shows how AI-generated information shapes patient understanding, expectations, and behaviour in AHP-led care

SECTION 4 OF 7: Using AI for personal health – learning from experience

Patients are increasingly using tools like ChatGPT and Microsoft Copilot to explain medical terms, summarise clinical notes, or translate technical language into everyday wording. First we ask a personal question (as it's a good way of learning about GAI)...

Q12. Have you used ChatGPT, Google Gemini, Microsoft Copilot or any other AI platform to explain anything to do with your own health or that of a family member or friend? Yes, I use it frequently; Yes, I have used it at least once; No, I have used it for other things but never used it for this purpose; No, never used these AI tools

SECTION 4 OF 7: Your views on patients' use of AI in the undergraduate curriculum for your discipline

Q13. In your view, how important is it to include something on AI use by patients in undergraduate (pre-registration) curriculum for your discipline? Not important; Useful; Quite important; Essential

Q14. Would you use any or all of the publications presented just now or would you take a more personal approach for undergraduate (pre-registration) students in your discipline?

The papers were

- (a) Armbruster, W., Wong, A., Gallagher, T.H., Shen, N., et al. (2024) Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, 184(5), pp. 513–520.
- (b) Chen C, Lam KT, Yip KM, et al. "Comparison of an AI Chatbot With a Nurse Hotline in Reducing

Anxiety and Depression Levels in the General Population: Pilot Randomized Controlled Trial.” JMIR Human Factors. 2025;12:e65785.

- (c) Bennett M, King D, Marshall S. Artificial intelligence in physiotherapy: A systematic review of current and emerging applications. *Physiotherapy*. 2022;114:1–12.

I would use paper (a) only; I would use paper (b) only; I would use paper (c) only; I would use all of the papers; I would use none of the papers; I would take a more personal approach; I have other ideas.

Q15. What specific knowledge and skills should students develop related to AI use by patients? (Open question).

Q16. Have you already taught patients’ AI use content in your programme? Yes at least one session on AI use by patients; Yes mentioned in passing other sessions; No but planning to; No, and not currently planning to; Unsure

SECTION 5 OF 7: Implementation in the curriculum

Q17. The curriculum is crowded. Given the other priorities of topics and competencies for inclusion, how important is patient online to records and their use of AI? Of minor importance – include only if there is time/space; Quite important – need to find time/space by trimming other subjects; Essential – need to find time/space even if means dropping something else.

Q18. Use this space to comment on if these topics should be taught, if so what should be taught, and how it should be implemented. Please make sure the answer does not reveal your institution. Answers to this question will be thematically analysed with assistance from AI. Free text answers will be checked for anonymity, randomised, uploaded to get provisional themes. These will be checked and agreed by the researchers. AI will then be asked to classify each answer with multiple themes. These will be checked by the PI. A quantitised approach to reporting the themes will be used. (Open question).

SECTION 6 OF 7: ANONYMITY VS NAMED CONTRIBUTIONS

CPD Certificate and Paper Acknowledgement: If you wish to remain anonymous we cannot send you a CPD certificate or acknowledge your contribution in any publication.

Q19. Would you like to receive a CPD certificate and further papers as pdfs, and/or to be added to the Smart

Patient Academic Network? (This will require leaving your email address). Yes I wish to give my email; No I wish to remain anonymous.

Q20. Your email address

Q21. Do you want to receive the CPD certificate and further papers? Yes; No

Q22. If we publish findings, may we list your name among an Appendix of contributors? Yes, include name; No, remain anonymous

Q23. What is your name for acknowledgement? (Open question)

Q24. Would you like your name and email added to the Smart Patient Academic Network? This will be a time-limited network in which the principal investigator (Ray Jones) will attempt to ‘join up’ like-minded researchers and educators on explainable health information for patients. All information on the network will be deleted by March 2027. Yes include me; No I am not interested.

SECTION 7 OF 7: REVIEW OF METHODS OF THIS STUDY

We realise that academics are busy and can also be overloaded with requests to complete surveys. We have tried to make participating in this survey useful – information exchange rather than just questionnaire.

Q25. As a method, do you think this works? Yes definitely; To some extent; No, not really; No, definitely not

Toward Robust Machine-Learning-Based Discrimination of a 1 Hz Difference in the Auditory Cortex Using fMRI Activation Patterns

Yoshitaka Ooyashiki

Kochi University of Technology
Tosayamada, Kami, Kochi, 782-8502, Japan
e-mail: ooyashikiyoshitaka@gmail.com

Kyoko Shibata

Kochi University of Technology
Tosayamada, Kami, Kochi, 782-8502, Japan
e-mail: shibata.kyoko@kochi-tech.ac.jp

Abstract - Brain decoding using functional magnetic resonance imaging (fMRI) has attracted considerable attention as a method for investigating how stimulus-related information is represented in patterns of human brain activity and for examining whether subtle differences in sensory input can be reflected in brain activation patterns. In this study, we investigated whether machine-learning-based classification could discriminate between two auditory stimuli differing by 1 Hz using fMRI activation images obtained from the auditory cortex. This study extends our earlier work by examining the reproducibility of the method across multiple participants and by comparing a three-dimensional convolutional neural network (3DCNN) with a support vector machine (SVM). In the individual analysis, the 3DCNN showed consistently high classification accuracy across participants under the present experimental setting, and its performance was higher than that of the SVM. In the group analysis, both the 3DCNN and the SVM showed more variable performance across participants, although several train-test combinations achieved accuracy levels around or above 60%. These results suggest that the proposed framework is particularly effective in participant-wise analysis, whereas cross-participant application remains more challenging under the current conditions. Overall, the findings support the feasibility of fine-grained auditory discrimination from fMRI activation patterns and provide additional motivation for further investigation using more rigorous evaluation procedures, improved model-selection strategies, and more diverse participant cohorts.

Keywords-fMRI; Brain decoding; Convolutional Neural Network; Support Vector Machine.

I. INTRODUCTION

Brain decoding is a technology that interprets physical and psychological states from measured brain activity, and functional magnetic resonance imaging (fMRI) has been widely used for this purpose in neuroscience. Because fMRI data consist of a large number of voxels, Multivoxel Pattern Analysis (MVPA) is commonly employed for analysis. In recent years, machine learning techniques have attracted considerable attention as part of the MVPA framework and have contributed to advances in brain decoding research. To date, many studies have investigated the use of fMRI to decode visual stimuli. For example, Liang et al. [2] examined emotion decoding from visual cortex images using a Support Vector Machine (SVM) and a Convolutional Neural Network (CNN). Horikawa et al. [3] also investigated the decoding of objects seen in dreams by applying Deep Neural Network (DNN) features to visual cortex data. In contrast, research on fMRI-based auditory cortex decoding has lagged behind that

on visual cortex decoding. One reason is that human perceptual information is predominantly visual, which has historically led to greater emphasis on the visual cortex. In addition, scanner noise during fMRI acquisition makes it more difficult to detect neural activity in the auditory cortex. However, improvements in fMRI technology, including noise-canceling mechanisms, have made auditory cortex research increasingly feasible.

In studies of auditory cortex decoding, many investigations have focused on music-related targets and have reported the identification or classification of genres and moods, such as upbeat songs, somber songs, or mood-lifting songs, from brain activity [4][5][6]. For example, Casey et al. [6] developed and validated an SVM-based classification method using fMRI data acquired during auditory stimulation with 25 songs from five musical genres. Many of these studies have focused on qualitative musical characteristics such as mood and genre. In contrast, relatively few studies have examined quantitative musical characteristics, particularly the three fundamental elements of sound: frequency, sound pressure, and timbre.

To better understand fine-grained auditory processing, it is important to investigate not only qualitative but also quantitative sound characteristics. Accordingly, our research group has focused on decoding quantitative musical characteristics. Shigemoto et al. [7], for example, focused on frequency and used deep learning to discriminate between two sounds differing by 124.5 Hz from brain activation images, achieving an accuracy of 75%. Building on this result, the present study aims to examine even smaller frequency differences. At present, this study is positioned as fundamental neuroscience, and possible clinical application remains a topic for future investigation.

Our research group has defined the auditory cortex as a Region of Interest (ROI) based on the concept of tonotopy in order to capture brain responses to auditory stimuli. Tonotopy refers to the spatial organization of frequency-specific responses within the auditory cortex. Tonotopy has been reported in many previous studies [8][9][10][11], particularly in Brodmann Areas (BA) 41 and 42, corresponding to the primary auditory cortex in Heschl's gyrus. These studies typically examined a wide frequency range and often focused on continuous frequency modulation. However, the possibility that tonotopy may remain informative even for very small frequency differences has not been sufficiently explored.

Therefore, to investigate smaller frequency differences than those examined by Shigemoto et al. [7], this study hypothesizes that tonotopy may still function even for

frequency differences that are effectively inaudible to humans. Healthy individuals can typically distinguish frequency differences of approximately 5 Hz, whereas differences of 2–3 Hz are more strongly influenced by musical experience and individual variation. In this study, a 1 Hz difference between two sounds is treated as a frequency difference that is difficult for humans to perceive.

Accordingly, the objective of this study is to examine a machine-learning-based method for discriminating between two sounds differing by 1 Hz from fMRI activation patterns. Machine learning methods were adopted because the difference in brain activation between such closely spaced frequencies is expected to be subtle. To capture these subtle activation patterns in high-dimensional fMRI data, machine learning techniques, as part of the MVPA framework, were employed. In addition, such techniques may be relevant to future methodological developments in medical and assistive technologies.

The present study builds on two lines of our previous work. First, the imaging design was examined in earlier studies [12][13]. Two major imaging designs are commonly used in fMRI experiments: the event-related design and the block design. The event-related design uses shorter stimulus durations and can provide a larger number of data points, whereas the block design uses longer stimulus durations and generally yields clearer activation images. Because the present study focuses on very small frequency differences, the higher image clarity of the block design was considered advantageous for feature extraction. At the same time, the limited number of data points obtainable with the block design raises concerns about the amount of data available for model training. Comparative experiments in our previous work [12][13] showed discrimination accuracies of 55.9% for the event-related design and 63.4% for the block design, suggesting that the block design was more suitable for 1 Hz difference discrimination.

Second, based on these findings, our previous study [1] re-examined the ROIs to improve discrimination accuracy. In the earlier studies [12][13], the ROIs were limited to BA41 and BA42. In the later study [1], the ROI definition was expanded based on the hypothesis that additional activation information would improve feature extraction for subtle activation patterns. In addition to BA41 and BA42, BA22 was included as an ROI because previous work has also suggested tonotopic organization in BA22 [14]. Using this expanded ROI definition, discrimination accuracy reached 60.4% with the event-related design and 100% with the block design in participant-wise analysis. These findings suggested that using a block design with ROIs in BA41, BA42, and BA22 was an effective configuration for 1 Hz difference discrimination using a 3DCNN.

Based on these previous findings, the objective of the present study is to further evaluate the 1 Hz difference discrimination framework from two perspectives.

The first objective is to examine reproducibility across multiple participants. Because the previous studies [1][12][13] involved only a single participant, it remained unclear whether the observed high accuracy was specific to that participant. Therefore, the first objective of this study is

to determine whether similar results can be obtained across multiple participants under the same experimental framework.

The second objective is to examine the suitability of a 3DCNN for 1 Hz difference discrimination by comparing it with an SVM. In general, studies employing machine learning methods often compare a proposed method with another learning algorithm in order to assess its relative usefulness [2][15][16]. In this study, the 3DCNN was selected as the main discrimination method because it can extract features while preserving the spatial structure of brain activation images. However, because fMRI requires a specialized experimental environment, the amount of data available for training is limited. Since deep learning methods generally require larger datasets, an SVM, which is often effective for smaller datasets, was selected as a baseline method for comparison.

To address this second objective, both individual and group analyses were conducted. In the individual analysis, training and testing were performed using data from the same participant, representing a participant-wise discrimination framework applicable when training data can be acquired in advance. In the group analysis, one participant was used for testing and the remaining participants were used for training, thereby providing an initial evaluation of cross-participant applicability in cases where participant-specific prior data may not be available.

The remainder of this paper is organized as follows. Section II describes the acquisition of fMRI data and the discrimination procedures. Section III presents the experimental results. Section IV discusses the findings from the perspectives of reproducibility and comparative model performance. Section V concludes the paper and outlines future work.

II. METHOD

The following procedure is to be followed. Acoustic stimuli consisting of between two sounds with a 1 Hz difference are presented using an fMRI scanner to acquire brain activation images. These images are annotated using SPM12 to provide input to the 3DCNN and SVM. The annotated brain activation images are converted into 3D data for the 3DCNN and into $1 \times N$ dimensional data for the SVM. The analysis is performed using the converted data for each of the four methods (3DCNN Method in Individual Analysis, SVM Method in Individual Analysis, 3DCNN Method in Group Analysis, and SVM Method in Group Analysis). A detailed explanation is provided in the subsequent section.

A. Participants

Six adult male participants without hearing impairments took part in this study. One of these participants was the same individual included in our previous study [1], for whom fMRI data had already been acquired; therefore, new experiments were conducted for the remaining five participants. This study was conducted in accordance with the ethical standards of the Ethics Review Committee of Kochi University of Technology, and informed consent was obtained from all participants after a full explanation of the experimental procedures.

B. fMRI Experiment

In this study, a SIEMENS 3-Tesla fMRI scanner (MAGNETOM Prisma 3T) [17] and a 64-channel head coil [17], both housed at the Brain Communication Research Center of Kochi University of Technology, were used, as shown in Figures 1 and 2. The functional imaging parameters are listed in Table I, and the structural imaging parameters are listed in Table II. During the fMRI experiment, each participant lay in a supine position inside the scanner bore, and the head was stabilized within the head coil to reduce motion artifacts during image acquisition. This arrangement was intended to maintain a consistent measurement condition throughout the experiment and to improve the reliability of the acquired activation images.

Auditory stimuli were presented to the participants through an RME FireFace UCX USB audio interface connected to a Windows PC. The OptoACTIVE headphones [18] used in this study, shown in Figure 3, were equipped with active noise cancellation. This auditory presentation setup was adopted to deliver the stimuli as consistently as possible in the MRI environment, where scanner noise can interfere with auditory perception and the stability of baseline brain activity.



Figure 1. 3-Tesla fMRI scanner (MAGNETOM Prisma 3T) [17].



Figure 2. 64-channel head coil [17].

TABLE I. FUNCTIONAL BRAIN IMAGING PARAMETERS.

Echo time (TE), ms	48
Repetition time (TR), ms	3000
Field of view (FOV), mm	192×192
Flip angle (FA), °	90
Matrix size, mm	2.0×2.0×3.0
Slice thickness, mm	3.0
Slice gap, mm	0.75
Number of slices	36
Slice acquisition order	Ascending

TABLE II. STRUCTURAL BRAIN IMAGING PARAMETERS.

Echo time (TE), ms	2.52
Repetition time (TR), ms	1900
Field of view (FOV), mm	250×250
Flip angle (FA), °	9
Matrix size, mm	1.0×1.0×1.0
Slice thickness, mm	0.5
Slice gap, mm	0.50
Number of slices	176
Slice acquisition order	Ascending



Figure 3. OptoACTIVE headphones [18].

The use of active noise-canceling headphones therefore played an important role in improving the listening condition during data acquisition. White noise was continuously presented throughout the experiment in order to stabilize the auditory background during scanning and to reduce fluctuations in the baseline auditory state caused by scanner noise and other environmental sounds. By maintaining a more consistent auditory background, the response evoked by the target stimuli could be captured more clearly as a relative change from that baseline condition. In addition, the onset of auditory stimulation was controlled in accordance with the experimental timing so that the task periods could be

presented under a stable acquisition schedule throughout the scanning session. This design was intended to reduce unnecessary variability in the measured responses and to improve the consistency of the data used for subsequent analysis. The auditory stimuli administered to the participants were generated using Steinberg Nuendo 10.3. The stimuli consisted of two pure tones at 523 Hz and 524 Hz, with sound pressure levels ranging from 78 to 85 dB. As in our previous studies [1][12][13], a block design was adopted. Each task period lasted 9 s, and each rest period lasted 15 s. The block design was selected because it was considered advantageous for obtaining clearer activation patterns from subtle auditory differences than an event-related design. Since the present study focused on discriminating a very small frequency difference of 1 Hz, the use of a design expected to provide relatively stable activation images was regarded as appropriate for subsequent machine-learning-based analysis.

C. Annotation

Brain activation images obtained by fMRI are typically stored in the Digital Imaging and Communications in Medicine (DICOM) format, which is widely used in medical imaging devices such as computed tomography (CT). In this study, preprocessing, statistical image creation, and ROI definition were performed using Statistical Parametric Mapping 12 (SPM12). For compatibility with SPM12, the fMRI data were converted from DICOM format to Neuroimaging Informatics Technology Initiative (NIFTI-1) format before analysis.

During preprocessing, three main steps were carried out: realignment, normalization, and smoothing. The first step was realignment. Although fMRI experiments are generally conducted over relatively short periods to reduce participant burden, head motion may still occur during scanning. Therefore, motion correction was required. Realignment corrected variations in the fMRI data caused by head movement by aligning all subsequent images to the first acquired fMRI image. The second step was normalization. Because brain structures differ among participants, the data were transformed into a standard brain space to reduce structural variability across individuals. SPM12 uses the Montreal Neurological Institute (MNI) standard brain model for this purpose. The third step was smoothing, which reduces noise introduced during the realignment and normalization procedures and helps stabilize the spatial distribution of activation signals.

During statistical image creation, brain activation features were extracted by selecting multiple preprocessed images, performing statistical analysis, and generating contrast images. In this study, each statistical image used for training was generated from two preprocessed brain activity scans. This setting was adopted on the basis of our previous studies, in which the use of two scans was found to provide a practical balance between image stability and the number of available training samples [1][12][13]. If only a single scan was used, the resulting activation image tended to be more sensitive to scan-to-scan fluctuation. In contrast, combining more than two scans would reduce the number of available samples for machine-learning-based analysis. Because four scans were

available within each condition, four different combinations of two scans could be constructed, and these combinations were used to increase the number of training samples while preserving the basic experimental structure.

The generated statistical images were then restricted to the ROIs, and both activation and spatial information were obtained in CSV format. Based on our previous findings [1], the ROIs were set to BA41, BA42, and BA22, which are closely related to auditory information processing. After the CSV data were obtained, the input data for the 3DCNN were converted into three-dimensional data of size $H41 \times W50 \times D15$ on the basis of spatial position information so that the spatial arrangement of the activation pattern could be preserved. In contrast, the input data for the SVM consisted of $1 \times N$ -dimensional activation vectors that could be directly used for classification. Thus, the 3DCNN and SVM differed not only in classifier structure but also in the way spatial information was incorporated into the analysis.

This annotation procedure yielded 320 training samples and 32 test samples for each participant. Table III summarizes the numbers of training and test samples used in the individual and group analyses. These data were subsequently used to construct the datasets for both the 3DCNN- and SVM-based discrimination experiments described in the following sections.

D. Learning Algorithms

In this study, a 3DCNN, a deep learning method, and a SVM, a conventional machine learning method, were used to discriminate brain activation images. These two methods were selected because they differ substantially in the way spatial information is handled. The 3DCNN can directly process volumetric activation images while preserving their spatial structure, whereas the SVM uses vectorized input data for classification.

The 3DCNN is an extension of the CNN architecture into three dimensions and is designed to process volumetric image data. It alternates convolution and pooling operations in the spatial directions, thereby enabling feature extraction while preserving the spatial structure of brain activation images. The structure of the 3DCNN used in this study is shown in Figure 4. It consisted of three sets of convolutional and pooling layers, followed by fully connected layers immediately before the output layer. Because the task involved discrimination between two classes, the output layer contained two units. The Softmax function was used as the activation function in the output layer so that the outputs could be interpreted as class probabilities.

The training parameters used for the 3DCNN are listed in Table IV. For each parameter, several values were tested, and a grid search was conducted over the combinations shown in Table IV. Training was terminated when the training error fell below 0.1, at which point the model was saved. The trained model was then evaluated using the test data in the two-sound discrimination task. In this study, the parameter search was performed in order to examine the behavior of the model under multiple settings within the present experimental framework.

TABLE III. NUMBER OF TRAINING AND TEST SAMPLES FOR EACH ANALYSIS METHOD (THE SECOND ROW SHOWS INDIVIDUAL ANALYSIS, AND THE THIRD ROW SHOWS GROUP ANALYSIS).

Analysis	Training samples	Test samples
Individual	320 samples	32 samples
Group	1600 samples	32 samples

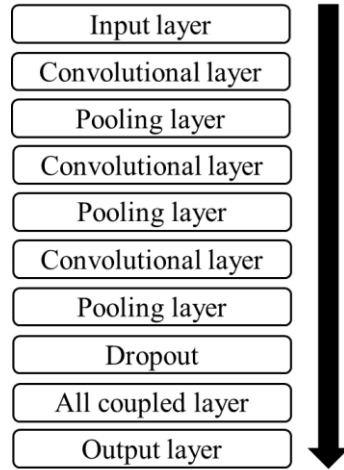


Figure 4. Architecture of the 3DCNN used in this study.

An advantage of the 3DCNN is that it can preserve the spatial structure of three-dimensional brain activation images during feature extraction. This characteristic was considered particularly relevant in the present study because the activation patterns were derived from localized regions in the auditory cortex. In contrast, the SVM treats the input as one-dimensional feature vectors and therefore does not directly use spatial structure in the same manner.

The SVM is a machine learning method that learns a decision boundary from the training data and then uses that boundary for classification. The decision boundary is determined by maximizing the margin with respect to the support vectors. In this study, the SVC (support vector classification) function [19] from the scikit-learn library was used to implement the SVM. Table V lists the hyperparameter combinations examined for the SVM. As in the 3DCNN experiments, multiple parameter settings were tested in order to compare model behavior under the present experimental conditions.

As described above, in the individual analysis, training and testing were performed separately for each participant using participant-specific annotated data. This setting was intended to evaluate whether the subtle 1 Hz frequency difference could be discriminated within each participant on the basis of that participant's own activation patterns. In the group analysis, one participant was designated as the test participant, while the remaining participants were used to construct the training model. The trained model was then evaluated using the data from the single test participant. This setting was used to examine how well the learned discrimination model generalized when applied to data from a

TABLE IV. TRAINING PARAMETERS FOR THE 3DCNN.

Layer			Set value
Input Layer			$41 \times 50 \times 15 \times 1$
Condition parameter	Number of Layer		6
	Convolutional layer	Stride	1
	Pooling layer	Filter size	$2 \times 2 \times 2$
		Stride	2
	Output layer	Class	2
	Learning rate		0.01
	Error rate		0.1
	Dropout		0.2
	Termination condition		Error rate
Hyperparameter	Convolutional layer	Kernel size (Ks)	3, 4, 5, 6, 7
		Filters (F)	16, 32
	Batch size (Bs)		8, 16, 32

TABLE V. HYPERPARAMETERS USED FOR THE SVM.

Kernel (K)	linear	rbf
C	0.01, 0.1, 1, 10, 100	0.01, 0.1, 1, 10, 100
Gamma (G)	—	0.01, 0.1, 1, 10, 100

participant who was not included in the training process.

III. RESULTS

Only models that completed training successfully were evaluated. For each trained model, the test data were input, and the model produced output probabilities for the two class labels (523 Hz or 524 Hz). The label with the higher probability was assigned as the predicted label and compared with the ground-truth label of the test data. This procedure was repeated for all 32 test samples, and the discrimination accuracy was defined as the number of correct predictions divided by 32.

Tables VI-IX present the discrimination accuracies and the corresponding hyperparameter combinations. Table VI shows the results for the 3DCNN method in individual analysis, Table VII shows those for the SVM method in individual analysis, Table VIII shows those for the 3DCNN method in group analysis, and Table IX shows those for the SVM method in group analysis. In all tables, the first column indicates the Participant ID used for training, and the second column indicates the Participant ID used for testing. The third column reports the highest discrimination accuracy observed in the grid search, and the fourth column lists the hyperparameter combinations that achieved this result. In some cases, the same discrimination accuracy was obtained with multiple hyperparameter combinations. For example, in Table VI for Participant A (Test A), a discrimination accuracy

of 100% was obtained with two combinations: $K_s = 4$, $F = 32$, $B_s = 8$, and $K_s = 7$, $F = 32$, $B_s = 8$.

Participant A in Table VI corresponds to the participant reported in the previous study [1], whereas Participants B-F represent the additional participants examined in the present study. In the individual-analysis setting, the 3DCNN generally showed high discrimination accuracy across participants, and high values were observed not only for Participant A but also for the additional participants. The SVM also achieved relatively high accuracies in several cases, but the overall tendency was less favorable than that of the 3DCNN. These results suggest that, under the present participant-specific setting, the 3DCNN was better able to capture the activation patterns associated with the subtle auditory difference. At the same time, the participant-wise results indicate that the favorable performance of the 3DCNN was not restricted to the participant examined in the previous study, but was also observed in the newly added participants under the same analytical framework.

A closer comparison of the individual-analysis results also suggests that the favorable performance of the 3DCNN was not limited to a single exceptionally good case. Rather, high accuracies were repeatedly observed across multiple participants, whereas the SVM showed a wider spread of values depending on the participant. This tendency is consistent with the summary statistics in Table X, in which the 3DCNN showed both a higher average accuracy and a smaller standard deviation in the individual-analysis setting. Thus, the participant-wise results support the view that preserving the spatial structure of activation images was advantageous under the present experimental conditions.

In the group analyses shown in Tables VIII and IX, the first column lists the five Participant IDs used for training. In contrast to the individual-analysis results, the group-analysis results showed substantial variation depending on the participant combination. Some train-test combinations produced discrimination accuracies above the chance level, whereas others yielded much lower values. This tendency suggests that the difficulty of discrimination increased when inter-participant variability was introduced into the training and test data. In addition, the variation observed in the group analyses was larger than that in the individual analyses, indicating that cross-participant discrimination was more sensitive to differences among participants. Thus, although certain participant combinations yielded relatively favorable results, the group-analysis setting as a whole showed less stable behavior than the individual-analysis setting.

The reported accuracies correspond to the highest observed results among the tested hyperparameter combinations under the present protocol.

Tables VI-IX each contain six train-test combinations. The average and standard deviation of the six discrimination accuracies in each table are summarized in Table X. Because the observed discrimination accuracies varied depending on both the hyperparameter settings and the train-test combinations, these average and standard deviation values should be interpreted as useful summary measures of the overall behavior of each method under the present experimental setting. In this study, these values are referred to

as the performance and stability of each method. A higher average generally indicates better overall discrimination performance, whereas a lower standard deviation indicates smaller variation across train-test combinations and therefore greater overall stability.

According to Table X, the 3DCNN showed better overall performance than the SVM in the individual-analysis setting, as indicated by its higher average discrimination accuracy. In addition, the standard deviation for the 3DCNN in the individual analysis was relatively small, suggesting that its performance was comparatively stable across participants under this setting. In the group-analysis setting, however, both methods showed lower overall performance and greater variability than in the individual analysis. These results are consistent with the tendency observed in Tables VIII and IX and indicate that generalization across participants remained more difficult than discrimination within individual participants. Taken together, the results summarized in Table X support the view that the 3DCNN was particularly advantageous in the individual-analysis setting, whereas neither method showed uniformly stable performance in the group-analysis setting.

IV. DISCUSSION

In this study, the discrimination task was formulated as a binary classification problem, for which the nominal chance level is 50%. In the neuroscience literature, classification performance above chance level is often regarded as a meaningful indication that stimulus-related information is represented in brain activity, and accuracy levels around 60% have also been discussed as practically informative in some decoding contexts [20][21]. From this perspective, the observed results, especially those obtained in the individual analysis, are of clear interest.

First, regarding the reproducibility of the 1 Hz difference discrimination method across multiple participants, the individual-analysis results showed that high classification accuracy was observed not only for Participant A, reported in the previous study [1], but also for Participants B-F. As shown in Table VI, the 3DCNN method in individual analysis achieved $100 \pm 0.0\%$ for Participants A, B, C, E, and F, and $95.6 \pm 3.2\%$ for Participant D. This tendency supports the view that the previously reported result was not limited to a single participant and that similar outcomes can be obtained across multiple participants under the same experimental framework. In this sense, the present findings strengthen the participant-wise reproducibility of the proposed method. They also indicate that the favorable result observed in the earlier study was not an isolated case limited to one particularly well-performing participant.

At the same time, the present results should be interpreted within the scope of the current evaluation protocol. Because only 32 test samples were used per participant and permutation testing or label-randomization testing was not conducted, the observed accuracies should not be treated as a fully independent statistical demonstration by themselves. Rather, they should be interpreted as encouraging results under the present experimental setting, while stricter statistical validation remains an important subject for future work. In

particular, the present findings should be viewed as supporting the promise of the framework under the current protocol, rather than as definitively establishing statistical robustness.

Next, to examine the suitability of the 3DCNN for 1 Hz difference discrimination, its performance was compared with that of the SVM in both individual and group analyses. As summarized in Table X, the 3DCNN method in individual analysis achieved $99.3 \pm 1.6\%$, whereas the SVM method in individual analysis achieved $85.9 \pm 16.5\%$. In the individual analysis, the 3DCNN showed higher average accuracy and smaller variation than the SVM. This result suggests that preserving the spatial structure of brain activation images may be advantageous for participant-wise discrimination. Because the training and test data were derived from the same participant in the individual analysis, structural and activation features were relatively consistent, which likely facilitated feature extraction by the 3DCNN. In contrast, the SVM handled the input as one-dimensional vectors and therefore could not directly exploit spatial structure in the same manner. This interpretation is consistent with previous findings such as those reported by Vu et al. [22]. Vu et al. examined fMRI volume classification using three methods, including a 3DCNN and an SVM, across several task conditions, and reported that the 3DCNN outperformed the comparison methods because it was able to preserve and utilize the spatial organization of brain activation patterns. Their interpretation is therefore consistent with the present finding that the preservation of spatial structure is likely to be advantageous in participant-wise analysis.

In the group analysis, however, the performance of both methods was lower and more variable than in the individual analysis. As shown in Tables VIII and IX, both the 3DCNN and the SVM showed participant combinations with accuracies above 60%, whereas other combinations remained at or near the chance level. The summary values in Table X also show that average group-analysis performance was much lower than that in individual analysis. Accordingly, the group-analysis results should not be interpreted as demonstrating uniformly reliable discrimination across participants. Rather, they indicate that cross-participant discrimination is possible in some cases under the present framework, but that its stability remains limited. This contrast between the individual and group analyses is one of the most important findings of the present study, because it suggests that participant-specific consistency plays a substantial role in successful discrimination.

One plausible reason for this difficulty is the lack of uniformity in participant-specific brain structure and activation patterns. Because the 3DCNN preserves structural information during feature extraction, it may learn not only the sound-related features necessary for discrimination but also participant-specific differences that are not directly relevant to the task. In individual analysis, this property is advantageous because the model is trained and tested on the same participant. In group analysis, however, inter-participant differences may interfere with stable discrimination performance. As described in Section II-C, normalization to a standard brain was applied to reduce anatomical differences across participants; however, some degree of residual

misalignment likely remained. As Kikuchi et al. [23] noted, complete alignment across participants is difficult to achieve, and such remaining differences may affect subsequent analysis. Such inter-participant variability may affect the spatial features extracted by the 3DCNN. In addition, differences in physiological response characteristics across participants may have influenced activation patterns, as discussed by Tavor et al. [24]. These factors may have reduced the consistency of the features relevant to the two auditory stimuli in the group-analysis setting. Thus, the lower stability in the group analysis may reflect not only limitations of the classifier itself but also the inherent difficulty of harmonizing participant-specific activation patterns within a single learning framework.

Another point requiring careful interpretation is the model-selection procedure. In this study, the reported accuracies correspond to the highest observed results among the tested hyperparameter combinations. This procedure was useful for exploratory comparison of model behavior under multiple settings; however, a stricter separation between hyperparameter selection and final evaluation, such as nested cross-validation or an independent held-out test set, would provide a more rigorous estimate of generalization performance. This remains an important direction for future work. Accordingly, the present results should be understood as reflecting the best observed behavior under the tested settings rather than as a definitive estimate of out-of-sample performance.

The issue of overfitting is also relevant to possible future improvements of the proposed framework. As many researchers have noted, common causes of overfitting include an excessive number of learnable parameters relative to the available data and insufficient training data [25][26]. In the present study, these conditions may have affected performance, particularly in group analysis. One possible countermeasure is stronger regularization. As Srivastava et al. [27] reported, Dropout is a representative technique for mitigating overfitting. Although Dropout was already used in the present study, it was fixed at 0.2, as shown in Table IV. Further examination of the dropout setting, together with additional training data and improved evaluation design, may help improve the stability of group-analysis performance. In addition, the stopping criterion for 3DCNN training in the present study was based on training error. Although this criterion enabled training to be completed consistently within the present framework, validation-based early stopping would offer a more rigorous approach to controlling overfitting. Together with stronger regularization and improved evaluation design, this may further improve the stability and interpretability of the results.

The participant cohort in the present study was also limited in size and diversity, consisting of six adult males in their twenties. Therefore, while the findings are meaningful as a methodological examination under controlled conditions, they should not yet be interpreted as establishing broad population-level robustness or practical applicability. Further validation using larger and more diverse cohorts will be necessary. Such expansion will also be important for determining whether the present framework remains effective under more

TABLE VI. DISCRIMINATION ACCURACY AND HYPERPARAMETERS FOR THE 3DCNN METHOD IN INDIVIDUAL ANALYSIS (THE SECOND ROW SHOWS THE RESULT REPORTED IN THE PREVIOUS STUDY [1]).

Participant ID (Training)	Participant ID (Testing)	Discrimination accuracy	Hyperparameters (Ks: kernel size; F: number of filters; Bs: batch size. See Table IV for details.)
A	A	100 ± 0.0%	Ks:4, 7, F:32, Bs:8
B	B	100 ± 0.0%	Ks:3, 4, 5, 6, 7, F:16, 32, Bs:8, 16, 32
C	C	100 ± 0.0%	Ks:3, 6, 7, F:32, Bs:8, 32 Ks:4, F:16, Bs:8 Ks:4, F:32, Bs:8, 16, 32 Ks:5, 6, 7, F:16, Bs:8, 16 Ks:5, F:32, Bs:16, 32
D	D	95.6 ± 3.2%	Ks:7, F:32, Bs:8
E	E	100 ± 0.0%	Ks:3, 4, 5, 6, 7, F:16, 32, Bs:8, 16, 32
F	F	100 ± 0.0%	Ks:3, 4, 5, 6, 7, F:16, 32, Bs:8, 16, 32

TABLE VII. DISCRIMINATION ACCURACY AND HYPERPARAMETERS FOR THE SVM METHOD IN INDIVIDUAL ANALYSIS.

Participant ID (Training)	Participant ID (Testing)	Discrimination accuracy	Hyperparameters (K: kernel; G: gamma. See Table V for details.)
A	A	84.4%	K:rbf, C:1, 10, 100, G:0.01
B	B	96.9%	K:rbf, C:1, 10, 100, G:0.01
C	C	90.6%	K:linear, C:0.1, 1, 10, 100
D	D	87.5%	K:linear, C:0.1, 1, 10, 100
E	E	50.0%	K:linear, C:0.01, 0.1, 1, 10, 100 K:rbf, C:1, 10, 100, G: 0.01, 0.1, 1, 10, 100 K:rbf, C:0.01, 0.1, G: 0.01, 1, 10, 100
F	F	100%	K:rbf, C:1, 10, 100, G:0.01

TABLE VIII. DISCRIMINATION ACCURACY AND HYPERPARAMETERS FOR THE 3DCNN METHOD IN GROUP ANALYSIS.

Participant ID (Training)	Participant ID (Testing)	Discrimination accuracy	Hyperparameters (Ks: kernel size; F: number of filters; Bs: batch size. See Table IV for details.)
BCDEF	A	50.0 ± 0.0%	Ks:4, 6, 7, F:16, Bs:8 Ks:5, 6, F:16, Bs:16 Ks:4, 5, F:32, Bs:32 Ks:4, 6, 7, F:16, Bs:8 Ks:6, F:16, Bs:32 Ks:6, F:32, Bs:8
ACDEF	B	85.9 ± 7.8%	Ks:5, F:16, Bs:16
ABDEF	C	57.8 ± 14.1%	Ks:7, F:32, Bs:32
ABCEF	D	50.0 ± 0.0%	Ks:3, 7, F:16, Bs:8, 16 Ks:3, 4, 5, 7, F:32, Bs:8 Ks:3, 5, 6, 7, F:32, Bs:16 Ks:4, 5, F:16, Bs:32 Ks:4, 6, 7, F:32, Bs:32 Ks:5, 6 F:16, Bs:8
ABCDF	E	60.9 ± 7.8%	Ks:7, F:16, Bs:32
ABCDE	F	50.0 ± 0.0%	Ks:3, F:16, Bs:8, 16 Ks:3, F:32, Bs:8 Ks:4, 5, 6, 7, F:16, 32, Bs:8, 16, 32

TABLE IX. DISCRIMINATION ACCURACY AND HYPERPARAMETERS FOR THE SVM METHOD IN GROUP ANALYSIS.

Participant ID (Training)	Participant ID (Testing)	Discrimination accuracy	Hyperparameters (K: kernel; G: gamma. See Table V for details.)
BCDEF	A	59.4%	K:linear, C: 0.1, 1, 10, 100
ACDEF	B	50.0%	K:rbf, C:0.01, 0.1, G: 0.01, 0.1, 1, 10, 100 K:rbf, C:1, 10, 100 G: 0.1, 1, 10, 100
ABDEF	C	59.4%	K:linear, C:1, 10, 100 K:rbf, C:10, 100, G:0.01
ABCEF	D	96.9%	K:rbf, C:1, 10, 100, G:0.01
ABCDF	E	50.0%	K:rbf, C:0.01, G:0.1, 1, 10, 100 K:rbf, C:0.1, 1, 10, 100, G:0.01, 0.1, 1, 10, 100
ABCDE	F	50.0%	K:rbf, C:0.01, 0.1, G:0.1, 1, 10, 100 K:rbf, C:1, 10, 100, G:1, 10, 100

TABLE X. AVERAGE DISCRIMINATION ACCURACY AND STANDARD DEVIATION FOR EACH METHOD, CALCULATED FROM THE SIX DATASET COMBINATIONS SHOWN IN TABLES VI–IX.

Method	Performance and stability (average $\pm$ standard deviation)
3DCNN Method in Individual Analysis (TABLE VI)	99.3 $\pm$ 1.6%
SVM Method in Individual Analysis (TABLE VII)	85.9 $\pm$ 16.5%
3DCNN Method in Group Analysis (TABLE VIII)	59.1 $\pm$ 12.7%
SVM Method in Group Analysis (TABLE IX)	60.6 $\pm$ 16.6%

heterogeneous real-world conditions.

The present findings are also meaningful from the viewpoint of methodological development in auditory brain decoding. In particular, the contrast between the individual and group analyses indicates that the choice of analytical framework substantially affects the observed performance when subtle auditory differences are examined. This suggests that future studies should not only seek higher classification accuracy, but also carefully consider how training-test design, participant variability, and feature representation influence the interpretability of the decoding results. From this perspective, the present study provides a useful basis for refining experimental and analytical strategies in future auditory fMRI studies. It also highlights the importance of evaluating not only whether subtle auditory information can be decoded, but also under what conditions such decoding remains stable across different analytical settings.

Taken together, the present findings indicate that the proposed framework is particularly promising for individual analysis, where the 3DCNN achieved high and comparatively stable performance across participants under the current setting. In contrast, the group-analysis results suggest that cross-participant application remains more challenging and requires further methodological refinement. Thus, although the present study does not yet establish robust generalization across participants, it provides additional support for the possibility that subtle frequency differences, even at the 1 Hz level, are reflected in fMRI activation patterns in the auditory cortex and can be captured to some extent by machine learning methods.

V. CONCLUSIONS AND FUTURE WORK

The aim of this study was to examine a machine-learning-based method for discriminating between two auditory stimuli differing by 1 Hz from fMRI activation patterns in the auditory cortex. To further evaluate the method beyond our earlier study, we investigated its reproducibility across multiple participants and compared the performance of a 3DCNN with that of a SVM.

In the individual analysis, high classification accuracy was observed across all participants, and the 3DCNN showed higher performance and smaller variation than the SVM under the present experimental setting. These results support the usefulness of the 3DCNN-based framework for participant-wise discrimination of very small frequency differences from fMRI activation patterns. More specifically, the present findings suggest that subtle frequency differences can be reflected in auditory-cortex activation patterns and can be captured under the present experimental setting when training and testing are performed within the same participant. The reproducibility observed across Participants A-F also suggests that the present framework has value as a participant-specific decoding approach under controlled conditions.

In the group analysis, both the 3DCNN and the SVM showed more variable performance across participants. While some train-test combinations achieved accuracies around or above 60%, several cases remained at the chance level. These results suggest that cross-participant discrimination is more difficult than participant-wise discrimination under the current framework and that further methodological improvement is needed before stronger conclusions can be

drawn regarding broader applicability. In particular, the present results indicate that stable generalization across participants remains a more demanding problem than within-participant discrimination of subtle auditory differences.

The present study also has several methodological limitations that should be recognized. The test size per participant was limited, permutation testing was not conducted, hyperparameter selection and final evaluation were not fully separated, and the participant cohort was small and homogeneous. Accordingly, the present findings should be understood as evidence obtained under restricted experimental conditions, while still providing meaningful support for the feasibility of fine-grained auditory discrimination from fMRI activation patterns. These limitations do not negate the interest of the present findings, but they do indicate that additional validation will be necessary before broader claims can be justified.

Future work will include more rigorous statistical validation, improved model-selection and stopping procedures, examination of regularization settings such as Dropout, and evaluation using larger and more diverse participant cohorts. In addition, extending the framework to other frequency ranges, environmental sounds, and musical stimuli may help clarify the broader scope of auditory decoding from human brain activity. Further investigation along these lines may also contribute to a better understanding of how subtle acoustic differences are represented in human brain activation patterns. In this respect, the present findings may serve as a useful reference point for future methodological refinement in auditory fMRI decoding studies.

Overall, the present study is meaningful not only because it examined discrimination of an extremely small auditory difference of 1 Hz, but also because it extended the previous work by evaluating reproducibility across multiple participants and by comparing two different classification frameworks. Although the current results should be interpreted within methodological limitations, they provide a useful basis for future studies aimed at clarifying how fine-grained auditory information is represented in human brain activity and how such information can be decoded more reliably from fMRI activation patterns. Such a line of research may also contribute to future efforts to establish more reliable frameworks for decoding fine-grained auditory information from human brain activity.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to all members of the laboratory for their valuable advice and support throughout this study. The authors also thank all participants for their kind cooperation in the experiments.

REFERENCES

- [1] Y. Ooyashiki, and K. Shibata, "Discrimination by Deep Learning of 1Hz Difference in Auditory Cortex Using fMRI Activation Patterns", The Tenth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing (HEALTHINFO 2025), ISBN: 978-1-68558-312-5, pp. 1-4, 2025.
- [2] Y. Liang, K. Bo, S. Meyyappan, and M. Ding, "Decoding fMRI data with support vector machines and deep neural networks", *Journal of Neuroscience Methods*, vol. 401, pp. 110004, 2024.
- [3] T. Horikawa and Y. Kamitani, "Hierarchical Neural Representation of Dreamed Objects Revealed by Brain Decoding with Deep Neural Network Features", *Frontiers in Computational Neuroscience*, vol. 11, pp. 00004, 2017.
- [4] M. Guilhem, M. D. L. Giovanni, and A. S. Shihab, "The Music of Silence: Part I: Responses to Musical Imagery Encode Melodic Expectations and Acoustics", *Journal of Neuroscience*, vol. 41 (35), pp. 7435-7448, 2021.
- [5] I. Daly, "Neural decoding of music from the EEG", *Nature, Scientific Reports*, 13, pp. 624, 2023.
- [6] M. A. Casey, "Music of the 7Ts: Predicting and Decoding Multivoxel fMRI Responses with Acoustic, Schematic, and Categorical Music Features", *Front. Psychol.*, vol. 8:1179, 2017.
- [7] N. Shigemoto, H. Satoh, K. Shibata, and Y. Inoue, "Study of Deep Learning for Sound Scale Decoding Technology from Human Brain Auditory Cortex", *IEEE, Global Conference on Life Sciences and Technologies*, pp. 212-213, 2019.
- [8] D. R. M. Langer, W. H. Backes, and P. V. Dijk, "Representation of lateralization and tonotopy in primary versus secondary human auditory cortex", *NeuroImage*, vol. 34, pp. 264-273, 2007.
- [9] W. Guo et al., "Robustness of Cortical Topography across Fields, Laminae, Anesthetic States, and Neurophysiological Signal Types", *The Journal of Neuroscience*, vol. 32(27), pp. 9159-9172, 2012.
- [10] V. A. Kalatsky, D. B. Polley, M. M. Merzenich, C. E. Schreiner, and M. P. Stryker, "Fine functional organization of auditory cortex revealed by Fourier optical imaging", *PNAS* vol. 102 No. 37, pp. 13325-13330, 2005.
- [11] S. Romero et al., "Cellular and Widefield Imaging of Sound Frequency Organization in Primary and Higher Order Fields of the Mouse Auditory Cortex", *Cerebral Cortex*, vol. 30, pp. 1603-1622, 2020.
- [12] Y. Ooyashiki, K. Shibata, and H. Satoh, "Identification of Small Frequency Differences in Primary Auditory Cortex of Human Brain by Deep Learning Using fMRI", *JSME Chugoku-Shikoku Student Association 54th Student Member Graduation Research Presentation Lecture*, pp. 01c2, 2023, (in Japanese).
- [13] Y. Ooyashiki, K. Shibata, and H. Satoh, "Discrimination fMRI Learning of 1Hz Difference in Primary Auditory Cortex of Human Brain using fMRI", *JSME Mechanical Engineering Congress*, pp. J162p-21, 2024, (in Japanese).
- [14] J. B. Issa et al., "Multiscale optical Ca²⁺ imaging of tonal organization in mouse auditory cortex", *Neuron*, vol. 83, pp. 944-959, 2014.
- [15] G. C. Oliveira et al., "Robust deep learning for eye fundus images: Bridging real and synthetic data for enhancing generalization", *Biomedical Signal Processing and Control*, vol. 94, pp. 106263, 2024.
- [16] J. Wang, Z. Wang, and G. Liu, "Recording brain activity while listening to music using wearable EEG devices combined with Bidirectional Long Short-Term Memory Networks", *Alexandria Engineering Journal*, vol. 109, pp. 1-10, 2024.
- [17] SIEMENS <<https://www.siemens-healthineers.com/jp/about>> 2026.05.27
- [18] OptoACTIVE <<https://www.optoacoustics.com/medical/optoactive-inflatable>> 2026.05.27
- [19] scikit-learn SVC <<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>> 2026.05.27
- [20] T. A. Carlson, T. Grootswagers, and K. A. Robinson, "An introduction to time-resolved decoding analysis for M/EEG", *arXiv, qbio.NC*, 1905.04820, 2019.
- [21] A. K. Robinson, G. L. Quek, and T. A. Carlson, "Visual Representations: Insights from Neural Decoding", *Annual Review of Vision Science*, vol. 9, pp. 313-335, 2023.
- [22] H. Vu, H.-C. Kim, M. Jung, and J.-H. Lee, "fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations", *NeuroImage*, vol. 223, pp. 117328, 2020.
- [23] Y. Kikuchi, and Y. Shiraishi, *Practical Analysis of Brain Activity and Functional Connectivity in fMRI Data: Mastering SPM, SnPM, and CONN*, Ishiyaku Publishers, 2019, (in Japanese).

- [24] I. Tavor et al. , "Task-free MRI predicts individual differences in brain activity during task performance", *Science*, vol.352, pp. 216-220, 2016.
- [25] D. M. Hawkins, "The Problem of Overfitting", *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 1, pp. 1–12, 2004.
- [26] K. Saito, *Deep Learning from Scratch: Theory and Implementation with Python*, O'Reilly Japan, pp.190-197, 2016.
- [27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.