frontiers | Research Topics

Artificial intelligence for smart health: learning, simulation, and optimization

Edited by
Bing Yao, Nathan Gaw and Hyo Kyung Lee

Published in
Frontiers in Physiology

ISSN 1664-8714
ISBN 978-2-8325-5890-4
DOI 10.3389/978-2-8325-5890-4

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Artificial intelligence for smart health: learning, simulation, and optimization

Topic editors

Bing Yao — The University of Tennessee, Knoxville, United States
Nathan Gaw — Air Force Institute of Technology, United States
Hyo Kyung Lee — Korea University, Republic of Korea

Citation

Yao, B., Gaw, N., Lee, H. K., eds. (2025). *Artificial intelligence for smart health: learning, simulation, and optimization*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-5890-4

Table of contents

TYPE Editorial
PUBLISHED 23 December 2024
DOI 10.3389/fphys.2024.1541057

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Raimond L. Winslow,
Northeastern University, United States

*CORRESPONDENCE
Bing Yao,
byao3@utk.edu

RECEIVED 06 December 2024
ACCEPTED 11 December 2024
PUBLISHED 23 December 2024

CITATION
Yao B, Gaw N and Lee HK (2024) Editorial: Artificial intelligence for smart health: learning, simulation, and optimization.
Front. Physiol. 15:1541057.
doi: 10.3389/fphys.2024.1541057

Editorial: Artificial intelligence for smart health: learning, simulation, and optimization

Bing Yao[1]*, Nathan Gaw[2] and Hyo Kyung Lee[3]

[1]Department of Industrial and Systems Engineering, The University of Tennessee, Knoxville, TN, United States, [2]Department of Operational Sciences, Air Force Institute of Technology, Wright-Patterson AFB, OH, United States, [3]School of Industrial Management and Engineering, Korea University, Seoul, Republic of Korea

KEYWORDS

machine learning and AI, health informatics, simulation optimization, smart health, data mining

Editorial on the Research Topic
Artificial intelligence for smart health: learning, simulation, and optimization

With rapid developments in medical sensing and imaging, we now live in an era of data explosion in which large amounts of data are readily available in clinical environments. The fast-growing biomedical and healthcare data provide unprecedented opportunities for data-driven scientific knowledge discovery and clinical decision support. Our Research Topic aims to catalyze synergies among biomedical informatics, machine learning, computer simulation, operations research, systems engineering, and other related fields with three specific goals: (1) develop cutting-edge data-driven models to accelerate scientific knowledge discovery in biomedicine using healthcare data collected from laboratory systems, imaging systems, and medical and sensing devices; (2) develop advanced simulation and calibration algorithms to build personalized digital twins by effectively assimilating patient-specific medical data with population-level computer models, facilitating precision medical planning; (3) develop innovative optimization algorithms for optimal medical decision making in the face of uncertainty factors, conflicting objectives, and complex trade-offs. This Research Topic, containing 10 articles, will offer a timely collection of information to benefit researchers and practitioners working in the broad fields of biomedical informatics, healthcare data analytics, medical image processing, and health-related AI.

Jiang et al. investigated the development and implementation of a high-fidelity simulation training course for fostering medical and nursing collaboration in China, guided by the Fink integrated curriculum design model. This training course was delivered to 14 nursing students and 8 clinical medicine students between March and July 2022. The results showed high satisfaction, increased self-confidence, and positive evaluations across various teaching practice dimensions. The study underscores the value of standardized simulation curricula in advancing healthcare education in China.

Rovati et al. evaluated the usability, workload, and acceptance of a digital twin application designed to simulate patient clinical trajectories based on EHR data for critical care education. Tested with 35 first-year internal medicine residents, the application demonstrated good usability and low to moderate workload. Residents expressed interest in using the digital twin application for ICU training and suggested improvements in clinical fidelity, interface design, learning experience, gaming elements, and implementation strategies.

Xie et al. developed a multi-branching ResNet model for atrial fibrillation detection from single-lead ECG signals. This method combines continuous wavelet transform for feature extraction with a multi-branching architecture to handle class imbalance in ECG datasets. Their framework was evaluated on two databases: PhysioNet/CinC challenge 2017 and private datasets from the University of Oklahoma Health Sciences Center. Their model achieved F1 scores of 0.8865 and 0. 7369 on the two datasets respectively, demonstrating strong performance in balancing precision and recall.

Patharka et al. provided a systematic review of research challenges in modeling biomedical temporal data, including missing values, capturing multi-dimensional correlations, and accounting for short- and long-term temporal patterns. This paper categorizes time series models into statistical, machine learning, and deep learning approaches, and further discusses their strengths and limitations. Strategies such as model enhancement, ensemble forecasting, and hierarchical models are examined for improving clinical predictions. It also explores implementation challenges in biomedical data modeling and outlines future directions for integrating AI in healthcare.

Kim et al. developed a Timely Early Warning System for Septic Shock (TEW3S), which emphasizes predicting the onset timing of septic shock to assist proactive clinical interventions. Utilizing machine learning and EHRs from the MIMIC-IV database, TEW3S achieved 94% accuracy in predicting all shock events with a maximum lead time of 8 h. By addressing the limitations of traditional risk-based prediction systems, this approach highlights the critical role of timeliness in improving patient outcomes during acute deterioration in hospital settings.

Rao et al. developed a multi-scale long short-term memory (LSTM) neural network trained with a variety of time scale data for classifying fetal heart rate patterns during labor. They employed preprocessing techniques to mitigate negative effects such as missing signals and artifacts on the model, and further utilized data augmentation techniques to address the data imbalance issue. Their framework was evaluated on the CTU-UHB dataset and achieved superior performance compared with traditional LSTM.

Stanik et al. developed a predictive model to identify stroke survivors at high risk of seizures following an infection, using data from the Long-Term Care Minimum Data Set. Data balancing techniques and feature selection methods are incorporated into machine learning models (Logistic Regression, Random Forest, XGBoost, Neural Network), achieving high accuracy in seizure prediction. Key factors contributing to seizure risk identified by this article included therapy hours, independence in daily activities, and mood.

Trevena et al. developed a graph-based patient simulation application designed to model critically ill patients with sepsis. The authors utilize directed acyclic graphs to represent the complex physiological and medication interactions during the first 6 h of critical illness. Their system consists of three core components: a cross-platform frontend for clinicians and trainees, a cloud-hosted simulation engine, and a graph database to determine the progression of each simulation. The simulation architecture demonstrates the potential to help train future generations of healthcare professionals and facilitate clinicians' bedside decision-making.

Wang et al. developed a three-phase methodology for emotion recognition from electroencephalography signals. Their framework addresses the challenges of capturing the complex, nonlinear, and nonstationary dynamics of brain activity by integrating manifold embedding, multilevel heterogeneous recurrence analysis, and ensemble learning. Evaluated on the SJTU-SEED IV database, their method demonstrates superior performance compared to existing commonly used techniques.

Meyers et al. investigated the sources of variability affecting operating room (OR) efficiency. The OR process was segmented into eight stages to quantify key process times, such as procedure duration and start time delay. The authors developed linear mixed models to evaluate the effects of factors such as the primary surgeon, anesthesia provider, and procedure type on OR efficiency. This study emphasizes the importance of segmenting the OR process into finer stages for better understanding of efficiency.

Finally, we extend our sincere gratitude to the reviewers for their thoughtful and constructive feedback on the manuscripts submitted to this Research Topic. Their insightful evaluations have significantly contributed to enhancing the quality and impact of this Research Topic.

Author contributions

BY: Writing–original draft, Writing–review and editing. NG: Writing–review and editing. HL: Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Author BY acknowledges the funding support from the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number R01HL172292. Author HL is funded by the National Research Foundation (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00212713). Author NG is funded by Air Force Office of Scientific Research (AFOSR) Grant No. 23AFCOR003 under the Dynamic Data and Information Processing (DDIP) program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

TYPE Original Research
PUBLISHED 11 January 2024
DOI 10.3389/fmed.2023.1336897

OPEN ACCESS

EDITED BY
Bing Yao,
The University of Tennessee, Knoxville,
United States

REVIEWED BY
Rui Zhu,
University of Oklahoma, United States
Eiji Kawamoto,
Mie University, Japan

*CORRESPONDENCE
Amos Lal
✉ Lal.Amos@mayo.edu

RECEIVED 16 November 2023
ACCEPTED 26 December 2023
PUBLISHED 11 January 2024

CITATION
Rovati L, Gary PJ, Cubro E, Dong Y, Kilickaya O, Schulte PJ, Zhong X, Wörster M, Kelm DJ, Gajic O, Niven AS and Lal A (2024) Development and usability testing of a patient digital twin for critical care education: a mixed methods study.
Front. Med. 10:1336897.
doi: 10.3389/fmed.2023.1336897

Development and usability testing of a patient digital twin for critical care education: a mixed methods study

Lucrezia Rovati[1,2], Phillip J. Gary[1], Edin Cubro[3], Yue Dong[4], Oguz Kilickaya[1], Phillip J. Schulte[5], Xiang Zhong[6], Malin Wörster[7], Diana J. Kelm[1], Ognjen Gajic[1], Alexander S. Niven[1] and Amos Lal[1]*

[1]Department of Medicine, Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, MN, United States, [2]School of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy, [3]Department of Information Technology, Mayo Clinic, Rochester, MN, United States, [4]Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, Rochester, MN, United States, [5]Department of Quantitative Health Sciences, Division of Clinical Trials and Biostatistics, Mayo Clinic, Rochester, MN, United States, [6]Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, United States, [7]Center for Anesthesiology and Intensive Care Medicine, Department of Anesthesiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Background: Digital twins are computerized patient replicas that allow clinical interventions testing *in silico* to minimize preventable patient harm. Our group has developed a novel application software utilizing a digital twin patient model based on electronic health record (EHR) variables to simulate clinical trajectories during the initial 6 h of critical illness. This study aimed to assess the usability, workload, and acceptance of the digital twin application as an educational tool in critical care.

Methods: A mixed methods study was conducted during seven user testing sessions of the digital twin application with thirty-five first-year internal medicine residents. Qualitative data were collected using a think-aloud and semi-structured interview format, while quantitative measurements included the System Usability Scale (SUS), NASA Task Load Index (NASA-TLX), and a short survey.

Results: Median SUS scores and NASA-TLX were 70 (IQR 62.5–82.5) and 29.2 (IQR 22.5–34.2), consistent with good software usability and low to moderate workload, respectively. Residents expressed interest in using the digital twin application for ICU rotations and identified five themes for software improvement: clinical fidelity, interface organization, learning experience, serious gaming, and implementation strategies.

Conclusion: A digital twin application based on EHR clinical variables showed good usability and high acceptance for critical care education.

KEYWORDS

critical care, medical education, patient-specific modeling, simulation training, patient safety, medical intensive care unit

1 Introduction

Medical errors remain a major cause of morbidity, mortality, and cost in the US healthcare system (1). The intensive care unit (ICU) is particularly prone to preventable adverse events due to the complexity of care delivery and the patient severity of illnesses (2). The fast pace and high acuity of critical care practice can also limit opportunities for trainee autonomy. Providing a safe environment to practice decision-making in this setting may improve the ICU educational experience, care processes, and patient-centered outcomes (3).

Digital twins are virtual models that simulate the behavior of real objects in a digital environment. With the increasing availability of electronic health record (EHR) and sensor-derived patient data, digital twins hold significant potential applications within the healthcare sector (4, 5). In particular, digital twin technology enables the creation of computerized patient replicas, simulating diverse clinical scenarios and intervention testing *in silico* to reduce avoidable risk in real patients (6).

Digital twins offer particular promise in critical care, where large quantities of data are continuously available, and the risk to patient safety posed by medical interventions is often significant (7, 8). The benefits of a digital twin patient model to inform clinical decision-making in critical illness have been previously proposed (9–12). Digital twins could also be adapted for critical care education, allowing learners to simulate the effects of various interventions and explore their potential outcomes in a controlled, virtual environment without negative patient impacts (13, 14). Compared to conventional virtual patient simulation models, digital twins provide users with a more authentic experience in complex illness management by incorporating real-time, EHR-derived patient data into comprehensive computational models (15–17).

Our group has previously described the design and validation of a novel digital twin based on EHR clinical variables to model critically ill patients with sepsis for bedside decision support (11, 13). In this model, major organ systems interact based on programmed expert rules to recreate and predict the future patient state in response to specific clinical interventions. In this work, we developed a novel application software utilizing this digital twin patient model to simulate clinical trajectories during the initial 6 h of critical illness. This study aimed to assess the usability, workload, and acceptance of the digital twin application software for critical care education in a cohort of internal medicine residents.

2 Materials and methods

Figure 1 provides an overview of the overall critical care patient digital twin project.

This study comprised three sequential phases:

1 Design and coding of a digital twin patient model based on EHR clinical variables and expert rules to simulate patient trajectories during the initial 6 h of critical illness.
2 Development of the user interface for an iOS digital twin application software designed for critical care education delivery.
3 Usability testing of the digital twin application software with a cohort of internal medicine residents and collection of user feedback for iterative software improvement.

2.1 Digital twin patient model design and coding

The digital twin patient model tested in this study focused on physiologic interactions and medication effects relevant to the initial 6 h, or golden hours, of critical illness (18). Variables included in the model comprised clinical data commonly displayed in the ICU EHR. Expert rules describing the interactions between the seven major organ systems (neurologic, respiratory, cardiovascular, gastrointestinal, renal, immunologic, and hematologic) were developed using available literature and current clinical practice guidelines and refined using a modified Delphi panel of international critical care experts (11, 13, 19, 20). Medication effects and pharmacokinetic rules were derived from publicly available drug databases. The model was based on 70 total expert rules and iteratively improved based on feedback from the investigator group. A detailed description of model design and coding, together with two examples of expert rules, are presented in the Supplementary Materials and Methods. The rules that describe the physiologic interactions between the organ system variables are represented graphically in Supplementary Figure 1.

2.2 Digital twin application software development

The digital twin pilot application software tested in this study was developed on iOS using Swift programming language and Xcode integrated development environment version 14.2. User testing sessions were performed with a tablet version of the iOS digital twin application.

The user interface of the digital twin application software consists of a case selection screen, a patient room screen, an EHR screen, and an order entry screen. Users can select a case from a list of virtual clinical scenarios that include urosepsis, chronic obstructive pulmonary disease exacerbation, acute respiratory failure due to pneumonia, acute liver failure, gastrointestinal bleeding, myocardial infarction, and acute decompensated congestive heart failure. Each clinical scenario incorporates specific organ system variable alterations into the initial virtual patient presentation. The user can review the patient's history and physical examination findings on the patient room screen. The EHR screen displays the most relevant data for critical care decision-making, organized by organ systems and color-coded based on the degree of abnormality (21). These data are divided into physical examination, laboratory testing, and other diagnostic findings. Clinical interventions performed by the user are displayed on the EHR screen, maintaining the organ system organization (Figure 2).

After using the order entry screen to initiate a diagnostic test or intervention, the user can advance the timeline (by 15-min intervals for the first hour, then by one-hour intervals until the 6-h endpoint of the simulation) to trigger the associated expert rules coded in the digital twin patient model. The expert rule engine determines which rules are executed based on the interventions ordered and the current value of each organ system variable, which defines the patient's clinical status. The effects of these rules are displayed as changes in the relevant clinical variables presented in the EHR, which reflect the patient's physiological response to the different interventions.

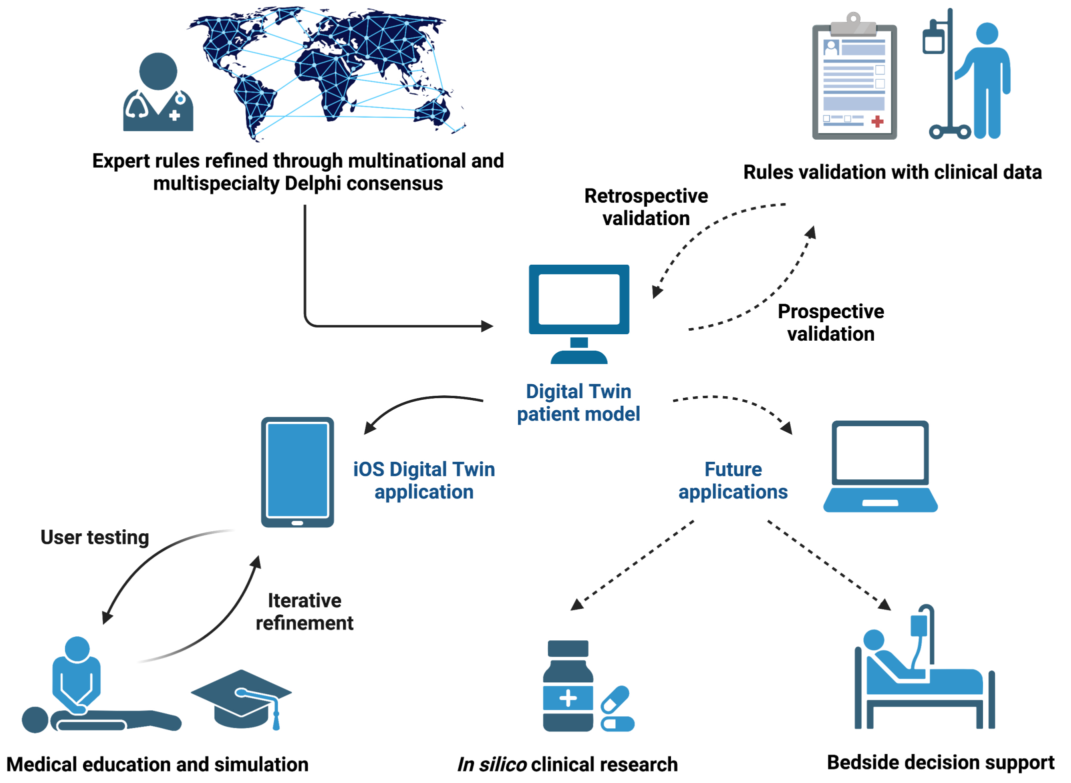

Expert rules refined through multinational and multispecialty Delphi consensus
Rules validation with clinical data
Retrospective validation
Prospective validation
Digital Twin patient model
iOS Digital Twin application
Future applications
User testing
Iterative refinement
Medical education and simulation
In silico clinical research
Bedside decision support

FIGURE 1
Overview of the critical care patient digital twin project. The digital twin patient model was designed based on expert rules and electronic health record clinical variables. In this study, we focused on the development and usability testing of an iOS digital twin application for critical care education (solid arrows). After further prospective and retrospective validation with clinical data, future applications of the digital twin model include *in silico* clinical trials and bedside decision support (dashed arrows). This figure was created with BioRender.com.

2.3 Usability testing of the digital twin application software

2.3.1 Study design and setting

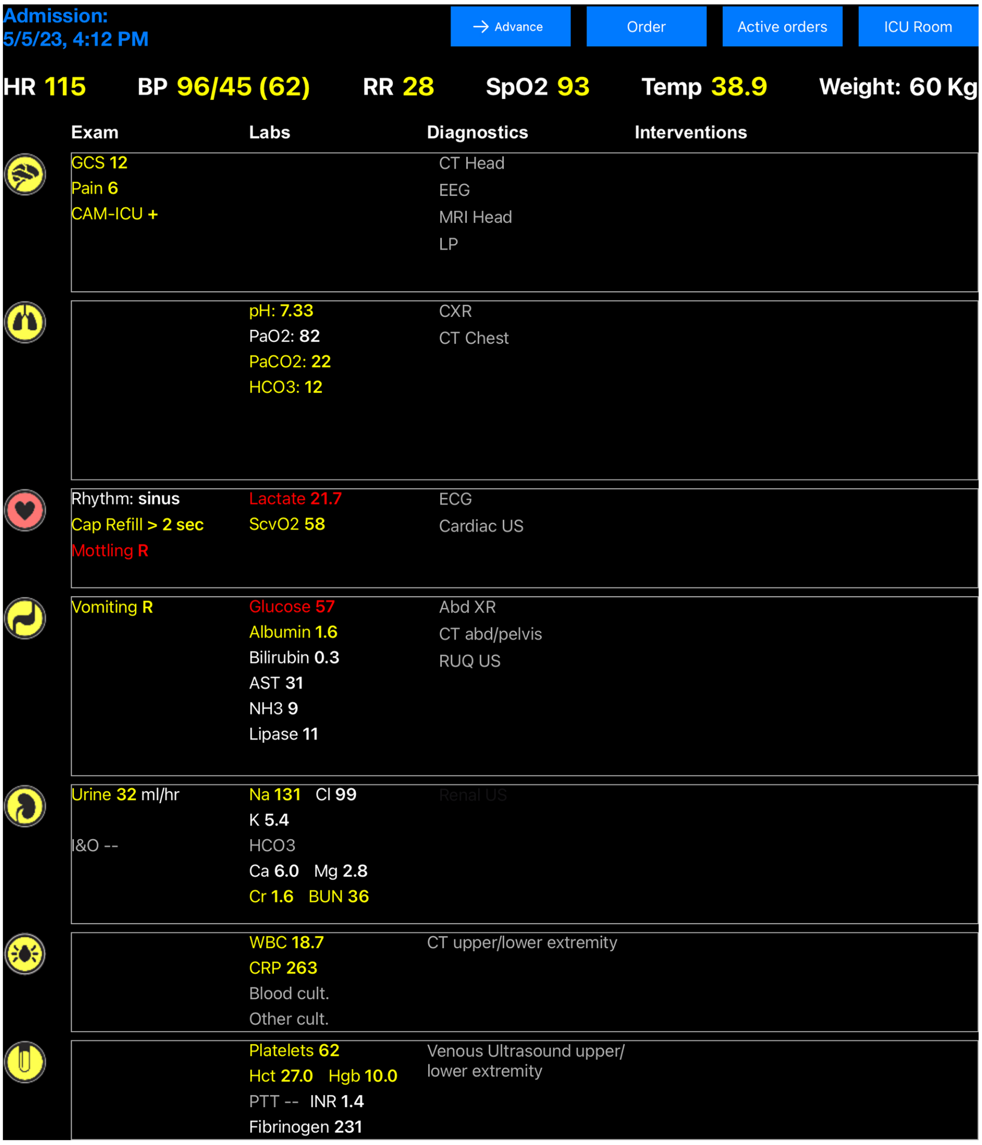
To explore the usability of the digital twin application software as an educational tool in critical care, we collected both quantitative and qualitative data during seven user testing sessions with internal medicine resident volunteers performed at the Mayo Clinic, Rochester, from August 2022 to June 2023. Participants were compensated for their time with a gift card. The study protocol was evaluated and approved as exempt by the Mayo Clinic Institutional Review Board (IRB 21-010982; study title “Critical Care Coaching with an Electronic Health Record Digital Twin”; approval date 11/8/2021) after review by the Mayo Clinic Education Research Committee and the Mayo Clinic Internal Medicine Research in Education Group. The study was conducted in accordance with the ethical standards of the responsible institutional committee on human experimentation and with the Helsinki Declaration of 1975, as most recently amended. Verbal consent was obtained from the participants before each testing session.

2.3.2 Qualitative data collection and analysis

During user testing sessions, residents interacted for 15 min with a simulated case, describing their experience using a think-aloud and semi-structured interview format. The urosepsis case was used for all the user testing sessions to ensure consistency. Each case scenario and debriefing session was recorded, de-identified, transcribed, and analyzed for common themes. Qualitative data were used to refine the software and identify possible digital twin application implementation strategies in the current critical care curriculum.

2.3.3 Quantitative data collection and analysis

The System Usability Scale (SUS), NASA Task Load Index (NASA-TLX), and two survey questions were administered to each user at the end of the simulation session to collect quantitative information on software usability, workload, and learner acceptance. SUS is a measure of usability consisting of 10 questions with five options each (22). The final score ranges from 0 (low usability) to 100 (high usability). NASA-TLX measures perceived workload and evaluates six domains: mental demand, physical demand, temporal demand, performance, effort, and frustration (23). Each domain is scored from 0 (low workload) to 100 (high workload) in 5-point steps, then the unweighted average of the subscale scores is obtained. The survey questions explored how residents would consider using the digital twin application to prepare for or as part of their medical ICU rotation. De-identified data were collected and managed using Research Electronic Data Capture version 8.11.11 (REDCap, Vanderbilt University, Nashville, Tennessee, USA). Statistical analysis was performed using GraphPad Prism version 9.0.0 (GraphPad Software, San Diego, California, USA). To summarize the results, median (interquartile range, IQR) and counts (%) were used.



Admission:
5/5/23, 4:12 PM

HR 115 BP 96/45 (62) RR 28 SpO2 93 Temp 38.9 Weight: 60 Kg

Exam	Labs	Diagnostics	Interventions
GCS 12 Pain 6 CAM-ICU +		CT Head EEG MRI Head LP	
	pH: 7.33 PaO2: 82 PaCO2: 22 HCO3: 12	CXR CT Chest	
Rhythm: sinus Cap Refill > 2 sec Mottling R	Lactate 21.7 ScvO2 58	ECG Cardiac US	
Vomiting R	Glucose 57 Albumin 1.6 Bilirubin 0.3 AST 31 NH3 9 Lipase 11	Abd XR CT abd/pelvis RUQ US	
Urine 32 ml/hr I&O --	Na 131 Cl 99 K 5.4 HCO3 Ca 6.0 Mg 2.8 Cr 1.6 BUN 36	Renal US	
	WBC 18.7 CRP 263 Blood cult. Other cult.	CT upper/lower extremity	
	Platelets 62 Hct 27.0 Hgb 10.0 PTT -- INR 1.4 Fibrinogen 231	Venous Ultrasound upper/ lower extremity	

FIGURE 2
Electronic health record interface of the digital twin application software. Clinical variables included in the digital twin patient model are represented in the electronic health record screen and updated based on expert rules triggered by clinical interventions or changes in the patient's clinical status. White color indicates that a clinical variable is in its normal range and no intervention is needed, while yellow or red colors indicate a variable disturbance that would require urgent or emergent action.

3 Results

Thirty-five post-graduate year one internal medicine residents participated in the user testing sessions of the digital twin application software. All residents were recruited during pre-planned central venous catheter procedural workshops conducted before the start of their medical ICU rotation.

3.1 Digital twin application software usability, workload, and acceptance

The average SUS score in our cohort was 70 (IQR 62.5–82.5), consistent with good software usability (22). The average NASA-TLX score was 29.2 (IQR 22.5–34.2), reflecting a low to moderate workload (24). The scores of each NASA-TLX domain are presented in Figure 3. The greatest perceived difficulty was the successful performance of required tasks, while physical and temporal demand and frustration levels were considered low. Mental demand and overall effort were rated as moderately high. More than 60% of residents indicated that they would use the digital twin application for a moderate amount or a great deal of time to prepare for and as part of their medical ICU rotation (Table 1).

3.2 User feedback for iterative software improvement

Resident comments for iterative software improvement were clustered in five domains, summarized in Table 2. Learners highlighted

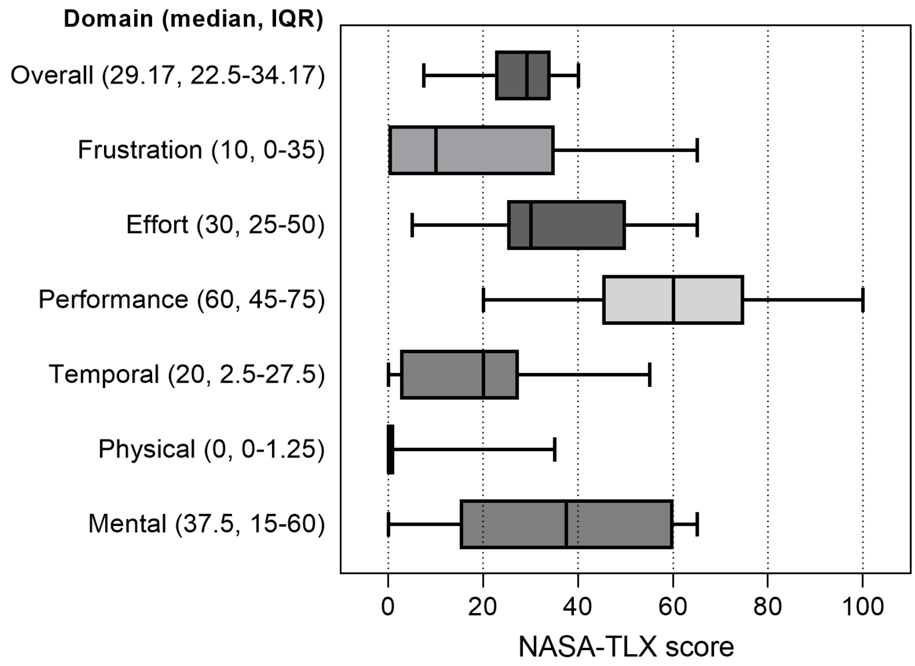


FIGURE 3
Perceived workload of the digital twin application software as measured by the NASA Task Load Index. Overall and single-domain NASA Task Load Index (NASA-TLX) scores were obtained for each resident during user testing sessions (*n* = 35). Box plots represent median values (solid bar), interquartile range (IQR, margins of the box), and minimum and maximum values (whiskers).

TABLE 1 Results from the survey questions assessing the willingness of residents to use the digital twin application for medical ICU orientation and education.

Responses (*n* = 35)	Would you use this tool to prepare for medical ICU rotation?	Would you use this tool as part of your medical ICU rotation?
Never	0 (0%)	0 (0%)
Rarely	2 (6%)	5 (14%)
Occasionally	11 (31%)	6 (17%)
A moderate amount	15 (43%)	17 (49%)
A great deal	7 (20%)	7 (20%)

the importance of the digital twin application delivering a realistic clinical experience, including interactions with the virtual patient and simulated clinical environment and a plausible timeline for scenario progression. Residents also suggested that the EHR interface of the application software should be similar to the commercial product they use in the clinical environment. This would help them to learn to gather and interpret results and navigate the ordering process efficiently. They felt the digital twin application was most helpful in learning medication dosing and effects, enhancing pattern recognition, and improving their understanding of current guidelines through practice managing common ICU scenarios. Learners were mainly interested in a serious gaming experience to test their clinical skills in a safe environment, with a final evaluation reflected by a performance score attributed at the end of each scenario. Residents expressed a willingness to utilize the digital twin application before and during medical ICU rotations; however, they highlighted that their busy clinical schedules pose a significant obstacle to the implementation of the application, as they have limited free time available to use it. To address this issue, the internal medicine residents proposed incorporating practice sessions utilizing the digital twin application software into the current critical care education curriculum.

4 Discussion

This study presents the development and usability testing of a novel application software for critical care education built upon a digital twin patient model based on EHR clinical variables. The digital twin application allows physicians-in-training to test clinical interventions on virtual patients, fostering autonomy and advancing clinical skills in a safe environment that does not expose real patients to preventable harm. Digital twin application testing in a cohort of internal medicine residents suggests high software usability and learner willingness to use this tool to enhance their medical ICU rotation experience.

Although simulation-based education can improve learner confidence and knowledge, evidence supporting superior learning outcomes over more traditional educational delivery methods has varied based on the learning goals (25–28). One notable advantage of simulation is its capacity to offer standardized, reproducible clinical scenarios within a risk-free learning environment, with clear patient safety benefits (29, 30). Emerging technologies, including medical simulation mobile applications and virtual reality, provide further opportunities for remote and on-demand training using simulated clinical cases, providing a consistent framework of residency training experiences that is more cost-effective than traditional high-fidelity simulation (31–33). In addition to providing flexible, efficient online opportunities for deliberate practice, digital twin technology can also integrate real-time patient data to create highly accurate and realistic

TABLE 2 Main themes identified during user testing sessions.

Theme	Sub-themes
Clinical fidelity	Interaction with the virtual patient
	Interaction with the virtual environment
	Virtual time progression similar to real life
Interface organization	Avoid information overload
	Reflect on what is used in daily clinical practice
Learning experience	Learn and practice using medications, including dosing and effects, in common ICU scenarios
	Blend simulation with formal explanations
	Accurate, up-to-date information reflecting current guidelines
Serious gaming	Test clinical skills in a safe environment
	Obtain a performance score at the end of the simulation
Implementation barriers and strategies	Limited free time to use the application software
	Integration of practice sessions with the digital twin application into the existing critical care education curriculum

virtual patient models (4). Indeed, residents underlined the importance of clinical fidelity during user testing sessions of our digital twin application software, including appropriate and realistic responses to clinical interventions. The major disadvantage of the digital twin and other virtual simulation applications is that they do not allow for hands-on practice of the clinical interventions being tested, for which traditional manikin-based simulation remains the gold standard.

Residents acknowledged the potential of the digital twin application to enhance their critical care educational experience. However, they identified clinical schedule demands as the primary obstacle to effectively implementing this tool. In addition to dedicating time within the current critical care curriculum to practice using the digital twin application, residents suggested incorporating additional gamification features, such as a point and badge system, to increase user engagement. Serious gaming has been utilized in various medical education settings, including critical care and emergency medicine, and has been shown to improve knowledge retention and clinical competence (34, 35). However, most studies to date have lacked well-defined control groups, and further research is needed to better understand the benefits of this educational delivery method on learning outcomes, together with the most appropriate learner group, educational context, and experience to achieve these goals (36, 37).

Clinical data display was an important theme raised during software development and user testing sessions. Residents must rapidly learn to identify and review a significant volume of data associated with each patient in the ICU setting. Reviewing this clinical information takes significant time, and this task can feel overwhelming for new trainees without an organized approach (21, 38, 39). To address these challenges, the digital twin application interface displays only the most relevant data for treating critical illness. These data are also organized by organ system and color-coded based on the degree of physiological disturbance and need for action (Figure 2). This user interface design has been shown to reduce time to clinical task completion, task load, and errors of cognition in the ICU when compared with standard EHR interfaces (40, 41). During user testing sessions, residents acknowledged the potential usefulness of the system-based interface organization in the ICU context. However, they also emphasized the differences between this data display and the interface they regularly encounter in their clinical duties. They specifically highlighted the importance of practicing navigation within standard EHR systems at the beginning of their training. This situation creates a dilemma between two distinct learning objectives: the need for clear data presentation to minimize cognitive load and support deliberate practice in critical care decision-making versus data presentation that closely resembles the clinical EHR interface to enhance order entry efficiency through practice but potentially hinders the development of clinical reasoning in typical critical care scenarios. The challenges of adapting to the new interface might also have contributed to the moderately high NASA-TLX scores recorded in the domains of mental demand, successful task performance, and overall effort recorded during testing sessions. Additionally, the significant variations observed in the performance, mental demand, effort, and frustration domains of the score could indicate differences among residents in terms of their critical care knowledge and problem-solving capabilities rather than being attributed solely to the interface itself (42). This subject will require more targeted studies to qualify further.

The digital twin application software offers a convenient, low-cost alternative to enhance the current delivery of critical care education to learners at various levels of experience. This is the first time that digital twin technology has been applied to critical care education. The major strength of our digital twin patient model resides in using transparent pathophysiological relationships to derive expert rules, which have been refined using multinational and multi-specialty Delphi consensus (11, 19). Digital twins can also be developed as purely data-driven models that do not consider causal pathways of diseases, but the lack of clarity in how these physiologic responses are derived creates significant barriers to their acceptance by bedside clinicians (43, 44). To provide clinicians with a better understanding of how the underlying model reaches its output state, future iterations of the digital twin application will offer visualization of pathophysiological relationships using directed acyclic graphs in the user interface (45, 46). The purpose of this methodology for digital twin model design and the user-centered software development process described in this work is to facilitate technology adoption and address the cognitive, emotional, and contextual concerns of clinicians who will utilize this tool (47–49). In the future, the digital twin model will be connected to the current EHR system, allowing continuous update based on real-time patient data to support clinical decision-making, clinical research, and medical education (Figure 1). This will allow clinicians at all experience levels to practice decision-making skills in a safe environment using actual, real-time cases encountered during daily ICU practice. When this step is accomplished, important ethical and regulatory issues must be considered before implementing this novel tool in daily clinical practice (44, 50).

This study has some limitations. First, the digital twin patient model described in this work has been tested on simulated clinical scenarios and on a relatively small cohort of patients with sepsis (11). We plan to prospectively validate this model on a larger cohort of critically ill patients importing real-time EHR data into the application software and further refine expert rules based on these and additional retrospective data. Second, only a limited number of users at a single center participated in the usability testing of the digital twin application software. In addition, all users belonged to a cohort of internal medicine residents with no previous ICU experience, which limit the generalizability of the results. We plan to continue the user testing sessions to iteratively improve the current digital twin application software, involving more senior residents, fellows, and staff intensivists with different experience levels to systematically validate this educational tool's performance and learning outcomes and compare it to more conventional educational techniques.

5 Conclusion

Our novel digital twin application software based on EHR clinical variables proved highly usable and well accepted by first-year internal medicine residents, and their feedback will inform further iterative improvement of its interface. The digital twin application software provides an attractive, realistic, low-cost option to teach critical care clinical decision-making. It offers opportunities for deliberate practice in a virtual environment, building experience and confidence on real-time ICU cases, which may result in greater opportunities for graduated learner autonomy at the bedside and reduced risk of medical errors.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Mayo Clinic Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

LR: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. PG: Conceptualization, Investigation, Methodology, Writing – original draft. EC: Conceptualization, Software, Writing – original draft. YD: Conceptualization, Investigation, Project administration, Supervision, Writing – review & editing. OK: Conceptualization, Investigation, Writing – review & editing. PS: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. XZ: Conceptualization, Funding acquisition, Methodology, Software, Supervision, Writing – review & editing. MW: Conceptualization, Writing – review & editing. DK: Conceptualization, Supervision, Writing – review & editing. OG: Conceptualization, Methodology, Supervision, Writing – review & editing. AN: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. AL: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Center for Advancing Translational Science (CTSA Grant UL1TR002377); the Mayo Clinic Endowment for Education Research Award (EERA) 2022; and the National Science Foundation (Grant IIS-2123900 and IIS-2123848).

Acknowledgments

The authors would like to thank Amila Fejzić, Haris Corbo, Harun Lugušić, and Irvin Kočan for the technical development of the iOS digital twin application software and Simon Zec, Nika Zorko Garbajs, John M. Litell, and William Trevena for their contribution to the development of expert rules.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1336897/full#supplementary-material

References

1. Anderson JG, Abrahamson K. Your health care may kill you: medical errors. *Stud Health Technol Inform*. (2017) 234:13–7. doi: 10.3233/978-1-61499-742-9-13

2. Ahmed AH, Giri J, Kashyap R, Singh B, Dong Y, Kilickaya O, et al. Outcome of adverse events and medical errors in the intensive care unit: a systematic review and meta-analysis. *Am J Med Qual*. (2015) 30:23–30. doi: 10.1177/1062860613514770

3. Herzog TL, Sawatsky AP, Kelm DJ, Nelson DR, Park JG, Niven AS. The resident learning journey in the medical intensive care unit. *ATS Scholar*. (2022) 4:177–90. doi: 10.34197/ats-scholar.2022-0103OC

4. Sun T, He X, Song X, Shu L, Li Z. The digital twin in medicine: a key to the future of healthcare? *Front Med*. (2022) 9:907066. doi: 10.3389/fmed.2022.907066

5. Sun T, He X, Li Z. Digital twin in healthcare: recent updates and challenges. *Digit Health*. (2023) 9:205520762211496. doi: 10.1177/20552076221149651

6. Sahal R, Alsamhi SH, Brown KN. Personal digital twin: a close look into the present and a step towards the future of personalised healthcare industry. *Sensors*. (2022) 22:5918. doi: 10.3390/s22155918

7. Chase JG, Preiser JC, Dickson JL, Pironet A, Chiew YS, Pretty CG, et al. Next-generation, personalised, model-based critical care medicine: a state-of-the art review of in silico virtual patient models, methods, and cohorts, and how to validation them. *Biomed Eng Online*. (2018) 17:24. doi: 10.1186/s12938-018-0455-y

8. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. (2018) 24:1716–20. doi: 10.1038/s41591-018-0213-5

9. Ang CYS, Lee JWW, Chiew YS, Wang X, Tan CP, Cove ME, et al. Virtual patient framework for the testing of mechanical ventilation airway pressure and flow settings protocol. *Comput Methods Prog Biomed*. (2022) 226:107146. doi: 10.1016/j.cmpb.2022.107146

10. Chakshu NK, Nithiarasu P. An AI based digital-twin for prioritising pneumonia patient treatment. *Proc Inst Mech Eng H*. (2022) 236:1662–74. doi: 10.1177/09544119221123431

11. Lal A, Li G, Cubro E, Chalmers S, Li H, Herasevich V, et al. Development and verification of a digital twin patient model to predict specific treatment response during the first 24 hours of sepsis. *Crit Care Explor*. (2020) 2:e0249. doi: 10.1097/CCE.0000000000000249

12. Lonsdale H, Gray GM, Ahumada LM, Yates HM, Varughese A, Rehman MA. The perioperative human digital twin. *Anesth Analg*. (2022) 134:885–92. doi: 10.1213/ANE.0000000000005916

13. Trevena W, Lal A, Zec S, Cubro E, Zhong X, Dong Y, et al. Modeling of critically ill patient pathways to support intensive care delivery. *IEEE Robot Automation Lett*. (2022) 7:7287–94. doi: 10.1109/LRA.2022.3183253

14. Zackoff MW, Rios M, Davis D, Boyd S, Roque I, Anderson I, et al. Immersive virtual reality onboarding using a digital twin for a new clinical space expansion: a novel approach to large-scale training for health care providers. *J Pediatr*. (2023) 252:7–10.e3. doi: 10.1016/j.jpeds.2022.07.031

15. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. *J Med Internet Res*. (2019) 21:e14676. doi: 10.2196/14676

16. Eissing T, Kuepfer L, Becker C, Block M, Coboeken K, Gaub T, et al. A computational systems biology software platform for multiscale modeling and simulation: integrating whole-body physiology, disease biology, and molecular reaction networks. *Front Physiol*. (2011) 2:4. doi: 10.3389/fphys.2011.00004

17. McDaniel M, Keller JM, White S, Baird A. A whole-body mathematical model of Sepsis progression and treatment designed in the BioGears physiology engine. *Front Physiol*. (2019) 10:1321. doi: 10.3389/fphys.2019.01321

18. Hauffe T, Krüger B, Bettex D, Rudiger A. Shock management for cardio-surgical ICU patients – the golden hours. *Card Fail Rev*. (2015) 1:75–82. doi: 10.15420/cfr.2015.1.2.75

19. Dang J, Lal A, Montgomery A, Flurin L, Litell J, Gajic O, et al. Developing DELPHI expert consensus rules for a digital twin model of acute stroke care in the neuro critical care unit. *BMC Neurol*. (2023) 23:161. doi: 10.1186/s12883-023-03192-9

20. Montgomery AJ, Litell J, Dang J, Flurin L, Gajic O, Lal A. Gaining consensus on expert rule statements for acute respiratory failure digital twin patient model in intensive care unit using a Delphi method. *Biomol Biomed*. (2023) 23:1108–17. doi: 10.17305/bb.2023.9344

21. Pickering BW, Herasevich V, Ahmed A, Gajic O. Novel representation of clinical information in the ICU: developing user interfaces which reduce information overload. *Appl Clin Inform*. (2010) 01:116–31. doi: 10.4338/ACI-2009-12-CR-0027

22. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. *Int J Hum Comput Interact*. (2008) 24:574–94. doi: 10.1080/10447310802205776

23. Hart SG. NASA-task load index (NASA-TLX); 20 years later. *Proc Hum Fact Ergon Soc Ann Meet*. (2006) 50:904–8. doi: 10.1177/154193120605000909

24. Grier RA. How high is high? A meta-analysis of NASA-TLX global workload scores. *Proc Hum Fact Ergon Soc Ann Meet*. (2015) 59:1727–31. doi: 10.1177/1541931215591373

25. Zendejas B, Brydges R, Wang AT, Cook DA. Patient outcomes in simulation-based medical education: a systematic review. *J Gen Intern Med*. (2013) 28:1078–89. doi: 10.1007/s11606-012-2264-5

26. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *JAMA*. (2011) 306:978–88. doi: 10.1001/jama.2011.1234

27. Schroedl CJ, Corbridge TC, Cohen ER, Fakhran SS, Schimmel D, McGaghie WC, et al. Use of simulation-based education to improve resident learning and patient care in the medical intensive care unit: a randomized trial. *J Crit Care*. (2012) 27:219.e7–219.e13. doi: 10.1016/j.jcrc.2011.08.006

28. Steadman RH, Coates WC, Huang YM, Matevosian R, Larmon BR, McCullough L, et al. Simulation-based training is superior to problem-based learning for the acquisition of critical assessment and management skills. *Crit Care Med*. (2006) 34:151–7. doi: 10.1097/01.CCM.0000190619.42013.94

29. Seam N, Lee AJ, Vennero M, Emlet L. Simulation training in the ICU. *Chest*. (2019) 156:1223–33. doi: 10.1016/j.chest.2019.07.011

30. Hester RL, Pruett W, Clemmer J, Ruckdeschel A. Simulation of integrative physiology for medical education. *Morphologie*. (2019) 103:187–93. doi: 10.1016/j.morpho.2019.09.004

31. Chandran VP, Balakrishnan A, Rashid M, Pai Kulyadi G, Khan S, Devi ES, et al. Mobile applications in medical education: a systematic review and meta-analysis. *PLoS One*. (2022) 17:e0265927. doi: 10.1371/journal.pone.0265927

32. Jiang H, Vimalesvaran S, Wang JK, Lim KB, Mogali SR, Car LT. Virtual reality in medical students' education: scoping review. *JMIR Med Educ*. (2022) 8:e34860. doi: 10.2196/34860

33. Haerling KA. Cost-utility analysis of virtual and mannequin-based simulation. *Simul Healthc*. (2018) 13:33–40. doi: 10.1097/SIH.0000000000000280

34. Dankbaar ME, Roozeboom MB, Oprins EA, Rutten F, van Merrienboer JJ, van Saase JL, et al. Preparing residents effectively in emergency skills training with a serious game. *Simul Healthc*. (2017) 12:9–16. doi: 10.1097/SIH.0000000000000194

35. Donovan CM, Cooper A, Kim S. Ready patient one: how to turn an in-person critical care simulation scenario into an online serious game. *Cureus*. (2021) 13:e17746. doi: 10.7759/cureus.17746

36. Gorbanev I, Agudelo-Londoño S, González RA, Cortes A, Pomares A, Delgadillo V, et al. A systematic review of serious games in medical education: quality of evidence and pedagogical strategy. *Med Educ Online*. (2018) 23:1438718. doi: 10.1080/10872981.2018.1438718

37. van Gaalen AEJ, Brouwer J, Schönrock-Adema J, Bouwkamp-Timmer T, Jaarsma ADC, Georgiadis JR. Gamification of health professions education: a systematic review. *Adv Health Sci Educ Theory Pract*. (2021) 26:683–711. doi: 10.1007/s10459-020-10000-3

38. Pickering BW, Litell JM, Herasevich V, Gajic O. Clinical review: the hospital of the future - building intelligent environments to facilitate safe and effective acute care delivery. *Crit Care*. (2012) 16:220. doi: 10.1186/cc11142

39. Nolan ME, Cartin-Ceba R, Moreno-Franco P, Pickering B, Herasevich V. A multisite survey study of EMR review habits, information needs, and display preferences among medical ICU clinicians evaluating new patients. *Appl Clin Inform*. (2017) 8:1197–207. doi: 10.4338/ACI-2017-04-RA-0060

40. Ahmed A, Chandra S, Herasevich V, Gajic O, Pickering BW. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. *Crit Care Med*. (2011) 39:1626–34. doi: 10.1097/CCM.0b013e31821858a0

41. Herasevich S, Pinevich Y, Lipatov K, Barwise AK, Lindroth HL, LeMahieu AM, et al. Evaluation of digital health strategy to support clinician-led critically ill patient population management: a randomized crossover study. *Crit Care Explor*. (2023) 5:e0909. doi: 10.1097/CCE.0000000000000909

42. Favre-Félix J, Dziadzko M, Bauer C, Duclos A, Lehot JJ, Rimmelé T, et al. High-Fidelity simulation to assess task load index and performance: a prospective observational study. *Turk J Anaesthesiol Reanim*. (2022) 50:282–7. doi: 10.5152/TJAR.2022.21234

43. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann Intern Med*. (2020) 172:59–60. doi: 10.7326/M19-2548

44. Armeni P, Polat I, De Rossi LM, Diaferia L, Meregalli S, Gatti A. Digital twins in healthcare: is it the beginning of a new era of evidence-based medicine? A critical review. *J Pers Med*. (2022) 12:1255. doi: 10.3390/jpm12081255

45. Kühne F, Schomaker M, Stojkov I, Jahn B, Conrads-Frank A, Siebert S, et al. Causal evidence in health decision making: methodological approaches of causal inference and health decision science. *Ger Med Sci*. (2022) 20:Doc12. doi: 10.3205/000314

46. Bray A, Webb JB, Enquobahrie A, Vicory J, Heneghan J, Hubal R, et al. Pulse physiology engine: an open-source software platform for computational modeling of human medical simulation. *SN Compr Clin Med*. (2019) 1:362–77. doi: 10.1007/s42399-019-00053-w

47. Dang J, Lal A, Flurin L, James A, Gajic O, Rabinstein AA. Predictive modeling in neurocritical care using causal artificial intelligence. *World J Crit Care Med*. (2021) 10:112–9. doi: 10.5492/wjccm.v10.i4.112

48. Lal A, Pinevich Y, Gajic O, Herasevich V, Pickering B. Artificial intelligence and computer simulation models in critical illness. *World J Crit Care Med*. (2020) 9:13–9. doi: 10.5492/wjccm.v9.i2.13

49. Linkov I, Galaitsi S, Trump BD, Keisler JM, Kott A. Cybertrust: from explainable to actionable and interpretable artificial intelligence. *Computer*. (2020) 53:91–6. doi: 10.1109/MC.2020.2993623

50. Lal A, Dang J, Nabzdyk C, Gajic O, Herasevich V. Regulatory oversight and ethical concerns surrounding software as medical device (SaMD) and digital twin technology in healthcare. *Ann Transl Med*. (2022) 10:950. doi: 10.21037/atm-22-4203

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 05 March 2024
DOI 10.3389/fmed.2024.1286582

OPEN ACCESS

EDITED BY
Nathan Gaw,
Air Force Institute of Technology,
United States

REVIEWED BY
Reem Kais Jan,
Mohammed Bin Rashid University
of Medicine and Health Sciences, United Arab
Emirates
Jessica Gonzalez-Vargas,
The Pennsylvania State University (PSU),
United States
Sylvain Bouchigny,
Commissariat à l'Energie Atomique et aux
Energies Alternatives (CEA), France

*CORRESPONDENCE
Cui-Xia Lin
✉ breezelinlin@163.com

RECEIVED 05 September 2023
ACCEPTED 22 February 2024
PUBLISHED 05 March 2024

CITATION
Jiang M-H, Dou L-W, Dong B, Zhang M, Li Y-P and Lin C-X (2024) Development and implementation of a high-fidelity simulation training course for medical and nursing collaboration based on the Fink integrated course design model.
Front. Med. 11:1286582.
doi: 10.3389/fmed.2024.1286582

Development and implementation of a high-fidelity simulation training course for medical and nursing collaboration based on the Fink integrated course design model

Meng-Han Jiang[1], Li-Wen Dou[1], Bo Dong[2], Man Zhang[3], Yue-Ping Li[4] and Cui-Xia Lin[1]*

[1]School of Health, Shandong University of Traditional Chinese Medicine, Jinan, Shandong, China, [2]School of Political Science and Public Administration, Wuhan University, Wuhan, Hubei, China, [3]Peking University First Hospital, Beijing, China, [4]School of Nursing, Shandong Modern University, Jinan, Shandong, China

Aim: The purpose of this study is to examine the design and implementation of a high-fidelity simulation training course for medical and nursing collaboration, based on the Fink integrated course design model. Additionally, the study aims to validate the teaching effectiveness of the course.

Background: Previous empirical studies have highlighted the effectiveness of collaborative healthcare education in institutional teaching and hospital training. However, the development of healthcare collaborative education in China has been slow to develop in China. In recent years, Chinese nursing educators and researchers have shown interest in utilizing high-fidelity simulators for healthcare collaborative education. These simulators help address the limitations of traditional nursing teaching and healthcare separation simulation. Nevertheless, a standardized simulation interprofessional education curriculum is still lacking. Therefore, nursing educators need to develop a standardized high-fidelity simulation training curriculum for healthcare collaboration, guided by established science curriculum development theories.

Methods: A high-fidelity simulation training course on healthcare collaboration was designed based on the Fink integrated curriculum design model. The course was taught to 14 nursing students and 8 clinical medicine students from March to July 2022. To comprehensively evaluate the effectiveness of the healthcare collaboration high-fidelity simulation training course, several assessment tools were used. These included course grades, satisfaction and self-confidence scales, simulation teaching practice scales, healthcare collaboration attitude scales, critical thinking skills scales, and semi-structured interviews.

Results: After the course was implemented, students demonstrated high overall scores (79.19 ± 5.12) and reported high satisfaction ratings (4.44 ± 0.37). They also exhibited increased self-confidence (4.16 ± 0.33). Additionally, students evaluated all four dimensions of the course teaching practice scale positively. Furthermore, the study demonstrated significant improvements in various aspects, such as attitudes toward medical and nursing collaboration ($t = -7.135$,

$P < 0.01$), shared education and teamwork ($t = -3.247$, $P = 0.002$), job autonomy for nurses ($t = -1.782$, $P = 0.000$), and reduced physician dominance ($t = -6.768$, $P = 0.000$). The critical thinking skills of the students showed significant improvement, with higher scores in truth-seeking ($t = -3.052$, $P = 0.004$), analytical ability ($t = -2.561$, $P = 0.014$), systematic ability ($t = -3.491$, $P = 0.001$), self-confidence in critical thinking ($t = -4.024$, $P = 0.000$), and curiosity ($t = -5.318$, $P = 0.000$) compared to their scores before the course (all $P < 0.05$). The interviews showed that the course's student-centered approach enabled active learning. Students suggested enhancing teaching cases and allocating more time for reflection and summarization.

Conclusion: The study successfully designed a high-fidelity simulation training course for healthcare collaboration by utilizing the Fink integrated curriculum design model. The findings provide valuable insights for the development of standardized curricula and healthcare collaboration education in China. Moreover, the course adheres to best practice principles, fostering improved attitudes toward healthcare collaboration and enhancing students' healthcare collaboration and clinical thinking skills.

KEYWORDS

Fink integrated curriculum design model, collaborative healthcare education, high fidelity simulation, curriculum design, collaborative healthcare attitudes, medical and nursing collaboration, critical thinking

1 Introduction

Modern medical personnel training models emphasize the need to strengthen teamwork and promote interprofessional education (1). Interprofessional education, which was first proposed in the United Kingdom during the 1960s, has gained continuous support and development by organizations such as the World Health Organization (WHO) (2). Currently, interprofessional education involves extensive collaboration between institutions and regions (3, 4).

Collaborative healthcare education is a type of interprofessional education where nursing and clinical medicine students learn from each other. The goal is to improve patient health outcomes by strengthening collaboration between healthcare professionals (5). Studies conducted overseas have confirmed the positive effects of collaborative healthcare education on improving students' skills and non-skills. For example, Oxelmark et al. (6) researchers used five clinically common scenarios of interprofessional collaboration scenarios, such as post-operative hemorrhage and allergic reactions, to improve the ability of clinical medical students and nursing students to collaborate during emergencies. Similarly, in a study conducted by Jakobsen et al. (7), nursing students, anesthesia nurses, and clinical medical students underwent interprofessional training. The results showed that the students were able to adapt to their team roles better and enhance their non-technical skills. Lau et al. (8) conducted a 2-day interprofessional advanced cardiovascular life support training for nursing and clinical medicine students. The results showed that the training improved students' team performance, communication skills, and ability to work effectively in acute and critical care situations. In contrast, collaborative healthcare education in China has only been reported in the early 21st century, with research still in its early stages (9).

Scenario-based simulation can provide a safe healthcare environment for collaborative healthcare education and enable students to improve their practical skills in real-life situations. In recent years, the development of situational simulation teaching has garnered attention from nursing educators and researchers in China, particularly in the realm of medical-nursing collaborative education based on high-fidelity simulators. Wang et al. (10) investigated the effectiveness of high-fidelity simulation in teaching operating room nursing collaboration. Other researchers have also applied this method in nursing planning and implementation (11) and emergency nursing courses (12). The results demonstrate that this teaching method can enhance students' interest in learning and improve their teamwork skills. Currently, China's high-fidelity simulation teaching of healthcare collaboration is still in the developmental stage. Most researchers design the teaching content based on the actual needs and available resources of their institutions. The teaching is mostly carried out by focusing on one or more trainings in a nursing specialty course (13–15). However, this approach may lack scientific rigor in the teaching process and make it difficult to compare teaching effects horizontally.

Since curriculum development is the initial step in implementing curriculum teaching, and its quality directly affects the curriculum's implementation, nursing educators must standardize the development of a high-fidelity simulation training course for healthcare collaboration under the guidance of scientific curriculum development theories. Studies have shown that educators, both domestic and international, have adopted curriculum development theories to guide the process. For instance, some have used the flexible learning model to design a health assessment course (16), while others have developed their own model based on competency-based education theory (17).

However, one integrated curriculum design model (Below is referred to as the "Fink model") that has emphasized the creation of meaningful learning experiences as a key aspect of quality education was developed by Fink (18). The model is holistic, comprehensive, and practical, focusing on both theoretical exploration and conceptual analysis, as well as concrete implementation to improve teaching effectiveness (19).

The Fink model has been successful in a variety of fields, including basic dental anatomy courses (20), health policy courses (21), and narrative nursing courses (22). In this study, the Fink model served as the theoretical basis for developing a high-fidelity simulation training course for healthcare collaboration, offering several benefits: (1) This tool assists educators in analyzing the course needs to clarify the course's nature and curriculum significance objectively. (2) Instead of traditional goal-setting, this tool employs meaningful learning objectives. (3) The course evaluation elements align with the formative and summative evaluation advocated by the simulation teaching evaluation method. (4) Analyzing whether the course elements can support each other to ensure the course's systematic nature; and (5) Predicting potential problems that may arise during the course implementation stage to ensure its feasibility.

Based on the need to improve curriculum development for collaborative education, a SimMan3G (SimMan3G is actually a high-fidelity mannequin from Laerdal) has been developed as an integrated simulator-based healthcare cooperation training curriculum using Fink's design model. This study aims to explore the development, implementation, and evaluation of the SimMan3G in teaching nursing and clinical medicine students. The findings will provide valuable insights for standardizing the development of healthcare collaboration curriculum, cultivating students' awareness of healthcare collaboration, and enhancing their healthcare collaboration skills.

2 Materials and methods

This study is divided into two parts: curriculum development and curriculum implementation. Firstly, we explored the process of developing a SimMan3G-based collaborative healthcare training course using the Fink model. Secondly, we implemented the curriculum with students from two specialties, clinical medicine and nursing, as research subjects and verified its teaching effectiveness.

2.1 Course development

2.1.1 Theoretical basis

The Fink integrated curriculum design model consists of three phases (18), outlined in **Table 1**. Each phase includes specific operational steps to guide educators through the curriculum development process. The initial stage is particularly important and serves as the foundation for designing a course. To guide the development of a SimMan3G-based healthcare collaboration training course using the Fink model, instructional designers should first analyze contextual factors to understand the current status of healthcare collaboration in the nursing field in China. Then, they should determine meaningful learning objectives for the course and select appropriate feedback assessment procedures and effective teaching activities based on the course objectives. The intermediate phase aims to integrate foundational elements into a dynamic and coherent whole. The final phase aims to enhance the curriculum design.

TABLE 1 Fink integrated curriculum design model content.

Stage	Main steps
Initial phase: determining the foundational elements of the course	Clarify contextual factors
	Define learning objectives
	Develop appropriate feedback and evaluation systems
	Designing teaching activities
	Integration of the identified basic components of the curriculum
Intermediate phase: integration of essential factors into the whole	Designing course structure
	Choosing effective teaching strategies
	Designing an overall learning activity plan
Final phase: completion of other important tasks	Establishment of a scoring system
	Identify issues that may arise
	Completing a course outline
	Planned assessment of curriculum and instruction

2.1.2 Course construction

This study presents the development of a high-fidelity training course for medical and nursing collaboration in three stages: initial, intermediate, and final. The Fink model was used as a basis for this construction. The analysis of each stage is presented below:

(1) Initial stage

The contextual factors of the course include six specific aspects. (1) External Expectations: The aim of this course is to address the issue of neglecting healthcare collaboration in nursing practical training courses and promote teaching reform in the nursing profession. (2) Specific Context: This course was proposed in the context of the new medical science background (23) and the specific context of China's relatively lagging development of education on healthcare collaboration. (3) Course Nature: The course is an interprofessional elective course on medical situational simulation, which emphasizes the cultivation of teamwork attitudes and abilities among nursing and clinical medical students. (4) Student Characteristics: The students are senior-level and possess professional knowledge and basic operational skills. They can analyze cases based on their own understanding. (5) Teacher characteristics: the teachers all have the title of associate senior and above and rich experience in simulation teaching, and they can instruct the students how to use SimMan3G for training. (6) Teaching special challenges: the SimMan3G integrated simulation system can't meet the actual needs of the teaching content of the course. As a result, the School of Nursing, the School of Clinical Medicine, and the teaching hospital collaborated in the preliminary stage to jointly prepare eight teaching cases based on certain case preparation principles and processes (24).

Fink emphasizes the importance of meaningful learning in teaching practices and has created six taxonomies to achieve this: basics, applications, synthesis, humanities, caring, and learning to learn. When determining the course's total learning objectives based on this taxonomy, teachers should focus not only on students' understanding of the basics but also on developing their application skills and other levels (25). The study developed the courses' learning objectives, which are listed in Table 2.

The course was evaluated using three methods: prospective assessment, self-assessment, and FIDeLity feedback (Frequent, Immediate, Discriminating based on criteria and standards, Delivered Lovingly or supportively). A questionnaire was used to assess changes in students' attitudes toward healthcare cooperation and critical thinking skills before and after the course implementation. The instructor conducted summative scoring of group-recorded case videos using a self-designed key competency checklist. The checklist includes 5 areas: team decision-making, communication, situational monitoring, mutual support, and first aid, with 20 points allocated to each area. The checklist was used to develop students' self-assessment skills. Additionally, the instructor utilized a Context-Content-Course (3C) guided feedback model (26) to encourage students' analysis and reflection during high-fidelity simulation training sessions. The course included various active learning activities such as independent review of theoretical knowledge and skills related to the case, role-playing, collaborative learning, high-fidelity simulation training, and guided feedback. The course facilitated student learning through three areas: gaining information and perspectives, experiencing, and reflecting.

TABLE 2 Total learning objectives of the medical-nursing collaborative high-fidelity simulation training course.

Dimensionality	Course objectives
Basic knowledge	Master the basic theoretical knowledge of case-related diseases and diagnostic and treatment (nursing) measures, familiar with the assessment, diagnosis, and treatment (nursing) plan development
Applications	Ability to perform specialized skills in related diseases and to work effectively with team members in the development of diseases
Synthesis	Ability to think about the connections between the 2 disciplines, the meaning of division of labor and collaboration, and how to apply collaborative thinking and skills in healthcare to future clinical work
Humanities	To be able to recognize the role of the learning process, to improve the attitude of cooperation between healthcare and nursing, and to take the code of professional ethics as the guiding code of conduct, reflecting the humanistic care for patients
Caring	Be curious and motivated by the phenomena, ideas, and learning process of the content being studied
Learning to Learn	Build knowledge through reflection and promote independent learning while strengthening the effect of simulation teaching

A review form based on the Fink design was used to examine a high-fidelity simulation training course on healthcare collaboration. The course addresses the learning objectives and selects appropriate feedback and assessment methods and instructional activities. The foundational elements were able to support each other and work together to promote meaningful learning.

(2) Intermediate stage

The course was an elective and consisted of two topics: introduction and case study. The introduction topic was allocated 2 h, while each of the 8 cases was assigned 4 h, resulting in a total of 34 h of instruction. The course employed a "team-based learning" strategy, leveraging the SimMan3G integrated simulator to simulate real clinical situations. Students worked in groups to engage in high-quality applied learning for the cases. The course design consisted of four components: course theme, teaching content, teaching activities, and credit hours, as shown in Table 3.

(3) Final stage

After identifying the course elements in the first two stages, the final stage involves determining the course's teaching assessment, grading system weighting, and completing the course outline. The course outline comprises eight sections: basic information (including course name, total hours, prerequisite courses, applicable target, and course leader), course objectives, teaching content and class schedule, teaching methods, performance assessment methods, recommended teaching materials, connection and division of labor with other courses, and course introduction.

2.2 Course implementation

2.2.1 Study population

In March 2022, a teaching class was formed for the study, consisting of students in the fourth year of a 5-year clinical medicine program and the third year of a 4-year nursing program at a university. The recruitment criteria are as follows: (1) Full-time undergraduate clinical medicine and nursing majors; (2) Completion of basic medical courses, including human anatomy, pathology, and physiology. Clinical medicine students have also completed professional courses such as surgery, internal medicine, obstetrics and gynecology, and pediatrics. Nursing students have completed courses such as surgical nursing, internal medicine nursing, obstetrics and gynecology nursing, and pediatric nursing; (3) No exposure to interprofessional-related content in daily practical training; (4) Experience in simulation learning; (5) Availability to participate in the course; and (6) Understanding of the purpose and significance of the course. Due to time constraints and limited manpower, 22 students were recruited for the initial course development. The participants included 14 nursing students and 8 clinical medicine students, with ages ranging from 20 to 23 years old (mean age of 20.73 ± 0.94 years). The group consisted of 2 male and 20 female participants. In order to further verify the reliability of the data, we have done a power analysis, which shows that the data has good reliability.

TABLE 3 Overall plan of the medical-nursing collaborative high-fidelity simulation training course.

Sessions	Topics	Teaching content	Teaching activities	Credit hours
1	Introduction	Introduction to the course (objectives, teaching arrangements, evaluation methods), explanation of the application of the simulation system (simulators, scene layout, simulation fidelity, how students observe during the simulation)	Lecture method	2
2	Case study	Acute myocardial infarction	Case studies, Roleplay, Collaborative learning, High-quality simulation training, Guided feedback	4
3		Diabetic ketoacidosis		4
4		Perforated duodenal ulcer		4
5		Thyroid Cancer		4
6		Postpartum bleeding in normal labor		4
7		Amniotic fluid embolization		4
8		Neonatal asphyxia resuscitation		4
9		Pediatric severe pneumonia		4

2.2.2 Study design

The course implementation is divided into 2 parts: pre-teaching preparation and teaching implementation. Pre-teaching preparation involves preparing the teachers, students, and learning environment. Teaching implementation follows the steps of scenario introduction, high-fidelity simulation training, and review. Take "acute myocardial infarction" for example, the details are described as follows:

(1) Pre-teaching preparation

Each case is taught by a team of instructors consisting of a nursing faculty member, a clinical medicine faculty member, a laboratory faculty member, and a teaching assistant. The instructors conduct an in-depth analysis of the case and prepare a lesson plan in advance. The lesson plan contains a schedule, training objectives, prerequisite knowledge for students, case overview, pre-course preparation (including scene setting, simulators, teaching aids, role division, consultation/nursing aids, and drugs), case trend chart, development process, and review outline. Furthermore, the case and learning tasks are provided to students beforehand. The laboratory instructor imports the case information into the instructor console for the teaching team to pilot. They work with the teaching assistant to provide the necessary equipment and items for the class according to the lesson plan. Before class, students form their own medical and nursing cooperative teams, determine their roles, familiarize themselves with the script, and review the relevant theoretical knowledge and operational skills.

(2) Teaching implementation

In the introduction scenario link, the teacher presents the students with a high-fidelity simulation training case of acute myocardial infarction healthcare collaboration, as shown in Table 4. The teacher addresses any questions the students may have encountered during their independent study before the class, confirms the role division of students, analyzes the simulation tasks with them, explains the presentation requirements, and encourages students to be fully prepared for the training. During the high-fidelity training session, the teacher initiates the program, and students assume their roles based on the disease progression and tasks in each scenario. This commences the high-fidelity training for medical and nursing collaboration, as shown in Figure 1. One group performs the simulation training while the other groups observe and record through live video in the observation room. During the review session, the teacher and students review the high-fidelity simulation training process together using video replay. The review session consists of two phases: (1) Introduction phase, during which the teacher explains the purpose and steps of the session to the students, and (2) Situational phase. The teacher prompts students to provide feedback on the performance of their peers during high-fidelity training. This is done by asking simple questions such as "How do you feel about the performance of this group of students just now?" (3) The content stage involves presenting objective facts, encouraging open discussion, and providing the teacher's perspective from the patient's point of view. (4) The expansion phase follows. Students are instructed to summarize their learning experiences and consider how they can apply what they have learned to their future clinical practice.

2.2.3 Evaluation methods

A mixed methods approach is suitable for comprehensively evaluating the SimMan3G collaboration training curriculum. When evaluating the effectiveness of nursing high-fidelity simulation teaching, researchers usually focus on various aspects, including student achievement, course satisfaction, student confidence, teamwork ability, and critical thinking ability (27–29). This study comprehensively assessed the teaching effectiveness of the course based on the following dimensions:

(1) Student Course Grades: The total score is graded out of 100 points. The weight of each assessment component was determined based on the course syllabus and the opinions of the interdisciplinary teaching team. The formative evaluation constitutes 60% of the total student course grade, with 10% for self-evaluation, 20% for peer evaluation, and 30% for teacher evaluation. The remaining 40% is allocated to teacher evaluation of the group recording video.

(2) Student Satisfaction and Self-confidence in Learning (SSS): The SSS scale, developed by the National League for Nursing

TABLE 4 Acute myocardial infarction healthcare cooperation high-fidelity simulation practical training case.

Case title	Acute myocardial infarction
Teaching goal	① Cognitive domain: recognize the etiology of acute myocardial infarction, associated risk factors, and clinical manifestations. ② Action skill domain: medical students need to apply the knowledge they have learned to skillfully implement the receiving process, body check, cardiopulmonary resuscitation, bedside electrocardiogram, defibrillation; nursing students need to skillfully implement indwelling catheterization, intravenous fluids, and collection of blood specimens; and medical and nursing students jointly master the resuscitation process. ③ Emotional domain: students embody humanistic care through good communication with patients and their families; through the implementation of treatment as well as nursing measures for patients, students develop a collaborative attitude toward healthcare.
Case description	Patient, male, 61 years old, chief complaint and history: the patient complained of chest pain that suddenly appeared 1 h ago with no obvious cause, the pain site is mainly in the precordial area, and the pain range is about the size of the palm, the pain is pressure-like pain, accompanied by profuse sweating, palpitation, radiating pain in the back of the shoulder and the pharynx, there is no nausea, vomiting, there is no tightness in the chest, shortness of breath, fatigue, there is no coughing, coughing up sputum, hemoptysis, Self-medication "fast-acting heart pills" after the symptoms did not relieve, and he called 120 and came to our hospital urgently. He underwent cardiopulmonary resuscitation and electrocardiogram showed "acute extensive anterior wall myocardial infarction," and was transferred to our department for thrombolytic therapy. Past history: 10 years history of hypertension and coronary heart disease. Physical examination: temperature 36.5°C, respiration 21 times/min, pulse 90 times/min, blood pressure 90/59 mmHg, clear, superficial lymph nodes are not palpable enlargement, lips and lips without cyanosis, no jugular veins; symmetry of the thorax, the lungs breath sounds thick, heard full lung wet rales, percussion of the cardiac boundary is not big; listening to the rhythm of the heart is synchronous, the valvular auscultation area did not hear a murmur; the abdominal flat and soft, no compression pain and rebound pain The abdomen was flat and soft, with no pressure or rebound pain. The liver and spleen were not palpable, and there was no edema in the lower limbs. The electrocardiogram showed that the V1-V5 ST segments were elevated about 0.3–0.5 mv.
Scenario setting	Scenario 1: out-of-hospital treatment ① Doctor's task: 120 telephone reception, instructing family members to perform cardiopulmonary resuscitation, bedside electrocardiogram measurement, decision-making, and completion of medical orders; ② Nurse's task: oxygen supply, establishment of intravenous access, and administration of medication in accordance with medical advice; ③ Medicine and nursing joint task: communication of the patient's vital signs, and comforting the patient's family members. Scenario 2: in-hospital emergency care ① Doctor's task: explain the patient's condition, bedside electrocardiogram, cardiopulmonary resuscitation, and defibrillation, to complete the doctor's orders; ② Anesthesiologist's task: endotracheal intubation, simple respiratory balloon ventilation; ③ nurse's task: the preparation of resuscitation supplies, coordination of various departments to do a good job of resuscitation preparations, blood sampling, resuscitation records; ④ healthcare common task: communication of the patient's vital signs Scenario 3: internal medicine treatment ① Doctor's task: physical examination, asking the family about the patient's medical history, decision-making about thrombolytic therapy, judgment of the condition; ② Nurse's task: blood sampling, thrombolytic operation, changing the patient's position, oxygenation, indwelling catheterization, resuscitation records; ③ Healthcare co-worker's task: explaining to the patient's family about the treatment and recommendation for transferring to a different hospital.

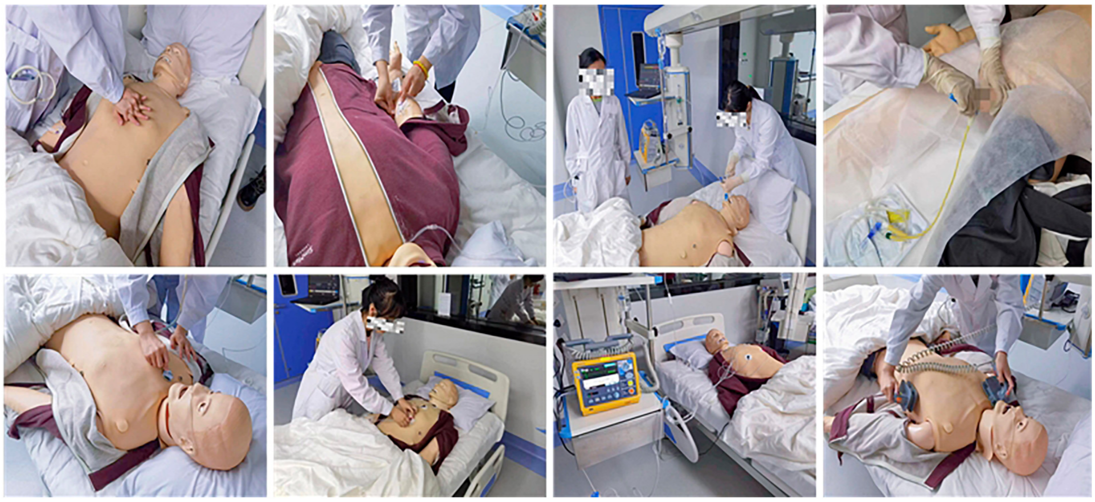


FIGURE 1
Students undergoing high-fidelity simulation training.

in collaboration with Laredal (30), is completed by students after the course implementation. It consists of two subscales: satisfaction and self-confidence, each comprising 13 items rated on a Likert 5-point scale. Higher scores indicate greater levels of satisfaction and self-confidence.

(3) Educational Practices in Simulation Scale (EPSS): The EPSS measures the extent to which best practice principles are applied in simulation instruction. It consists of four dimensions: self-directed learning, collaboration, learning styles, and high expectations, with a total of 16 items. The

scale used is a Likert 5-point scale, and the total score ranges from 16 to 80, with higher scores indicating a higher degree of application of best practice principles in the simulation. The Cronbach's alpha coefficient of the EPSS is 0.91 (31). The Chinese version of Wang et al. (32) from 2013 was used in this study, with a Cronbach's alpha coefficient of 0.94.

(4) Jefferson Health Care Cooperation Attitude Scale: This scale, developed by Hojat et al. (33), measures physicians' and nurses' attitudes toward healthcare cooperation. The Chinese version by Yang et al. (34) was used in this study. It consists of four dimensions: shared education and teamwork (7 items), nursing vs. treatment (3 items), nurses' work autonomy (3 items), and physician domination (2 items), with a total of 15 items. The Likert 4-point scale is used, and the total score ranges from 15 to 60, with higher scores indicating a more positive attitude toward healthcare cooperation. A score between 45.01 and 60.00 was considered a high level of healthcare cooperation attitude, while a score between 30.01 and 45.00 was considered moderate, and a score between 15.01 and 30.00 was considered low. Hojat et al. (35) assessed the structural validity, content validity, and reliability of the scale. The Chinese version of the Jefferson Health Care Cooperation Attitude Scale had a Cronbach's alpha coefficient of 0.848 and a content validity index of 0.893.

(5) Critical Thinking Disposition Inventory-Chinese Version (CTDI-CV): The impact of the curriculum before and after its implementation was assessed using the CTDI-CV, which was translated and revised by Peng et al. (36). The inventory consisted of 70 items, categorized into 7 dimensions: truth-seeking, open-mindedness, analytical ability, systematic ability, self-confidence in critical thinking, intellectual curiosity, and cognitive maturity. Each dimension comprised 10 items. A 6-point scale was used to measure critical thinking ability, ranging from 1 (strongly disagree) to 6 (strongly agree). Some items were reverse scored. The total score ranged from 70 to 420. Scores of 70–210 indicated negative critical thinking ability, 211–279 represented unclear meaning, 280–349 reflected positive critical thinking ability, and 350–420 denoted strong performance. The scale exhibited strong internal consistency, as demonstrated by a Cronbach's alpha coefficient of 0.90, and content validity with an index of 0.89.

(6) Semi-structured interview: The study conducted one-to-one semi-structured interviews using an interview outline as a basis, as shown in **Figure 2**. The researcher developed the outline based on a literature review, the study's purpose, and input from the teaching team. Two students were then selected for pre-interviews to ensure the outline met the research questions' needs. The final version of the interview outline was formed by the researcher after correcting any misrepresentations of the pre-interviews. The outline included specific elements such as inquiring about the most helpful aspect of the course for personal professional development and identifying strengths and weaknesses in the program's design and implementation. What suggestions do you have for improving the implementation of the course in the future? The instructor conducted interviews with the students at the end of the course instruction in July 2022. After analyzing the profiles of eight students, no new themes emerged, indicating that data saturation had been reached. The interviews continued with two additional students, resulting in a sample size of ten students.

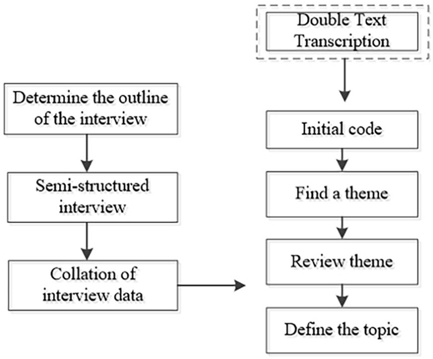

Double Text Transcription
Determine the outline of the interview
Initial code
Semi-structured interview
Find a theme
Collation of interview data
Review theme
Define the topic

FIGURE 2
Semi-structured interview research process.

2.3 Statistical analysis

All raw data were entered into an Excel sheet and imported into SPSS 25.0 statistical software for analysis (37). Descriptive statistics, specifically the mean ± standard deviation, were employed to depict the students' age. Two independent samples *t*-tests were conducted to compare the scores of the attitude toward healthcare cooperation scale and the critical thinking skills scale before and after the course (both scale scores followed a normal distribution). The scores of the simulated teaching practice scale, student learning satisfaction, and self-confidence scales were examined for normality and demonstrated conformity to a normal distribution, thus described using the mean ± standard deviation.

This study employed a phenomenological research methodology (38) to fully comprehend the students' experience of the course, a widely used approach in the fields of nursing education, nursing administration, and clinical nursing. The collection, transcription, and analysis of interview data were conducted simultaneously. Each respondent's audio-recorded interview data was transcribed into text within 48 h by a team consisting of Menghan Jiang and Bo Dong. The interview text data were managed, analyzed, and coded using the Colaizzi seven-step analysis method and NVivo 12.0 software (39, 40). Using the above analytical procedures, this study initially labeled the initial data of the ten students as A1–A10 (A1–A3 for clinical medical students, A4–A10 for nursing students). The initial data was then refined and summarized to form sub-themes, denoted as B1–Bn. These sub-themes were further generalized to form the themes of this study, denoted as C1–Cn.

2.4 Ethical procedures

The study was approved by the Ethics Committee of Shandong University of Traditional Chinese Medicine before

TABLE 5 Student achievement scores.

	Minimum value	Maximum value	Score (X ± S)
Formative evaluation	38.80	53.30	46.80 ± 3.51
Self-esteem	6.40	9.30	7.87 ± 0.70
Others' evaluations	13.00	17.60	15.62 ± 1.20
Teacher evaluation	18.00	26.40	23.32 ± 1.87
Summative evaluation	28.80	36.80	32.38 ± 2.01
Totals	69.20	90.10	79.19 ± 5.12

TABLE 6 Student satisfaction, self-confidence, and teaching practice scale scores.

Scale	Dimensionality	Score (X ± S)
Satisfaction and self-confidence scales	Satisfaction	4.44 ± 0.37
	Self-confidence	4.16 ± 0.33
Simulation of teaching practice scale	Independent learning	4.19 ± 0.27
	Cooperation	4.39 ± 0.34
	Multiple learning styles	4.41 ± 0.40
	High expectations	4.23 ± 0.34

data collection. The researcher provided a comprehensive explanation of the study's purpose, methods, and significance to the prospective participants, who were given the freedom to decide whether or not to participate after being fully informed. The questionnaire was collected anonymously, and the researcher assured the participants that the personal data collected would be strictly utilized for academic research purposes only. Moreover, the video recordings of the teaching process and the interview content would be treated with utmost confidentiality.

3 Results

3.1 Student course grades

At the end of the course, the average score of the 22 students ranged from 69.2 to 90.1, with a mean of 79.19 ± 5.12. Out of these, one student scored 90.01 or above, seven students scored between 80.01 and 90, twelve students scored between 70.01 and 80, and two students scored between 60.01 and 70. The scores for each specific subdimension are detailed in Table 5.

3.2 Student satisfaction, self-confidence, and teaching practice scale scores

Table 6 displays the results of the survey on students' satisfaction with course teaching, self-confidence, and feelings about teaching practice. The mean score for students' satisfaction with course teaching was 4.44 ± 0.37 (maximum average score of 5), with 21 students (95.45%) scoring 4 or higher, and no students scoring below 3. The mean score for self-confidence was 4.16 ± 0.33 (maximum average score of 5), with 15 students (68%). All students scored 3 or higher, with 18% scoring 4 or higher. Students reported positive perceptions of the teaching practice experience, with all four dimensions of the teaching practice scale receiving high ratings: independent learning, cooperation, multiple learning styles, and high expectations. The dimension with the highest score was multiple learning styles, with a mean score of 4.41 ± 0.40.

3.3 Comparison of students' attitudes toward healthcare cooperation scores before and after the implementation of the curriculum

Table 7 illustrates the changes in students' scores on the HealthCare Cooperation Attitude Scale before and after the course. The scores and total scores for the dimensions of shared education and teamwork, job autonomy of nurses, and physicians' domination were significantly higher after the course, demonstrating statistically significant differences ($P < 0.01$). However, there were no significant differences in the control dimensions of nursing and treatment.

3.4 Comparison of student's critical thinking skills scores before and after the implementation of the curriculum

Table 8 presents the differences in students' scores and total scores for each dimension of the Critical Thinking Skills Scale before and after the course. Statistically significant differences were observed in the scores and total scores for each dimension, indicating a significant improvement in critical thinking skills after the course. Notably, the comparative differences in scores for the open-mindedness and cognitive maturity dimensions were not statistically significant.

3.5 Results of interviews

This study constructed 11 sub-themes (B1–B11) and 4 themes (C1–C4) by coding, organizing, and analyzing the content of the interviews. C1–stimulating interest in learning

TABLE 7 Comparison of students' attitudes toward healthcare cooperation before and after the implementation of the curriculum (X ± S).

Projects	Pre-teaching	After teaching	*t*	*P*
Shared education and teamwork	22.82 ± 1.99	24.55 ± 1.50	−3.247	0.002***
Comparison of nursing and treatment	9.91 ± 0.87	10.45 ± 1.14	−1.782	0.082
Job autonomy of nurses	9.64 ± 1.36	10.91 ± 0.92	−3.626	0.000***
Physicians' domination	4.59 ± 1.14	6.68 ± 0.89	−6.768	0.000***
Total score	46.95 ± 2.87	52.59 ± 2.34	−7.135	0.000***

***$P < 0.01$.

TABLE 8 Comparison of students' critical thinking skills before and after the implementation of the curriculum (X ± S).

Projects	Pre-teaching	After teaching	*t*	*P*
Searching for the truth	34.05 ± 4.99	38.32 ± 4.27	−3.052	0.004***
Open-mindedness	41.86 ± 3.54	42 ± 3.82	−0.123	0.903
Analytical skills	42.41 ± 4.89	45.5 ± 2.86	−2.561	0.014**
Systematic capabilities	38.73 ± 5.16	42.95 ± 2.38	−3.491	0.001***
Self-confidence in critical thinking	39.55 ± 4.81	45.14 ± 4.40	−4.024	0.000***
Desire for knowledge	42.41 ± 3.49	48.14 ± 3.66	−5.318	0.000***
Cognitive maturity	39.27 ± 6.48	41.82 ± 3.70	−1.600	0.117
Total score	278.27 ± 21.85	303.86 ± 13.90	−4.635	0.000***

**$P < 0.05$, ***$P < 0.01$.

and promoting active learning; C2–collaborative learning and improving healthcare collaboration; C3–student-centeredness and promoting the development of clinical thinking skills; and C4–students' suggestions for curriculum optimization and improvement. The levels and information of specific nodes are shown in Table 9.

4 Discussion

This study developed a simulation training course for medical and nursing collaboration based on the Fink model. The course's teaching effectiveness was evaluated, and the results showed that all students passed the assessment with a mean grade of 79.19 ± 5.12. The course grades were calculated by combining formative and summative evaluations. Formative evaluations included self-evaluation, peer evaluation, and teacher evaluation. Self-evaluation and peer evaluation promote effective student participation in class. Teacher evaluation, based on group members' performance, helps teachers focus on individual performance. Video evaluation serves as the summative review for the teacher after teaching the course. This assessment approach is multifaceted, focusing not only on student learning outcomes but also on capturing changes in the learning process.

According to research, best practices in undergraduate education involve seven principles. These include developing reciprocity and cooperation among students, honoring diverse talents and learning styles, and providing timely feedback (41). The Simulated Teaching Practices Scale used in this study can assess the extent to which these principles are implemented. The study results indicate that all dimensions scored above 4, similar to Liu et al's study (42), suggesting that the course adhered to best practice principles. The course objectives are clearly stated and emphasize independent learning and active participation. This allows for effective communication and idea exchange between students and teachers, with the latter providing guidance to address individual student needs. As a result, students express high satisfaction with the course's teaching methods, scoring it (4.44 ± 0.37) which is higher than in other studies (43).

Self-confidence is an essential trait for healthcare professionals to possess, as it can greatly impact their clinical decision-making ability and response to emergencies. Research has shown that individuals with higher levels of self-confidence are better equipped to handle the challenges they encounter, particularly in the realm of patient safety (44). Therefore, it is crucial to cultivate self-confidence in medical and nursing students. The study found that the curriculum significantly contributed to the students' confidence levels, as evidenced by their self-confidence score of (4.16 ± 0.33). This can be attributed to the hands-on opportunities provided by the course, where students were able to apply their knowledge and skills in completing case tasks alongside their team members during high-fidelity simulation training. Such experiences fostered confidence in their abilities and knowledge (45).

According to a study (46), a standardized interprofessional collaborative education program has a positive impact on developing students' teamwork skills and overall competence. The study found that completing the course significantly improved students' attitudes toward healthcare cooperation and their scores in three dimensions: shared education and teamwork, job autonomy of nurses, and physicians' domination ($P < 0.05$). In interviews, students emphasized that the curriculum improved

TABLE 9 Interview results nodes.

Topics	Sub-topic	Example of coding (from the original words of the interviewee)
C1: stimulating interest in learning and promoting active learning	B1: concentration	A2: I can concentrate more than before in class and work with other students to reorganize the theoretical and operational knowledge I had learned and apply it to my training.
	B2: review of knowledge and skills	A7: before the start of each class, we review the theoretical knowledge and operational steps related to the case in advance.
C2: collaborative learning to improve healthcare collaboration	B3: self-perception of role	A8: in this class, I learned what doctors and nurses should do, respectively, in a specific situation in the atmosphere of healthcare collaboration, and had a clearer understanding of the roles they assume.
	B4: leadership	A1: the course is team-based learning, in each class, I can gain, in addition to the case of relevant theories and operational skills more familiar, give me a great feeling is to recognize the power of team leadership, in the face of emergencies, the nurse in charge or attending doctors need to find the condition promptly and report to the superiors, then around the patient-centered team leader needs to be accurate, timely and make the right decision, only then a team can effectively organize and implement the resuscitation.
	B5: medical and nursing communication	A3: I learned some communication strategies in the class, for example, in the class on postpartum hemorrhage in normal labor, I learned how to use the SBAR communication model to report the patient's condition to the doctor effectively and accurately. I believe that the communication strategies I learned in the class will be very practical in my future clinical work.
	B6: situational awareness	A9: the high-fidelity simulation training session in each class is very tense, I sometimes forget what I am going to do next, but the team members will kindly give me some small reminders so that I can finish the operation smoothly. This shows that when working in a team, we not only need to do our job well but also improve our ability to monitor the situation.
C3: student-centeredness for clinical thinking skills development	B7: identifying and solving problems	A5: the guided feedback was an accomplished session in which I realized that I had many shortcomings, but the teacher and my classmates did not make fun of me, and at the same time, through a few explanations and pointers from the teacher, I was able to know what to do to correct my mistakes.
	B8: adaptability	A10: the complexity of the case scenarios and the progression of the disease in this course gave me a deeper and more systematic understanding of the disease itself as well as the difficulties of clinical work, and greatly enhanced my resilience so that I believe I won't be alarmed when I encounter situations similar to those in the cases in the future.
	B9: critical thinking	A6: this course has made me bold in expressing my ideas, honed my analytical skills, and improved my logic skills a lot.
C4: student suggestions for course optimization and improvement	B10: increase reflection time	A8: I think the teacher-guided reflection activity can make me better at identifying mistakes, but this session sometimes the teacher imparts a little too much knowledge and speaks a little too fast for me to keep up with the pace, so I hope I can increase the time for reflection and summarization.
	B11: rich case study	A4: I hope the instructor can design more emergencies or rare clinical cases and conduct more of these courses so that we can build a stronger foundation for entering the clinic.

their leadership abilities, communication skills, and ability to work collaboratively. These findings suggest that the curriculum effectively enhanced students' attitudes toward healthcare cooperation and their collaborative skills, which is consistent with previous research (15, 19). Effective communication and collaboration among healthcare professionals are essential for patient-centered care. However, healthcare professionals may have varying concerns when treating the same patients due to different specialties. Therefore, it is essential to foster teamwork awareness and skills among healthcare professionals. The institutional education stage plays a crucial role in cultivating mutual respect and cooperation among medical students from various disciplines. In this study, students were trained in a high-fidelity simulation through role-playing and group work. This allowed students to understand that nurses are not solely assistants to doctors and that healthcare professionals have equal importance in enhancing patient health outcomes. Additionally, students learned how to follow the process of division of labor among their team members and work collaboratively to complete practical training tasks. This teaching method can enhance students' attitudes toward healthcare collaboration and help them internalize the concept of interprofessionalism. This, in turn, can lead to effective collaboration in future clinical work (47).

Additionally, the study results revealed no noteworthy distinction in the students' scores regarding the dimension of "care vs. treatment." This outcome could be attributed to the students' regular education in professional knowledge and skills. They already comprehended that healthcare aims to provide quality services to patients. Consequently, they were able to offer physical and mental health education to patients while monitoring the effectiveness of treatment during nursing interventions.

High-fidelity simulation for healthcare collaboration can exercise students' critical thinking skills. Some studies have measured the level of students' critical thinking skills by using teachers' subjective evaluation, which was categorized as excellent, good and fair (48). Whereas many studies assessed students' critical thinking skills by means of a scale (49, 50), which is more objective. The study utilized the latter approach. The results indicated a positive increase in the total score of students' critical thinking skills scale after the curriculum was taught (303.86 ± 13.90) compared to before the teaching. Additionally, significant differences ($P < 0.05$) were observed in the scores for the five dimensions of finding the truth, analytical ability, systematic ability, self-confidence in critical thinking, and curiosity. The interview results revealed that students exhibited increased confidence in emergency handling and improvement in clinical thinking skills, such as problem identification and problem-solving, after the implementation of the curriculum.

The enhancement of students' critical thinking skills in this study can be attributed to the positive learning atmosphere created during the course. Through high-fidelity simulation training sessions, clinical and nursing students collaborated to complete tasks related to teaching cases. This allowed them to effectively provide treatment and care in clinical practice when faced with similar situations, improving their understanding of disease progression and routine management processes. During the review sessions, students had the opportunity to exchange and discuss ideas with teachers and classmates, express their opinions, and exercise their logical thinking and analytical abilities. Self-reflection helped students identify their own shortcomings, motivating them to address gaps in theoretical knowledge and operational skills in a timely manner.

Furthermore, the study revealed no notable distinction in scores regarding the aspects of open-mindedness and cognitive maturity. Two factors may affect students' perception of simulators vs. real patients: psychological differences and limited opportunities to integrate classroom learning with clinical practice due to lack of hospital internships. To improve integration, students should focus on developing medical and nursing communication skills as well as emergency resuscitation techniques. Their insight and psychological cognition may still be developing, and further observation is needed as they gain more experience.

Although the high-fidelity training course on healthcare cooperation has demonstrated a positive impact on students' attitudes, abilities in healthcare cooperation, and clinical thinking skills, there are several limitations to consider. Firstly, since this course is the first interprofessional course conducted at our university, there is room for improvement in terms of teaching faculty and their skills. Future efforts should focus on providing further training for faculty in interprofessional education and simulation teaching. Secondly, the sample size was relatively small, and only the initial effects of the course were tested. To objectively analyze the impact of the medical-nursing cooperation training course on students' performance, future studies should expand the sample size and establish control groups. Furthermore, to enhance the evaluation process, it may be beneficial to include a high-quality scale for assessing students' medical and nursing cooperation abilities and resilience. Thirdly, a comparative analysis of the attitudes toward healthcare cooperation between clinical medical students and nursing students was not conducted. Further exploration is needed to examine potential differences in attitudes toward healthcare cooperation between these two specialties.

5 Conclusion

In this study, we developed a high-fidelity simulation training course on healthcare collaboration based on the Fink model. We implemented the course and verified its teaching effectiveness. The course improved students' attitudes toward healthcare collaboration and enhanced their critical thinking abilities, promoting cross-fertilization of nursing disciplines and curriculum reform. This provides a reference for the development of healthcare collaboration education.

However, this study still has limitations: Firstly, since this course is the first interprofessional course conducted at our university, there is room for improvement in terms of teaching faculty and their skills. Future efforts should focus on providing further training for faculty in interprofessional education and simulation teaching. Secondly, the sample size was relatively small, and only the initial effects of the course were tested. In the future, as the course progresses, the sample size can be expanded, and control groups can be established to objectively analyze the impact of the medical-nursing collaboration training course on students' performance. Additionally, incorporating a high-quality scale to assess students' medical and nursing collaboration ability and resilience would further enhance the evaluation process. Thirdly, a comparative analysis of the attitudes toward healthcare collaboration between clinical medical students and nursing students was not conducted. Further exploration is needed to examine potential differences in attitudes toward healthcare collaboration between these two specialties.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by the Medical Ethics Committee of School of Nursing, Shandong University of Traditional Chinese Medicine. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

M-HJ: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review and editing. L-WD: Data curation, Formal Analysis, Writing – review and editing. BD: Data curation, Formal Analysis, Writing – original draft, Writing – review and editing. MZ: Methodology, Writing – original draft. Y-PL: Data curation, Writing – original draft. C-XL: Funding acquisition, Writing – review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the 2022 Education and Teaching Research Project (Experimental Teaching Special Project) of Shandong University of Traditional Chinese Medicine (SYJX2022028).

Acknowledgments

We thank all students and other faculties for their generous contribution to this research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Frenk J, Chen L, Bhutta ZA, Cohen J, Crisp N, Evans T, et al. Health professionals for a new century: Transforming education to strengthen health systems in an interdependent world. *Lancet.* (2010) 376:1923–58. doi: 10.1016/S0140-6736(10)61854-5

2. Thistlethwaite J. Interprofessional education: A review of context, learning and the research agenda. *Med Educ.* (2012) 46:58–70. doi: 10.1111/j.1365-2923.2011.04143.x

3. Gilbert JH, Yan J, Hoffman SJ. A WHO report: Framework for action on interprofessional education and collaborative practice. *J Allied Health.* (2010) 39(Suppl. 1):196–7.

4. Defenbaugh N, Chikotas NE. The outcome of interprofessional education: Integrating communication studies into a standardized patient experience for advanced practice nursing students. *Nurse Educ Pract.* (2016) 16:176–81. doi: 10.1016/j.nepr.2015.06.003

5. Feather RA, Carr DE, Garletts DM, Reising DL. Nursing and medical students teaming up: Results of an interprofessional project. *J Interprof Care.* (2017) 31:661–3. doi: 10.1080/13561820.2017.1322563

6. Oxelmark L, Nordahl Amorøe T, Carlzon L, Rystedt H. Students' understanding of teamwork and professional roles after interprofessional simulation-a qualitative analysis. *Adv Simul (Lond).* (2017) 2:8. doi: 10.1186/s41077-017-0041-6

7. Jakobsen RB, Gran SF, Grimsmo B, Arntzen K, Fosse E, Frich JC, et al. Examining participant perceptions of an interprofessional simulation-based trauma team training for medical and nursing students. *J Interprof Care.* (2018) 32:80–8. doi: 10.1080/13561820.2017.1376625

8. Lau Y, Chee DGH, Ab Hamid ZB, Leong SH, Lau ST. Interprofessional simulation–based advanced cardiac life support training: Video-based observational study. *Clin Simul Nurs.* (2019) 30:16–24. doi: 10.1016/j.ecns.2019.03.001

9. Cui PP, Bie WQ, Wang PP, Chen CY. Application of physician-nurse collaboration simulation-based learning in nursing education. *Chin Nurs Manag.* (2018) 18:922–7. doi: 10.3969/j.issn.1672-1756.2018.07.013

10. Wang RM, Shi NK, Zhao Y. Application of medical-nursing students collaboration combined with scenario simulation teaching in the operating room nursing. *Chin J Nurs.* (2015) 50:336–9. doi: 10.3761/j.issn.0254-1769.2015.03.020

11. Liu Q, Ouyang YQ, Li SY, Xu J, Li L, Xu AJ, et al. Application of interprofessional simulation teaching in the course of nursing planning and implementation. *J Nurs Sci.* (2020) 35:69. doi: 10.3870/j.issn.1001-4152.2020.15.069

12. Hu FQ, Liu FP, Liu DM, Hu TT, Qin HZ. Study on the teaching model of medical-nursing cooperation high simulation first aid nursing training. *J Anhui Health Vocat Techn Coll.* (2019) 18:102–4.

13. Li Y, Yang YF. Collection of clinical medical scenario simulation teaching plans. *Beijing Peoples Health Publish House.* (2022) 2022:1–8.

14. Li HW, Cui YN, Li YF, Wang XQ, Liu LJ, Li ZD, et al. Application of interprofessional collaboration in emergency simulation skill training for undergraduates. *Chin J Nurs Educ.* (2020) 17:785–9. doi: 10.3761/j.issn.1672-9234.2020.09.003

15. Zhuang QL, Wang T, Ye JF. Application of interprofessional simulation learning in the course of comprehensive nursing skill training. *J Nurs Sci.* (2021) 36:83–8. doi: 10.3870/j.issn.1001-4152.2021.10.083

16. Fitzgerald L, Wong P, Hannon J, Solberg Tokerud M, Lyons J. Curriculum learning designs: teaching health assessment skills for advanced nursing practitioners through sustainable flexible learning. *Nurse Educ Today.* (2013) 33:1230–6. doi: 10.1016/j.nedt.2012.05.029

17. Kim HS. Outcomes-based curriculum development and student evaluation in nursing education. *J Korean Acad Nurs.* (2012) 42:917–27. doi: 10.4040/jkan.2012.42.7.917

18. Fink LD. *Creating significant learning experiences: An integrated approach to designing college courses.* 2nd ed. San Francisco, CA: Jossey-Bass (2013). p. 3–27.

19. Marrocco GF. Fostering significant learning in graduate nursing education. *J Nurs Educ.* (2014) 53:177–9. doi: 10.3928/01484834-20140223-02

20. Uribe Cantalejo JC, Pardo MI. Fink's integrated course design and taxonomy: The impact of their use in a "basics of dental anatomy" course. *J Dent Educ.* (2020) 84:964–73. doi: 10.1002/jdd.12183

21. Krueger KP, Russell MA, Bischoff J. A health policy course based on Fink's taxonomy of significant learning. *Am J Pharm Educ.* (2011) 75:14. doi: 10.5688/ajpe75114

22. Yu HR, Liu L, Zhang J, Shen J, Jiang AL. Development and application of the humanistic course "narrative nursing". *Nurs J Chin Peoples Liber Army.* (2018) 35:18–22. doi: 10.3969/j.issn.1008-9993.2018.22.004

23. General Office of the State Council. Guiding opinions of the general office of the state council on accelerating the innovative development of medical education. *Gazette State Council Peoples Republic China.* (2020) 28:27–31.

24. Jiang MH, Shi XP, Zhang M, Zhao RW, Lin CX, Li YP, et al. Development of teaching cases of high simulation training of physician-nurse cooperation in obstetrics and gynecology nursing. *Chin J Nurs Educ.* (2023) 20:146–50. doi: 10.3761/j.issn.1672-9234.2023.02.003

25. Branzetti J, Gisondi MA, Hopson LR, Regan L. Aiming beyond competent: The application of the taxonomy of significant learning to medical education. *Teach Learn Med.* (2019) 31:466–78. doi: 10.1080/10401334.2018.1561368

26. Gross Forneris S, Fey MK. Critical conversations: The NLN Guide for teaching thinking. *Nurs Educ Perspect.* (2016) 37:248–9. doi: 10.1097/01.NEP.0000000000000069

27. Yuan HB, Williams BA, Fang JB, Ye QH. A systematic review of selected evidence on improving knowledge and skills through high-fidelity simulation. *Nurse Educ Today.* (2012) 32:294–8. doi: 10.1016/j.nedt.2011.07.010

28. Shinnick MA, Woo MA. The effect of human patient simulation on critical thinking and its predictors in prelicensure nursing students. *Nurse Educ Today.* (2013) 33:1062–7. doi: 10.1016/j.nedt.2012.04.004

29. Wang ZP, Wang J, Li Y, Chen ZQ, Yue SJ, Su CX. The progress of high fidelity simulation applied in nursing courses. *Chin J Nurs Educ.* (2018) 15:698–702. doi: 10.3761/j.issn.1672-9234.2018.09.013

30. Jeffries PR. A framework for designing, implementing, and evaluating simulations used as teaching strategies in nursing. *Nurs Educ Perspect.* (2005) 26:96–103.

31. Jeffries PR, Rogers KJ. *Using simulations in Nursing Education: From conceptualization to evaluation*. New York, NY: The National League for Nursing (2007). p. 1–10.

32. Wang AL, Fitzpatrick JJ, Petrini MA. Use of simulation among Chinese nursing students. *Clin Simul Nurs.* (2013) 9:e311–7. doi: 10.1016/j.ecns.2012.03.004

33. Hojat M, Nasca TJ, Cohen MJ, Fields SK, Rattner SL, Griffiths M, et al. Attitudes toward physician-nurse collaboration: A cross-cultural study of male and female physicians and nurses in the United States and Mexico. *Nurs Res.* (2001) 50:123–8. doi: 10.1097/00006199-200103000-00008

34. Yang XL, Lv HY, Li SG. Comparison of attitudes of physicians and nurses toward physician-nurse collaboration. *Chin J Nurs.* (2006) 41:466–9.

35. Hojat M, Fields SK, Veloski JJ, Griffiths M, Cohen MJ, Plumb JD. Psychometric properties of an attitude scale measuring physician-nurse collaboration. *Eval Health Prof.* (1999) 22:208–20. doi: 10.1177/01632789922034275

36. Peng MC, Wang GC, Chen JL, Chen MH, Bai HH, Li SG, et al. Validity and reliability of the Chinese critical thinking disposition inventory. *Chin J Nurs.* (2004) 39:644–7. doi: 10.1016/s0020-7489(01)00019-0

37. Ma C, Zhou W. Effects of unfolding case-based learning on academic achievement, critical thinking, and self-confidence in undergraduate nursing students learning health assessment skills. *Nurse Educ Pract.* (2022) 60:103321. doi: 10.1016/j.nepr.2022.103321

38. Peters K, Halcomb E. Interviews in qualitative research. *Nurse Res.* (2015) 22:6–7. doi: 10.7748/nr.22.4.6.s2

39. Michalik B, Kulbat M, Domagała A. Factors affecting young doctors' choice of medical specialty-a qualitative study. *PLoS One.* (2024) 19:e0297927. doi: 10.1371/journal.pone.0297927

40. Xia WY, Wang TL, Shan SX, Yang XY. A qualitative study of patients' expectations of needling based on focus group interview methodology. *J Chin Med.* (2023) 64:992–8. doi: 10.13288/j.11-2166/r.2023.10.005

41. Chickering AW, Gamson ZF. Seven principles for good practice in undergraduate education. *Biochem Educ.* (1989) 17:140–1.

42. Liu Q, Yang BX, Yu SH. Application of simulated teaching part replacing clinical internship in undergraduate nursing teaching. *Chin Nurs Res.* (2015) 29:2513–5. doi: 10.3969/j.issn.1009-6493.2015.20.030

43. Wang H, Peng XH, Wang J, Luo YY, Jiang X. Application of case design combined with scenario simulation exercise in teaching critical care nursing for interns. *Health Vocat Educ.* (2021) 39:91–3.

44. Lyons K, McLaughlin JE, Khanova J, Roth MT. Cognitive apprenticeship in health sciences education: A qualitative review. *Adv Health Sci Educ Theory Pract.* (2017) 22:723–39. doi: 10.1007/s10459-016-9707-4

45. Qin F, He XF, Shi L, Zhang YW, Fang YX. Application of companion pilot guided feedback in high simulation scenario simulation teaching. *J Nurs Sci.* (2023) 38:68–71+75. doi: 10.3870/j.issn.1001-4152.2023.24.068

46. Liu XL, Ni XL, Chen J. The attitudes toward physician-nurse collaboration among medical students and nursing students after clinical practice. *Chin J Nurs.* (2013) 48:701–3. doi: 10.3761/j.issn.0254-1769.2013.08.009

47. Homeyer S, Hoffmann W, Hingst P, Oppermann RF, Dreier-Wolfgramm A. Effects of interprofessional education for medical and nursing students: Enablers, barriers and expectations for optimizing future interprofessional collaboration – a qualitative study. *BMC Nurs.* (2018) 17:13. doi: 10.1186/s12912-018-0279-x

48. Wang YH, Cao Y. Research on the application of situational simulation based on interprofessional cooperation in surgical nursing teaching. *Theory Pract Innov Enterpreneurship.* (2020) 21:113–5.

49. Zhou ZX, Li F, Liu Y, Zhu H, Xia LP. The impact of scenario based case-based scenario simulation teaching on the critical thinking ability of nursing students. *Modern Med Hyg.* (2018) 34:1748–50. doi: 10.3969/j.issn.1009-5519.2018.11.051

50. Deng FF, Deng H. The influence of high simulation teaching method on the competence and critical thinking ability of nursing students in higher vocational college. *Chin Nurs Res.* (2017) 31:4198–201. doi: 10.3969/j.issn.1009-6493.2017.33.007

TYPE Original Research
PUBLISHED 09 April 2024
DOI 10.3389/fphys.2024.1362185

Check for updates

OPEN ACCESS

EDITED BY
Kuanquan Wang,
Harbin Institute of Technology, China

REVIEWED BY
Laura Burattini,
Marche Polytechnic University, Italy
Dongping Du,
Texas Tech University, United States
Muhammad Adib Uz Zaman,
University of Cincinnati, United States
Yang Liu,
Baidu, China

*CORRESPONDENCE
Bing Yao,
✉ byao3@utk.edu

RECEIVED 27 December 2023
ACCEPTED 25 March 2024
PUBLISHED 09 April 2024

CITATION
Xie J, Stavrakis S and Yao B (2024), Automated identification of atrial fibrillation from single-lead ECGs using multi-branching ResNet.
Front. Physiol. 15:1362185.
doi: 10.3389/fphys.2024.1362185

Automated identification of atrial fibrillation from single-lead ECGs using multi-branching ResNet

Jianxin Xie[1], Stavros Stavrakis[2] and Bing Yao[3]*

[1]School of Data Science, University of Virginia, Charlottesville, VA, United States, [2]Health Sciences Center, University of Oklahoma, Oklahoma City, OK, United States, [3]Department of Industrial and Systems Engineering, University of Tennessee at Knoxville, Knoxville, TN, United States

Introduction: Atrial fibrillation (AF) is the most common cardiac arrhythmia, which is clinically identified with irregular and rapid heartbeat rhythm. AF puts a patient at risk of forming blood clots, which can eventually lead to heart failure, stroke, or even sudden death. Electrocardiography (ECG), which involves acquiring bioelectrical signals from the body surface to reflect heart activity, is a standard procedure for detecting AF. However, the occurrence of AF is often intermittent, costing a significant amount of time and effort from medical doctors to identify AF episodes. Moreover, human error is inevitable, as even experienced medical professionals can overlook or misinterpret subtle signs of AF. As such, it is of critical importance to develop an advanced analytical model that can automatically interpret ECG signals and provide decision support for AF diagnostics.

Methods: In this paper, we propose an innovative deep-learning method for automated AF identification using single-lead ECGs. We first extract time-frequency features from ECG signals using continuous wavelet transform (CWT). Second, the convolutional neural networks enhanced with residual learning (ReNet) are employed as the functional approximator to interpret the time-frequency features extracted by CWT. Third, we propose to incorporate a multi-branching structure into the ResNet to address the issue of class imbalance, where normal ECGs significantly outnumber instances of AF in ECG datasets.

Results and Discussion: We evaluate the proposed Multi-branching Resnet with CWT (CWT-MB-Resnet) with two ECG datasets, i.e., PhysioNet/CinC challenge 2017 and ECGs obtained from the University of Oklahoma Health Sciences Center (OUHSC). The proposed CWT-MB-Resnet demonstrates robust prediction performance, achieving an F1 score of 0.8865 for the PhysioNet dataset and 0.7369 for the OUHSC dataset. The experimental results signify the model's superior capability in balancing precision and recall, which is a desired attribute for ensuring reliable medical diagnoses.

KEYWORDS

convolutional neural network, residual network, wavelet transform, multi-branching outputs, ECG signal analysis, imbalanced data, atrial fibrillation

1 Introduction

Cardiovascular diseases have been the leading cause of mortality globally. The World Health Organization (WHO) states that about 17.9 million people perish due to cardiovascular disease each year (World Health Organization, 2024), contributing 32% to the worldwide death toll (University of Washington, 2024). Atrial fibrillation (AF) is the most common cardiac arrhythmia caused by uncoordinated electrical activities in the atria (Nesheiwat et al., 2023). Although AF itself does not lead to a lethal condition, it will substantially increase the risk of catastrophic diseases such as heart failure, stroke, and sudden death (Lubitz et al., 2013; Bernstein et al., 2021). The prevalence of AF plagues over 2.7 million people in the United States, and this number is estimated to rise to 12.1 million in 2030, as the population ages (Colilla et al., 2013). In healthcare practice, the electrocardiogram (ECG) is a cost-effective and noninvasive medical approach to record the electrical signals on the body surface as a reflection of cardiac health conditions (Yao and Yang, 2016; Yao and Yang, 2020; Yao et al., 2021; Xie and Yao, 2023).

Historically, the utilization of ECG for cardiac monitoring has been substantially constrained by the need for expensive equipment and the involvement of specialized medical doctors to interpret complex ECG recordings. However, recent advancements in portable ECG sensors, such as the AliveCor (aliveCor, 2024), AD8232 (Analog Devices, 2024), and consumer-grade devices like the smartwatch (Isakadze and Martin, 2020), have revolutionized the way to detect heart abnormalities. These portable devices now enable the capture of high-fidelity ECG signals outside of traditional clinical settings. While multi-lead ECGs provide comprehensive cardiac activity information, single-lead ECGs make cardiac monitoring more accessible and less obtrusive for long-term rhythm surveillance or frequent measurements (Abdou and Krishnan, 2022). This is especially valuable in ambulatory settings, home monitoring, and situations where rapid and non-invasive monitoring is desired. Single-lead ECGs offer a simplified yet effective method for the early detection of AF and other cardiac anomalies (Boriani et al., 2021).

In conjunction with advanced sensing technologies, there has been a parallel development in machine learning methodologies. Given the prevalence of AF, a significant number of machine learning models have been developed specifically for the task of distinguishing AF from normal heart rhythms. Traditional machine learning models focus on extracting morphological features and heart rate variability from ECG signals to detect AF, which depends heavily on manual feature engineering (Ye et al., 2012; Da Silva-Filarder and Marzbanrad, 2017; Athif et al., 2018). Deep Neural Network (DNN), which does not require explicit feature engineering, is another powerful tool that has achieved promising results in data-driven disease detection. Various DNN-based models such as convolutional and recurrent neural networks (i.e., CNNs, RNNs) have been designed for AF detection and outperformed conventional machine learning methods (Andreotti et al., 2017; Schwab et al., 2017; Gao et al., 2021). Despite the performance improvement achieved by DNNs in detecting AF with single-lead ECG, there remains potential for further prediction enhancements. Four major challenges remain to be tackled: 1) ECG recordings collected from clinics are often in Protable Document Format (PDF). An effective preprocessing procedure is needed to retrieve digital ECG signals from the PDFs before being fed to the machine learning models. 2) ECG signals are generally composed of a wide spectrum of frequency components. DNN models built upon raw ECG time series may not fully exploit the time-frequency information inherent in the signals. 3) Note that the learning capacity for a DNN often increases when the network goes deeper. However, the deeper structure can result in gradient dissipation problems, leading to unsatisfactory prediction performance. 4) Data-driven AF detection also suffers from the common issue of imbalanced data in machine learning (e.g., AF samples are much less compared to normal ECGs). The classifier directly built from the imbalanced data will generate biased and inaccurate predictions.

In this paper, we develop an automatic AF detector based on continuous wavelet transform (CWT) and 18-layer Residual Neural Network (ResNet18) with a multi-branching structure (CWT-MB-ResNet). We first develop a preprocessing procedure to extract ECG signals from ECG PDFs and leverage the CWT to transform the extracted signals into the time-frequency domain. Second, ResNet18 is engaged to alleviate the gradient dissipation problem in deep-structured networks, allowing it to learn deeper features from 2D time-frequency images and achieve better performance. Finally, we propose to incorporate a multi-branching output structure adapted from our prior work (Wang and Yao, 2021) into the ResNet to deal with the issue induced by the imbalanced dataset in AF identification. The multi-branching technique exempts artificial data augmentation and does not require any preassumptions in solving the imbalanced data issue. The performance of the proposed framework is evaluated by two real-world datasets: PhysioNet/CinC challenge 2017 (Goldberger et al., 2000; Clifford et al., 2017) and ECG data obtained from the University of Oklahoma Health Sciences Center (OUHSC). Experimental results show that our CWT-MB-ResNet significantly outperforms existing methods commonly used in current practice.

The rest of this paper is organized as follows: Section 2 presents the literature review of existing data-driven methods for AF detection. Section 3 introduces the data processing details and the proposed prediction method. Section 4 shows the experimental results in AF identification. Section 6 concludes the present investigation.

2 Research background

Traditional machine learning approaches focus on the extraction of ECG morphological features (De Chazal et al., 2004) and heart rate variability information (Park et al., 2009) to identify AF conditions. Those methods are mostly in light of two aspects of AF-altered ECG characteristics: 1) the absence of distinct P waves, which are replaced by irregular fibrillatory waves or F waves as oscillations in low amplitude around the baseline (Ladavich and Ghoraani, 2015); 2) irregular R-R intervals (Oster and Clifford, 2015). Multiple feature-based automation techniques have been proposed to classify AF-altered ECGs, such as linear discriminant analysis (De Chazal et al., 2004), support vector machine (Billeci et al., 2017; Islam et al., 2017), independent component analysis (Ye et al., 2012). When there exists a high level of noise or faulty detection, the performance of feature-extraction methods that solely study the P wave deteriorates significantly due to the chaotic signal baseline introduced by the noise (Larburu et al., 2011). Most R-R interval-based methods (Tateno and Glass, 2001; Lian et al., 2011) usually require long ECG segments to detect AF episodes, and become ineffective when it comes to short ECG signals (less than 60s) or in the presence of significant sinus arrhythmia or frequent premature atrial contractions (Xia et al., 2018). Moreover, traditional methods require a separate feature extraction process before feeding the data into the classifier, as well as manually establishing the detection rules and threshold. This can be computationally expensive and may not generalize well when applied to a larger population.

In the past few decades, deep learning or deep neural network (DNN) has emerged as a powerful tool for pattern recognition that can learn the abstracted features from complex data and yield state-of-the-art predictions (Mousavi et al., 2019; Xie and Yao, 2022a; Xie and Yao, 2022b; Chen et al., 2022; Wang et al., 2022). As opposed to traditional machine learning, deep learning presents strong robustness and fault tolerance to uncertain factors, which makes it suitable for beat and rhythm classification from ECGs (Tutuko et al., 2021). Moreover, existing research has indicated that deep learning methods demonstrate more efficient and more potent predictive power than classical machine learning methods for AF identification (Cai et al., 2020; Murat et al., 2021). There has been a significant surge in leveraging deep learning for AF detection using single-lead ECGs, showing promising potential in enhancing diagnostic accuracy. We summarized four commonly used network structures in discerning AF samples using single-lead ECGs:

1) **Convolutional neural networks (CNNs):** CNNs, specifically 1-dimensional CNNs (1D-CNNs), have been widely applied to extracting hierarchical features from ECG data for distinguishing AF from normal heart rhythms (Andreotti et al., 2017; Fan et al., 2018; Lai et al., 2019; Phukan et al., 2023). For example, Andreotti et al. Andreotti et al. (2017) balanced the PhysioNet/CinC 2017 dataset by augmenting AF samples from various sources to address the class imbalance issue. They employed a ResNet model with 34 convolutional layers for AF detection, achieving a final F1 score of 0.79. Lai et al. Lai et al. (2019) developed a streamlined two-stream CNN with each stream containing only 8 layers. This model achieved a sensitivity of 89.5% and a specificity of 82.7% on the PhysioBank dataset (PhysioBank, 2000). The extracted cardiac rhythm features, specifically RR intervals and F-wave frequency spectra, served as dual inputs for the neural network. Similarly, Fan et al. Fan et al. (2018) developed a multi-scaled two-stream network with different filter sizes at each stream to capture features of different scales using single-lead ECGs from PhysioNet/Cinc 2017, achieving an F1 score of 0.8355. Phukan et al. Phukan et al. (2023) did a systematic experiment on selections of filter size, number of layers, and activation function on multiple standard datasets. They concluded that the best 5-layer CNN with activation function of exponential linear unit and kernel size 4 × 1 provides the highest accuracy of 99.84% for 5s ECG segments.
2) **Recurrent Neural Networks (RNNs):** An RNN is a type of neural network designed to effectively process sequential data by maintaining a memory of previous inputs, making it suitable for classifying time-series signals, e.g., AF detection. For example, Schwab et al. Schwab et al. (2017) built an ensemble of RNNs to jointly distinguish AF from normal ECGs, resulting in 0.79 of F1 score on the PhysioNet/Cinc 2017 dataset. Faust et al. Faust et al. (2018) utilized RNNs, specifically the long short-term memory (LSTM) architecture, to analyze ECGs from the MIT-BIH Atrial Fibrillation Database, achieving an accuracy rate of 99.77% for AF detection. Wang et al. Wang et al. (2023a) proposed a dual-path RNN which includes the intra- and inter-RNN modules to study the global and local aspects for end-to-end AF recognition. They used the PhysioNet/Cinc 2017 dataset to validate their model and achieved an F1 score of 0.842. More recently, bidirectional long short-term memory (Bi-LSTM), a type of RNN architecture capable of capturing both past and future context in sequential data, has been used to discern AF. Ramkumar et al. Ramkumar et al. (2022) created an auto-encoder and Bi-LSTM-based network to detect AF among others. This method integrated a reconstruction error from the auto-encoder into the total loss function, leading to a sensitivity of 92% and specificity of 97% on the PhysioNet/Cinc 2017 dataset.
3) **CNN-RNNs:** CNN-RNN hybrids combine the morphological feature extraction capabilities of 1D-CNNs with the temporal pattern recognition strengths of RNNs to address complex tasks such as AF detection from ECG signals. For example, Limam et al. Limam and Precioso, (2017) used dual CNNs to process the inputs consisting of both ECGs and heart rates independently, and then the processed features were merged into RNN to learn the temporal patterns, achieving a validated F1 score of 0.856 on the PhysioNet/CinC 2017 dataset. Wang et al. Wang and Li, (2020) combined CNN with Bi-LSTM, exploring two concatenation strategies: a parallel concatenation of CNN and Bi-LSTM, and a sequential one where the CNN output feeds into the Bi-LSTM. They evaluated the methods on the MIT-BIH dataset, reporting a final F1 score of 0.82 for the sequential strategy. Zhang et al. developed a model that merges a multi-branch CNN (MCNN) with Bi-LSTM to improve AF detection from short ECG recordings (Zhang et al., 2022). Unlike our multi-branching approach for addressing the imbalanced data issue, their model extracted features from various segments of a single-lead ECG, which were then processed by the Bi-LSTM. They tested the model on the PhysioNet/CinC 2017 dataset, achieving an F1 score of 0.7894.
4) **Attention-based networks:** The attention mechanism (Bahdanau et al., 2014; Vaswani et al., 2017) in deep learning dynamically weighs the importance of different input features, allowing models to focus more on relevant data while processing a task. This special capability can facilitate pattern recognition in ECG signals, enhancing the accuracy and efficiency of AF detection. For example, Gao et al. Gao et al. (2021) designed a residual-based temporal attention CNN, generating temporal informative features related to AF, so as to consider the semantic information to achieve better performance. This model achieved an accuracy of 85.43% on the PhysioNet/CinC 2017 dataset. Nankani et al. Nankani and Baruah, (2022) investigated the transformer network for AF detection and underscored clinically relevant signal timestamps triggering the diagnosis, achieving an F1 score of 0.87 on the PhysioNet/Cinc 2017 dataset. Rohr et al.

Rohr et al. (2022) explored and assessed two advanced models for AF detection: a transformer-based DualNet architecture and a CNN-LSTM hybrid model, achieving F1 scores of 0.9127 and 0.9072, respectively, on the PhysioNet/CinC 2017 dataset.

As highlighted above (Andreotti et al., 2017; Fan et al., 2018; Lai et al., 2019; Phukan et al., 2023), 1D-CNNs have exhibited their effectiveness in identifying morphological features and comprehending temporal variations in time series data, demonstrating superior capability in AF detection using single-lead ECG signals. However, despite the promising utility of 1D-CNNs in time series analysis, comparative studies in the literature Ullah et al. (2021) and Wu et al. (2018) indicate that 1D-CNNs often yield lower prediction accuracies than their 2D counterparts under similar network configurations for ECG classification tasks.

This discrepancy can be attributed to the richer, more comprehensive information encapsulated in 2D input data, coupled with the inherently superior capacity of 2D CNNs for feature extraction and interpretation.

Owing to the outstanding performance and strong ability in pattern recognition, 2D CNN has been explored for ECG classification by virtue of its capacity to smartly suppress measurement noises and extract pertinent feature maps using convolutional and pooling layers (Huang et al., 2019). For example, Izci et al. Izci et al. (2019) engaged a 2D CNN model to investigate ECG signals for arrhythmia identification. They segmented the ECG signals by heartbeats and directly converted each heartbeat into grayscale images, which served as the input of the 2D CNN model. Similarly, Jun et al. Jun et al. (2018) proposed to combine 2D CNN and data augmentation with different image cropping techniques to classify 2D grayscale images of ECG beats. However, these end-to-end 2D CNNs are directly fed with original ECG beat segments without considering the possible noise contamination. Moreover, the 2D input data were created by directly plotting each ECG beat as a grayscale image with unavoided redundant information residing in the image background. This procedure requires extra storage space for training data and increases the computational burden without extracting relative features inherent in the ECG beats.

ECG signals generally consist of various frequency components, which can be used to identify disease-altered cardiac conditions. Wavelet transform (WT) (Daubechies, 1990; Yao et al., 2017; van Wyk et al., 2019) has been proven to be a useful technique for extracting critical time-frequency information pertinent to disease-altered ECG patterns (Kutlu and Kuntalp, 2012; He et al., 2018). As such, WT is favored as a feature-preprocessing procedure that converts 1D ECG signals into 2D images containing time-frequency features. The resulting 2D feature images then serve as the input of CNNs for ECG classification instead of the original 2D ECG plots. For instance, Xia et al. Xia et al. (2018) engaged the short-term Fourier transform (STFT) and stationary wavelet transform to convert ECG segments into 2D matrices which were then fed into a three-layer CNN for AF detection. Wang et al. Wang et al. (2021) combined the time-frequency features extracted by Continuous Wavelet Transform (CWT) and R-interval features to train a 2D CNN model for ECG signal classification. Wu et al. Wu et al. (2019) built a 2D CNN based on time-frequency features of short-time single-lead ECGs extracted from three methods, i.e., STFT, CWT, and pseudo Wigner-Ville distribution, to detect arrhythmias. Huang et al. Huang et al. (2019) developed an ECG classification model by transforming ECG signals into time-frequency spectrograms using STFT and feeding them into a three-layer 2D CNN. Li et al. Li et al. (2019) included three different types of wavelet transform (i.e., Morlet wavelet, Paul wavelet, Gaussian Derivative) to create 2D time-frequency images as the input data to the 2D CNN-based ECG classifier. The above literature unequivocally demonstrates that incorporating frequency information through the WT can significantly enhance the efficacy of ECG classification, underscoring the vital role of frequency domain analysis in AF identification.

In addition to effective information extraction from ECG time series, the realization of the full data potential is heavily reliant on advanced analytical models. Although the abovementioned works have validated the superiority of 2D CNN-based approaches, the shallow network structures with a limited number of layers can potentially hinder the extraction of deeper features. Naturally, the capacity for a neural network to learn is enhanced by an increase in the number of layers. However, having a deeper network structure can result in a gradient dissipation problem, which impedes convergence during network training, leading to suboptimal prediction performance. To cope with this issue, the residual neural network (ResNet) has been developed with an important modification, i.e., identity mapping, induced by the skip connection technique (He et al., 2016), which has wide applications in classifying the ECG signals. For example, Jing et al. Jing et al. (2021) developed an improved ResNet with 18 layers for single heartbeat classification. Park et al. Park et al. (2022) used a squeeze-and-excitation ResNet with 152 layers and compared the model performance trained by ECGs from a 12-lead ECG system and single-lead ECG data. Guan et al. Guan et al. (2022) proposed a hidden attention ResNet to capture the deep spatiotemporal features using 2D images converted from ECG signals.

Automated ECG classification also suffers from the long-standing issue of imbalanced data in machine learning. Diverse sampling and synthetic strategies have been proposed to address the imbalanced data issue, which focuses on creating a balanced training dataset out from the original imbalanced data to mitigatethe potential bias introduced by imbalanced data distribution during model training (He and Garcia, 2009). Frequently employed techniques consist of random over-sampling and under-sampling, informed adaptive undersampling, and synthetic minority over-sampling technique (SMOTE) (Gao et al., 2019; Wang and Yao, 2021; Qiu et al., 2022). For example, Luo et al. Luo et al. (2021) engaged SMOTE to synthesize minority samples and create a balanced training dataset for automated arrhythmia classification. Ramaraj et al. Ramaraj and Clement Virgeniya, (2021) incorporated an adaptive synthetic sampling process into the training of deep learning models built with gated recurrent units to address the class imbalance problem for ECG pattern recognition. Nurmaini et al. Nurmaini et al. (2020) compared sampling schemes of SMOTE and random oversampling with RNN and concluded that the balanced dataset created by SMOTE significantly improved the classification performance. In addition to fabricating balanced ECG datasets, Gao et al. Gao et al. (2019) and Petmezas et al. Petmezas et al. (2021) proposed to engage dynamically-scaled focal loss function to suppress the weight of loss corresponding to the majority class, so that their contribution to the total loss is reduced to alleviate the class imbalance problem. However, this method requires the preassumption of a focusing parameter to modulate the effect of the majority class on the total loss. Existing methods mainly focus on using sampling and synthetic strategies or modifying the loss function, little has been done to create new network structures without making extra assumptions and feature engineering to cope with the imbalanced data issue in AF identification from ECG signals.

3 Materials and methods

3.1 Dataset

In this study, two AF databases from different sources, i.e., ECG recordings from PhysioNet/CinC challenge 2017 (Goldberger et al., 2000; Clifford et al., 2017) and ECG PDFs from OUHSC, are used to evaluate the performance of data-driven detection methods. Both databases consist of short single-lead ECG recordings for AF and non-AF patients. PhysioNet/CinC Challenge 2017 is an open database including 8,528 single-lead ECG signals and their annotations. Among them, 5050 ECG recordings are labeled as normal sinus rhythm while 738 signals are annotated as AF. The sampling frequency of recordings is 300 Hz and the duration of ECG signals varies from 9s to 30s. The OUHSC database contains ECG signals in PDF format with 33 recordings from AF subjects and 227 normal samples, which are annotated by cardiologists from OUHSC. Each recording has a duration of around 30s with a sampling frequency of 60 Hz.

3.2 ECG signal preprocessing

Note that the original ECG recordings from OUHSC are in PDF format, as shown in Figure 1A. It is necessary to accurately extract the numerical ECG readings from the PDF files for further data preprocessing and analysis, which is achieved by the following procedure:

- *Transforming PDF files into gray-scale images represented by 2D-pixel matrices*: We discretize the 2D image into a pixel matrix. Then, each pixel is converted to a fixed number of bits to represent the gray-scale intensity of the corresponding point in the image. As shown in Figure 1A, the ECG signals are displayed in the darkest color on the plot with the color intensity of 1, i.e., $h(m,n) = 1$, while the grid lines appear in a lighter color, i.e., $0 < h(m,n) < 1$, where $h(m,n)$ denotes the color intensity of the pixel at column m and row n. Note that the background color intensity is 0.
- *Removing grid lines from the ECG plot*: We replace the pixel shade values of the grid lines with the background color value: i.e., $h(m,n \mid h(m,n) < 1) = 0$. This allows the ECG signals to distinguishably stand out, as illustrated in Figure 1B. The quantized image is thus encoded into a binary digital format, i.e., black as "1" and white as "0". As such, the entire ECG image is transformed into a binary digital matrix without the grid lines.
- *Extracting the digital ECG time series*: The positions of black pixels (i.e., ECG signal) in the binary matrix are further extracted, which are represented as a set of (m,n) pairs:

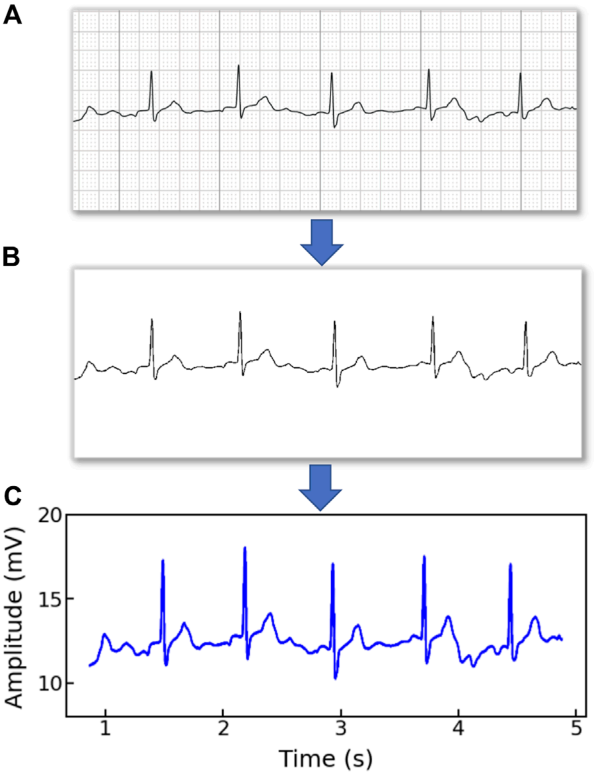


FIGURE 1
An example of **(A)** a raw image recording of an ECG segment in PDF format, **(B)** the ECG image that filters out the grid background, **(C)** the digitalized ECG time series signal.

$$S = \{(m,n) \,|\, h(m,n) = 1\}$$

The resulting S is then used to reconstruct the digital ECG time series, where m stands for the time course, and n corresponds to the magnitude of the ECG signal. As such, we are able to extract the ECG recordings from the PDFs to digitalized ECG time series signals (Figure 1C), which will be used for further processing and model training.

Raw ECG recordings are often contaminated by noises, such as baseline wandering, electromyography disturbance, and power-line interference (Mian Qaisar, 2020), which will negatively impact the information extraction and model performance. In this work, we engage BioSPPy, a toolbox for biosignal processing written in Python, for ECG signal denoising. The BioSPPy library provides comprehensive functions for processing ECG signals including functions for importing ECGs, filtering out interfering components, and correcting baseline wandering (PIA-Group, 2021). Specifically, after loading the ECG data, we apply a high-pass filter to remove the low-frequency noise (e.g., baseline wandering), a notch filter to remove power-line interference, and a low-pass filter to filter out the high-frequency noise.

3.3 Continuous wavelet transform

ECG signals encompass multiple feature components in both the time and frequency domains. In this study, we engage the continuous wavelet transform (CWT) to extract time-frequency features from ECGs due to its excellent performance in the analysis of transient and non-stationary time series signals (Keissar et al., 2009). CWT is the most popular tool for time-frequency analysis that reflects the frequency components of data changing with time. CWT is verified to outperform the traditional STFT due to its ability to provide multi-resolution decompositions of the signal, which allows for a trade-off between time and frequency resolution, i.e., higher frequency resolution for signals with sharp transients and higher time resolution for signals with slow-varying frequency content (Dokur and Ölmez, 2001). Additionally, compared to discrete wavelet transform (DWT), CWT remedies non-stationarity and coarse time-frequency resolution defects and supports the extraction of arbitrarily high-resolution features in the time-frequency domain (Addison, 2005).

The CWT of the ECG time-series signal denoted as $x(t)$ is achieved according to:

$$T(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\,\psi\left(\frac{t-b}{a}\right)\mathrm{d}t \tag{1}$$

where $T(a,b)$ stands for the intensity of transformed signals, $\psi(\cdot)$ is the wavelet basis (also known as the mother wavelet), a is the scale factor quantifying the compressed or stretched degree of a wavelet, and b is the time shift parameter defining the location of the wavelet. The scale can be used to derive the characteristic frequency of the wavelet as (Wu et al., 2019):

$$F = \frac{F_c \times f_s}{a} \tag{2}$$

where F_c is the center frequency of the mother wavelet and f_s is the sampling frequency of the signal. This relationship shows that smaller (larger) values of a correspond to higher (lower) frequency components. In CWT, the mother wavelet plays a critical role in time-frequency analysis, the choice of which depends on its similarity with the original signal (Ngui et al., 2013). Here, the Mexican hat wavelet (mexh) is chosen to serve as the mother wavelet because its shape is similar to the QRS waves and it is commonly used in ECG signal analysis (Wang et al., 2021). Specifically, the mexh is the second derivative of a Gaussian function (Addison, 2005), which is defined as:

$$\psi(t) = \frac{2}{\sqrt{3}\sqrt[4]{\pi}}\exp\left(-\frac{t^2}{2}\right)\left(1-t^2\right) \tag{3}$$

Continuously changing the scale factor a and time shift parameter b generates the 2D wavelet coefficients $T(a,b)$, which can be viewed as a 2D scalogram of the ECG signal in both the time and frequency domain (Wang et al., 2021).

Figures 2A–D show the healthy and AF examples of the raw ECG signals obtained from PhysioNet and their 2D time-frequency patterns after CWT transformation with mexh wavelet, respectively. The colors in the scalogram indicate the energy density of the signal component at the corresponding frequency and time (Addison, 2005; He et al., 2018). According to Figure 2A,C, two general differences can be observed: 1) The AF ECG signal lacks a distinct P wave, while it shows a fast and chaotic F wave due to the atrial fluttering (Figure 2C), in comparison to a normal ECG signal (Figure 2A); 2) Irregular RR intervals are observed in AF ECG (Figure 2C) caused by a non-synchronized ventricular response to the abnormal atrial excitation (He et al., 2018). The discriminative information in the time domain can also be captured by the CWT scalograms shown in Figures 2B,D. By using a 2D CNN to analyze the visual representation of 2D time-frequency scalograms, we can better understand the features that distinguish AF from normal heart rhythms and make more accurate predictions.

3.4 Convolutional neural network

We engage CNN to build a data-driven classifier for differentiating AF samples from normal ECG samples. CNN is a type of network architecture specifically designed to process data that has a grid-like structure such as images (Khan et al., 2020). As opposed to traditional multilayer perceptron networks (MLPs), where the input of each neuron consists of the outputs of all the neurons from the previous layer, the neuron in CNN only receives its input from a localized region of the previous layer, known as its receptive field. The main building blocks of a CNN are convolutional layers, pooling layers, and fully connected layers.

Convolutional layers are responsible for performing a convolution operation on the input data, using a set of filters to extract local features in the data, and producing a feature map that summarizes such local information. Let θ and X denote the filter (also known as the kernel) and the input. The convolution operation works as follows:

$$(X \otimes \theta)_{ij} = \sum_{m=0}^{s_1-1}\sum_{n=0}^{s_2-1} X(i+m, j+n)\,\theta(m,n) \tag{4}$$

where s_1 and s_2 denote the size of the 2D kernel, and (i,j) denotes the location on the 2D input (e.g., image). After being applied with the activation function, the feature map of the input is obtained as (LeCun and Bengio, 1995; Jing et al., 2021):

$$X_q^l = \sigma\left(\sum_p \theta_{pq}^l \otimes X_p^{l-1} + b_q^l\right) \tag{5}$$

where X_q^l is the qth feature at layer l, X_p^{l-1} is the pth input feature map of the previous $(l-1)$-th layer, σ denotes the activation function to induce the non-linearity in the functional mapping, and b_q represents the bias. This procedure is repeated by applying multiple filters to generate multiple feature maps to capture different characteristics of the input. Note that kernels are shared across all the input positions, which is also called weight sharing, the key feature of CNN. The weight-sharing technique guarantees the extracted local patterns are translation invariant and increases computational efficiency by reducing the model parameters to learn compared with fully connected neural networks.

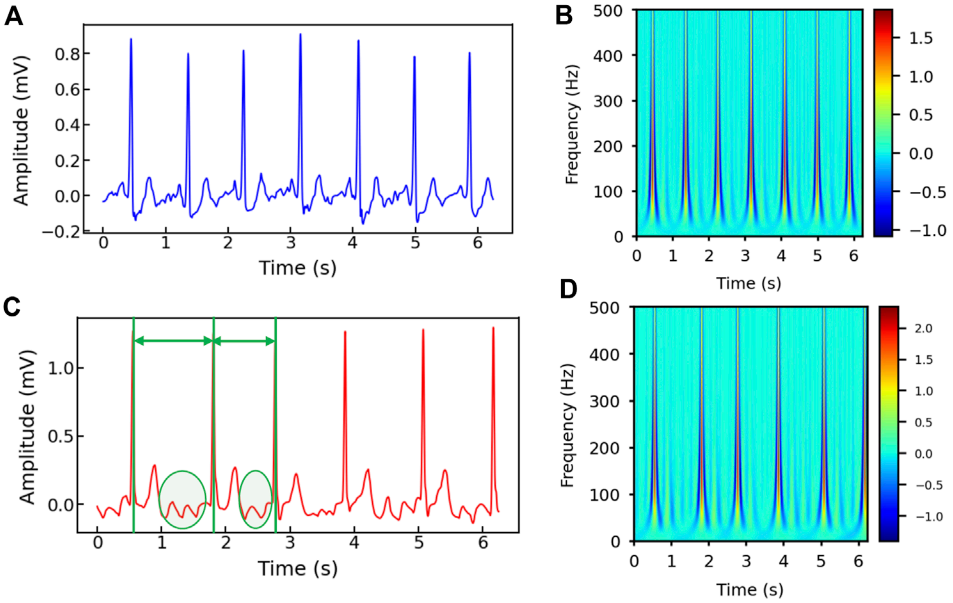


FIGURE 2
(A) The raw ECG signal from Physionet labeled as normal and **(B)** its corresponding 2D CWT scalogram. **(C)** The raw ECG signal from Physionet labeled as AF and **(D)** its corresponding 2D CWT scalogram. Note that the RR intervals are different in the AF sample and irregular F waves (circled) appear in **(C)**.

The pooling layer mimics the human visual system by combining the outputs of multiple neurons (i.e., clusters) into a single neuron in the next layer, effectively creating a condensed representation of the input. The pooling significantly reduces the spatial resolution and only focuses on the prominent patterns of the feature maps, making the network more robust to small translations and distortion in the input data (Xia et al., 2018). Popular pooling techniques include maximum pooling, average pooling, stochastic pooling, and adaptive pooling. They are typically performed on the values in a sub-region of the feature map (Akhtar and Ragavendran, 2020).

The fully-connected layers form a dense network that can learn complex non-linear relationships between the inputs and outputs. It takes the output of the previous layer, which is typically a high-dimensional tensor containing discriminant features extracted by convolutional and pooling layers, and flattens it into a one-dimensional vector. This vector is then used as the input to a fully connected layer. The fully-connected layer is similar to an MLP in that every neuron in one layer is connected to every neuron in the next layer. By using a proper activation function, the neural network is able to produce classification decisions (Nurmaini et al., 2020). By stacking these building blocks (convolutional layers, pooling layers, and fully connected layers) in various combinations, CNN is able to learn complex features in the input data, allowing them to effectively solve a wide range of image and signal processing tasks (Andreotti et al., 2017).

3.5 2D CNN with ResNet

We propose to engage 2D CNN to investigate the 2D time-frequency scalograms converted from denoised ECG signals by CWT for AF identification. It has been demonstrated that the substantial depth of the convolutional network is beneficial to the network performance (Simonyan and Zisserman, 2014). However, as the number of convolutional layers increases, the training loss stops further decreasing and becomes saturated because of the gradient dissipation issue. As such, a CNN with a deeper architecture, counterintuitively, sometimes incurs a larger training error compared to its shallow counterpart upon convergence (He et al., 2016). To solve such network degradation and gradient vanishing problems, the residual network (ResNet) has been developed to improve the accuracy of CNNs with considerably increased depth.

The core of ResNet is the residual learning technique (He et al., 2016). Specifically, instead of using the stacked convolutional layers to directly fit the underlying mapping from the input to the output, ResNet focuses on fitting a residual mapping. Figure 3 shows a ResNet building block with input X and its corresponding output mapping Y. The residual block engages a shortcut connection that bypasses one or more convolutional layers and allows the information to flow directly from the input to the output. As such, the input X is added to the output of the block $F(X)$ (enclosed by the dashed circle in Figure 3, allowing the network to learn the residual mapping represented as $Y = F(X) + X$ instead of learning the direct mapping as $Y = F(X)$. This design mitigates the gradient vanishing problem and allows for deeper networks to be trained effectively.

In our study, we engage the ResNet with 18 layers (ResNet18) to build the AF classifier because ResNet18 has been proven to be able to generate a comparable result with a faster convergence compared to a deeper counterpart (He et al., 2016). Figure 4 shows the detailed structure of ResNet18. Note that the notation of 2DConv$(n_{input}, n_{output}, n_{fdim1} \times n_{fdim2})$ denotes that, in the current 2D convolutional layer, there are n_{input} input

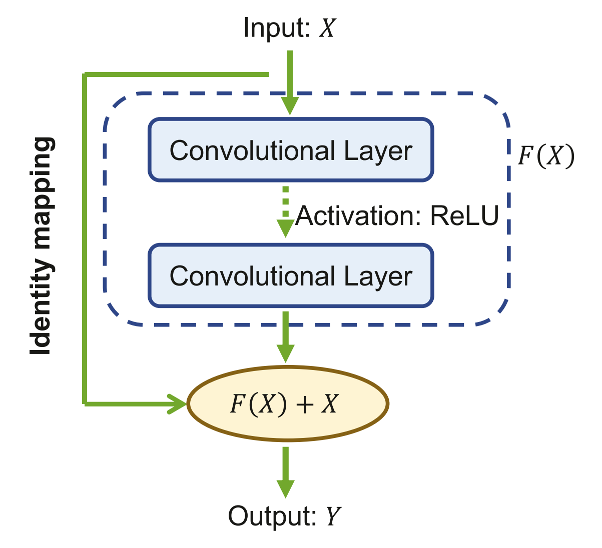


FIGURE 3
A building block of the ResNet.

channels, n_{output} output channels (i.e., number of filters) with the 2D filter size of $n_{fdim1} \times n_{fdim2}$. For example, $(64, 128, 3 \times 3)$ indicates that this convolutional layer is composed of 128 filters with the filter size of 3×3 applied on the input data with 64 channels.

3.6 Multi-branching convolutional network

Data-driven identification of AF from ECG recordings generally suffers from imbalanced data issues. Figure 5A presents the distribution of AF and normal samples in Physionet/CinC 2017 and OUHSC datasets, illustrating a normal to AF sample ratio of approximately 7:1 for both. To address the data imbalance issue, we create N_b balanced datasets from the original data $D = \{D_{-}, D_{+}\}$, where D_{-} denotes the majority normal ECG samples and D_{+} stands for the minorityset, i.e., the entire AF training samples. D_{-} is partitioned into multiple subsets $D_{-} = \cup_{i=1}^{N_b} D_{-}^{i}$, where each subset D_i is roughly equivalent in size to D_{+}. The normal subsets D_i for $i = 1, \ldots, N_b$ are then paired with D_{+} to formulate balanced sub-datasets. Each balanced subset, denoted as $D_i = \{D_{-}^{i}, D^{+}\}$ for $i = 1, \ldots, N_b$, is processed through the ResNet core, with individual branches trained on their respective balanced sub-datasets. Figure 5B visualizes this method of partitioning the original dataset D into N_b balanced sub-datasets, i.e., D_i for $i = 1, \ldots, N_b$, which serve as the balanced input in Figure 6. This strategic partitioning and training approach ensures a comprehensive model learning from a balanced representation of AF and normal ECG samples (Wang and Yao, 2021; Wang et al., 2022; Wang et al., 2023b).

In the current investigation, we aim to identify AF samples from normal ECG samples. The neural network is expected to produce high probabilities (close to 1) for AF samples and low probabilities (close to 0) for normal ECG samples. We choose the binary cross-entropy as the loss function for MB-ResNet, which is defined as:

$$\mathcal{L}(\boldsymbol{\omega}; D) = -\sum_{j=1}^{N_d}\sum_{i=1}^{N_b}\mathcal{I}(j \in D_i)\Big(y_j \log\left(\hat{P}_i\left(\boldsymbol{\omega}; X^j\right)\right) + \left(1 - y_j\right)\log\left(1 - \hat{P}_i\left(\boldsymbol{\omega}; X^j\right)\right)\Big) \qquad (6)$$

where $\boldsymbol{\omega}$ denotes the neural network parameter set, X^j and y_j stand for one input sample and its corresponding true label respectively, $\mathcal{I}(\cdot)$ denotes the indicator function, N_d is the total number of the training samples, and $\hat{P}_i\left(\boldsymbol{\omega}; X^j\right)$ represents the predicted probability for AF at the ith branching output given the input signal X^j.

The adaptive momentum method (Adam) (Kingma and Ba, 2014) is adopted to minimize the loss function and update the network parameters. In the inference stage, the MB network generates N_b predictions for AF probability, which correspond to the N_b branching outputs. The final predicted probability for AF ($\hat{P}$) is determined by taking the average of the N_b outputs:

$$\hat{P} = \frac{1}{N_b}\sum_{i=1}^{N_b}\hat{P}_i$$

where $\hat{P}_i$ is the predicted probability of ith branching output.

4 Experimental design and results

4.1 Experimental design

We validate and evaluate the performance of the proposed CWT-MB-ResNet framework using both OUHSC and Physionet Challenge datasets. In this study, the training and testing datasets are split interpatiently for both data sources. This ensures that no overlap exists between the patients in the training set and those in the testing set. We allocate 80% of the total samples for the training purpose and the remaining 20% for testing, applied on both datasets.

We first explore the impact of the learning rate on the training outcomes of the proposed CWT-MB-ResNet. We then conducted a comparison study to showcase the significance of ECG digitalization for the proposed multi-branching ResNet (MB-ResNet) model in identifying the AF samples. Next, we compare the performance of our CWT-MB-ResNet with 1D-CNN (Figure 7A), 1D-CNN with the multi-branching network (1D-MB-CNN) (Figure 7B), and ResNet with CWT features (CWT-ResNet). Note that the input of 1D-CNN and 1D-MB-CNN consists of the denoised ECG time series. The detailed 1D-CNN architecture is illustrated in Figure 8, including three convolutional layers followed by pooling layers to reduce the dimensionality of the data, a batch-normalization layer to stabilize the network training, and one fully connected layer to make the final prediction. Note that the notation of 1DConv($n_{input}, n_{output}, n_{fdim}$) indicates that, in the current 1D convolutional layer, there are n_{input} input channels and n_{output} output channels (i.e., number of filters) with a 1D filter size of n_{fdim}.

The classification performance will be evaluated with three metrics: Receiver-Operating-Characteristic (ROC) Curve, Precision-Recall (PR) Curve, and F1 score, which will be calculated using the test set. The ROC provides the graphic representation of the trade-off between the true positive rate (TPR) and the false

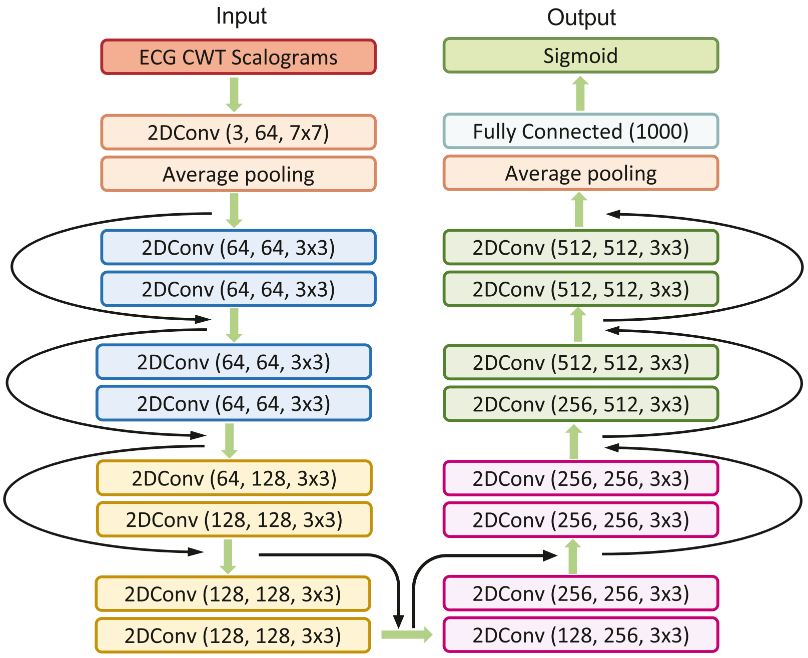


FIGURE 4
The detailed architecture of ResNet18.

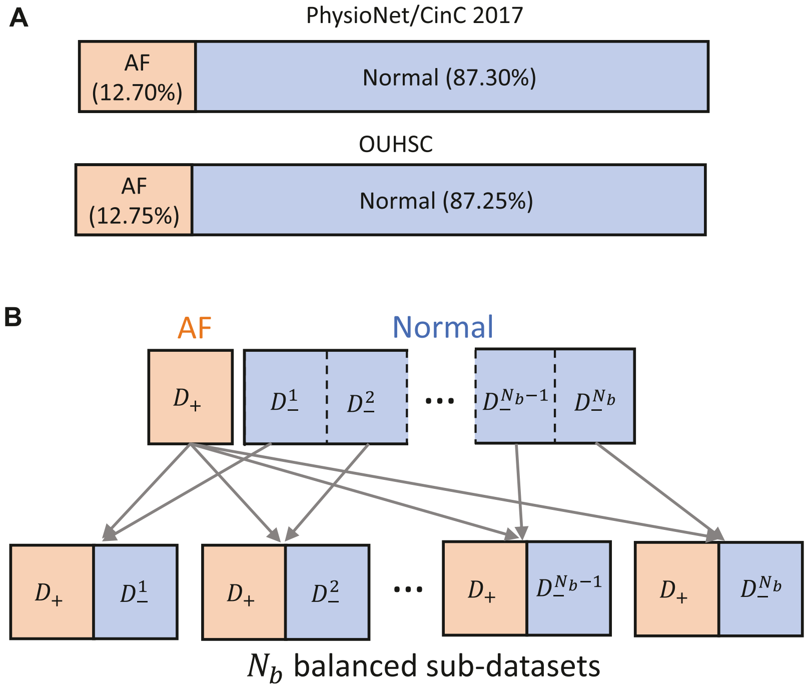


FIGURE 5
(A) Class distribution in PhysioNet/Cinc 2017 and OUHSC datasets. **(B)** Illustration of creating N_b balanced sub-datasets to train our MB-ResNet model.

positive rate (FPR) for different threshold settings. The area under ROC (AUROC) is often used as a metric to compare different models, with a larger AUROC indicating a better-performing classifier. A good model typically has a ROC curve that is situated toward the top-left corner of the graph. The PRC illustrates the interplay between a predictive model's precision and recall metrics across a range of probability thresholds. A good classifier has the PR curve towards the top-right corner. A higher area under PRC

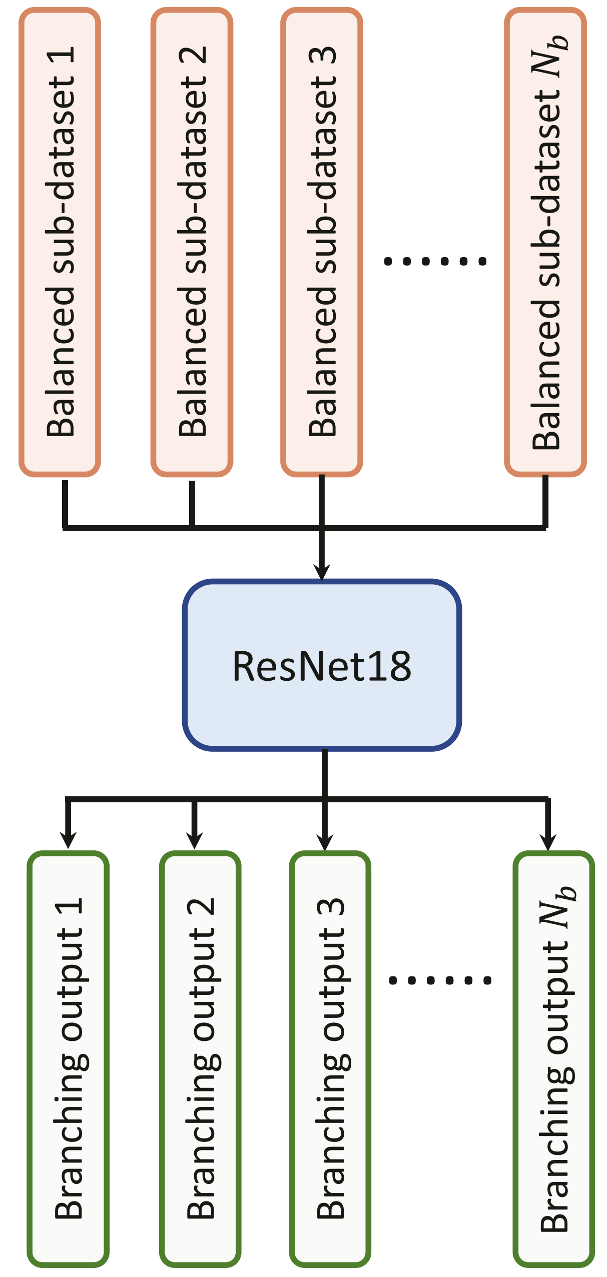


FIGURE 6
Illustration of the multi-branching architecture.

(AUPRC) value suggests a more effective model. The F1 score quantifies the equilibrium between a model's precision and recall for a binary classifier system by computing their harmonic mean, which is defined as

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Note that the F1 score ranges from 0 to 1, where a score of 1 indicates a perfect balance between precision and recall and a good overall prediction performance.

4.2 The effect of the learning rate on CWT-MB-ResNet

In this study, we initiate the analysis by transforming ECG time series data into 2D scalograms utilizing CWT. These scalograms encapsulating both time and frequency information are input into our tailored MB-ResNet model. Specifically, we employ ResNet18 due to its proven efficacy in achieving results comparable to those of its deeper counterparts, while also ensuring faster convergence rates (He et al., 2016). The architecture of ResNet18, as adopted from He et al. (2016) and illustrated in Figure 4, comes with a predefined set of network architecture parameters, including number of layers, kernel size, and number of residual blocks.

In addition to selecting ResNet18 for its balance between efficiency and performance, the learning rate has a critical influence on the training outcomes. To further optimize our model, we conducted an experiment specifically focused on assessing the impact of various learning rates on the model's performance, particularly looking at the F1 score on the test set across both datasets used in our study. Table 1 summarizes the performance of the MB-ResNet given different learning rates. For both datasets, the highest F1 score achieved is 0.8865 for PhysioNet/CinC 2017 and 0.7396 for OUHSC datasets when the learning rate is set as 0.001. This indicates that a learning rate of 0.001 is the most effective in training our MB-ResNet model.

4.3 The effect of ECG digitalization from PDFs on CWT-MB-ResNet

We carry out a comparative analysis to demonstrate the importance of digitizing ECG records from their original PDF format. Specifically, we transform the original ECG PDFs into image files (i.e., Portable Network Graphic (.PNG) files) and apply segmentation to augment the sample sizes. Figure 9 illustrates examples of the resulting ECG images from normal and AF categories, which directly serve as inputs for our MB-ResNet without further preprocessing.

Figure 10 displays the ROC and PR curves generated by two variants of the MB-ResNet model: one trained on 2D scalograms derived from digitalized ECGs after undergoing denoising and CWT (referred to as CWT-MB-ResNet), and the other trained on pure ECG images converted directly from raw PDF files (denoted as PDF-MB-ResNet). Utilizing the same MB-ResNet model, we observed a substantial increase in the area under both ROC and PR curves when the model inputs were 2D scalograms processed from digitalized ECGs compared with using raw ECG images directly. Specifically, our CWT-MB-ResNet model demonstrates superior performance with an AUROC of 0.9351, AUPRC of 0.7930, and an F1 score of 0.7396. This performance significantly surpasses that of the PDF-MB-ResNet trained by raw ECG images with an AUROC of 0.8683, AUPRC of 0.6462, and an F1 score of 0.6257, highlighting the efficacy of our digitalization and preprocessing procedure. The enhanced performance of the MB-ResNet model trained with 2D scalograms from digitalized ECGs, as compared to training with raw ECG images, is be attributed to several factors:

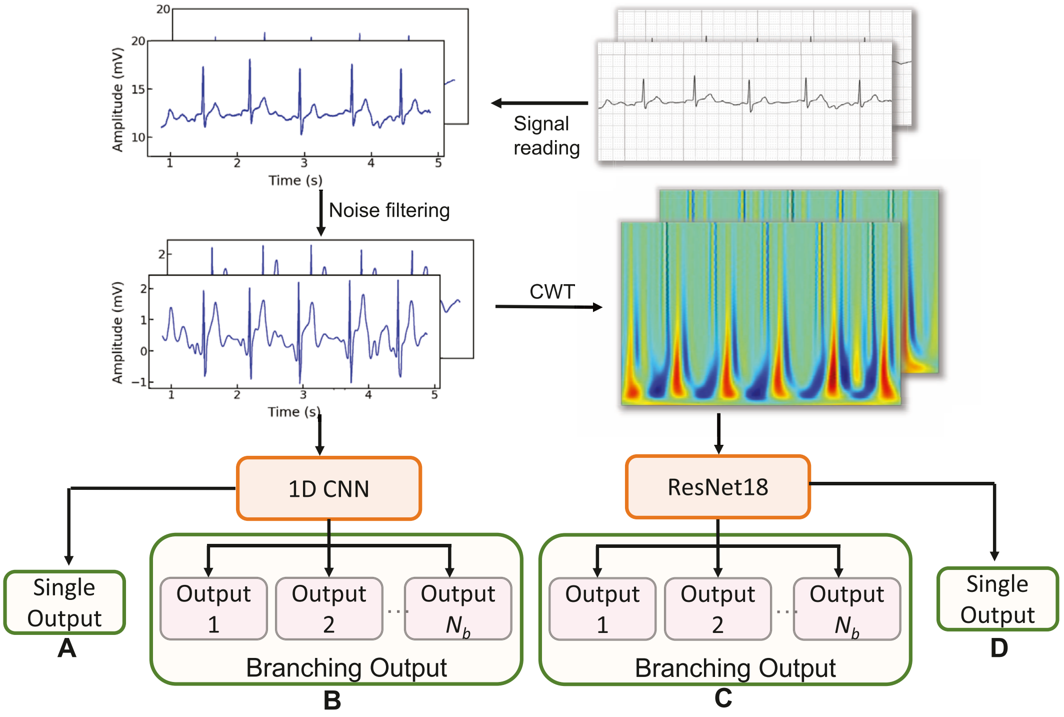


FIGURE 7
The flowchart of the experimental design: **(A)** 1D-CNN; **(B)** 1D-MB-CNN; **(C)** CWT-MB-ResNet; **(D)** CWT-ResNet.

TABLE 1 F1 scores on the testing set given different learning rates for MB-ResNet training.

learning rate	0.0001	0.0005	0.001	0.005	0.01
F1 (PhysioNet)	0.8493	0.8759	0.8865	0.8652	0.8580
F1 (OUHSC)	0.6854	0.7273	0.7396	0.7385	0.7151

- The 2D scalograms provide a rich representation of temporal and frequency features, offering a more comprehensive dataset for the model to learn from.
- The raw ECG segmentation images contain large blank areas devoid of any ECG-related information, which do not contribute to learning discriminative features.
- The superimposed gridlines in the area could introduce noise into the data, potentially hindering the model's training efficiency.

By training with 2D scalograms, the abovementioned issues are mitigated, allowing the MB-ResNet to focus on more relevant ECG features, leading to significant improvement in overall model performance.

4.4 Experimental results from the OUHSC dataset

Figure 11 displays the ROC and PR curves of all four models using the OUHSC dataset. The 2D ResNet models (i.e., CWT-ResNet and CWT-MB-ResNet), which use 2D scalograms transformed from ECG signals as the input, produce a larger area under the curves (both ROC and PR) compared to their 1D counterparts (i.e., 1D-CNN and 1D-MB-CNN). This demonstrates the efficacy of using the CWT to extract time-frequency features in the ECG signal analysis. Additionally, the models with an MB architecture (i.e., 1D-MB-CNN and CWT-MB-ResNet) produce a larger AUROC and AUPRC compared to models without MB outputs (i.e., 1D-CNN and CWT-ResNet), which highlights the effectiveness of using the MB structure in addressing imbalanced data issues. The ROC and PR plots demonstrate the superiority and robustness of the proposed CWT-MB-ResNet framework for identifying the AF samples.

Table 2 shows AUROC, AUPRC, and F1 scores generated from the four methods using the OUHSC dataset. The proposed CWT-MB-ResNet method generates the best AUROC, AUPRC, and F1 scores with values of 93.51%, 79.30%, and 0.7396. Note that the MB technique demonstrates its effectiveness on both 1D-CNN and CWT-ResNet as the AUROC, AUPRC, and F1 scores provided by the MB-based neural network models are higher than their non-MB counterparts. Moreover, the AF classifier using 2D-CNN-based ResNet18 supported by the time-frequency transformation of ECG

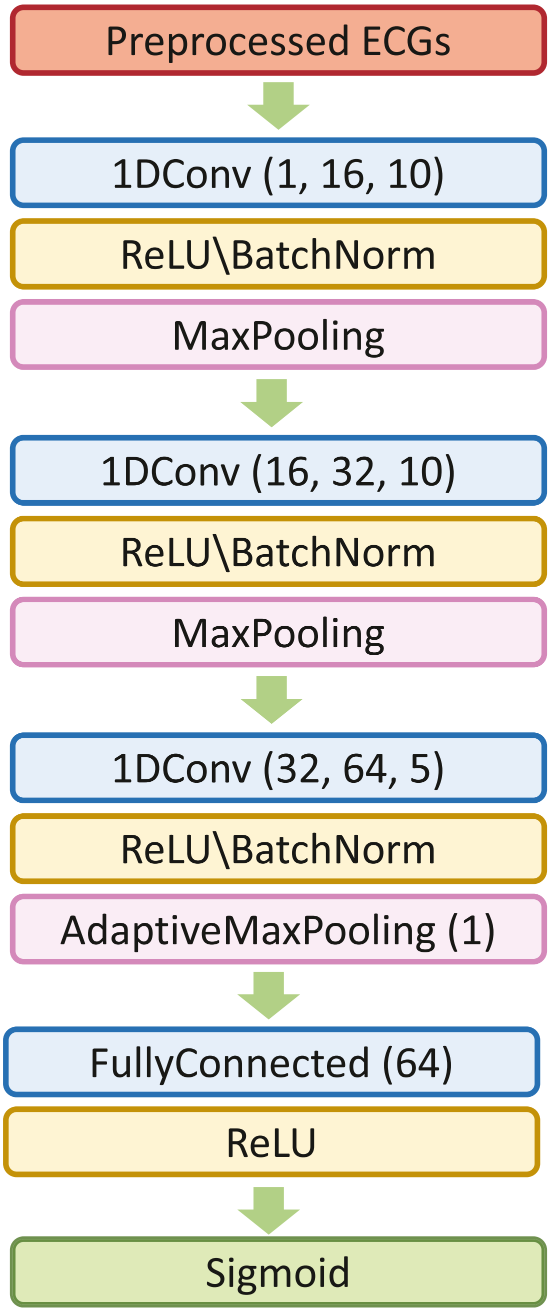


FIGURE 8
The 1D-CNN architecture.

time series presents a more potent predictive power than time sequence classification using 1D CNN. For example, CWT-MB-ResNet improves the AUROC, AUPRC, and F1 scores from 87.55%, 69.97%, and 0.6384% to 93.51%, 79.30%, and 0.7396 respectively compared with the 1D-MB-CNN.

4.5 Experimental results from the Physionet/CinC 2017 challenge dataset

Figure 12 further shows the ROC and PRC analysis for the Physionet/Cinc 2017 challenge dataset. Similar to the results from the OUHSC dataset, the 2D ResNet models (CWT-ResNet and CWT-MB-ResNet) outperform their 1D counterparts (1D-CNN and 1D-MB-CNN) in both the ROC and PR spaces. Furthermore, the MB-based models (1D-MB-CNN and CWT-MB-ResNet) effectively account for the imbalanced data issues, exhibiting better performance compared to the non-MB-based models (1D-CNN and CWT-ResNet). Table 3 demonstrates the comparison of AUROC, AUPRC, and F1 scores provided by 1D-CNN, 1D-MB-CNN, CWT-ResNet, and CWT-MB-ResNet. Our CWT-MB-ResNet yields the best classification performance among the four methods, generating the highest AUROC, AUPRC, and F1 scores of 97.41%, 93.53%, and 0.8865. Especially, our CWT-MB-ResNet model improves the F1 score by 46.2% percent compared to the pure 1D-CNN with no CWT transform or MB structure.

5 Discussion

5.1 Strengths of the proposed pipeline

This paper proposes a pipeline of CWT-MB-ResNet to identify the AF condition. The unique strengths of the proposed framework are:

1) **Digitalization of ECG readings in PDF:** This pipeline designed an ECG preprocessing method that can automatically convert ECG PDFs into digitalized, ready-to-use ECG time series data. This step is crucial for integrating machine learning models into clinical workflows, where ECGs are often archived in non-digitalized formats.
2) **Effectiveness of CWT representation:** The integration of CWT enhances feature extraction, enabling the model to better identify AF characteristics that might be missed by directly learning from raw time-series analysis alone. The resulted 2D ECG scalograms offer a rich representation of ECG data by encapsulating both time series and frequency components. The CWT-based feature reformulation can significantly enhance the model's performance by providing more comprehensive information for classifying ECG signals.
3) **Advantage of the network design:** The use of ResNet18 as the foundation allows our model to benefit from the strengths in deep residual learning, enabling it to learn from significantly deepened convolutional layers with improved accuracy. The ResNet18 has demonstrated comparable results to its deeper counterparts, meanwhile keeping its computational efficiency. This is further enhanced by our innovative multi-branching design, which addresses the class imbalance issue by training each branch on a balanced subset of the original dataset while the core network is exposed to the entire range of samples. This approach ensures that both AF and normal class is adequately represented and learned during the training process, significantly enhancing the network's ability to generalize across the imbalanced classes.

5.2 Discussion on the limitations

The proposed CWT-MB-ResNet framework, while effective, is not devoid of limitations. In our study, ECG segments were

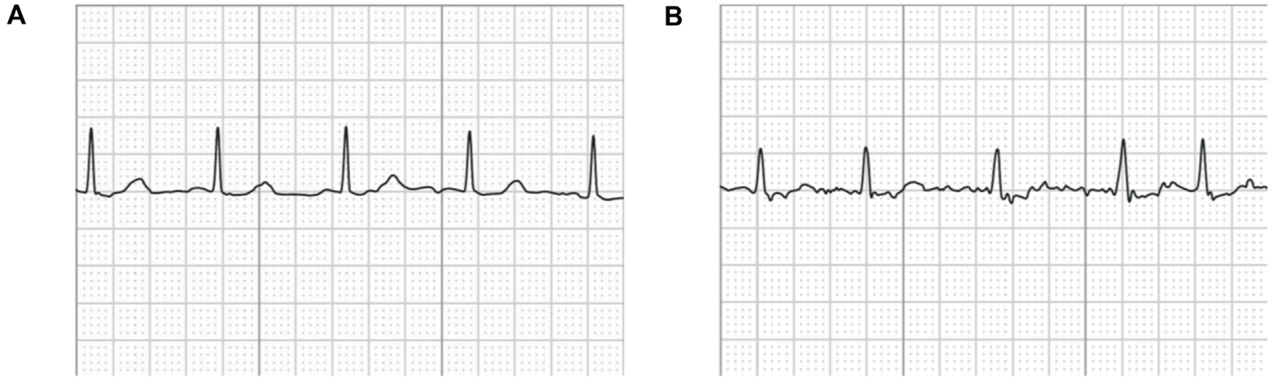


FIGURE 9
Examples of ECG segments: **(A)** normal sample; **(B)** AF sample.

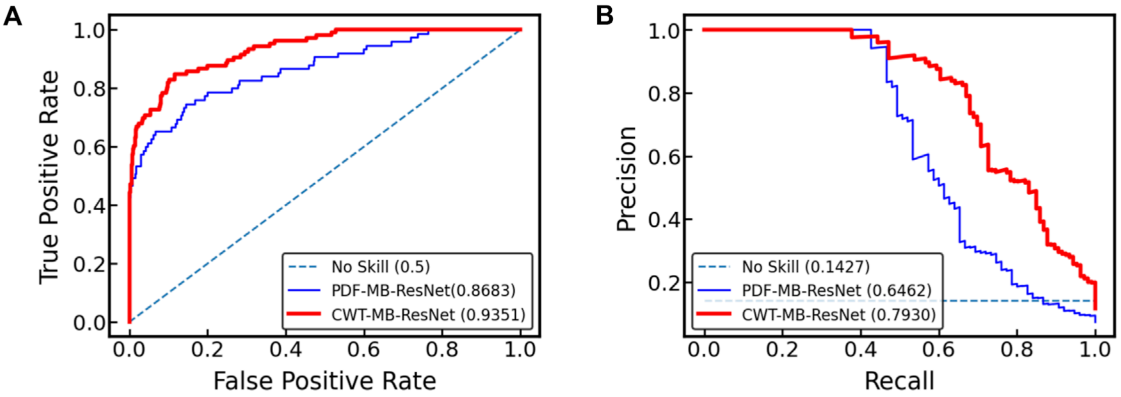


FIGURE 10
Comparison of **(A)** ROC and **(B)** PR curves for the MB-ResNet model trained with two different data preparation techniques: one involving 2D scalograms derived from digitalized ECGs which are denoised and processed through CWT (CWT-MB-ResNet), and the other using unprocessed ECG images directly from raw PDF files (PDF-MB-ResNet).

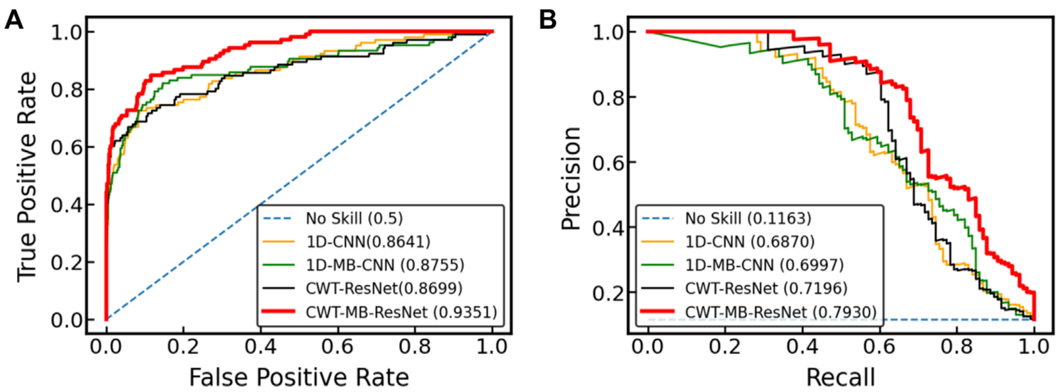


FIGURE 11
The comparison of **(A)** ROC and **(B)** PRC among different models using the OUHSC data.

TABLE 2 The comparison of AUROC, AUPRC, and F1 scores generated from 1D-CNN, 1D-MB-CNN, CWT-ResNet, and the proposed CWT-MB-ResNet using OUHSC data.

	1D-CNN	1D-MB-CNN	CWT-ResNet	CWT-MB-ResNet
AUROC	86.41%	87.55%	86.99%	93.51%
AUPRC	68.70%	69.97%	71.96%	79.30%
F1	0.6370	0.6384	0.7150	0.7396

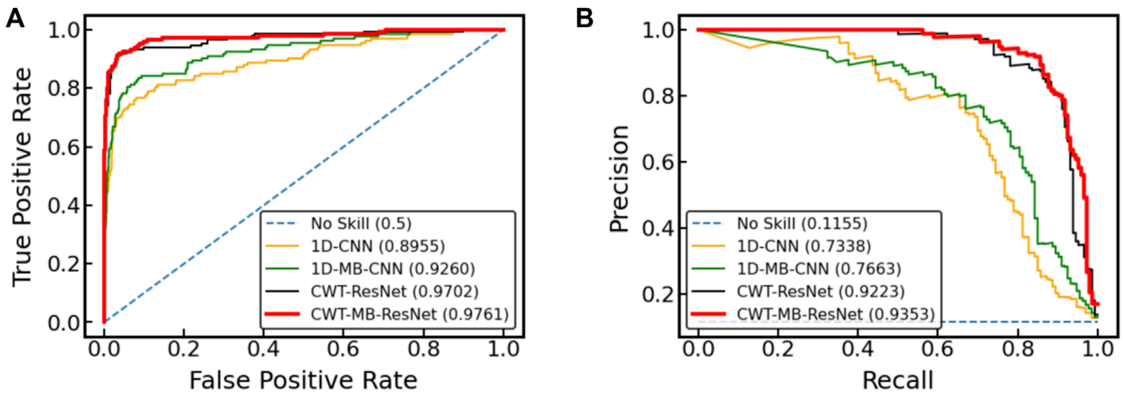


FIGURE 12
The comparison of **(A)** ROC and **(B)** PRC between different models using data from Physionet/Cinc 2017 challenge.

TABLE 3 The comparison of AUROC, AUPRC, and F1 scores generated from 1D-CNN, 1D-MB-CNN, CWT-ResNet, and the proposed CWT-MB-ResNet using data from Physionet/CinC 2017 challenge.

	1D-CNN	1D-MB-CNN	CWT-ResNet	CWT-MB-ResNet
AUROC	89.55%	92.60%	97.02%	97.61%
AUPRC	73.38%	76.63%	92.23%	93.53%
F1	0.7219	0.7380	0.8690	0.8865

around 5 s long. However, analyzing longer ECG recordings will significantly increase computational complexity. This is due to the CWT method of processing data across both time and frequency domains at various scales, demanding more computational resources. Additionally, while our method effectively addresses class imbalance, its performance remains influenced by the quality and diversity of the training data, which is a long-lasting limitation of most data-driven machine learning models. This is evident from the differing performances on the PhysioNet and OUHSC datasets. Specifically, PhysioNet, with its larger and more diverse pool of 5,788 subjects, provides a richer training environment compared to OUHSC, which is limited to ECG samples from only 260 subjects. Despite utilizing segmentation to expand the sample size of the OUHSC dataset to 5,809, notable differences in performance metrics remain, as detailed in Tables 2 and 3. This suggests that merely increasing the sample size by segmentation cannot fully address the limitations posed by data diversity and quality. Additionally, deep learning models, including the proposed CWT-MB-ResNet, are often criticized for their “black box” nature. This means that while those models can make accurate predictions, the reasoning behind the predictions is not always clear or understandable to humans. This lack of interpretability can be a significant hurdle in clinical settings, making clinicians less confident in implementing machine learning models for automated diagnosis. One of our future research directions will focus on the development of interpretable models for AF detection.

5.3 Comparison with existing work

The direct comparison of our results with the values of performance metrics reported in other studies mentioned in Section 2 is neither fair nor feasible due to several factors: 1) variations in ECG duration used for training/testing data; 2) employment of non-unified metrics for evaluating model performance across studies; 3) variations in the proportions of

TABLE 4 The comparison of F1 scores between the proposed CWT-MB-ResNet method with existing literature using data from Physionet/CinC 2017 and OUHSC.

Authors	Methods	F1 (PhysioNet)	F1 (OUHSC)
Andreotti et al. Andreotti et al. (2017)	ResNet	0.8405	0.7054
Limam et al. Limam and Precioso, (2017)	CRNN	0.8310	0.7323
Wang et al. Wang and Li, (2020)	CNN-Bi-LSTM	0.7094	0.6996
Gao et al. Gao et al. (2021)	Residual-based temporal attention	0.8172	0.7368
This paper	CWT-MB-ResNet	0.8865	0.7396

training/testing data splits; 4) the model implementation on different databases. To enable a fairer and more meaningful comparison, we applied the ECG data from both the PhysioNet/CinC 2017 database and OUHSC to four deep learning models reviewed in Section 2, ensuring that the comparison is based on consistent data and preprocessing steps.

Table 4 summarizes the comparison results in terms of F1 score. Even though the proposed CWT-MB-ResNet model does not resort to complex neural network designs, it demonstrates the best F1 score compared with the other network structures developed in Andreotti et al. (2017); Limam and Precioso, (2017); Wang and Li, (2020); Gao et al. (2021). Specifically, the utilization of CWT distills both frequency and temporal insights from ECG signals, converting them into an image data format that significantly enriches the input information. We integrate the widely recognized image model, ResNet18 to achieve a robust interpretation of image data and meanwhile circumvent the gradient vanishing problem. Furthermore, the multi-branching structure is meticulously designed to address issues of data imbalance, ensuring that our model remains sensitive and accurate for both normal and AF classes.

6 Conclusion

In this paper, we develop a novel framework based on Continous Wavelet Transform (CWT) and multi-branching ResNet for AF identification. We first transform the 1D ECG time series into 2D time-frequency scalograms to take into account various frequency components, which can serve as the input to the 2D CNN-based classifier. Second, we leverage the ResNet architecture to cope with the gradient dissipation problems in deep 2D CNN and increase the effectiveness of network training. Moreover, a multi-branching architecture is incorporated into the ResNet to mitigate the possible prediction bias caused by the imbalanced data issue. Finally, we implement the proposed CWT-MB-ResNet to predict AF using the ECG recordings from PhysioNet/CinC Challenge 2017 and the ECG PDFs from OUHSC. Experimental results show that the proposed CWT-MB-ResNet achieves the best prediction performance for both datasets in AF detection. The CWT-MB-ResNet framework has great potential to be applied in clinical practice to improve the accuracy in ECG-based diagnosis of heart disease.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: In this study, two AF databases from different sources, i.e., ECG recordings from PhysioNet/CinC challenge 2017 and ECG PDFs from OUHSC, are used to evaluate the performance of data-driven detection methods. PhysioNet/CinC challenge 2017 is open-source database. The ECG PDFs from OUHSC are provided by our cardiologist collaborators. Requests to access these datasets should be directed to Stavros-Stavrakis@ouhsc.edu.

Author contributions

JX: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. SS: Conceptualization, Data curation, Resources, Supervision, Validation, Visualization, Writing–review and editing. BY: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Research reported in this publication was supported by the National Heart, Lung, And Blood Institute of the National Institutes of Health under Award Number R01HL172292. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdou, A., and Krishnan, S. (2022). Horizons in single-lead ecg analysis from devices to data. *Front. Signal Process.* 2, 866047. doi:10.3389/frsip.2022.866047

Addison, P. S. (2005). Wavelet transforms and the ecg: a review. *Physiol. Meas.* 26, R155–R199. doi:10.1088/0967-3334/26/5/R01

Akhtar, N., and Ragavendran, U. (2020). Interpretation of intelligence in cnn-pooling processes: a methodological survey. *Neural Comput. Appl.* 32, 879–898. doi:10.1007/s00521-019-04296-5

AliveCor (2024). AliveCor website. Available at: https://www.alivecor.com/ (Accessed February 26, 2024).

Analog Devices (2024). AD8232 single-lead, heart rate monitor front end. Available at: https://www.analog.com/media/en/technical-documentation/data-sheets/ad8232.pdf (Accessed February 26, 2024).

Andreotti, F., Carr, O., Pimentel, M. A., Mahdi, A., and De Vos, M. (2017). "Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ecg," in *2017 computing in cardiology (CinC)* (IEEE), 1–4.

Athif, M., Yasawardene, P. C., and Daluwatte, C. (2018). Detecting atrial fibrillation from short single lead ecgs using statistical and morphological features. *Physiol. Meas.* 39, 064002. doi:10.1088/1361-6579/aac552

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bernstein, R. A., Kamel, H., Granger, C. B., Piccini, J. P., Sethi, P. P., Katz, J. M., et al. (2021). Effect of long-term continuous cardiac monitoring vs usual care on detection of atrial fibrillation in patients with stroke attributed to large-or small-vessel disease: the stroke-af randomized clinical trial. *Jama* 325, 2169–2177. doi:10.1001/jama.2021.6470

Billeci, L., Chiarugi, F., Costi, M., Lombardi, D., and Varanini, M. (2017). "Detection of af and other rhythms using rr variability and ecg spectral measures," in *2017 computing in cardiology (CinC)* (IEEE), 1–4.

Boriani, G., Palmisano, P., Malavasi, V. L., Fantecchi, E., Vitolo, M., Bonini, N., et al. (2021). Clinical factors associated with atrial fibrillation detection on single-time point screening using a hand-held single-lead ecg device. *J. Clin. Med.* 10, 729. doi:10.3390/jcm10040729

Cai, W., Chen, Y., Guo, J., Han, B., Shi, Y., Ji, L., et al. (2020). Accurate detection of atrial fibrillation from 12-lead ecg using deep neural network. *Comput. Biol. Med.* 116, 103378. doi:10.1016/j.compbiomed.2019.103378

Chen, S., Wang, Z., Yao, B., and Liu, T. (2022). "Prediction of diabetic retinopathy using longitudinal electronic health records," in 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE) (IEEE), 949–954.

Clifford, G. D., Liu, C., Moody, B., Li-wei, H. L., Silva, I., Li, Q., et al. (2017). "Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017," in *2017 computing in cardiology (CinC)* (IEEE), 1–4.

Colilla, S., Crow, A., Petkun, W., Singer, D. E., Simon, T., and Liu, X. (2013). Estimates of current and future incidence and prevalence of atrial fibrillation in the us adult population. *Am. J. Cardiol.* 112, 1142–1147. doi:10.1016/j.amjcard.2013.05.063

Da Silva-Filarder, M., and Marzbanrad, F. (2017). "Combining template-based and feature-based classification to detect atrial fibrillation from a short single lead ecg recording," in *2017 computing in cardiology (CinC)* (IEEE), 1–4.

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Inf. theory* 36, 961–1005. doi:10.1109/18.57199

De Chazal, P., O'Dwyer, M., and Reilly, R. B. (2004). Automatic classification of heartbeats using ecg morphology and heartbeat interval features. *IEEE Trans. Biomed. Eng.* 51, 1196–1206. doi:10.1109/TBME.2004.827359

Dokur, Z., and Ölmez, T. (2001). Ecg beat classification by a novel hybrid neural network. *Comput. methods programs Biomed.* 66, 167–181. doi:10.1016/s0169-2607(00)00133-4

Fan, X., Yao, Q., Cai, Y., Miao, F., Sun, F., and Li, Y. (2018). Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ecg recordings. *IEEE J. Biomed. health Inf.* 22, 1744–1753. doi:10.1109/JBHI.2018.2858789

Faust, O., Shenfield, A., Kareem, M., San, T. R., Fujita, H., and Acharya, U. R. (2018). Automated detection of atrial fibrillation using long short-term memory network with rr interval signals. *Comput. Biol. Med.* 102, 327–335. doi:10.1016/j.compbiomed.2018.07.001

Gao, J., Zhang, H., Lu, P., and Wang, Z. (2019). An effective lstm recurrent network to detect arrhythmia on imbalanced ecg dataset. *J. Healthc. Eng.* 2019, 6320651. doi:10.1155/2019/6320651

Gao, Y., Wang, H., and Liu, Z. (2021). An end-to-end atrial fibrillation detection by a novel residual-based temporal attention convolutional neural network with exponential nonlinearity loss. *Knowledge-Based Syst.* 212, 106589. doi:10.1016/j.knosys.2020.106589

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation* 101, e215–e220. doi:10.1161/01.cir.101.23.e215

Guan, Y., An, Y., Xu, J., Liu, N., and Wang, J. (2022). Ha-resnet: residual neural network with hidden attention for ecg arrhythmia detection using two-dimensional signal. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20, 3389–3398. doi:10.1109/TCBB.2022.3198998

He, H., and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. data Eng.* 21, 1263–1284. doi:10.1109/tkde.2008.239

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778.

He, R., Wang, K., Zhao, N., Liu, Y., Yuan, Y., Li, Q., et al. (2018). Automatic detection of atrial fibrillation based on continuous wavelet transform and 2d convolutional neural networks. *Front. physiology* 9, 1206. doi:10.3389/fphys.2018.01206

Huang, J., Chen, B., Yao, B., and He, W. (2019). Ecg arrhythmia classification using stft-based spectrogram and convolutional neural network. *IEEE access* 7, 92871–92880. doi:10.1109/access.2019.2928017

Isakadze, N., and Martin, S. S. (2020). How useful is the smartwatch ecg? *Trends Cardiovasc. Med.* 30, 442–448. doi:10.1016/j.tcm.2019.10.010

Islam, S., Ammour, N., and Alajlan, N. (2017). "Atrial fibrillation detection with multiparametric rr interval feature and machine learning technique," in 2017 International Conference on Informatics, Health & Technology (ICIHT) (IEEE), 1–5.

Izci, E., Ozdemir, M. A., Degirmenci, M., and Akan, A. (2019). "Cardiac arrhythmia detection from 2d ecg images by using deep learning technique," in *2019 medical technologies congress (TIPTEKNO)* (IEEE), 1–4.

Jing, E., Zhang, H., Li, Z., Liu, Y., Ji, Z., and Ganchev, I. (2021). Ecg heartbeat classification based on an improved resnet-18 model. *Comput. Math. Methods Med.* 2021, 6649970. doi:10.1155/2021/6649970

Jun, T. J., Nguyen, H. M., Kang, D., Kim, D., Kim, D., and Kim, Y.-H. (2018). Ecg arrhythmia classification using a 2-d convolutional neural network. arXiv preprint arXiv:1804.06812

Keissar, K., Davrath, L. R., and Akselrod, S. (2009). Coherence analysis between respiration and heart rate variability using continuous wavelet transform. *Philosophical Trans. R. Soc. A Math. Phys. Eng. Sci.* 367, 1393–1406. doi:10.1098/rsta.2008.0273

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* 53, 5455–5516. doi:10.1007/s10462-020-09825-6

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*

Kutlu, Y., and Kuntalp, D. (2012). Feature extraction for ecg heartbeats using higher order statistics of wpd coefficients. *Comput. methods programs Biomed.* 105, 257–267. doi:10.1016/j.cmpb.2011.10.002

Ladavich, S., and Ghoraani, B. (2015). Rate-independent detection of atrial fibrillation by statistical modeling of atrial activity. *Biomed. Signal Process. Control* 18, 274–281. doi:10.1016/j.bspc.2015.01.007

Lai, D., Zhang, X., Bu, Y., Su, Y., and Ma, C.-S. (2019). An automatic system for real-time identifying atrial fibrillation by using a lightweight convolutional neural network. *IEEE access* 7, 130074–130084. doi:10.1109/access.2019.2939822

Larburu, N., Lopetegi, T., and Romero, I. (2011). "Comparative study of algorithms for atrial fibrillation detection," in *2011 computing in cardiology* (IEEE), 265–268.

LeCun, Y., and Bengio, Y. (1995). "Convolutional networks for images, speech, and time series," in *The handbook of brain theory and neural networks*, 3361, 1995.

Li, Q., Liu, C., Li, Q., Shashikumar, S. P., Nemati, S., Shen, Z., et al. (2019). Ventricular ectopic beat detection using a wavelet transform and a convolutional neural network. *Physiol. Meas.* 40, 055002. doi:10.1088/1361-6579/ab17f0

Lian, J., Wang, L., and Muessig, D. (2011). A simple method to detect atrial fibrillation using rr intervals. *Am. J. Cardiol.* 107, 1494–1497. doi:10.1016/j.amjcard.2011.01.028

Limam, M., and Precioso, F. (2017). "Atrial fibrillation detection and ecg classification based on convolutional recurrent neural network," in *2017 computing in cardiology (CinC)* (IEEE), 1–4.

Lubitz, S. A., Moser, C., Sullivan, L., Rienstra, M., Fontes, J. D., Villalon, M. L., et al. (2013). Atrial fibrillation patterns and risks of subsequent stroke, heart failure, or death in the community. *J. Am. Heart Assoc.* 2, e000126. doi:10.1161/JAHA.113.000126

Luo, X., Yang, L., Cai, H., Tang, R., Chen, Y., and Li, W. (2021). Multi-classification of arrhythmias using a hcrnet on imbalanced ecg datasets. *Comput. Methods Programs Biomed.* 208, 106258. doi:10.1016/j.cmpb.2021.106258

Mian Qaisar, S. (2020). Baseline wander and power-line interference elimination of ecg signals using efficient signal-piloted filtering. *Healthc. Technol. Lett.* 7, 114–118. doi:10.1049/htl.2019.0116

Mousavi, S., Afghah, F., Razi, A., and Acharya, U. R. (2019). "Ecgnet: learning where to attend for detection of atrial fibrillation with deep visual attention," in 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (IEEE), 1–4.

Murat, F., Sadak, F., Yildirim, O., Talo, M., Murat, E., Karabatak, M., et al. (2021). Review of deep learning-based atrial fibrillation detection studies. *Int. J. Environ. Res. public health* 18, 11302. doi:10.3390/ijerph182111302

Nankani, D., and Baruah, R. D. (2022). "Atrial fibrillation classification and prediction explanation using transformer neural network," in 2022 International Joint Conference on Neural Networks (IJCNN) (IEEE), 01–08.

Nesheiwat, Z., Goyal, A., and Jagtap, M. (2023). Atrial fibrillation. Available at: https://www.ncbi.nlm.nih.gov/books/NBK526072/ (Accessed February 22, 2024).

Ngui, W. K., Leong, M. S., Hee, L. M., and Abdelrhman, A. M. (2013). Wavelet analysis: mother wavelet selection methods. *Appl. Mech. Mater.* 393, 953–958. 10.4028/www.scientific.net/amm.393.953.

Nurmaini, S., Tondas, A. E., Darmawahyuni, A., Rachmatullah, M. N., Partan, R. U., Firdaus, F., et al. (2020). Robust detection of atrial fibrillation from short-term electrocardiogram using convolutional neural networks. *Future Gener. Comput. Syst.* 113, 304–317. doi:10.1016/j.future.2020.07.021

Oster, J., and Clifford, G. D. (2015). Impact of the presence of noise on rr interval-based atrial fibrillation detection. *J. Electrocardiol.* 48, 947–951. doi:10.1016/j.jelectrocard.2015.08.013

Park, J., An, J., Kim, J., Jung, S., Gil, Y., Jang, Y., et al. (2022). Study on the use of standard 12-lead ecg data for rhythm-type ecg classification problems. *Comput. Methods Programs Biomed.* 214, 106521. doi:10.1016/j.cmpb.2021.106521

Park, J., Lee, S., and Jeon, M. (2009). Atrial fibrillation detection by heart rate variability in poincare plot. *Biomed. Eng. online* 8, 38–12. doi:10.1186/1475-925X-8-38

Petmezas, G., Haris, K., Stefanopoulos, L., Kilintzis, V., Tzavelis, A., Rogers, J. A., et al. (2021). Automated atrial fibrillation detection using a hybrid cnn-lstm network on imbalanced ecg datasets. *Biomed. Signal Process. Control* 63, 102194. doi:10.1016/j.bspc.2020.102194

Phukan, N., Manikandan, M. S., and Pachori, R. B. (2023). Afibri-net: a lightweight convolution neural network based atrial fibrillation detector. *IEEE Trans. Circuits Syst. I Regul. Pap.* 70, 4962–4974. doi:10.1109/tcsi.2023.3303936

PhysioBank, P., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, e215–e220. doi:10.1161/01.cir.101.23.e215

PIA-Group (2021). Biosppy: a toolbox for biosignal processing in python. Available at: https://github.com/PIA-Group/BioSPPy.

Qiu, J., Zhu, J., Rosenberg, M., Liu, E., and Zhao, D. (2022). Optimal transport based data augmentation for heart disease diagnosis and prediction. arXiv preprint arXiv:2202.00567.

Ramaraj, E., and Clement Virgeniya, S. (2021). A novel deep learning based gated recurrent unit with extreme learning machine for electrocardiogram (ecg) signal recognition. *Biomed. Signal Process. Control* 68, 102779. doi:10.1016/j.bspc.2021.102779

Ramkumar, M., Kumar, R. S., Manjunathan, A., Mathankumar, M., and Pauliah, J. (2022). Auto-encoder and bidirectional long short-term memory based automated arrhythmia classification for ecg signal. *Biomed. Signal Process. Control* 77, 103826. doi:10.1016/j.bspc.2022.103826

Rohr, M., Reich, C., Höhl, A., Lilienthal, T., Dege, T., Plesinger, F., et al. (2022). Exploring novel algorithms for atrial fibrillation detection by driving graduate level education in medical machine learning. *Physiol. Meas.* 43, 074001. doi:10.1088/1361-6579/ac7840

Schwab, P., Scebba, G. C., Zhang, J., Delai, M., and Karlen, W. (2017). "Beat by beat: classifying cardiac arrhythmias with recurrent neural networks," in *2017 computing in cardiology (CinC)* (IEEE), 1–4.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.*1556*

Tateno, K., and Glass, L. (2001). Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of rr and δrr intervals. *Med. Biol. Eng. Comput.* 39, 664–671. doi:10.1007/BF02345439

Tutuko, B., Nurmaini, S., Tondas, A. E., Rachmatullah, M. N., Darmawahyuni, A., Esafri, R., et al. (2021). Afibnet: an implementation of atrial fibrillation detection with convolutional neural network. *BMC Med. Inf. Decis. Mak.* 21, 216–217. doi:10.1186/s12911-021-01571-1

Ullah, A., Rehman, S. u., Tu, S., Mehmood, R. M., and Ehatisham-Ul-Haq, M. (2021). A hybrid deep cnn model for abnormal arrhythmia detection based on cardiac ecg signal. *Sensors* 21, 951. doi:10.3390/s21030951

University of Washington (2024). Learn more – global cardiovascular health program. Available at: https://depts.washington.edu/globalcardio/about/learn-more/ (Accessed February 22, 2024).

van Wyk, F., Khojandi, A., Williams, B., MacMillan, D., Davis, R. L., Jacobson, D. A., et al. (2019). A cost-benefit analysis of automated physiological data acquisition systems using data-driven modeling. *J. Healthc. Inf. Res.* 3, 245–263. doi:10.1007/s41666-018-0040-y

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. neural Inf. Process. Syst.* 30.

Wang, J., and Li, W. (2020). Atrial fibrillation detection and ecg classification based on cnn-bilstm. arXiv preprint arXiv:2011.06187

Wang, M., Rahardja, S., Fränti, P., and Rahardja, S. (2023a). Single-lead ecg recordings modeling for end-to-end recognition of atrial fibrillation with dual-path rnn. *Biomed. Signal Process. Control* 79, 104067. doi:10.1016/j.bspc.2022.104067

Wang, T., Lu, C., Sun, Y., Yang, M., Liu, C., and Ou, C. (2021). Automatic ecg classification using continuous wavelet transform and convolutional neural network. *Entropy* 23, 119. doi:10.3390/e23010119

Wang, Z., Liu, C., and Yao, B. (2022). "Multi-branching neural network for myocardial infarction prediction," in 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE) IEEE, 2118–2123.

Wang, Z., Stavrakis, S., and Yao, B. (2023b). Hierarchical deep learning with generative adversarial network for automatic cardiac diagnosis from ecg signals. *Comput. Biol. Med.* 155, 106641. doi:10.1016/j.compbiomed.2023.106641

Wang, Z., and Yao, B. (2021). Multi-branching temporal convolutional network for sepsis prediction. *IEEE J. Biomed. Health Inf.* 26, 876–887.

World Health Organization (2024). Cardiovascular diseases. Available at: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (Accessed February 22, 2024).

Wu, Y., Yang, F., Liu, Y., Zha, X., and Yuan, S. (2018). A comparison of 1-d and 2-d deep convolutional neural networks in ecg classification. arXiv preprint arXiv:1810.07088

Wu, Z., Lan, T., Yang, C., and Nie, Z. (2019). A novel method to detect multiple arrhythmias based on time-frequency analysis and convolutional neural networks. *IEEE Access* 7, 170820–170830. doi:10.1109/access.2019.2956050

Xia, Y., Wulan, N., Wang, K., and Zhang, H. (2018). Detecting atrial fibrillation by deep convolutional neural networks. *Comput. Biol. Med.* 93, 84–92. doi:10.1016/j.compbiomed.2017.12.007

Xie, J., and Yao, B. (2022a). Physics-constrained deep active learning for spatiotemporal modeling of cardiac electrodynamics. *Comput. Biol. Med.* 146, 105586. doi:10.1016/j.compbiomed.2022.105586

Xie, J., and Yao, B. (2022b). Physics-constrained deep learning for robust inverse ecg modeling. *IEEE Trans. Automation Sci. Eng.* 20, 151–166. doi:10.1109/tase.2022.3144347

Xie, J., and Yao, B. (2023). Hierarchical active learning for defect localization in 3d systems. *IISE Trans. Healthc. Syst. Eng.*, 1–15. doi:10.1080/24725579.2023.2233992

Yao, B., Chen, Y., and Yang, H. (2021). Constrained markov decision process modeling for optimal sensing of cardiac events in mobile health. *IEEE Trans. Automation Sci. Eng.* 19, 1017–1029. doi:10.1109/tase.2021.3052483

Yao, B., and Yang, H. (2016). Physics-driven spatiotemporal regularization for high-dimensional predictive modeling: a novel approach to solve the inverse ecg problem. *Sci. Rep.* 6, 39012–39013. doi:10.1038/srep39012

Yao, B., and Yang, H. (2020). Spatiotemporal regularization for inverse ecg modeling. *IISE Trans. Healthc. Syst. Eng.* 11, 11–23. doi:10.1080/24725579.2020.1823531

Yao, B., Zhu, R., and Yang, H. (2017). Characterizing the location and extent of myocardial infarctions with inverse ecg modeling and spatiotemporal regularization. *IEEE J. Biomed. health Inf.* 22, 1445–1455. doi:10.1109/JBHI.2017.2768534

Ye, C., Kumar, B. V., and Coimbra, M. T. (2012). Heartbeat classification using morphological and dynamic features of ecg signals. *IEEE Trans. Biomed. Eng.* 59, 2930–2941. doi:10.1109/TBME.2012.2213253

Zhang, H., Gu, H., Gao, J., Lu, P., Chen, G., and Wang, Z. (2022). An effective atrial fibrillation detection from short single-lead electrocardiogram recordings using mcnn-blstm network. *Algorithms* 15, 454. doi:10.3390/a15120454

TYPE Original Research
PUBLISHED 30 May 2024
DOI 10.3389/fphys.2024.1399374

OPEN ACCESS

EDITED BY
Nathan Gaw,
Air Force Institute of Technology, United States

REVIEWED BY
Teresa Wu,
Arizona State University, United States
Elif Konyar,
University of Florida, United States

*CORRESPONDENCE
Nan Kong,
✉ nkong@purdue.edu

RECEIVED 11 March 2024
ACCEPTED 10 May 2024
PUBLISHED 30 May 2024

CITATION
Stanik M, Hass Z and Kong N (2024), Seizure prediction in stroke survivors who experienced an infection at skilled nursing facilities—a machine learning approach.
Front. Physiol. 15:1399374.
doi: 10.3389/fphys.2024.1399374

Seizure prediction in stroke survivors who experienced an infection at skilled nursing facilities—a machine learning approach

Madeleine Stanik[1], Zachary Hass[2] and Nan Kong[1]*

[1]Purdue University, Department of Engineering, Weldon School of Biomedical Engineering, West Lafayette, IN, United States, [2]Purdue University, Schools of Industrial Engineering and Nursing, West Lafayette, IN, United States

Background: Infections and seizures are some of the most common complications in stroke survivors. Infections are the most common risk factor for seizures and stroke survivors that experience an infection are at greater risk of experiencing seizures. A predictive model to determine which stroke survivors are at the greatest risk for a seizure after an infection can be used to help providers focus on prevention of seizures in higher risk residents that experience an infection.

Methods: A predictive model was generated from a retrospective study of the Long-Term Care Minimum Data Set (MDS) 3.0 (2014–2018, n = 262,301). Techniques included three data balancing methods (SMOTE for up sampling, ENN for down sampling, and SMOTEENN for up and down sampling) and three feature selection methods (LASSO, Recursive Feature Elimination, and Principal Component Analysis). One balancing and one feature selection technique was applied, and the resulting dataset was then trained on four machine learning models (Logistic Regression, Random Forest, XGBoost, and Neural Network). Model performance was evaluated with AUC and accuracy, and interpretation used SHapley Additive exPlanations.

Results: Using data balancing methods improved the prediction performances of the machine learning models, but feature selection did not remove any features and did not affect performance. With all models having a high accuracy (76.5%–99.9%), interpretation on all four models yielded the most holistic view. SHAP values indicated that therapy (speech, physical, occupational, and respiratory), independence (activities of daily living for walking, mobility, eating, dressing, and toilet use), and mood (severity score, anti-anxiety medications, antidepressants, and antipsychotics) features contributed the most. Meaning, stroke survivors who received fewer therapy hours, were less independent, had a worse overall mood were at a greater risk of having a seizure after an infection.

Conclusion: The development of a tool to predict seizure following an infection in stroke survivors can be interpreted by providers to guide treatment and prevent complications long term. This promotes individualized treatment plans that can increase the quality of resident care.

KEYWORDS

stroke, seizure, infection, machine learning, binary classification, minimum data set, skilled nursing facility

1 Introduction

For the past decade, stroke has ranked in the top five leading causes of death in the United States (US) (Ahmad and Anderson, 2021; Heron, 2021; Shiels et al., 2022). Stroke related deaths account for 4.7% of deaths across all age groups and 6.1% of deaths in aging populations classified as age 65 and older (Shiels et al., 2022). However, not all strokes are fatal and 60% of ischemic stroke patients and 38% of hemorrhage stroke patients survive the first year (Smajlović et al., 2006). Patients that survive often face serious complications or disabilities. In fact, stroke is the leading cause of serious long-term disability in the United States and each year accounts for about $56.5 billion dollars (CDC, 2023).

Within the last few years, the number of stroke related deaths has been decreasing (Chohan et al., 2019). With this increased survival rate, there has been an increase in the number of patients with complications. The major complications include recurrent stroke (9% of patients), epileptic seizure (3%), urinary tract infection (24%), chest infection (22%), other infections (19%), falls (25%), shoulder pain (9%), other pain (34%), depression (16%), anxiety (14%), emotionalism (12%), and confusion (56%) (Langhorne et al., 2000). By focusing on the prevention of these complications, the long-term survival rate and quality of life for stroke survivors can be improved.

It has been well documented that infections are a leading risk factor for seizures and epilepsy (Vezzani et al., 2015). However, there has not been extensive research into how infections impact seizure risk in stroke survivors. Stroke survivors are especially prone to both bacterial and viral infections, and having these infections may consequently increase their seizure risk (Langhorne et al., 2000). Having frequent infections and seizures could severely postpone the patient's recovery process and possibly result in death. Exploring this coupling of complications could help prevent adverse effects by placing a stronger emphasis on limiting infection and preventing seizure in patients who have already had an infection.

To help prevent infection and subsequent seizure, focusing on the patient's recovery through their rehabilitation plan is a promising pathway. When a stroke survivor is discharged from the hospital or other treatment facilities to a skilled nursing facility (SNF), they will begin rehabilitation following a set plan (Bindawas and Vennu, 2016). The effectiveness of this set plan at the SNF relies heavily on the team of professionals that goes into making it (Lenze et al., 2012). In fact, it has been shown that rehabilitation plans made by a group of professionals are more effective than those made by a single professional (Graham, 2013). In addition, if the team takes the time to specialize the plan, it has been shown that the patient will have a faster recovery rate and yield better functional outcomes (Bindawas and Vennu, 2016). Other studies have also shown that specialized plans yield greater participant engagement with activities being completed at higher intensities (Lenze et al., 2012). These specialized rehabilitation plans are typically variations of a standardized version and vary depending on the patient's severity of complications and response to the therapy (Bernhardt et al., 2016). However, plans are adjusted by healthcare professionals using intuition rather than numerical feedback, which leads to plans that fail to help patients reach their recovery goal (Levinson, 2013). If the plans were individualized and a patient's response to changes in the plan could be measured with concrete numerical evidence, then the outcome of recovery could improve for stroke survivors.

Additionally, it has been shown that stroke survivors at nursing facilities receive fewer hours of rehabilitation compared to hospital settings (Koopmans et al., 2010). This is typically a result of the reduction in staffing and intensity of care, but receiving more therapy hours has been associated with increased independence (Jette et al., 2005), greater likelihood of discharge from SNF to community (Jette et al., 2004; Jung et al., 2016), and greater functional improvements (Chen et al., 2002). This means that residents at SNFs could benefit from an increase in therapy hours as part of their rehabilitation. With stroke survivor rehabilitation plans typically lasting between a few months to a few years (Bindawas and Vennu, 2016), this is considered long-term rehabilitation (IHCP, 2023). Assessing the relationship between the number of therapy hours in a rehabilitation plan and the risk of seizure following an infection could yield beneficial results in resident recovery.

This study used the Long-Term Care Minimum Data Set (MDS) 3.0 (2014–2018) in a midwestern US state to retrospectively investigate the risk of seizure following an infection both short term and long term. By focusing on the stroke to infection to seizure pathway, this study seeks to identify risk factors for seizure after an infection to then help limit seizures in stroke survivors who have experienced an infection. The model is fit to predict the risk, return an individualized resident risk estimate, and interpret which factors contribute the most to this risk estimate. Uncovering which factors contribute the most to seizure risk may aid healthcare professionals in adjusting rehabilitation plans to improve resident outcomes.

Other studies have predicted the risk of seizure in stroke surviving patients (Bunney et al., 2022; Looti et al., 2023; Lekoubou et al., 2024); however, none exist that include the infection to seizure pathways in stroke survivors. Another novel aspect is the use of the MDS data set for prediction of seizures in stroke survivors, which has not even been used for seizure prediction. The third novel aspect of this study is the use of SHapley Additive exPlanations (SHAP) for model interpretations, and though this technique was developed a number of years ago, its application to the healthcare space is relatively novel.

2 Materials and methods

This study had two specific aims for investigating the risk of seizure following infection in stroke survivors at skilled nursing facilities (SNFs).

1. Determine the risk ratio of stroke survivors experiencing a seizure after an infection short term (within 14 days) and long term (within 1 year).
2. Interpret the resultant predictive models to identify risk factors for a stroke surviving nursing home resident experiencing a seizure within 14 days following an infection.

The data initially includes all individuals admitted to a Medicare and Medicaid licensed SNF between 1 January 2014 to 20 April 2018 in Indiana taken from the Long Term Care Minimum Data Set. All assessments during the time period were evaluated. The main

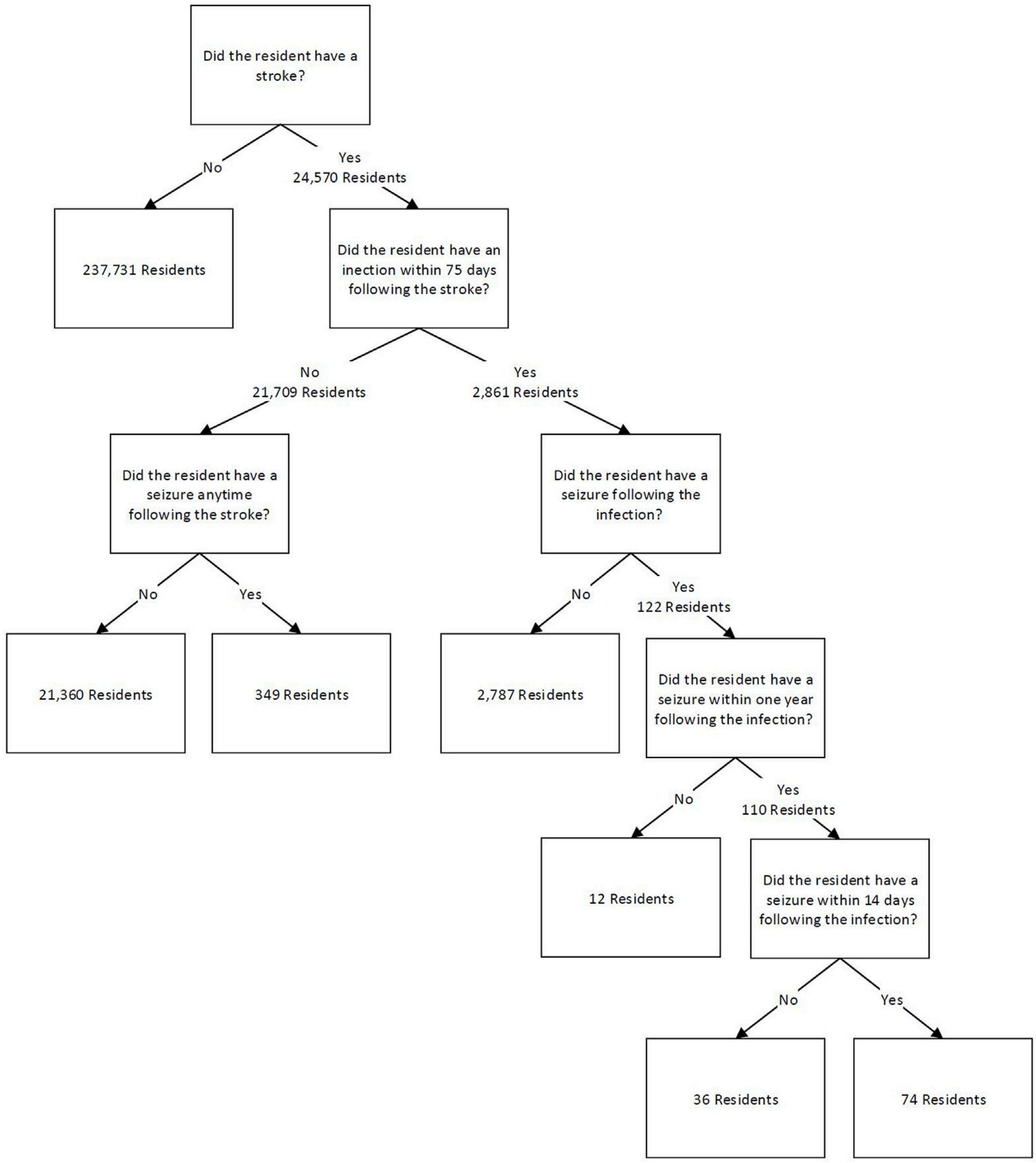

Did the resident have a stroke?
No
Yes
24,570 Residents
237,731 Residents
Did the resident have an inection within 75 days following the stroke?
No
21,709 Residents
Yes
2,861 Residents
Did the resident have a seizure anytime following the stroke?
Did the resident have a seizure following the infection?
No
Yes
No
Yes
122 Residents
21,360 Residents
349 Residents
2,787 Residents
Did the resident have a seizure within one year following the infection?
No
Yes
110 Residents
12 Residents
Did the resident have a seizure within 14 days following the infection?
No
Yes
36 Residents
74 Residents

FIGURE 1
Resident categorization flow chart.

data features include demographic, diagnosis, activities of daily living (ADL), pain, treatment, mobility, and therapy. Residents with a previous history of seizure and epilepsy disorder were excluded in order to establish the temporal association between stroke and seizure occurrence.

2.1 Risk ratio

Prior to modeling, a preliminary analysis was conducted to verify the relationship between stroke survivors, infections, and seizures. Stroke survivors considered were nursing facility residents with the stroke diagnosis code who remained in a skilled nursing facility (SNF) after the incident. This included both residents admitted with a stroke diagnosis and those who had a stroke while in the SNF. An infection was said to have occurred if any of the urinary tract infection, pneumonia, sepsis, tuberculosis, viral hepatitis, wound infection, and multidrug resistant organism diagnosis codes were noted in an assessment after the stroke noted assessment. A seizure was said to occur if the diagnosis code for seizure and epilepsy disorder was noted in an assessment after the infection noted assessment.

Assessments from stroke survivors were used to count the number of unique residents for four mutually exclusive categories. Divisions were based on the occurrence of an infection and/or stroke. The initial data were first split on the occurrence of stroke in resident assessments, which yielded 24,570 stroke survivor residents. The data was then split on whether a resident had an infection within 75 days following the stroke, a time period associated with increased risk of disability and

TABLE 1 Risk ratio of seizure in stroke survivors with infections.

In Stroke Survivors	Relative Risk Ratio	95% Confidence Interval
Experiencing a seizure within 14 days after an infection	1.1968	[0.9344, 1.5330]
Experiencing a seizure within a year after an infection	2.4168	[1.9604, 2.9795]

TABLE 2 Risk ratio calculations of seizure in stroke survivors with infections.

In Stroke Survivors	Relative Risk Ratio	95% Confidence Interval
Experiencing a seizure within 14 days after an infection	$\frac{\frac{74}{(74+2787+12+36-135)}}{\frac{349}{(349+21360-6051)}}$	$e^{\ln(1.5832)\pm1.96\sqrt{\frac{1}{74}+\frac{1}{2700}+\frac{1}{349}+\frac{1}{15309}}}$
Experiencing a seizure within a year after an infection	$\frac{\frac{110}{(110+2787+12-867)}}{\frac{349}{(349+21360-6051)}}$	$e^{\ln(2.6525)\pm1.96\sqrt{\frac{1}{110}+\frac{1}{1932}+\frac{1}{349}+\frac{1}{15309}}}$

death (Finlayson et al., 2011; Ulm et al., 2012; Learoyd et al., 2017). It was found that 2,861 unique residents had an infection within 75 days following a stroke. These could have been a resident who had a stroke in the SNF and then had an infection within 75 days following, or a resident who was admitted to the SNF with the stroke diagnosis and had an infection within 75 days from their first assessment. For the latter, this means the resident entered the facility with the stroke diagnosis and the infection threshold was set to within 75 days from the first assessment date. An additional 21,709 residents did not have an infection following a stroke.

These two groups were each split into two groups based on whether residents had a seizure following their stroke. The next two categories are sub-divisions of the first category based on the timing of the appearance of the seizure diagnosis. For the stroke survivors that had an infection within 75 days following their stroke, it was further evaluated if the resident had a seizure anytime following the infection or within 14 days following the infection. This value of 14 days was obtained from a study that found that seizures usually occur within one to 2 weeks following an infection, so 14 days was chosen based on the 2-week mark (Vezzani et al., 2015). Other studies have also found that seizures can occur after a stroke over 5 years later (Naess et al., 2004; Myint et al., 2006), so a long-term value after infection was also assessed for comparison as part of this risk ratio. For this study, this value was 1 year after the infection. The long-term follow-up period of 1 year had 110 residents experience a seizure following an infection, and the short term, 14 day follow up period, had 74 residents experience a seizure following an infection. For stroke survivors that did not experience an infection, it was determined that 349 residents had a seizure any amount of time following a stroke with no reported infection prior to the seizure. Figure 1 demonstrates these groups and their breakdown as a flow chart.

Residents admitted near the end of the dataset were removed if they did not experience a post-infection seizure and there was not an adequate number of days to observe the full follow-up period (right censoring of data). As an example, for 14-day post-infection seizure, a resident who had only 12 days of follow up in the data, but experienced an infection then had a seizure during that follow-up was kept in the data. A resident with only 12 days of follow up who did not have a seizure before the end of the dataset, however, was removed due to right censoring of the data. For 14-day follow-up, 135 residents were right censored, and for 1-year follow-up, 867 residents were right censored. Residents were also right censored following the same method for the 75-day follow-up period between stroke and infection. This latter group had 6,051 residents with right censoring of their data. These censored residents were subtracted from risk ratio calculations and were removed from the predictive models.

The categorized residents and their corresponding prevalence were used to calculate risk ratios based on the number of unique residents who experienced a seizure. Using unique residents reduced the possibility of carrying forward diagnoses in the data between assessments that could have artificially inflated occurrences. Therefore, the number of unique residents is a more robust method compared to the number of occurrences for calculating the risk ratio here. For a more detailed explanation, please see the discussion section. The ratios in Table 1 indicate that having an infection within 75 days after a stroke increases a resident's risk of having a seizure within 14 days post infection by 1.20-fold. Having an infection within 75 days after a stroke increases a resident's risk of having a seizure 1 year post infection by 2.42-fold. Risk ratios were calculated by comparing the population of individuals who experienced a seizure following an infection to those who experienced a seizure without first experiencing an infection (Table 2). The 95% confidence intervals did not contain one for the risk ratio of the 1-year follow-up period, indicating that the relative risk ratio was found to be statistically significant. This is consistent with current literature that indicates that infections increase the risk of seizures (Langhorne et al., 2000). The 14-day follow-up period's 95% confidence interval for the risk ratio did contain 1, so this risk ratio was not found to be statistically significant. However, a large proportion of the post-infection seizures occurred within this time frame and adjustment for additional features can be informative, so modeling was also completed for prediction of seizure over the 14-day follow-up period.

2.2 Data processing and modeling

The MDS data collection instrument includes 23 sections that contain information such as demographics, diagnoses, independence in

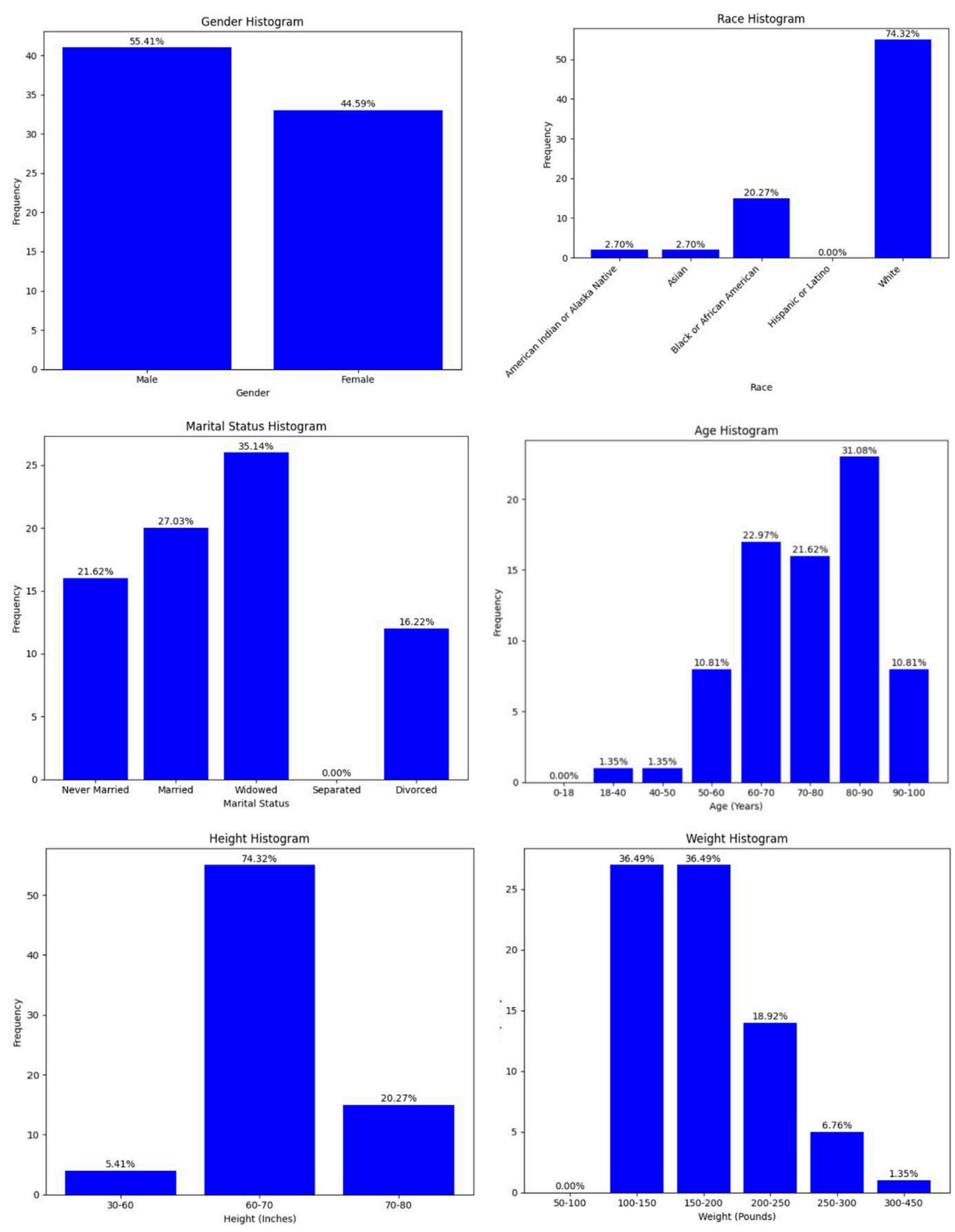


FIGURE 2
Demographic breakdown for fourteen-day risk.

performing activities of daily living, mood assessment, therapy, and medications by class. Each section contains data on all residents, and most residents had multiple entries in the dataset. These multiple entries were a result of periodic assessments (e.g., 5-day, 14-day, 30-day, 60-day, or 90-day post admission for Medicare Part A stays; admission, quarterly, and significant change in status for other stay types) that varied by resident when information would be updated. The date of the assessment was noted, and a de-identified person number was used to associate residents to all their assessments. The data was structured with the same number of rows appearing across all sections, but each section

TABLE 3 Model performance metrics for fourteen-day risk prediction.

Parameter	Logistic Regression	XGBoost	Neural Network	Random Forest
AUC	0.8380	0.9999	0.9988	0.9999
Accuracy	0.7654	0.9999	0.9991	0.9998
Recall	0.7838	0.9999	0.9999	0.9997
True Positive Rate (TPR)	0.7838	0.9999	0.9999	0.9997
True Negative Rate (TNR)	0.7468	1.0000	0.9981	1.0000
Sensitivity	0.7838	0.9999	0.9999	0.9997
Specificity	0.7468	1.0000	0.9981	1.0000
Positive Predictive Values (PPV)	0.7571	1.0000	0.9981	1.0000
Negative Predictive Values (NPV)	0.7743	0.9999	0.9999	0.9997
Precision	0.7571	1.0000	0.9981	1.0000

had a variable number of columns. The rows for each section match up directly by row index, so any row across all sections were the same assessment for the same resident. The columns were different sets of features broken up by sections, and within each section columns were related. For example, the therapy section contains columns for speech, occupational, physical, and other types of therapy. For this analysis, 149 features were selected from the thousands of features across the 23 sections.

These features included demographics (age, gender, marital status, race, height, and weight), treatments (physical therapy, occupational therapy, speech therapy, recreational therapy, psychological therapy, and medications), physical condition (daily activities, mobility, balance), and behavior (mood, pain, and delirium). These were the main feature groups, and all features were composed of more specific subgroups within these main groups. For example, in the category of physical therapy, there were variables on weekly individual minutes, concurrent minutes, group minutes, and number of days of therapy per week. The selection of these features followed other stroke survivor outcomes studies (Kelly-Hayes et al., 1998; Gittins et al., 2020).

Features with more than 70% missing values were removed (29 features removed). With 149 features to start, removing these 29 features reduced the total to 120. The remaining missing values were imputed using a two-step process. First, the resident's most recent value from a prior or future assessment was carried forward or backwards. For example, if age was missing but a resident's record from the previous month contained their age to be 65, the missing record was filled in with 65. For some features, no records were present for any entries, so as the second step, these remaining missing values were imputed with the k nearest neighbors method using five nearest neighbors. Missing values in diagnosis codes such as stroke, seizure, and infection were imputed with a zero indicating that event did not occur to prevent possible misdiagnosis or error carried forward. Dropping features with 70% missingness and using kth nearest neighbors with five neighbors is relatively common in healthcare datasets where missingness is relatively high (Wells et al., 2013; Salgado et al., 2016; Jäger et al., 2021). The use of 70% is on the higher end of what is found in the literature but was used as a matter of practicality. If the missingness cut off is set too low, a large proportion of data will be removed. Imputation using the resident's most recent value from another assessment and imputation of diagnostics with zeros was author determined. Imputation using the resident's other assessment could cause slight discrepancies, such as when imputing age, the method does not consider the resident's birthday (data element not available for this work), but the imputed value is still likely to be very near the true value and resident specific.

Following imputation of missing data, the data was balanced using three methods. These methods were the Synthetic Minority Oversampling Technique (SMOTE) for up sampling, Edited Nearest Neighbor (ENN) for down sampling, and SMOTEENN for up and down sampling. Applying each balancing technique resulted in three sets of balanced data that then underwent three feature selection methods: Least Absolute Shrinkage and Selection Operator (LASSO), Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA). These methods were selected from other studies that aimed to predict seizures post-stroke (Bunney et al., 2022; Looti et al., 2023; Lekoubou et al., 2024). These studies did not consider infections post-stroke; however, incorporation of post-stroke infections is not expected to significantly impact the results of feature selection methods. A balancing and feature selection technique was then chosen to apply to four different modeling methods: Logistic Regression, Random Forest, XGBoost, and Neural Network. Logistic regression was chosen for its distinction as one of the most fundamental modeling methods due to its linearity assumption, low computational intensity, and parametric interpretability. XGBoost and Random Forest were chosen due to their non-linear nature and ability to guard against underfitting and overfitting respectively. Neural Network was chosen because it is also non-linear and is not a tree-based model making for more interesting model comparisons and it has a strong ability to handle more complex relationships.

Hyper parameters for Logistic Regression (penalty: L1, L2 and C: 0.01, 0.1, 1, 10), XGBoost (learning rate: 0.1, 0.01, 0.001; and maximum depth: 1, 5, 10, 20), Random Forest (maximum depth:

TABLE 4 K fold cross validation scores for fourteen-day risk prediction.

Cross Validation Scores	Logistic Regression	XGBoost	Neural Network	Random Forest
Score 1	0.7687	0.9997	0.9971	0.9993
Score 2	0.7796	0.9998	0.9956	0.9998
Score 3	0.7623	0.9997	0.9978	0.9998
Score 4	0.7707	0.9998	0.9970	0.9998
Score 5	0.7745	0.9996	0.9986	0.9994
Average CV Score	0.7721	0.9997	0.9972	0.9996

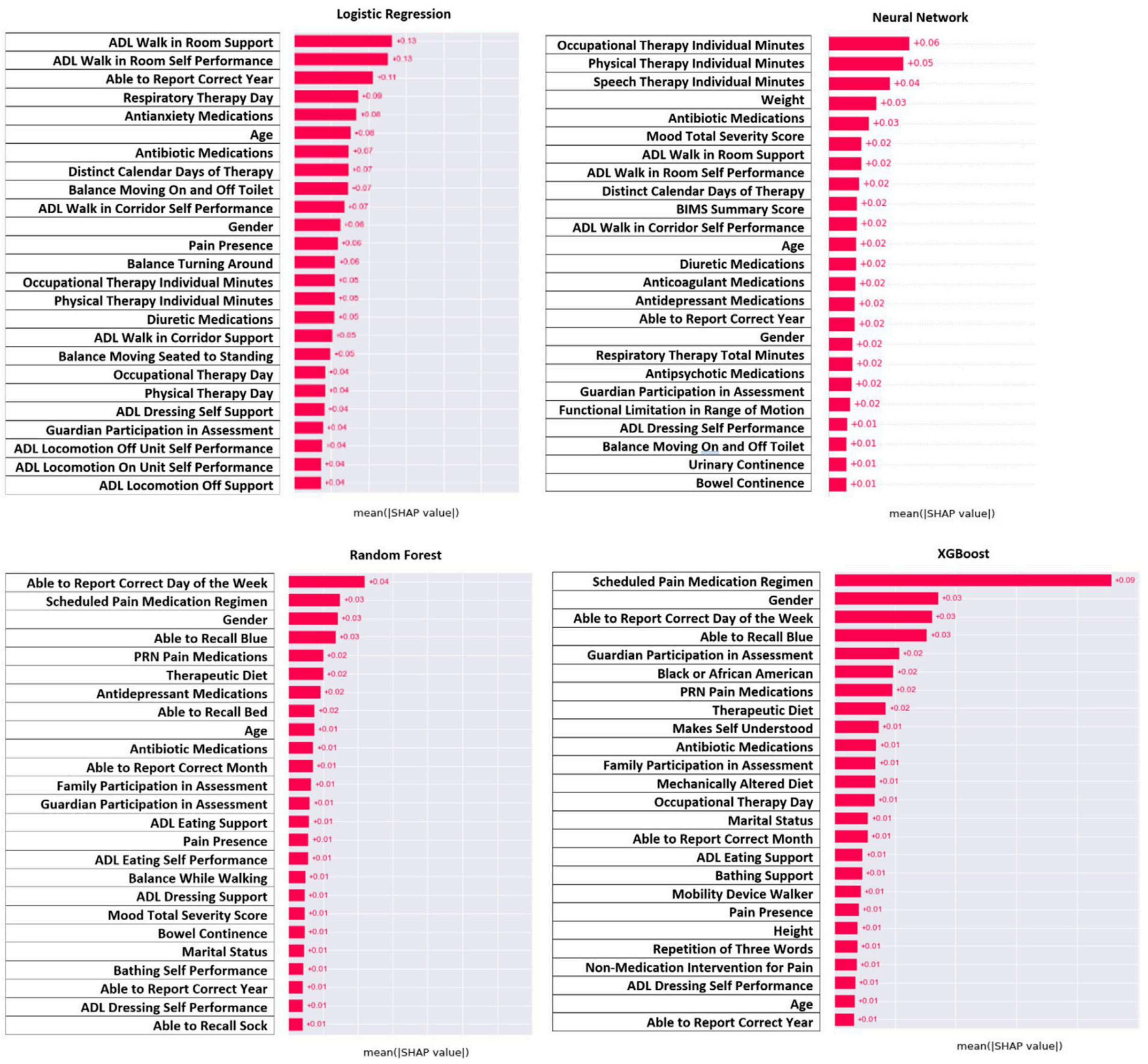


FIGURE 3
Shap values for top features to explain contribution to the model.

1, 5, 10, 20; and n estimators: 200, 1,000, 10,000), and Neural Network (maximum iterations: 100, 200; activation layer: logistic, tanh; and number of hidden layers: 2, 8, 64, 128) were tuned using GridSearchCV. This method used all combinations of hyperparameters within each model then chooses the one with the best specified metric, which in this case was ROC and AUC.

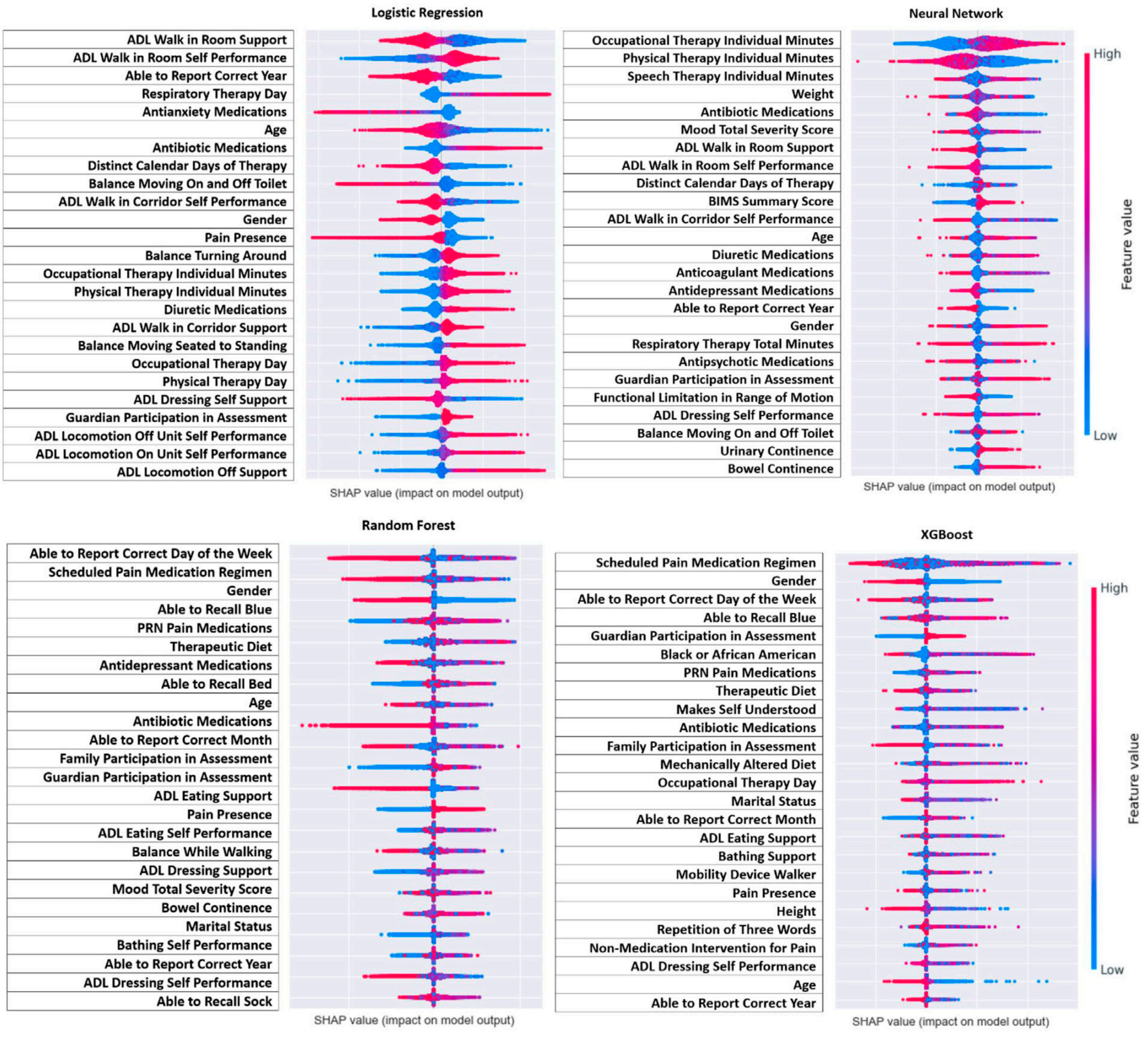


FIGURE 4
Direction of SHAP values for top features.

For example, the Logistic Regression method tested a total of six different models and chose the hyperparameters that yielded the greatest AUC of the six. For Logistic Regression, the selected hyperparameters were a penalty of L2 and a C of 0.1. The XGBoost model yielded 0.1 for the learning rate and 10 for maximum depth. Random Forest resulted in 10 for maximum depth and 1,000 for n estimators. Lastly, Neural Network chose 100 for maximum iterations, tanh for the activation layer, and 64 for the number of hidden layers.

Model performance was evaluated using prediction metrics such as Receiver Operator Curve Area Under the Curve (AUC), accuracy, recall, true positive rate, true negative rate, sensitivity, specificity, positive predictive values, negative predictive values, and precision. Data was split 80% and 20% for the training and testing set. Within the training set, 5-fold validation was used for tuning hyperparameters. The testing set was used to evaluate the model performance and interpretation. Model interpretation was evaluated with SHapley Additive exPlanations (SHAP).

3 Results

3.1 Demographics

Figure 2 shows the demographic breakdown of residents who suffered a seizure within 14 days post infection. From these demographics, older residents were more prevalent. The figure also shows that post-infection seizures were more prevalent in men, despite the MDS dataset containing primarily female residents. For the other demographic features, the trend follows those of the overall MDS dataset, so they are not as significant.

3.2 Model

The selected models used SMOTEENN for data balancing and all four models (Logistic Regression, XGBoost, Random

Forest, and Neural Network) were assessed for prediction quality and feature interpretation. The data balancing method was chosen based on the breakdown of the classes. For ENN, the distribution of no seizures after infection to seizures following infection was 99% and 1%, which was likely the result of highly imbalanced data that interrupts the methodology of ENN and did not allow for down sampling. For SMOTE, the distribution was 49% and 51%, and for SMOTEENN the distribution was 50% and 50%. Because SMOTEENN yielded the most equal distribution, this data balancing method was selected.

For feature selection, PCA was adjusted to a range of component numbers and used the same set of features as the other methods, but this method was not selected since it did not allow for the same degree of interpretability. For RFE, results suggested that no features needed to be removed from the model. RFE worked by fitting a model with all the features then ranked each feature depending on the contribution to the model. Features were then removed based on if they meaningfully contributed to the scoring metric, which was Area Under the Receiver Operator Curve (AUC) for this study. Results indicated no features were to be removed so all features meaningfully contributed to the AUC. For LASSO, 17 features were suggested to be removed from the model. The LASSO method worked by assigning a coefficient to each feature based on its contribution to the model then shrinking the coefficients using the selected regularization parameter alpha, in this case alpha was 0.00001. A cut off was set of 0.001, meaning any features with a coefficient smaller than this value (features whose coefficient shrunk to zero) would be removed. If no coefficients were reduced to zero, then no features were removed. However, we took the conservative approach of siding with RFE which gauged all features as important keeping all features in the models.

For modeling methods, Logistic Regression, XGBoost, Neural Network, and Random Forest all yielded accurate prediction results (Table 3). Overfitting was assessed through K-fold cross validation with five folds on the training set, and the result of the cross validation returned five scores also close in value and confirmed that these models were not overfit (Table 4). However, the high accuracy generated by XGBoost, Neural Network, and Random Forest may have been a result of data balancing, where up sampling created more distinct entries that were easier to predict.

3.3 Interpretation

For model interpretation, the features that contributed the most to the model were those with the greatest absolute value of the SHAP values. Figure 3 demonstrates the features with the greatest contribution (absolute SHAP value) whereas Figure 4 demonstrates the direction of that contribution (positive or negative). It was important to interpret results from all four models since all models had a strong prediction ability, and comparison between models could identify similar features. Across all four models, it can be seen that the amount of therapy a person receives, their ability to be independent, and their overall mood contributed the most to predicting seizure following infection (Figure 3). For therapy, this was in the form of the number of minutes for speech, occupational, and physical therapy as well as the distinct calendar days and frequency of the therapy. For independence, this was in the form of activities for daily living for walking, mobility, eating, and dressing. Finally, mood was categorized based on mood severity score and the use of medications like antidepressants, antianxiety, and antipsychotics. Other notable features include antibiotic medications, diuretic medications, therapeutic diet, continence, and recall ability. Demographics also contributed to the model with age and gender being the most prominent.

Figure 4 demonstrates the direction of impact for each feature. Some features present in one model indicate the opposite effect in another model, or the direction is challenging to distinguish. However, across all four models it appears that residents who receive more therapy (speech, occupational, and physical), had lower ADL scores (more independent), and had a lower mood severity score (more positive mood), and took mood related medications (antidepressant or antianxiety) had lower risk of post-infection seizure. Lower age and male gender were associated with higher risk, but this was not as consistent across models as other findings.

4 Discussion

The models achieved a prediction accuracy between 76.5% and 99.9% for whether a stroke survivor will experience a seizure after an infection. It is plausible that data imputation and synthetic data created by up sampling artificially improved these metrics leading to an overly optimistic view of model performance. In other words, with balancing having a focus primarily on up sampling, the number of entries in the dataset was synthetically increased. These synthetic entries could have caused the dataset entries to become more distinct, making it easier to predict post-infection seizures. Up sampling also caused there to be more data and was thus more computationally intensive for future steps. However, up sampling is meant to reduce the bias of the majority class by up sampling the minority class, so the computational intensity is a tradeoff for reduced bias. Testing the model on larger, national populations would help to minimize adverse effects caused by up sampling and validate the resulting high accuracy. By focusing on all four models and their interpretation, the goal is to make the results more generalizable to future nursing home residents. The short-term contribution of the model is the use of SHAP values which allow for model interpretation, furthering the understanding of the relative importance of risk factors. Understanding feature importance through the SHAP values can guide the development of strategies to mitigate the effect of seizure risk for high-risk stroke survivors experiencing an infection.

The three main types of features that contributed the most to predictions were therapy, independence in locomotion and activities of daily living, and overall mood. The features best used to interpret the model are those related to therapy minutes (speech, occupational, and physical), distinct calendar days of therapy, the independence score for activities of daily living, mood severity score, use of antidepressant medications, use of antianxiety medications, and use of antipsychotic medications. The SHAP values indicated that stroke survivors who received more therapy, were able to be more independent, and had a better overall mood were at a lower risk of seizure following an infection. These results align with literature that has suggested that adults with epilepsy who exercise regularly reduce their risk of seizures (Nakken et al., 1990; Mario Arida et al., 2010). This is also true of adults with epilepsy who remain in a better mood and experience less stress to reduce their seizure risk (Jackson and Turkington, 2005; Sawyer and Escayg, 2010; McKee and Privitera, 2016). Regarding independence in stroke survivors, a decrease in independence leads to decreased mood during the recovery process (Albanese et al., 2020). This could indicate that as stroke survivors recover and become more independent, they would improve their mood and subsequently reduce their seizure risk. Although stroke survivors are not the same as people with epilepsy, stroke survivors still experience neurological complications thus have a risk of seizures. Identifying these features of therapy, independence, and mood allows healthcare providers and researchers potential levers for those residents at greater risk.

These features can be determined early, within the first 2 weeks of a resident's admission to the SNF. When a resident enters a SNF, a therapy plan is set in place including the number of minutes of therapy they will receive each week. As time goes on, this plan will be updated to reflect their treatment needs, but from their admission assessment, physicians can estimate resident risk based on the number of minutes in the plan. Additionally, the resident receives scores for their mood severity and their independence during their first 14-day assessment. This would mean that residents who begin to show signs of less positive mood (have a high mood severity score), are more dependent (higher ADL scores), and receiving less therapy would be categorized as high risk. Providers could then identify these patients and determine if additional care is appropriate. Ultimately, the decision relies on the provider to take action to improve resident care, but this study helps contribute to the field of known risk factors. However, as this study is correlational and the studied population often have complex multi-morbid conditions, it is difficult to know whether the occurrence of therapy reduces the risk of seizure or if individuals with a risk of seizure are less able to receive therapy, or perhaps both. Disentangling this relationship will better inform resident care.

Other features that meaningfully contribute are those for antibiotic medications, antianxiety medications, PRN (pro re nata) pain medications, diuretic medications, gender, and therapeutic diet. Antibiotics are a less useful contribution since the use of antibiotics indicates an infection, which was already pre-established. Antianxiety, pain, and diuretic medications could once again be indicative of patient severity, but they could also indicate the possible presence of acute drug intoxication. Drug intoxication from antidepressants and pain medications has been found to cause seizures in patients with epilepsy (Chen et al., 2016), and it is possible that stroke survivors in SNF could also suffer from the same outcome. For the demographic factors, gender was previously discussed with the finding that males had a greater proportion of post infection seizure, and gender is also shown as a top contributing feature across models. Therapeutic diet was another feature found to contribute and represent the resident receiving altered meals to promote recovery. This feature likely represents a modifiable factor since lifestyle changes like diet tend to be important for promoting resident wellbeing.

In a simple reading of the results, if a resident is high risk, healthcare providers could enroll the resident in additional therapy time, encourage more independence, and focus on improving mood to reduce seizure risk. However, it is imperative to note that many important features may be signals of the severity of the resident's condition rather than levers that can be pulled to improve their condition. For example, the results show that more therapy minutes are associated with reduced risk of seizure following infection. However, it may be the case that residents who are physically able to have therapy have less severe complications following their stroke. The severity of a resident's condition following the stroke may be the influential factor underlying both the number of therapy minutes and the likelihood of seizure. Nevertheless, identifying these associations is valuable in furthering the discussion around improving post-stroke care in SNFs. That the model establishes associations rather than causal relationships should be considered as a limitation.

Aside from the resident post-stroke severity limitation, another limitation of the model is the way in which dates of the strokes, infections, and seizures are established. For determining the date of the event occurrence, the assessment date was used (or the date in which the SNF filled out the MDS data form). This is typically the practice for diagnosis dates for the MDS (Hua et al., 2021). However, it is likely that there is a lag between the time the event occurred and the assessment date. This means that a seizure could have occurred days prior but was not noted in the MDS until the date of the assessment. Therefore, the assumption was made that the lag time for all events to assessment was approximately the same. This would mean that the exact date of the event occurrence was not accurate, but the relation of the events to one another would be reasonably accurate. This assumption can be validated by the events in the MDS being chronological for individuals, such that the relation of events to one another is accurate (Mor et al., 2011). Thus, the time thresholds of 75 days between stroke and infection and 14 days between infection and seizure would be chronologically accurate with the exact time difference having some undetermined degree of error.

Similar to this relation between dates, there is also the possibility of error carried forward between assessments. For example, it was seen in the dataset that once a resident experienced a seizure, it often appeared in later consecutive assessments. This was caused by SNF staff using previous assessment data instead of reassessing the resident each time. The reason for this can be a variety of factors like understaffing, overcrowding, and distraction that cause staff to try to save time in completing assessments (Bowman, 2013). This made it challenging to distinguish repeated infections that then led to seizures in individual residents. As a result, unique residents were used for modeling, meaning that the residents could only have the stroke to infection to seizure pathway once. This caused the risk ratio to potentially indicate a smaller risk than if occurrences were evaluated. Additionally, this could have compromised some prediction and interpretation ability of the model. By only looking at one occurrence for each resident, it is possible that risk factors that contributed to a second or third occurrence would have been overlooked. This may also be the reason the model predicted so accurately if a first occurrence was easier to predict than subsequent occurrences. Although using unique residents may underestimate the risk ratio and miss risk factors in later occurrences, prevention of false post-infection seizures from error carried forward is more important. Risk factors can still be obtained by looking at unique residents, but using false post-infection seizures could skew results.

Even with the limitations of the model, it still serves as an effective tool for interpretation. Infections were associated with an increase in the risk of a seizure in stroke survivors through the calculation of a risk ratio and in the predictive model. The interpretability aspect of this model with SHAP allowed for the main factors that contribute to risk to be identified. These main factors that contribute to risk can help guide resident care. Looking into the future, this model and others in this research space could eventually be established to run in the background to continuously assess resident risk. Currently, the prediction aspect is not at the desired level for implementation into care, but with more iterations, the technology could eventually reach a high level. Hence these goals would be more suitable for long-term progress across the entire healthcare research field rather than individual study improvement. More realistically, implementation of the models on other datasets to confirm model performance and evaluate generalizability would be a more obtainable short-term goal. For an even longer-term goal bordering science fiction, having risk assessment across all resident diagnoses and all outcomes would be the greatest improvement for healthcare. For now, this individual resident risk and the interpretability of the model can help guide resident treatment to generate better outcomes for stroke survivors.

5 Conclusion

This machine learning model demonstrated a high degree of accuracy in predicting the occurrence of a seizure within 14 days following an infection in the population of stroke survivors at skilled nursing facilities. The interpretability of the model allowed for specific therapy, independence, and mood related features to be identified that are associated with the risk of seizure occurrence. This interpretability of the model can be used by healthcare providers to guide treatment decisions to prevent seizures in residents who suffered an infection.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Data Sharing Agreement. Requests to access these datasets should be directed to nkong@purdue.edu.

Ethics statement

The studies involving humans were approved by the Purdue University Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

MS: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. ZH: Conceptualization, Writing–review and editing. NK: Conceptualization, Data curation, Supervision, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Purdue Libraries Open Access Publishing Fund.

Acknowledgments

We thank Zhouyang Lou for his input and insight on this research. We also thank the Purdue Libraries Open Access Publishing Fund.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmad, F. B., and Anderson, R. N. (2021). The leading causes of death in the US for 2020. *JAMA* 325 (18), 1829–1830. doi:10.1001/jama.2021.5469

Albanese, A., Bartz-Overman, C., Parikh, T., and Thielke, S. (2020). Associations between activities of daily living independence and mental health status among Medicare managed care patients. *J. Am. Geriatr. Soc.* 68 (6), 1301–1306. doi:10.1111/jgs.16423

Bernhardt, J., Churilov, L., Ellery, F., Collier, J., Chamberlain, J., Langhorne, P., et al. (2016). Prespecified dose-response analysis for A very early rehabilitation trial (AVERT). *Neurology* 86 (23), 2138–2145. doi:10.1212/WNL.0000000000002459

Bindawas, S. M., and Vennu, V. S. (2016). Stroke rehabilitation. A call to action in Saudi Arabia. *Neurosci. J.* 21 (4), 297–305. doi:10.17712/nsj.2016.4.20160075

Bowman, S. (2013). Impact of electronic health record systems on information integrity: quality and safety implications. *Perspect. Health Inf. Manag.* 10, 1c.

Bunney, G., Murphy, J., Colton, K., Wang, H., Shin, H. J., Faigle, R., et al. (2022). Predicting early seizures after intracerebral hemorrhage with machine learning. *Neurocrit Care* 37, 322–327. doi:10.1007/s12028-022-01470-x

CDC (2023) *Stroke facts*. United States: Centers for Disease Control and Prevention.

Chen, C., Heinemann, A., Granger, C., and Linn, R. (2002). Functional gains and therapy intensity during subacute rehabilitation: a study of 20 facilities. *Arch. Phys. Med. Rehabil.* 83 (11), 1514–1523. doi:10.1053/apmr.2002.35107

Chen, H. Y., Albertson, T. E., and Olson, K. R. (2016). Treatment of drug-induced seizures. *Br. J. Clin. Pharmacol.* 81 (3), 412–419. doi:10.1111/bcp.12720

Chohan, S. A., Venkatesh, P. K., and How, C. H. (2019). Long-term complications of stroke and secondary prevention: an overview for Primary Care Physicians. *Singap. Med. J.* 60 (12), 616–620. doi:10.11622/smedj.2019158

Finlayson, O., Kapral, M., Hall, R., Asllani, E., Selchen, D., Saposnik, G., et al. (2011). Risk factors, inpatient care, and outcomes of pneumonia after ischemic stroke. *Neurology* 77 (14), 1338–1345. doi:10.1212/WNL.0b013e31823152b1

Gittins, M., Lugo-Palacios, D., Vail, A., Bowen, A., Paley, L., Bray, B., et al. (2020). Stroke impairment categories: a new way to classify the effects of stroke based on stroke-related impairments. *Clin. Rehabil.* 35 (3), 446–458. doi:10.1177/0269215520966473

Graham, L. (2013). Organization of rehabilitation services. *Handb. Clin. Neurol.* 110, 113–120. doi:10.1016/B978-0-444-52901-5.00010-1

Heron, M. (2021). Deaths: leading causes for 2019. *Natl. Vital Stat.* 70 (9), 1–114.

Hua, C., Thomas, K., Bunker, J., Gozalo, P., Bélanger, E., Mitchell, S., et al. (2021). Dementia diagnosis in the hospital and outcomes among patients with advanced dementia documented in the Minimum Data Set. *J. Am. Geriatr. Soc.* 70 (3), 846–853. doi:10.1111/jgs.17564

IHCP (2023) *Indiana health coverage programs: long-term care*. Indiana Family and Social Services Administration.

Jackson, M., and Turkington, D. (2005). Depression and anxiety in epilepsy. *J. Neurol. Neurosurg. Psychiatry* 76 (1), i45–i47. doi:10.1136/jnnp.2004.060467

Jäger, S., Allhorn, A., and Bießmann, F. (2021). A benchmark for data imputation methods. *Front. Big Data.* 4, 693674. doi:10.3389/fdata.2021.693674

Jette, D., Warren, R., and Wirtalla, C. (2004). Rehabilitation in skilled nursing facilities: effect of nursing staff level and therapy intensity on outcomes. *Am. J. Phys. Med. Rehabil.* 83 (9), 704–712. doi:10.1097/01.phm.0000137312.06545.d0

Jette, D., Warren, R., and Wirtalla, C. (2005). The relation between therapy intensity and outcomes of rehabilitation in skilled nursing facilities. *Arch. Phys. Med. Rehabil.* 86 (3), 373–379. doi:10.1016/j.apmr.2004.10.018

Jung, H.-Y., Trivedi, A., Grabowski, D., and Mor, V. (2016). Does more therapy in skilled nursing facilities lead to better outcomes in patients with hip fracture? *Phys. Ther.* 96 (1), 81–89. doi:10.2522/ptj.20150090

Kelly-Hayes, M., Robertson, J., Broderick, J., Duncan, P., Hershey, L., Roth, E., et al. (1998). The American heart association stroke outcome classification: executive summary. *Circulation* 97 (24), 2474–2478. doi:10.1161/01.cir.97.24.2474

Koopmans, R., Lavrijsen, J., Hoek, F., Went, P., and Schols, J. (2010). Dutch elderly care physician: a new generation of nursing home physician specialists. *J. Am. Geriatr. Soc.* 58 (9), 1807–1809. doi:10.1111/j.1532-5415.2010.03043.x

Langhorne, P., Stott, D., Robertson, L., MacDonald, J., Jones, L., McAlpine, C., et al. (2000). Medical complications after stroke: a multicenter study. *AHA J.* 31, 1223–1229. doi:10.1161/01.str.31.6.1223

Learoyd, A., Woodhouse, L., Shaw, L., Sprigg, N., Bereczki, D., Berge, E., et al. (2017). Infections up to 76 days after stroke increase disability and Death. *Transl. Stroke Res.* 8 (6), 541–548. doi:10.1007/s12975-017-0553-3

Lekoubou, A., Petucci, J., Ajala, T. F., Katoch, A., Sen, S., and Honavar, V. (2024) Large datasets from Electronic Health Records predict seizures after ischemic strokes: a Machine Learning approach. medRxiv. doi:10.1101/2024.01.24.24301755

Lenze, E., Host, H., Hildebrand, M., Morrow-Howell, N., Carpenter, B., Freedland, K., et al. (2012). Enhanced medical rehabilitation increases therapy intensity and engagement and improves functional outcomes in postacute rehabilitation of older adults: a randomized-controlled trial. *J. Am. Med. Dir. Asso* 13 (8), 708–712. doi:10.1016/j.jamda.2012.06.014

Levinson, D. (2013). Skilled nursing facilities often fail to meet care planning and discharge planning requirements. Office of Inspector General.

Looti, A., Petucci, J., Katoch, A., and Honavar, V. (2023). Machine learning prediction of seizures after ischemic strokes. (S30.007). *Neurology* 100 (17). doi:10.1212/WNL.0000000000203063

Mario Arida, R., Alexandre Scorza, F., Gomes da Silva, S., Schachter, S., and Abrão Cavalheiro, E. (2010). The potential role of physical exercise in the treatment of epilepsy. *Epilepsy Behav.* 17 (4), 432–435. doi:10.1016/j.yebeh.2010.01.013

McKee, H., and Privitera, M. (2016). Stress as a seizure precipitant: identification, associated factors, and treatment options. *Seizure* 44, 21–26. doi:10.1016/j.seizure.2016.12.009

Mor, V., Intrator, O., Unruh, M. A., and Cai, S. (2011). Temporal and geographic variation in the validity and internal consistency of the nursing home resident assessment Minimum data set 2.0. *BMC Health Serv. Res.* 11, 78. doi:10.1186/1472-6963-11-78

Myint, P., Staufenberg, E., and Sabanathan, K. (2006). Post-stroke seizure and post-stroke epilepsy. *Postgrad. Med. J.* 82 (971), 568–572. doi:10.1136/pgmj.2005.041426

Naess, H., Nyland, H., Thomassen, L., Aarseth, J., and Myhr, K. (2004). Long-term outcome of cerebral infarction in young adults. *Acta Neurol. Scand.* 110 (2), 107–112. doi:10.1111/j.1600-0404.2004.00273.x

Nakken, K., Bjørholt, P., Johannessen, S., LoSyning, T., and Lind, E. (1990). Effect of physical training on aerobic capacity, seizure occurrence, and serum level of antiepileptic drugs in adults with epilepsy. *Epilepsia* 31 (1), 88–94. doi:10.1111/j.1528-1157.1990.tb05365.x

Salgado, C. M., Azevedo, C., Proença, H., and Vieira, S. M. (2016) *Secondary analysis of electronic health records*. Cham: Springer.

Sawyer, N., and Escayg, A. (2010). Stress and epilepsy: multiple models, multiple outcomes. *J. Clin. Neurophysiol.* 27 (6), 445–452. doi:10.1097/WNP.0b013e3181fe0573

Shiels, M. S., Haque, A. T., González, A. B., and Freedman, N. D. (2022). Leading causes of death in the US during the COVID-19 pandemic, march 2020 to october 2021. *JAMA Intern. Med.* 182 (8), 883–886. doi:10.1001/jamainternmed.2022.2476

Smajlović, D., Kojić, B., and Sinanović, O. (2006). Five-year survival after first-ever stroke. *Bosn. J. Basic Med. Sci.* 6 (3), 17–22. doi:10.17305/bjbms.2006.3138

Ulm, L., Harms, H., Ohlraun, S., Reimnitz, P., and Meisel, A. (2012). Impact of infections on long-term outcome after severe middle cerebral artery infarction. *J. Neurol. Sci.* 319 (1-2), 15–17. doi:10.1016/j.jns.2012.05.042

Vezzani, A., Fujinami, R., White, H. S., Preux, P.-M., Blümcke, I., Sander, J., et al. (2015). Infections, inflammation and epilepsy. *Acta Neuropathol.* 131 (2), 211–234. doi:10.1007/s00401-015-1481-5

Wells, B. J., Chagin, K. M., Nowacki, A. S., and Kattan, M. W. (2013). Strategies for handling missing data in electronic health record derived data. *EGEMS (Wash DC).* 1 (3), 1035. doi:10.13063/2327-9214.1035

frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 12 June 2024
DOI 10.3389/fphys.2024.1398735

Automatic classification of fetal heart rate based on a multi-scale LSTM network

Lin Rao[1,2†], Jia Lu[1,2†], Hai-Rong Wu[3†], Shu Zhao[1,2], Bang-Chun Lu[1,2*] and Hong Li[1,2*]

[1]International Peace Maternity and Child Health Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [2]Shanghai Key Laboratory of Embryo Original Diseases, Shanghai, China, [3]Key Laboratory of System Control and Information Processing, Ministry of Education of Shanghai Jiao Tong University, Shanghai, China

Check for updates

OPEN ACCESS

EDITED BY
Hyo Kyung Lee,
Korea University, Republic of Korea

REVIEWED BY
Chenxi Yang,
Southeast University, China
Shier Nee Saw,
University of Malaya, Malaysia

*CORRESPONDENCE
Bang-Chun Lu,
✉ lubangchun@163.com
Hong Li,
✉ 2750074682@qq.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 10 March 2024
ACCEPTED 02 May 2024
PUBLISHED 12 June 2024

CITATION
Rao L, Lu J, Wu H-R, Zhao S, Lu B-C and Li H (2024), Automatic classification of fetal heart rate based on a multi-scale LSTM network. *Front. Physiol.* 15:1398735.
doi: 10.3389/fphys.2024.1398735

Introduction: Fetal heart rate monitoring during labor can aid healthcare professionals in identifying alterations in the heart rate pattern. However, discrepancies in guidelines and obstetrician expertise present challenges in interpreting fetal heart rate, including failure to acknowledge findings or misinterpretation. Artificial intelligence has the potential to support obstetricians in diagnosing abnormal fetal heart rates.

Methods: Employ preprocessing techniques to mitigate the effects of missing signals and artifacts on the model, utilize data augmentation methods to address data imbalance. Introduce a multi-scale long short-term memory neural network trained with a variety of time-scale data for automatically classifying fetal heart rate. Carried out experimental on both single and multi-scale models.

Results: The results indicate that multi-scale LSTM models outperform regular LSTM models in various performance metrics. Specifically, in the single models tested, the model with a sampling rate of 10 exhibited the highest classification accuracy. The model achieves an accuracy of 85.73%, a specificity of 85.32%, and a precision of 85.53% on CTU-UHB dataset. Furthermore, the area under the receiver operating curve of 0.918 suggests that our model demonstrates a high level of credibility.

Discussion: Compared to previous research, our methodology exhibits superior performance across various evaluation metrics. By incorporating alternative sampling rates into the model, we observed improvements in all performance indicators, including ACC (85.73% vs. 83.28%), SP (85.32% vs. 82.47%), PR (85.53% vs. 82.84%), recall (86.13% vs. 84.09%), F1-score (85.79% vs. 83.42%), and AUC(0.9180 vs. 0.8667). The limitations of this research include the limited consideration of pregnant women's clinical characteristics and disregard the potential impact of varying gestational weeks.

KEYWORDS

fetal heart rate, long short-term memory, classification, multi time scale, artificial intelligence

1 Introduction

Fetal heart rate (FHR) serves as an indicator of the fetal heart and central nervous system's reaction to factors such as blood pressure, blood gases, and acid–base balance. In a clinical setting, FHR analysis can aid in the identification of fetal distress, placental abruption, chorioamnionitis, and other medical conditions (Sykes et al., 1983; Newton, 1993; Usui et al., 2007). FHR monitoring during labor is a valuable tool for detecting alterations in fetal heart rate patterns indicative of insufficient fetal oxygenation, enabling timely intervention by obstetricians to mitigate the risk of hypoxic injury or mortality. Electronic fetal monitoring (EFM) is currently recognized as a crucial modality for evaluating intrauterine fetal wellbeing and oxygenation levels (Sweha et al., 1999), owing to its ease of use and non-invasive nature. Consequently, EFM has emerged as an essential adjunctive screening method in obstetrics, with its utilization expanding in both antenatal and intrapartum settings.

The recording of dynamic changes in fetal heart rate can serve as an indirect indicator of fetal oxygen supply *in utero*, facilitating early detection of acute and chronic intrauterine hypoxia or asphyxia, thereby enhancing clinical efficiency. The cardiotocography (CTG) generated by EFM displays both FHR and uterine contractions, providing insights into their interplay (Alfirevic et al., 2017). Presently, three widely utilized clinical criteria exist for evaluating FHR monitoring. The first method of FHR interpretation discussed in academic literature is the nonstress test (NST) categorization outlined in the guidelines of the Society of Obstetricians and Gynecologists of Canada (SOGC), which classifies FHR as normal, atypical, and abnormal (Liston et al., 2007). The second approach is the three-tier FHR system jointly developed by the American College of Obstetricians and Gynecologists (ACOG), the Society for Maternal-Fetal Medicine (SMFM), and the National Institute of Children's Health and Human Development (NICHD), which divides FHR into categories I, II, and III according to established criteria (Macones et al., 2008). The third source of guidance is the consensus guidelines on intrapartum fetal monitoring by the International Federation of Gynecology and Obstetrics (FIGO) and the National Institute for Health and Clinical Excellence (NICE), which categorize fetal monitoring into three classes: normal, suspicious, and pathological (Ayres-de Campos et al., 2015). The assessment of CTG basic features for each classification focuses on baseline, baseline variability, accelerations, and decelerations. However, despite standardized guidelines, discrepancies in recommendations and variations in obstetrician expertise contribute to significant diversity in observer interpretation of FHR.

In recent years, there has been an increasing integration of artificial intelligence (AI) technology in the healthcare sector, particularly in domains necessitating multifaceted inputs for evaluation and prompt decision-making. One notable application is in the realm of electronic fetal heart monitoring during labor and delivery. Using AI can minimize the variability among observers, enabling real-time interpretation of FHR data to prevent overlooking necessary interventions and enhance neonatal outcomes. Furthermore, AI provides a more standardized interpretation of the analysis of FHR monitoring findings.

Numerous researchers have endeavored to categorize FHR utilizing a blend of feature extraction and machine learning techniques. Georgoulas et al. (2006) conducted feature extractions in both time and frequency domains in conjunction with morphological features and applied a support vector machine (SVM) to classify the features. Spilka et al. (2012) utilized three types of features for classification, including 11 FIGO-like features, 14 heart rate variability-based features, and eight nonlinear features. Following dimensionality reduction, the classification model was trained using naive Bayes, SVM, and the C4.5 decision tree algorithm. Dash et al. (2014) incorporated additional features related to FHR responses to uterine contractions and subsequently conducted a comparative analysis of three generative models using SVM methods. Comert et al. (2016) utilized software to extract 21 features and implemented an extreme learning machine for data analysis. Spilka et al. (2017) advocated for sparse SVM classification, which offered the advantage of selecting a reduced number of features to detect various FHR patterns. In addition to traditional FHR features, techniques such as short-time Fourier transform (STFT), gray Level Co-occurrence matrix (GLCM) (Comert and Kocamaz, 2018), wavelet transform (Comert and Kocamaz, 2017), and common spatial pattern (CSP) (Alsaggaf et al., 2020) were employed to enhance classification performance.

All these methods were hindered by the requirement for feature extraction, which was typically done manually or with computer assistance. In response to this challenge, researchers introduced deep learning techniques to facilitate automatic feature extraction and classification. Convolutional neural networks (CNNs) have shown exceptional performance in image classification and have been extensively utilized in the medical field. Given that FHR signals are one-dimensional, researchers have explored various approaches to transform FHR signals into two-dimensional images, including STFT (Comert et al., 2019), continuous wavelet transform (CWT) (Zhao et al., 2019a), and recurrent plot (RP) (Zhao et al., 2019b). FHR analysis can be conducted using one-dimensional convolutional neural networks (1D-CNN) (Ismail Fawaz et al., 2019) as a time series method. Li et al. (2019) segmented 20-min FHR signals into 1–16 segments and applied 1D-CNN to analyze each segment, aggregating results through a voting mechanism. Cao et al. (Cao et al., 2023) employed a multimodal deep learning architecture (MMDLA) that integrates a CNN to extract high-level features from preprocessed cardiotocographic signals and maternal clinical data, thereby improving model performance. Zhou et al. (2023) proposed the trend-guided long convolution network (TGLCN), a deep learning methodology that integrates convolution kernel selection, residual structures, and attention mechanisms. Baghel et al. Baghel et al. (2022) utilized a Gaussian Butterworth band pass filter in conjunction with the CNN for the diagnosis of fetal acidosis. Furthermore, recurrent neural networks (RNNs), specifically long short-term memory (LSTM) networks, are crucial in FHR classification. Gao and Lu (2019) employed bidirectional LSTM (BiLSTM) for the segmental classification of FHR.

Although previous studies have made significant advances, certain challenges also persist, including imbalanced datasets affecting model performance and limited research on features at various time scales. To address these issues, this article introduces a multi-scale LSTM network. The article makes three key contributions: 1) Introducing a data augmentation methodology for time series to enhance datasets and address data imbalance. 2) Training LSTM models at different time scales through finetuning. 3) Proposing multi-scale LSTM networks to enhance model performance.

The subsequent sections of this article are organized as follows: Section 2 outlines the database utilized, the processing procedures applied, and the proposed methodology. Section 3 presents the experimental findings and compares them with previous studies. Section 4 provides a summary of the research and outlines potential future directions.

TABLE 1 Patient and labor outcome statistics for the CTU-UHB cardiotocography database.

	Mean	Min	Max
Maternal age (years)	29.8	18	46
Parity	0.43	0	7
Gravidity	1.3	1	11
Gestational age (weeks)	40	37	43
pH	7.23	6.85	7.47
Base excess (BE, mmol/L)	−6.36	−26.8	−0.2
Base deficit in extracellular fluid (BDecf, mmol/L)	4.60	−3.40	
Apgar 1 min	8.26	1	10
Apgar 5 min	9.06	4	10
Neonatal weight(g)	3408	1970	4750

2 Methods

2.1 Dataset description

The dataset utilized in this study is the CTU-UHB database (Chuda´cõek et al., 2014), an open-access repository comprising 552 recordings obtained at University Hospital in Brno (UHB) during the period of 2010–2012. Each recording is composed of two components: the cardiotocography (CTG) and clinical data. The CTG data are captured using three distinct methods: ultrasound Doppler probe, direct scalp measurement, or a hybrid approach. The CTG data encompass FHR and uterine contractions sampled at a rate of 4 Hz, resulting in four data points per second for each parameter.

The clinical data include information regarding fetal status and parameters concerning puerperal and newborn infants. Table 1 displays a portion of the clinical statistics obtained from the CTU-UHB database. Umbilical artery pH serves as a recognized marker for fetal acidemia, a condition associated with neonatal complications, such as multiple organ dysfunction in newborns (Sehdev et al., 1997; van den Berg et al., 1996). Studies have shown a relationship between FHR and variations in umbilical artery pH (Singh et al., 2021). Consequently, we employed the umbilical artery pH values from the clinical data to classify our dataset into two separate groups in Figure 1. In accordance with the established criterion that a pH value exceeding 7.15 signifies a normal condition, a total of 439 samples were classified as normal, and 113 samples were categorized as pathological based on their pH value (Comert et al., 2018).

2.2 Data preprocessing

During the data collection process, missing signals and artifacts may arise in the original data due to external factors such as limitations in data acquisition by ultrasound probe and maternal and fetal movement, necessitating the preprocessing of data. The process is as follows:

(1) The original data are divided into 1-min segments, each containing 240 points. Then, the number of zero-value points f_0 are counted, and the data loss rate LR is calculated according to Eq. 1.

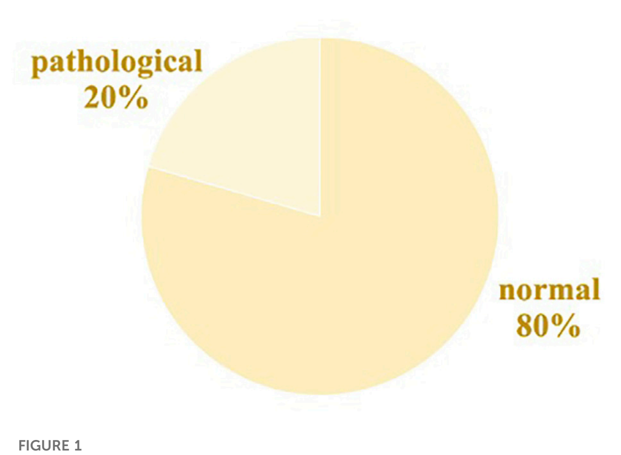


FIGURE 1
Class distribution.

$$LR = \frac{f_0}{240} \times 100\%, \tag{1}$$

if $LR \geq 40\%$, this data segment will be discarded.

(2) When the FHR value is greater than 220 times per minute or less than 60 times per minute, it is treated as an abnormality due to poor contact with the acquisition device. The linear random interpolation method is used to replace the abnormal data. The formula of linear random interpolation is displayed according to Eq. 2.

$$f_{in} = \lambda f_{before} + (1-\lambda) f_{after}, \tag{2}$$

where λ is a random factor, and f_{before} and f_{after} are values before and after the missing point.

Due to too many missing signals in some recordings, the number of recordings in the dataset decreased to 550, with 439 normal recordings and 111 pathological recordings.

There are only 550 recordings in the dataset, and the ratio of normal recordings and pathological recordings is 4:1. The limited number of recordings and the ratio of normal to pathological readings can easily cause model overfitting. The length of recordings varies from 60 to 90 min. Under the instruction of obstetricians, we take 20-min signals to do further analysis. Thus, the dataset can be augmented by window slicing (Liang and Lu, 2023). The specific process is given as follows:

Step 1: For an FHR time series $T = \{t_1, t_2, \ldots, t_n\}$, choose the length of slicing window s and step length k;

Step 2: Obtain the first slice with a window $T_1 = \{t_1, \ldots, t_s\}$;

Step 3: Move the window to get $T_2 = \{t_{k+1}, \ldots, t_{k+s}\}$, ..., $T_m = \{t_{mk+1}, \ldots, t_{mk+s}\}$ and stop the process when $mk + s > n$.

Figure 2 shows the signals before and after preprocessing. In this article, we chose $s = 4800$ and $k = 600$, which implies generating 20-min samples with the beginnings of two adjacent samples that are 2.5 min apart. An example of a slice operation is shown in Figure 3.

After data augmentation, the number of normal samples increased to 6382 from 439, and the number of pathological

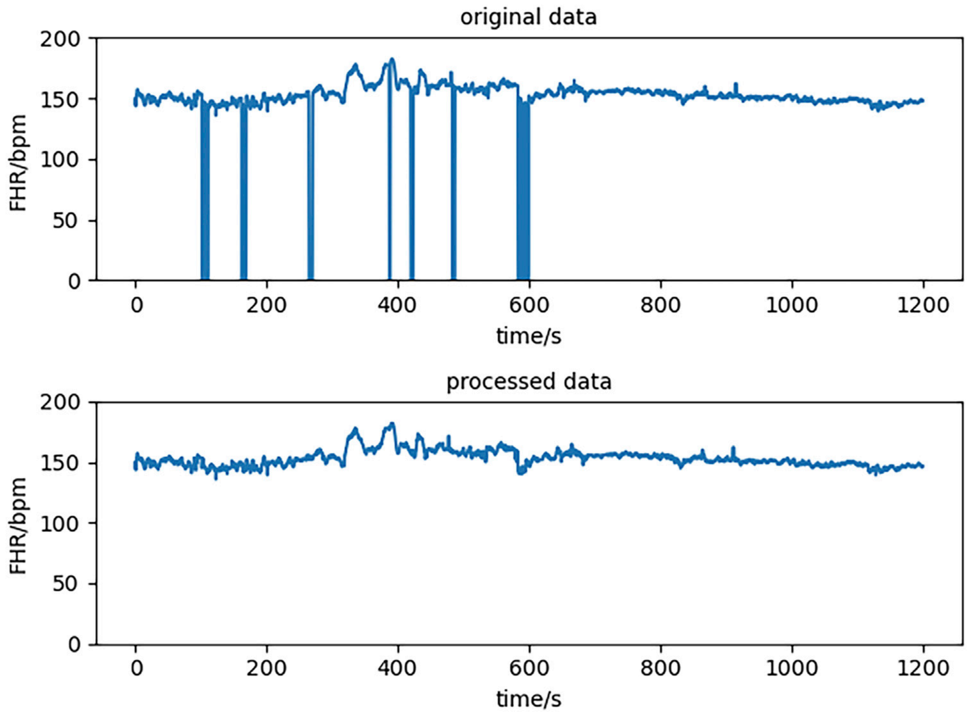


FIGURE 2
Data before and after preprocessing.

samples increased to 1615 from 111. Because the two classes were still imbalanced, we chose 1,615 from 6,382 normal samples randomly to create a new dataset with all pathological samples.

2.3 LSTM networks

An LSTM is a special kind of RNN designed to solve the problem of long-term dependency (Hochreiter and Schmidhuber, 1997).

The workflow of the LSTM cell at time t is as follows: the hidden state of the previous moment and the input of the current moment enter the forget gate, input gate, and output gate for calculation and then update the cell state and hidden state. The input gate can decide what new information can be stored in the cell state, and the output gate determines what information can be output based on the cell state. The forget gate can decide what information will be discarded from the cell state. The calculation process is according to Eqs 3–8.

$$f_t = \sigma\left(W_{fh}h_{t-1} + W_{fx}x_t + b_f\right). \tag{3}$$

$$i_t = \sigma\left(W_{ih}h_{t-1} + W_{ix}x_t + b_i\right). \tag{4}$$

$$\tilde{c_t} = tanh\left(W_{\tilde{c}h}h_{t-1} + W_{\tilde{c}x}x_t + b_{\tilde{c}}\right). \tag{5}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c_t}. \tag{6}$$

$$o_t = \sigma\left(W_{oh}h_{t-1} + W_{ox}x_t + b_o\right). \tag{7}$$

$$h_t = o_t \cdot tanh\left(c_t\right). \tag{8}$$

The architecture of LSTM cell is shown in Figure 4.

2.4 Multi-scale LSTM networks and voting mechanism

In clinical practice, obstetricians primarily utilize nonstress testing (NST) as the main modality for evaluating prenatal FHR. The SOGC (Liston et al., 2007) guidelines stipulate that interpreting NST results requires assessing various parameters, including baseline FHR, baseline variability, accelerations, and deceleration, each of which must be evaluated across different time intervals. For instance, the baseline FHR denotes the mean level of FHR over a 10-min period, excluding any accelerations, decelerations, or notable variability, and requires a minimum of 2 min of uninterrupted observation.

In contrast, acceleration and deceleration are typically evaluated within a time frame of less than 30 s. Consequently, the model must possess the capability to encompass both enduring characteristics that signify the general pattern in FHR data and fleeting characteristics that indicate minor fluctuations in specific areas. In accordance with this principle, we adopt the strategy of training numerous models by downsampling the data at varying frequencies. Downsampling is a prevalent technique in the processing of time series data. Downsampling facilitates the hybrid model in extracting data features across various time scales, thereby mitigating computational expenses and eliminating data redundancy (Liu et al., 2021).

Subsequently, each dataset undergoes downsampling by distinct sampling intervals before being inputted into diverse time-scale LSTM models. These outputs of multi-scale models are aggregated using weights to yield the ultimate result, represented by the final result vector y denoting the

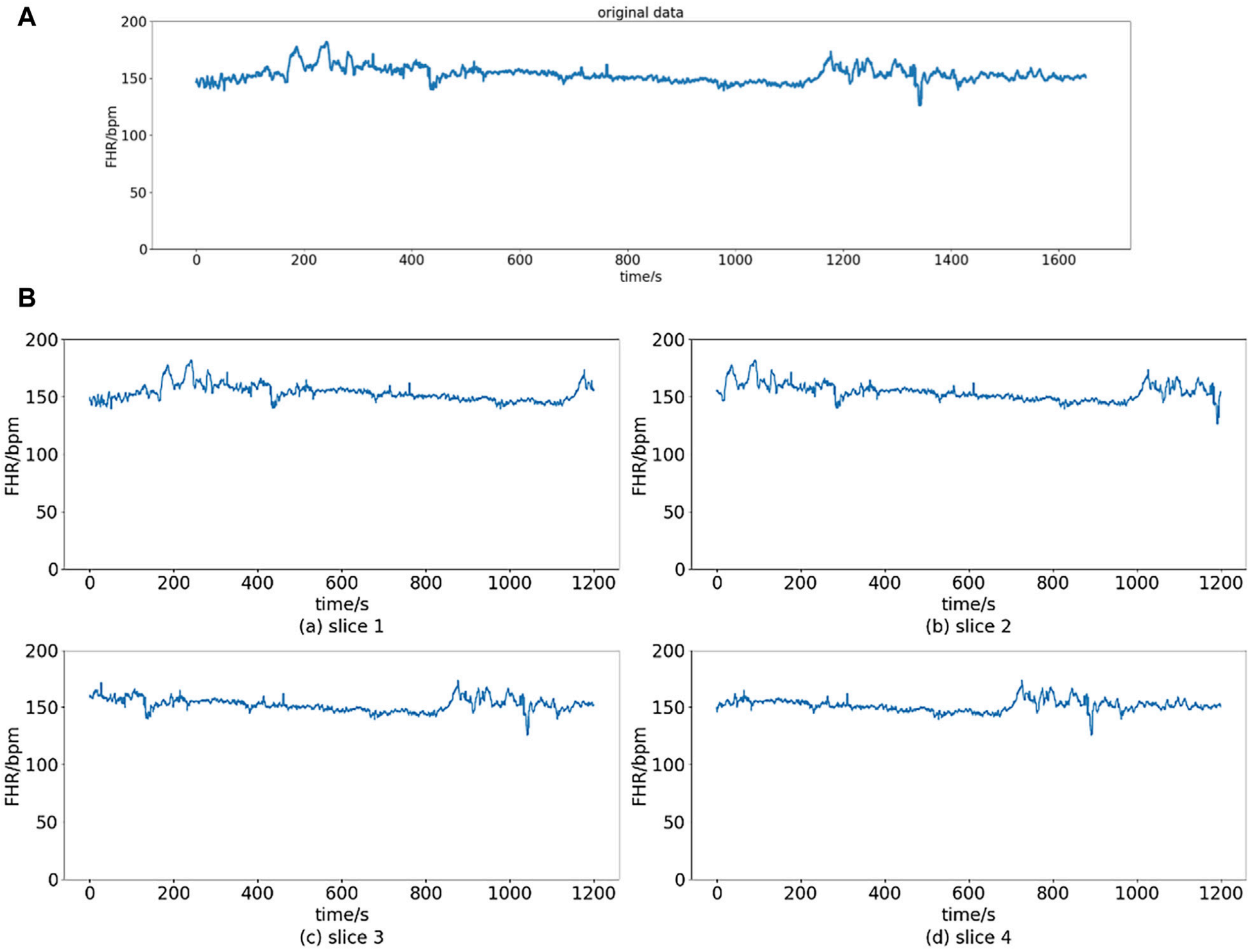


FIGURE 3
Slice operation. **(A)** Original data. **(B)** Slices of original data.

probability of data belonging to each category. The computation process is according to Eqs 9, 10.

$$y = \Sigma_{i=1}^{n} \omega_i y_{i.} \tag{9}$$

$$\Sigma_{i=1}^{n} \omega_i = 1, \tag{10}$$

where y_i is the output vector of the i-th model and ωi is the corresponding weight value of i-th model. The architecture of multi-scale LSTM networks is shown in Figure 5.

2.5 Evaluation index

The confusion matrix is a commonly utilized tool for assessing the efficacy of models in classification tasks (James et al., 2013). In the context of the binary classification discussed in this article, a confusion matrix with dimensions of two rows and two columns represents the frequency of four distinct prediction outcomes.

The metrics employed in our study include accuracy (ACC), specificity (SP), precision (PR), recall, F1-score, and area under the curve (AUC). ACC provides a comprehensive measure of the accuracy of predictions, while SP emphasizes the proportion of accurately identified negative samples. The constraints of electronic fetal monitoring contribute to a notable false positive rate in obstetric diagnoses. Inaccurate identification of pathological conditions may result in unwarranted medical interventions (Li et al., 2019). Therefore, it is imperative to consider precision and recall metrics, which evaluate the accuracy of positive predictions and the proportion of successfully detected positive samples. The F1-score represents the harmonic mean of PR and recall, while the quality index is calculated as the geometric mean of SP and sensitivity. The metrics mentioned above are calculated according to Eqs 11–15.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN.} \tag{11}$$

$$SP = \frac{TN}{TN + FP}. \tag{12}$$

$$PR = \frac{TP}{TP + FP}. \tag{13}$$

$$Recall = \frac{TP}{TP + FN}. \tag{14}$$

$$F1 - score = 2 \cdot \frac{PR \cdot Recall}{PR + Recall}. \tag{15}$$

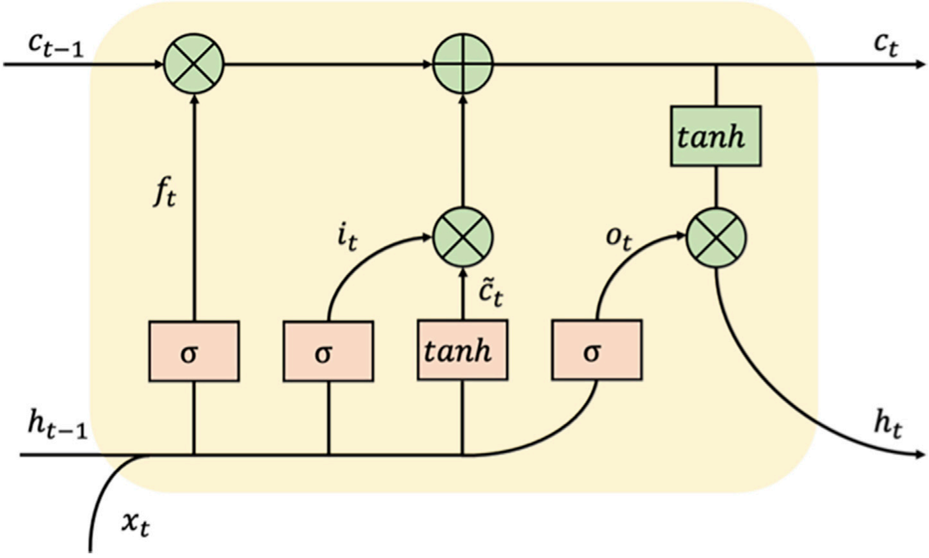


FIGURE 4
The architecture of an LSTM cell.

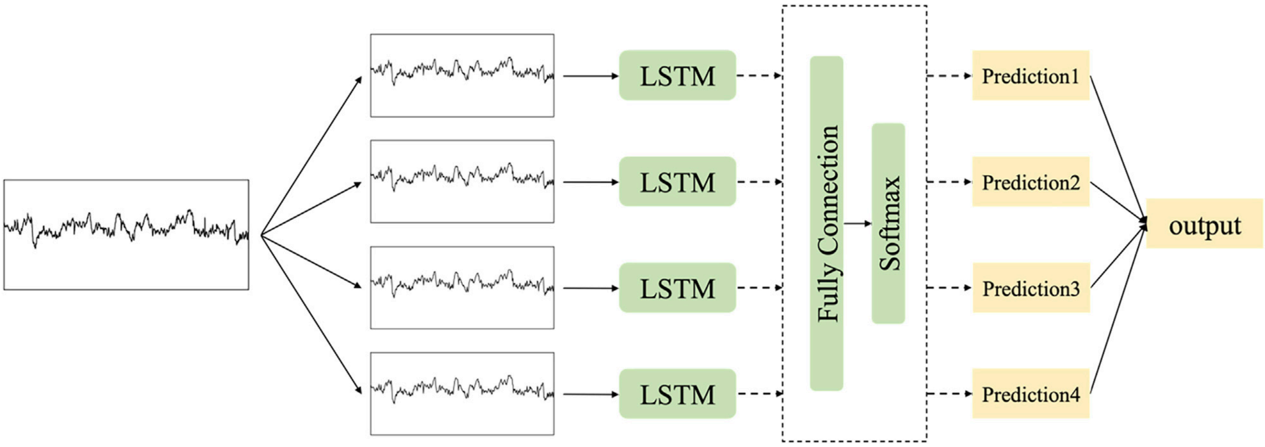


FIGURE 5
The architecture of multi-scale LSTM networks.

3 Experiments and results

3.1 Experimental settings

The experiment was carried out utilizing the PyTorch deep learning framework in Python, along with additional packages such as Numpy and Scikit-learn. The hardware configuration includes an Intel(R) Core (TM) i9-10900X CPU @ 3.70 Hz and an NVIDIA GeForce RTX 2080Ti.

The hybrid model is composed of two LSTM layers, three full connection layers, and an output layer, with each LSTM layer containing 512 hidden units. In order to address overfitting, a dropout rate of 0.2 is applied before the full connection layer. The output dimension is reduced to 2 through the full connection layers, with the final activation function being softmax for classification. The optimizer used is Adam, and the loss function employed is cross-entropy. To enhance the convergence of the network, we implemented a learning rate decay strategy during the training process consisting of 2,000 epochs. The initial learning rate was set at 0.001 and decreased by a factor of 10 after 500 and 1000 epochs.

The models were trained using a 10-fold cross-validation approach, where the dataset was partitioned into 10 subsets, each containing 323 samples. Nine subsets were utilized to train the model, while the remaining subset was used to test its performance. Following the training and testing of 10 models on the test set, the mean and standard deviation of the results were calculated.

TABLE 2 Comparison of the performance of different models.

Model	ACC (%)	SP (%)	PR (%)	Recall (%)	F1-score (%)	AUC
Sampling Rate = 4	74.49 ± 5.15	73.93 ± 4.33	74.15 ± 4.56	75.05 ± 6.54	74.57 ± 5.42	0.7699 ± 0.0552
Sampling Rate = 6	75.05 ± 4.39	73.68 ± 4.95	74.43 ± 4.53	76.41 ± 4.64	75.38 ± 4.34	0.7854 ± 0.0443
Sampling Rate = 8	78.39 ± 5.87	77.95 ± 6.51	78.22 ± 6.04	78.83 ± 6.56	78.47 ± 5.9	0.8193 ± 0.0626
Sampling Rate = 10	83.28 ± 4.37	82.47 ± 5.24	82.84 ± 4.68	84.09 ± 4.69	83.42 ± 4.24	0.8667 ± 0.0479
Multi-scale Model 1	85.73 ± 2.5	85.32 ± 3.68	85.53 ± 3.19	86.13 ± 3.1	85.79 ± 2.43	0.918 ± 0.0278
Multi-scale Model 2	84.92 ± 3.67	84.51 ± 5.06	84.78 ± 4.42	85.33 ± 4.01	85 ± 3.54	0.914 ± 0.0316
Multi-scale Model 3	84.18 ± 3.5	87.86 ± 5.15	87.11 ± 4.67	80.5 ± 5.12	83.56 ± 3.67	0.8992 ± 0.0375

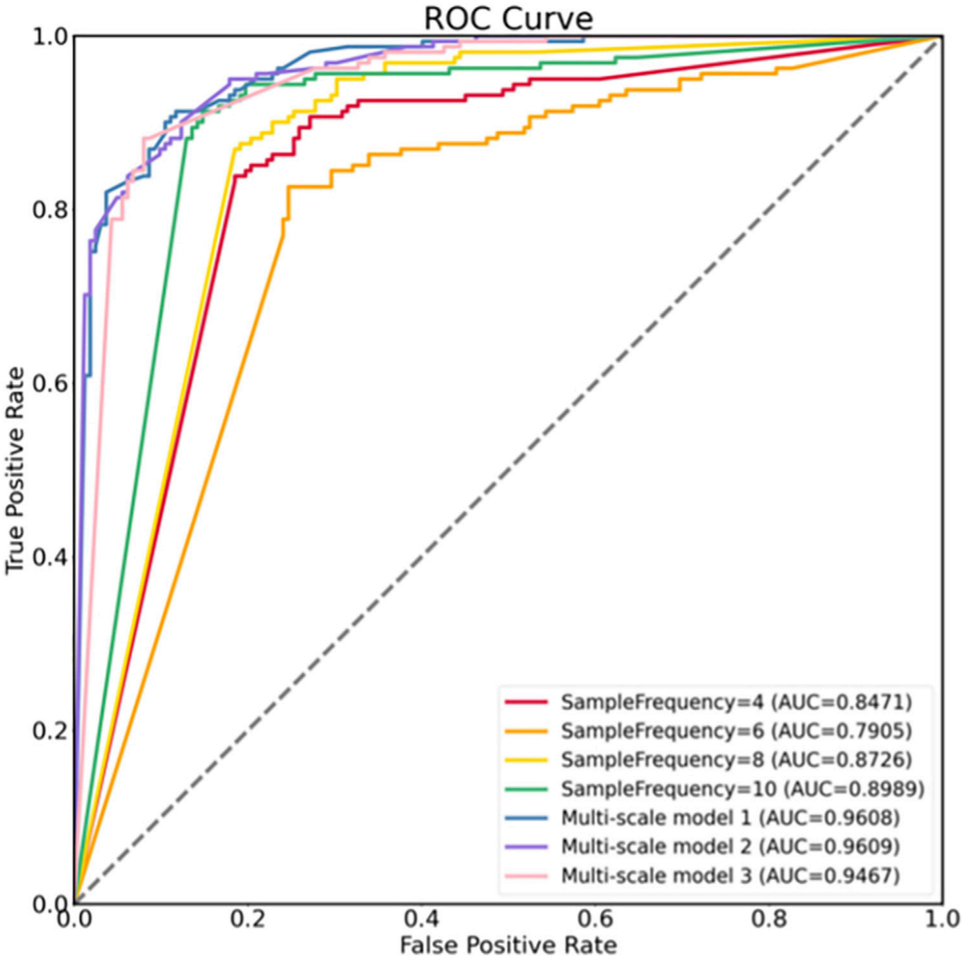


FIGURE 6
ROCs of different models.

3.2 Results analysis

Initially, the experiments were conducted to examine the impact of varying sampling rates on the efficacy of the model. The results presented in Table 2 indicate that the model exhibits optimal performance at a sampling rate of 10. Specifically, ACC and F1-score metrics demonstrate an improvement of approximately 5% compared to the next highest-performing model, while the SP and PR metrics show an enhancement of approximately 4.5%. The model's performance improves with increasing sampling intervals, potentially due to its enhanced ability to discern between normal and pathological data by capturing long-term features. Furthermore, larger sampling intervals serve to diminish the impact of noise signals within the data.

TABLE 3 Comparison of the proposed model with previous work.

References	Method	ACC (%)	SP (%)	PR (%)	Recall (%)	F1-score (%)	AUC
Comert et al. (2018)	EMD + DWT + SVM	67.00	67.26	\	57.42	\	\
O'Sullivan et al. (2021)	ARMA + SVM	83.3	77.7	\	82.6	\	0.809
Liu et al. (2021)	CNN-BiLSTM + Attention, DWT	71.71 ± 8.61	70.81 ± 12.20	\	75.23 ± 9.58	\	\
Singh et al. (2021)	HoloViz + CNN	69.6	\	63	70	66	\
Ben Barek et al. (2023)	LR	\	\	\	\	\	0.74
Ours	Multi-scale LSTM	85.73 ± 2.5	85.32 ± 3.68	85.53 ± 3.19	86.13 ± 3.1	85.79 ± 2.43	0.918 ± 0.0278

Three different multi-scale models were constructed by manipulating the quantity and magnitude of the component models. Multi-scale Model 1 comprises four sampling rates: 4, 6, 8, 10. Multi-scale Model 2 utilizes models with sampling rates of 4, 8, and 10, whereas multi-scale Model 3 exclusively integrates models with sampling rates of 8 and 10. The superior performance of all multi-scale models over the single models is evident in Table 2, indicating that the incorporation of multi-scale features aids in mitigating overfitting to some degree and enhances categorization accuracy. Multi-scale Model 1 demonstrates superior performance on ACC, recall, F1-score, and AUC, suggesting that incorporating diverse time-scale features enhances classification accuracy. Conversely, Model 3 exhibits higher SP and PR but comparatively lower performance on other evaluation criteria. The ROC curve depicted in Figure 6 illustrates the discriminative capabilities of single models *versus* multi-scale models, with the latter showcasing an enhanced ability to distinguish between two classes.

3.3 Discussion

In this research, we introduce a multi-scale LSTM model integrated with models that target various time scales. Experimental analyses were carried out on both single and multi-scale models. The results demonstrate that multi-scale LSTM models outperform regular LSTM models in various performance metrics. Specifically, among the single models tested, the model with a sampling rate of 10 exhibited the highest classification accuracy. Incorporating alternative sampling rates into the model resulted in enhancements across all performance indicators, including ACC (85.73% vs. 83.28%), SP (85.32% vs. 82.47%), PR (85.53% vs. 82.84%), recall (86.13% vs. 84.09%), F1-score (85.79% vs. 83.42%), and AUC (0.9180 vs. 0.8667).

To illustrate the importance of our model, the outcomes of both machine learning (Comert et al., 2018; O'Sullivan et al., 2021; Ben Barek et al., 2023) and deep learning approaches (Liu et al., 2021; Singh et al., 2021) utilizing the identical dataset are presented in Table 3. Our model exhibits superior performance in terms of ACC, SP, PR, recall, and AUC compared to the aforementioned machine learning methods (Liu et al., 2021; Singh et al., 2021). Furthermore, when compared to a specific model (Liu et al., 2021), our model demonstrates notably higher levels of ACC, SP, and recall. It is worth noting that the model discussed (Singh et al., 2021) achieves an ACC of 69.6%, potentially attributed to the limitations of CNNs in capturing temporal features effectively. This observation suggests that our model possesses enhanced classification capabilities.

In conclusion, the proposed model demonstrates enhanced performance in the classification of FHR. This model offers several advantages, including directly classifying FHR signals without the need for complex feature extraction processes and ensuring immediate discrimination. Additionally, incorporating various time-scale signals enables the model to effectively learn both long-term and short-term features, thereby optimizing overall performance.

4 Conclusion

In this study, a multi-scale LSTM model was developed for the automatic classification of FHR. The publicly available CTU-UHB database was utilized for this purpose. Following data preprocessing and enhancement, FHR signals were employed as input for the models. The proposed model demonstrated the ability to identify pathological FHR patterns. Experimental results indicate that our model outperforms common LSTM models and previous research efforts in terms of various metrics. Specifically, the model achieved an accuracy, specificity, and precision of 89.78%, 91.36%, and 91.03%, respectively. Our work presents significant contributions in utilizing the LSTM model for extracting hidden features from FHR signals, eliminating the need for manual feature extraction. Additionally, incorporating various time-scale features enhances the performance of the models. Ultimately, our model facilitates intelligent recognition of FHR, aiding obstetricians in identifying abnormal FHR patterns and supporting timely treatment interventions.

Nevertheless, it is important to acknowledge the limitations of our research. First, the clinical characteristics of pregnant women, including maternal age and weight, can significantly influence the classification results and should be taken into consideration. Second, the data in the CTU-UHB dataset were gathered 90 min prior to delivery, potentially overlooking the impact of varying gestational weeks on fetal heart rate patterns, particularly around 32 weeks. Moving forward, we plan to establish partnerships with medical facilities to expand our dataset by incorporating additional fetal heart rate, uterine contraction, and clinical information. Further analysis of additional features should be conducted during the model construction process, and adjustments to the model structure should be made in order to enhance classification accuracy.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the International Peace Maternity and Child Health Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

LR: conceptualization, methodology, project administration, writing–review and editing. JL: visualization, writing–review and editing. H-RW: conceptualization, data curation, writing–original draft. SZ: software, writing–review and editing. LB-C: writing–review and editing. HL: conceptualization, methodology, writing–original draft, writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the Construction of Nursing Discipline Project of Shanghai Jiao Tong University of Medicine under Grant SJTUHLXK 2021, in part by the Cultivation of Interdisciplinary Projects (Translation) of Shanghai Jiao Tong University under Grants YG2021QN139, YG2022QN113, and YG2019QNB05, and in part by Shanghai Municipal Health Commission under Grant 202040232.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alfirevic, Z., Gyte, G. M., Cuthbert, A., and Devane, D. (2017). Continuous cardiotocography (ctg) as a form of electronic fetal monitoring (efm) for fetal assessment during labour. *Cochrane Database Syst. Rev.* 2019 (5), CD006066. doi:10.1002/14651858.CD006066.pub3

Alsaggaf, W., Comert, Z., Nour, M., Polat, K., Brdesee, H., and Toğac¸ar, M. (2020). Predicting fetal hypoxia using common spatial pattern and machine learning from cardiotocography signals. *Appl. Acoust.* 167, 107429. doi:10.1016/j.apacoust.2020.107429

Ayres-de Campos, D., Spong, C. Y., Chandraharan, E., and Panel, F. I. F. M. E. C. (2015). Figo consensus guidelines on intrapartum fetal monitoring: cardiotocography. *Int. J. Gynecol. Obstetrics* 131 (1), 13–24. doi:10.1016/j.ijgo.2015.06.020

Baghel, N., Burget, R., and Dutta, M. K. (2022). 1d-fhrnet: automatic diagnosis of fetal acidosis from fetal heart rate signals. *Biomed. Signal Process. Control* 71, 102794. doi:10.1016/j.bspc.2021.102794

Ben Barek, I., Jauvion, G., Vitrou, J., Holmstro, E., Koskas, M., and Ceccaldi, P.-F. (2023). DeepCTG® 1.0: an interpretable model to detect fetal hypoxia from cardiotocography data during labor and delivery. *Front. Pediatr.* 11, 1190441. doi:10.3389/fped.2023.1190441

Cao, Z., Wang, G., Xu, L., Li, C., Hao, Y., Chen, Q., et al. (2023). Intelligent antepartum fetal monitoring via deep learning and fusion of cardiotocographic signals and clinical data. *Health Inf. Sci. Syst.* 11 (1), 16. doi:10.1007/s13755-023-00219-w

Chuda cek, V., Spilka, J., Bursa, M., Janku, P., Hruban, L., Huptych, M., et al. (2014). Open access intrapartum ctg database. *BMC Pregnancy Childbirth* 14 (1), 16. doi:10.1186/1471-2393-14-16

Comert, Z., Yang, Z., Velappan, S., Boopathi, A. M., and Kocamaz, A. F. (2018). "Performance evaluation of empirical mode decomposition and discrete wavelet transform for computerized hypoxia detection and prediction," in 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, May, 2018, 1–4.

Comert, Z., and Kocamaz, A. F. (2017). "Using wavelet transform for cardiotocography signals classification," in 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, May, 2017, 1–4.

Comert, Z., and Kocamaz, A. F. (2018). Open-access software for analysis of fetal heart rate signals, Biomedical. *Signal Process. Control* 45, 98–108. doi:10.1016/j.bspc.2018.05.016

Comert, Z., and Kocamaz, A. F. (2019). "Fetal hypoxia detection based on deep convolutional neural network with transfer learning approach," in *Software engineering and algorithms in intelligent systems*. Editor R. Silhavy (Cham: Springer International Publishing), 239–248.

Comert, Z., Kocamaz, A. F., and Gu¨ngo¨r, S. (2016). "Cardiotocography signals with artificial neural network and extreme learning machine," in 2016 24th Sig- nal Processing and Communication Application Conference (SIU), Zonguldak, Turkey, May, 2016, 1493–1496.

Comert, Z., Kocamaz, A. F., and Subha, V. (2018). Prognostic model based on image-based time-frequency features and genetic algorithm for fetal hypoxia assessment. *Comput. Biol. Medicine99* 99, 85–97. doi:10.1016/j.compbiomed.2018.06.003

Dash, S., Quirk, J. G., and Djuric´, P. M. (2014). Fetal heart rate classification using generative models. *IEEE Trans. Biomed. Eng.* 61 (11), 2796–2805. doi:10.1109/TBME.2014.2330556

Gao, W., and Lu, Y. (2019). "Fetal heart baseline extraction and classification based on deep learning," in 2019 International Conference on Information Technology and Computer Application (ITCA), Guangzhou, China, December, 2019, 211–216.

Georgoulas, G., Stylios, D., and Groumpos, P. (2006). Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines. *IEEE Trans. Biomed. Eng.* 53 (5), 875–884. doi:10.1109/TBME.2006.872814

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data Min. Knowl. Discov.* 33 (4), 917–963. doi:10.1007/s10618-019-00619-1

Ito, E. H., Nagasaki, S., Kotaki, H., Shimabukuro, M., Sakuma, J., Takano, M., et al. (2022). Optimal duration of cardiotocography assessment using the ipreface score to predict fetal acidemia. *Sci. Rep.* 12, 13064. doi:10.1038/s41598-022-17364-z

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An introduction to statistical learning*. New York: Springer.

Li, J., Chen, Z.-Z., Huang, L., Fang, M., Li, B., Fu, X., et al. (2019). Automatic classification of fetal heart rate based on convolutional neural network. *IEEE Internet Things J.* 6 (2), 1394–1401. doi:10.1109/jiot.2018.2845128

Liang, H., and Lu, Y. (2023). A cnn-rnn unified framework for intrapartum cardiotocograph classification. *Comput. Methods Programs Biomed.* 229, 107300. doi:10.1016/j.cmpb.2022.107300

Liston, R., Sawchuck, D., Young, D., Brassard, N., Campbell, K., Davies, G., et al. (2007). Fetal health surveillance: antepartum and intrapartum consensus guideline. *J. Obstetrics Gynaecol. Can.* 29 (9), S3–S56. doi:10.1016/S1701-2163(16)32615-9

Liu, M., Lu, Y., Long, S., Bai, J., and Lian, W. (2021). An attention-based cnn-bilstm hybrid neural network enhanced with features of discrete wavelet transformation for fetal acidosis classification. *Expert Syst. Applications186* 186, 115714. doi:10.1016/j.eswa.2021.115714

Macones, G. A., Hankins, G. D. V., Spong, C. Y., Hauth, J., and Moore, T. (2008). The 2008 national institute of child health and human development workshop report on electronic fetal monitoring: update on definitions, interpretation, and research guidelines. *J. Obstetric, Gynecol. &Neonatal Nurs.* 37 (5), 510–515. doi:10.1111/j.1552-6909.2008.00284.x

Newton, E. (1993). Chorioamnionitis and intraamniotic infection. *Clin. obstetrics Gynecol.* 36 (4), 795–808. doi:10.1097/00003081-199312000-00004

O Sullivan, M., Gabruseva, T., Boylan, G., O'Riordan, M., Lightbody, G., and Marnane, W. (2021). "Classification of fetal compromise during labour: signal processing and feature engineering of the cardiotocograph," in 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, August, 2021, 1331–1335.

Sehdev, H. M., Stamilio, D. M., Macones, G. A., Graham, E., and Morgan, M. A. (1997). Predictive factors for neonatal morbidity in neonates with an umbilical arterial cord pH less than 7.00. *Am. J. Obstetrics Gynecol.* 177 (5), 1030–1034. doi:10.1016/s0002-9378(97)70008-5

Singh, H. D., Saini, M., and Kaur, J. (2021). Fetal distress classification with deep convolutional neural network. *Curr. Women's Health Rev.* 17 (1), 60–73. doi:10.2174/1573404816999200821162312

Spilka, J., Chuda´cõek, V., Koucky´, M., Lhotska´, L., Huptych, M., Janku˚, P., et al. (2012). Using nonlinear features for fetal heart rate classification. *Biomed. Signal Process. Control* 7 (4), 350–357. doi:10.1016/j.bspc.2011.06.008

Spilka, J., Frecon, J., Leonarduzzi, R., Pustelnik, N., Abry, P., and Doret, M. (2017). Sparse support vector machine for intrapartum fetal heart rate classification. *IEEE J. Biomed. Health Inf.* 21 (3), 664–671. doi:10.1109/JBHI.2016.2546312

Sweha, A., Hacker, T., and Nuovo, J. (1999). Interpretation of the electronic fetal heart rate during labor. *Am. Fam. physician* 59 (9), 2487–2500.

Sykes, G. S., Molloy, P. M., Johnson, P., Stirrat, G. M., and Turnbull, A. C. (1983). Fetal distress and the condition of newborn infants. *BMJ* 287 (6397), 943–945. doi:10.1136/bmj.287.6397.943

Usui, R., Matsubara, S., Ohkuchi, A., Kuwata, T., Watanabe, T., Izumi, A., et al. (2007). Fetal heart rate pattern reflecting the severity of placental abruption. *Archives Gynecol. Obstetrics* 277 (3), 249–253. doi:10.1007/s00404-007-0471-9

van den Berg, P. P., Nelen, W. L., Jongsma, H. W., Nijland, R., Kolle´e, L. A., Nijhuis, J. G., et al. (1996). Neonatal complications in newborns with an umbilical artery pH < 7.00. *Am. J. Obstetrics Gynecol.* 175 (5), 1152–1157. doi:10.1016/s0002-9378(96)70021-2

Zhao, Z., Deng, Y., Zhang, Y., Zhang, Y., Zhang, X., and Shao, L. (2019a). Deepfhr: intelligent prediction of fetal acidemia using fetal heart rate signals based on convolutional neural network. *BMC Med. Inf. Decis. Mak.* 19 (1), 286. doi:10.1186/s12911-019-1007-5

Zhao, Z., Zhang, Y., Comert, Z., and Deng, Y. (2019b). Computer-aided diagnosis system of fetal hypoxia incorporating recurrence plot with convolutional neural network. *Front. Physiology* 10, 255. doi:10.3389/fphys.2019.00255

Zhou, Z., Zhao, Z., Zhang, X., Zhang, X., Jiao, P., and Ye, X. (2023). Identifying fetal status with fetal heart rate: deep learning approach based on long convolution. *Comput. Biol. Med.* 159, 106970. doi:10.1016/j.compbiomed.2023.106970

frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 25 July 2024
DOI 10.3389/fphys.2024.1425582

A novel methodology for emotion recognition through 62-lead EEG signals: multilevel heterogeneous recurrence analysis

Yujie Wang[1], Cheng-Bang Chen[1]*, Toshihiro Imamura[2,3], Ignacio E. Tapia[4], Virend K. Somers[5], Phyllis C. Zee[6] and Diane C. Lim[7,8]

[1]Department of Industrial and Systems Engineering, University of Miami, Coral Gables, FL, United States, [2]Division of Sleep Medicine, Department of Medicine, University of Pennsylvania, Phialdelphia, PA, United States, [3]Division of Pulmonary and Sleep Medicine, Children's Hospital of Philadelphia, Phialdelphia, PA, United States, [4]Division of Pediatric Pulmonology, Miller School of Medicine, University of Miami, Miami, FL, United States, [5]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States, [6]Center for Circadian and Sleep Medicine, Department of Neurology, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States, [7]Department of Medicine, Miami VA Medical Center, Miami, FL, United States, [8]Department of Medicine, Miller School of Medicine, University of Miami, Miami, FL, United States

Check for updates

OPEN ACCESS

EDITED BY
Bing Yao,
The University of Tennessee, United States

REVIEWED BY
Pranjali Gajbhiye,
Nirvesh Enterprises Private Limited, India
Hongtao Wang,
Wuyi University, China

*CORRESPONDENCE
Cheng-Bang Chen,
✉ cxc1920@miami.edu

RECEIVED 30 April 2024
ACCEPTED 27 June 2024
PUBLISHED 25 July 2024

CITATION
Wang Y, Chen C-B, Imamura T, Tapia IE, Somers VK, Zee PC and Lim DC (2024), A novel methodology for emotion recognition through 62-lead EEG signals: multilevel heterogeneous recurrence analysis.
Front. Physiol. 15:1425582.
doi: 10.3389/fphys.2024.1425582

Objective: Recognizing emotions from electroencephalography (EEG) signals is a challenging task due to the complex, nonlinear, and nonstationary characteristics of brain activity. Traditional methods often fail to capture these subtle dynamics, while deep learning approaches lack explainability. In this research, we introduce a novel three-phase methodology integrating manifold embedding, multilevel heterogeneous recurrence analysis (MHRA), and ensemble learning to address these limitations in EEG-based emotion recognition.

Approach: The proposed methodology was evaluated using the SJTU-SEED IV database. We first applied uniform manifold approximation and projection (UMAP) for manifold embedding of the 62-lead EEG signals into a lower-dimensional space. We then developed MHRA to characterize the complex recurrence dynamics of brain activity across multiple transition levels. Finally, we employed tree-based ensemble learning methods to classify four emotions (neutral, sad, fear, happy) based on the extracted MHRA features.

Main results: Our approach achieved high performance, with an accuracy of 0.7885 and an AUC of 0.7552, outperforming existing methods on the same dataset. Additionally, our methodology provided the most consistent recognition performance across different emotions. Sensitivity analysis revealed specific MHRA metrics that were strongly associated with each emotion, offering valuable insights into the underlying neural dynamics.

Significance: This study presents a novel framework for EEG-based emotion recognition that effectively captures the complex nonlinear and nonstationary dynamics of brain activity while maintaining explainability. The proposed methodology offers significant potential for advancing our understanding of emotional processing and developing more reliable emotion recognition systems with broad applications in healthcare and beyond.

KEYWORDS

heterogeneous recurrence analysis, emotion recognition, multi-channel EEG, dynamic system, ensemble learning

1 Introduction

The brain, one of the most intricate systems of the body, has been a subject of great interest for researchers aiming to unravel its complexities (Wolpaw and Birbaumer, 2006). The complexity of underlying nature (genetics) and the effect of nurture (life choices and experiences) creates an infinite number of possible stimuli and interactions, resulting in an evolving dynamic system within the brain. Understanding this dynamic system is crucial due to its pivotal role in various domains, including cognition, behavior, sleep, neurological disorders, and emotion (Lindquist et al., 2012; Akhand et al., 2023). To thoroughly explore this dynamic system, advanced technologies like functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) have been employed to measure brain activity and study interactions with the environment (Jellinger, 2003; Haynes and Rees, 2006; Tong and Pratte, 2012). Recently, EEG has become available as a wearable technology, making it an ideal choice for continuous monitoring of neural processes and brain activity.

Emotions are complex psychophysiological processes, yet universally, they are experienced similarly by all people. Thus, the study of emotion recognition has garnered significant attention in various fields, such as neurology, computer science, cognitive science, and psychology (Lindquist et al., 2012; Akhand et al., 2023). Prior research has leveraged the time-domain, (Liu et al., 2021; Chen D. et al., 2023), frequency-domain, (Gao et al., 2019; Houssein et al., 2022; Akhand et al., 2023), or time-frequency domain methods (Yuvaraj et al., 2023) to extract the features within EEG signals to identify emotions. Recent research (Chang et al., 2022; Yang et al., 2022) has focused on leveraging artificial intelligence and neural network models to enhance the accuracy and efficiency of emotion classification based on EEG data (Li J. et al., 2021; Tian et al., 2021). Dan et al. introduced a clustering-promoting semi-supervised method to enhance the performance of emotion recognition (Dan et al., 2021). Wang et al. established a convolutional neural network (CNN) framework for emotion recognition (Wang et al., 2020). These advancements not only contributed to the field of neuroscience but also have practical applications in human-computer interaction and mental health diagnoses (Chai et al., 2018). Thus, EEG has become an important technology for objective emotion recognition (Peng et al., 2023).

Recent developments in EEG-based emotion recognition have focused on improving classification accuracy and robustness through various techniques such as feature fusion, dynamic functional connectivity analysis, and deep learning architectures. Fusing frequency-domain features and brain connectivity features has shown promising results in cross-subject emotion recognition (Chen et al., 2022a). Dynamic functional connectivity analysis has also been employed to capture the time-varying characteristics of brain networks during emotional states (Liu et al., 2019). Novel deep learning architectures, such as deep CNNs (Chen J. et al., 2019), multi-scale masked autoencoders (Pang et al., 2024), transformer- and attention-based CNNs (Li C. et al., 2021; Si et al., 2023) have been proposed to enhance emotion recognition performance. Domain adaptation techniques have also been explored to facilitate the transfer of emotion recognition models across different subjects (Chen et al., 2022b). In addition to emotion recognition, EEG-based approaches have been applied to related fields, such as P300 wave detection, driving fatigue detection, and biometric authentication, where self-attentive channel-connectivity capsule networks (Chen C. et al., 2023; Wang et al., 2023) and attention-based multiscale CNN with dynamical graph convolutional network (GCN) (Wang et al., 2021) have demonstrated improved performance. Systems like E-Key (Xu et al., 2023a) combine biometric authentication with driving fatigue detection. EEG studies have also examined the effects of aging, task difficulty, and training on working memory capacities, highlighting EEG's diverse applications in cognitive research (Xu et al., 2023b).

Despite the progress made in EEG-based emotion recognition, several challenges remain. First, the nonlinear and nonstationary characteristics of EEG signals pose significant difficulties (Bazgir et al., 2018). Most machine learning based methodologies, such as linear discriminant analysis (Chen DW. et al., 2019), generalized linear regression (Li et al., 2019a), or Fast Fourier Transform (FFT) (Murugappan and Murugappan, 2013), often rely on linear assumptions, which fail to capture the nuanced nonlinear and nonstationary characteristics of EEG. Second, the complexity of multiple EEG electrodes capturing the interaction of brain activity and large volumes of data is another challenge. Deep learning models can address this complexity; however, they suffer from the "black box" problem while requiring substantial computational resources. Third, EEG signals present challenges in both temporal and spatial domains. While many studies focus on the temporal aspects of emotions (Liu et al., 2010; Zheng et al., 2019a), spatial information is equally important when adapting these methodologies in the future to neurological, sleep, or psychological disorders. Lastly, emotions are interconnected over time, with current emotional states being influenced by past emotions and potentially impacting future experiences (Thornton and Tamir, 2017). These transitions, between past, present, and future, have not been well studied using EEG signals.

To tackle these challenges, this paper presents an innovative three-phase methodology that characterizes and quantifies complex dynamic transitions of brain activities in multiple granularities while retaining high resolution to detect emotions from multi-channel EEG. In the first phase, manifold learning techniques are utilized to embed the dimensionality of high-dimensional 62-lead EEG signals into a more manageable lower-dimensional space. This embedding preserves the complex spatiotemporal characteristics of the signals, offering rich insights into brain activity while enhancing computational efficiency. In the second phase, we propose a novel multilevel heterogeneous recurrence analysis to characterize the nuanced, nonlinear, and nonstationary dynamic characteristics of the EEG signals at different granularities within the state-space domain. Our approach results in a quantification of dynamic patterns characterizing underlying brain activity, which cannot be achieved by other methods. The final phase employs ensemble supervised learning models that utilize metrics that quantify dynamic features and patterns within the EEG to classify each emotion. Ensemble learning not only improves overall performance but also provides a robust framework to prevent potential overfitting and account for variability in EEG data. This phase explains the decision-making processes underlying emotion classification. Experimental results show that our proposed methodology achieved accuracy and area under the receiver operating characteristic (ROC) curve (AUC) values of 0.7885 and 0.7552, respectively. These results surpass state-of-the-art studies using the same dataset. Moreover, our methodology provides the most consistent performance across different emotions compared to other models. Lastly, our method provides subtle quantifications and rich insights into the dynamic features of brain activity related to emotions.

In summary, this research introduces a novel recurrence analysis-based methodology for EEG-based emotion recognition that effectively captures the complex nonlinear and nonstationary dynamics of brain activity while maintaining explainability. The rest of this paper is organized as follows: Section 2 is a brief background relevant to our methodology; Section 3 describes the dataset employed to formulate our approach; Section 4 outlines the proposed methodology, structured in three distinct phases; Section 5 details the outcomes of our study; and Section 6 offers an in-depth discussion of the insights gained and conclusions drawn from our investigation.

2 Research background

In this section, we introduce the foundational concepts and background of our novel methodology, multilevel heterogeneous recurrence analysis (MHRA). We begin by discussing the basic principles of recurrence analysis (RA) and its evolution into heterogeneous recurrence analysis (HRA). Then, we review the development and application of HRA to complex transitions, which is further developed and refined into MHRA.

2.1 Recurrence analysis

Recurrence, defined as a situation where the state of a system at a certain time is very similar to its state at one or more previous times, is a fundamental feature of complex systems (Hatami et al., 2019). From Poincaré's initial descriptions of recurrence in the 1890s and the subsequent introduction of Recurrence Analysis (RA) by Webber and Zbilut in the 1980s (Khoo et al., 1996), the development of this analytical method has continuously evolved. In the early 2000s, Norbert Marwan and his colleagues made significant contributions to refining and applying RA, thereby enhancing its use across a variety of scientific fields, including geophysics (Eroglu et al., 2014; Lucarini et al., 2016), physiology (Khoo et al., 1996; Webber and Zbilut, 2005), meteorology (Bouabdelli et al., 2020), economics (Mosavi et al., 2020), and engineering (Shu et al., 2021). Consequently, RA has become one of the most widely used tools for analyzing dynamic complex systems. Note that the recurrence can be mathematically defined as $R_{i,j}$ in Eq. 1, indicating whether a recurrence exists between system states $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. If the proximity of $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, measured by $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$, is smaller than a predefined threshold ϵ, then a recurrence exists between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ (Eckmann et al., 1987; Marwan et al., 2007a; Marwan, 2008).

$$R_{i,j} = \mathbb{H}\left(\epsilon - \left\|\boldsymbol{s}_i - \boldsymbol{s}_j\right\|\right) \tag{1}$$

where $\mathbb{H}(x)$ is a Heaviside function, in which $\mathbb{H}(x) = 1$ if $x \geq 0$, and $\mathbb{H}(x) = 0$ otherwise; (Eckmann et al., 1987) $\boldsymbol{s}_t$ is the system state at time t. The recurrence of the system over a period of observation window is then represented as a symmetric matrix $\mathbf{R} = \left\{R_{ij}, \forall i, j\right\}$, which can be geometrically visualized as a Recurrence Plot (RP), typically shown as a dot plot where each axis represents the entire observation period and a dot plotted in the coordinate (i, j) indicates a recurrence exists between time i and j. This visualization not only highlights the frequency of recurrence but also reveals patterns and structures indicative of the dynamical behavior of the system, such as stability, periodicity, or chaotic dynamics (Eckmann et al., 1987). With analyzing the sophisticated geometric patterns in the RP, the nonlinear, nonstationary, and dynamic system characteristics are then quantified and characterized, known as Recurrence Quantification Analysis (Webber and Zbilut, 2005; Webber and Marwan, 2015). Notably, Marwan et al. generalized RP from a two dimensional matrix to a four dimensional tensor to capture the recurrence patterns within spatial data (Marwan et al., 2007b). RA has achieved tremendous success in various fields, for instance, it has been used to improve the normalization of electromyography signals (Avdan et al., 2023) detect series arc faults in photovoltaic systems (Amiri et al., 2022) and analyze histopathological images (Wang and Chen, 2022). Additionally, Donner et al. leveraged network topology to interpret the recurrence matrix $\mathbf{R}$, thereby developing a novel analytical framework known as the recurrence network (RN). This approach provides another perspective for effectively parsing the dynamic features of complex systems (Donner et al., 2010; Donner et al., 2011). Notably, our previous work developed an innovative RN to analyze the complex patterns in spatial data, which has already been implemented in characterizing surface roughness in ultra-precision machining (Chen et al., 2018) and in detecting invasive ductal carcinoma in breast cancer (Chen CB. et al., 2023).

2.2 Heterogeneous recurrence analysis

Traditional RA, including RP and RN, treats recurrence homogeneously, which presents limitations when characterizing nuanced dynamic features. To improve RA, Yang et al. developed HRA, which addresses the heterogeneity of recurrence and dramatically enhances the analytical capabilities (Yang and Chen, 2014; Chen and Yang, 2015; Chen and Yang, 2016). HRA differentiates recurrences based on the properties of system states, categorizing each state $\boldsymbol{s}_t$ into K different groups, denoted as $\mathcal{L}(\boldsymbol{s}_t) = k \in \{1, 2, \ldots, K\}$ for all t. It is crucial to note that the states within one category share similar system properties, while states in different categories exhibit distinct system properties. Heterogeneous recurrence is mathematically represented as Eq. 2:

$$\Omega_{ij} = \mathcal{L}(\boldsymbol{s}_i) \cdot \mathbb{H}\left(0 - \left\|\mathcal{L}(\boldsymbol{s}_i) - \mathcal{L}\left(\boldsymbol{s}_j\right)\right\|\right) \tag{2}$$

where $\mathcal{L}(\boldsymbol{s}_t)$ indicates the category of state $\boldsymbol{s}_t$ for all t, $\|\cdot\|$ represents the norm, and $\mathbb{H}(\cdot)$ denotes a Heaviside function. This approach means that if $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ belong to the same category $\mathcal{L}(\boldsymbol{s}_t)$, a recurrence exists between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ in category $\mathcal{L}(\boldsymbol{s}_i)$. This method not only enhances the resolution of single-state recurrences but also reveals the sophisticated dynamics of transitions, which are often limited by conventional RA. Furthermore, HRA employs the Iterated Function System (IFS), an iterative projection function used to construct fractals, to project a sequence of transitions into a fractal space. This utilization of a fractal structure's geometric features allows for a detailed characterization of complex dynamic properties associated with transitions (Yang and Chen, 2014). The analysis and quantification of these geometric structures, termed Heterogeneous Recurrence Quantification Analysis (HRQA), enable HRA to provide greater resolution in characterizing complex dynamic patterns. HRA has been successfully implemented to characterize complex systems in various fields, including finance (Zhang et al., 2023) medicine (Chen and Yang, 2015; Chen and Yang, 2016; Cheng et al., 2016; Chen et al., 2020; Avdan et al., 2024) physics (Yang and Chen, 2014) and engineering (Kan et al., 2016; Yang et al., 2020; Peng and Chen, 2023). Notably, Chen et al. extended the HRA to develop Spatial HRA (SHRA) for investigating complex recurrence patterns in spatial data. SHRA has been implemented in medical imaging (Yang et al., 2020; Van Booven et al., 2024a; Van Booven et al., 2024b) and additive manufacturing (Chen R. et al., 2019; Chen, 2019). However, while HRA can effectively characterize subtle nonlinear dynamic properties including complex transitions of a system, there has been little development of systematically investigating system dynamics across multiple scales, which could reveal additional system characteristics (Chen et al., 2017; Chen C-B. et al., 2019). To address this gap, we developed a novel HRA-based methodology to more precisely define multilevel transitions.

3 Data: 62-lead EEG signals

We utilized the Shanghai Jiao Tong University (SJTU) Emotion EEG Dataset for Four Emotions (SEED-IV), a specific subset of the broader SJTU Emotion EEG Dataset (available at https://bcmi.sjtu.edu.cn/~seed/), to develop our methodology for emotion recognition (Zheng et al., 2019b). The SEED-IV dataset includes both EEG and eye movement signals associated with four distinct emotions, neutral, sadness, fear, and happiness, collected from 15 college-aged participants (seven males and eight females, aged 20–24, all right-handed). Each participant was outfitted with a 62-channel EEG cap (Compumedics Neuroscan, Australia) and eye-tracking glasses (SensoMotoric Instruments, Germany). The data were gathered while participants watched 72 carefully selected film clips, each designed to elicit one of the target emotions. Each clip had a duration of approximately 2 minutes and was shown only once to avoid the effects of repetition. Participants attended three separate sessions on different days, each comprising 24 trials with six trials per emotion. Each trial began with a 5-s introductory hint, followed by a 45-s period for self-assessment, during which participants rated their emotional experience. Data from participants who either did not experience the intended emotion or exhibited weak emotional arousal were excluded from the analysis. The primary objective of this research is to identify these four emotions using dynamic features extracted from multi-channel EEG signals. For the purposes of this study, we focused exclusively on the raw EEG data from 62 channels, capturing the complex brain dynamics associated with each emotional state, while the eye movement data were not utilized in the analysis.

4 Multilevel heterogeneous recurrence analysis for emotion recognition

This study aims to identify four emotions by analyzing the complex spatiotemporal dynamics within high-dimensional EEG signals. We developed a novel three-phase methodology, named MHRA methodology, summarized in Figure 1, to accomplish this goal. The methodology comprises the following phases: Phase 1. Manifold Embedding: To preserve the intricate nonlinear spatiotemporal characteristics of raw EEG data while minimizing computational demands, we employed a manifold learning technique. This method projects the high-dimensional EEG data into a lower-dimensional space, thereby simplifying the dataset while retaining its essential features. Phase 2. MHRA: To capture the complex dynamic brain activity reflected in EEG signals, we developed a novel MHRA. This approach systematically portrays the multilevel dynamic characteristics of EEG data using fractal structures and quantifies the geometric features of these fractals to extract dynamic features for emotion recognition. Phase 3, Supervised Ensemble Learning: To differentiate emotions based on the dynamic properties extracted from EEG signals, we utilized various advanced ensemble learning techniques, including Random Forest, XGBoost, and Adaboost. The high accuracy achieved by our proposed model highlights the crucial role these dynamic properties play in effectively recognizing emotions. Further details of each phase are discussed in the remainder of this section.

4.1 Phase 1: manifold embedding

Massive data sizes and high dimensionality are two notorious obstacles in the field of data analytics. Effectively retaining data properties while efficiently processing data is crucial. This study analyzes data from 62-lead EEG signals, which presents significant challenges due to their massive data size and high dimensionality. Although these high-dimensional data offer superior spatiotemporal resolution, the inherent complexities of these EEG signals significantly increase the difficulties of data processing and analysis. Particularly in terms of the highly computational

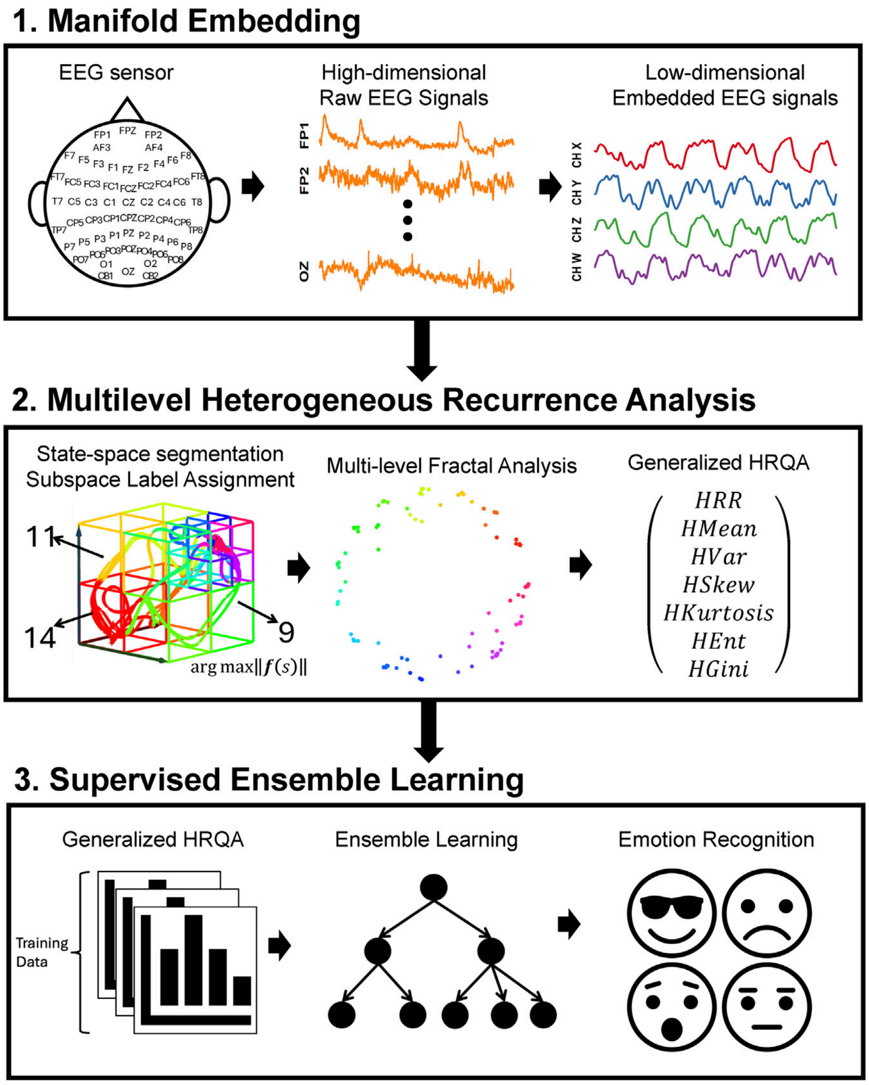

1. Manifold Embedding
EEG sensor
High-dimensional Raw EEG Signals
Low-dimensional Embedded EEG signals
FP1
FP2
OZ
CH X
CH Y
CH Z
CH W
2. Multilevel Heterogeneous Recurrence Analysis
State-space segmentation Subspace Label Assignment
11
14
9
arg max||f(s)||
Multi-level Fractal Analysis
Generalized HRQA
HRR
HMean
HVar
HSkew
HKurtosis
HEnt
HGini
3. Supervised Ensemble Learning
Generalized HRQA
Training Data
Ensemble Learning
Emotion Recognition

FIGURE 1
Overview of three-phase methodology, MHRA methodology, applied to EEG for emotion recognition. Phase 1. Manifold Embedding: A manifold learning method is applied to high-dimension EEG data to embed subtle nonlinear spatiotemporal characteristics into lower dimensions, reducing computational demands. Phase 2. MHRA: We developed the MHRA to quantify dynamic transitions using fractal representation at multiple levels. Phase 3. Supervised Ensemble Learning: Advanced ensemble learning methods are leveraged to analyze MHRA metrics for emotion recognition.

demands they impose. Therefore, reducing analytical and computational efforts to a manageable level while retaining the original data's spatiotemporal characteristics is essential. Traditional dimensionality reduction techniques, such as principal component analysis and singular value decomposition, often fall short with large, complex datasets. They tend to overlook the nonstationary, nonlinear features of the data, leading to extended computation times and ineffective dimension reduction outcomes that do not accurately reflect the original data's information (Roweis and Saul, 1979; Elgamal and Hefeeda, 2015; Pouyet et al., 2018).

To address these challenges, we have utilized manifold embedding, a technique within manifold learning that is particularly effective at uncovering the low-dimensional manifold structure embedded in high-dimensional spaces. It allows us to map high-dimensional data onto a lower-dimensional space efficiently, retaining the data's intrinsic and nonlinear properties. This simplification of the dataset preserves essential spatiotemporal information, facilitating further analysis (Turchetti and Falaschetti, 2019). Notably, manifold embedding encompasses various techniques collectively known as Nonlinear Dimensionality Reduction (NLDR). Common methods within NLDR include Uniform Manifold Approximation and Projection (UMAP), which constructs a high-dimensional graph representation of the data and then optimizes a low-dimensional graph to be as structurally similar as possible; Locally Linear Embedding (LLE), which preserves local properties of the data; Spectral Embedding, which uses the eigenvalues of the graph Laplacian to perform dimensionality reduction; Isomap, which preserves

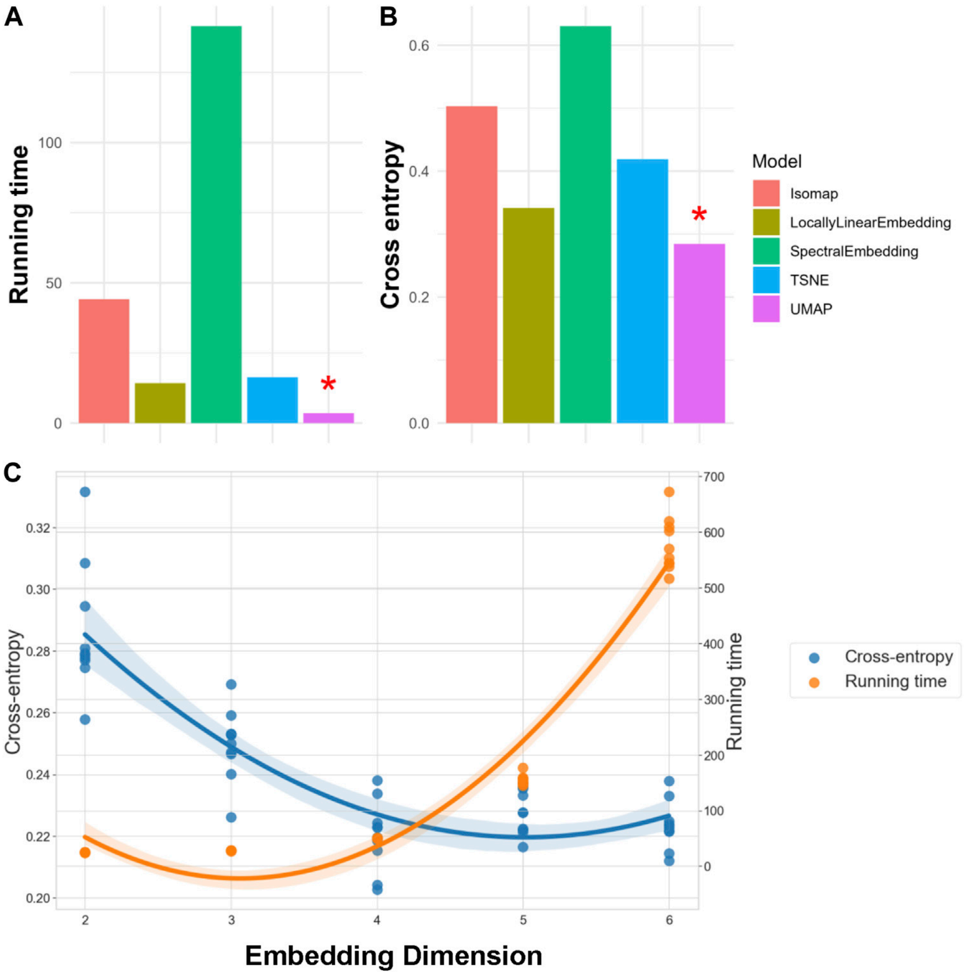


FIGURE 2
Evaluation of NLDR Methods and selecting the optimal number of Embedding Dimension. Panels **(A, B)** compare five manifold embedding candidates by running time and cross-entropy, respectively, indicating that UMAP is the best method for our specific dataset. Panel **(C)** illustrates how running time and cross-entropy were used to identify four as the optimal number of embedding dimensions to preserve critical spatiotemporal features within the dataset.

geodesic distances between data points; and t-distributed Stochastic Neighbor Embedding (t-SNE), which minimizes the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding (Meilă and Zhang, 2024).

To select the most appropriate NLDR method, we consider both the quality of dimensional reduction and computational efficiency. For assessing reduction quality, we utilize cross-entropy to compare the differences between the original and reduced signals. Cross-entropy is expressed as Eq. 3:

$$C(l) = -\sum_t \|l(s_t)\| \log(\|s_t\|) \tag{3}$$

where $l(\boldsymbol{s}_t)$ is the lower-dimensional projection of signals $\boldsymbol{s}_t$ converted by function $l(\cdot)$, and $\|\cdot\|$ takes L_2-norm of multi-channel signals. The NLDR technique with the best retention of original signals within the reduced signals will have the lowest cross-entropy value, indicating they contain a similar amount of information.

We evaluated each NLDR technique by analyzing a 10% random sample of SEED-IV data across ten replications. The performance of these manifold embeddings is presented in Figure 2. Panel A displays the average running time, while Panel B shows the average cross-entropy. Note that a lower running time indicates better efficiency, and a lower cross-entropy signifies higher information retention. For our 62-lead EEG data, UMAP not only achieved the lowest cross-entropy but also the best performance in terms of running time (Mcinnes et al., 2020), as indicated by a red asterisk. We used the same criteria, running time and cross-entropy, to determine the optimal number of embedding dimensions, referring to the number of dimensions in the lower-dimensional space. Our findings reveal that as embedding dimensions increase, the running time grows exponentially, while the improvement in cross-entropy diminishes. Figure C demonstrates these trends in UMAP, and it shows that the optimal performance, both in terms of running time and cross-entropy, occurs at four embedding dimensions. Notably, we also fine-tuned hyperparameters for all the manifold learning methods to optimize embedding performance. For our final selected method, UMAP, these hyperparameters included the number of neighbors (set to 5), the minimum distance between points in the low-dimensional space (set to 0.1), and the spread of the data points (set to 1.0). These settings were chosen to balance the retention of the data's intrinsic structure and computational efficiency. Consequently, UMAP was selected to embed the 62-lead EEG signals into four dimensions, effectively balancing critical spatiotemporal feature retention with computational efficiency.

4.2 Phase 2: multilevel heterogeneous recurrence analysis

After embedding the 62-lead EEG signals into a low-dimensional space, we deployed the proposed MHRA to characterize the dynamic spatiotemporal characteristics of brain activity. The MHRA is a state-space domain method comprising three major steps: 1. Heterogeneous state-space representation, 2. Fractal representation, and 3. Generalized HRQA. These steps outline a systematic and comprehensive approach to characterizing complex dynamic systems.

4.2.1 Heterogeneous state-space representation

To capture and delineate the recurrence dynamics of a system, we first transform time series data into a trajectory within a state space, $\mathbb{S}$, representing all possible states of the system. Notably, each point of a d-dimensional time series is projected as a corresponding point in the d-dimensional state-space, denoted by $\boldsymbol{s}_t = (x_t^1, x_t^2, \ldots, x_t^d) \in \mathbb{S}$, where each dimension of the state space corresponds to a different measure of the system. Consequently, the evolution of the time series data forms a trajectory in this space, denoted as $\boldsymbol{s} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_t\}$, and the geometric properties of this trajectory reveal the dynamic characteristics of the system.

Subsequently, to achieve a higher resolution of the recurrence properties, we constructed a heterogeneous state-space by dividing the original state-space, $\mathbb{S}$, into K subspaces, $\mathbb{S}_k$, denoted as $\mathbb{S} = \bigcup_{k\in\{1,\ldots,K\}} \mathbb{S}_k$. This segmentation helps differentiate recurrences, as system states within the same subspace exhibit similar system properties, and states in different subspaces display distinctly different system properties. Notably, there are many space segmentation methods that serve different purposes. This study utilizes one of widely used space segmentation method, Voronoi tessellation (Asghar et al., 2020), focusing on the similarity within each subspace when segmenting heterogeneous state-spaces. Therefore, by assigning the system states within the same subspace the same category label, denoted as $\mathcal{L}(\boldsymbol{s}_t) = k, \forall \boldsymbol{s}_t \in \mathbb{S}_k, \forall k \in \mathcal{K} = \{1, \ldots, K\}$, where $\mathcal{L}$ is a label assignment function maps each state $\boldsymbol{s}_t$ to a categorical variable k, the trajectory of evolution forms a categorical sequence that reveals the dynamic transitions within the system. To ensure that the trajectory retains sufficient patterns to accurately represent sophisticated emotions, a 20-s window was employed to capture the characteristics of brain activity in this study. Figure 3 conceptually illustrates the process of heterogeneous state-space representation used in this study. Initially, the embedded EEG signals are transformed into a trajectory within the state space (shown in three dimensions for better visualization). Subsequently, a space segmentation method, Voronoi tessellation, is employed to create a heterogeneous state-space representation, where each Voronoi cell represents a distinct subspace. By assigning a category to each subspace, the EEG signals are converted into a categorical sequence that reveals the dynamic evolution of brain activity.

Notably, Voronoi tessellation, typically a semi-supervised method, requires specifying the number of subspaces in advance. Selecting an inappropriate number of subspaces can significantly impact the effectiveness of information extraction. Determining the optimal number of subspaces is thus crucial for accurately representing the heterogeneous state-space. This research utilized the Davies-Bouldin Index, a measure of clustering quality, to find the optimal number of subspaces. Initially, as illustrated in Figure 4, we divided the original state-space into 10 subspaces and incrementally evaluated up to 100 subspaces. The black line represents the Davies-Bouldin Index, the smooth blue line indicates a fitted curve of the index values, and the grey shading denotes the confidence interval. A lower Davies-Bouldin Index indicates more effective clustering, with clear separation between subspaces. The index stabilized after 45 subspaces, identifying this number as optimal for our dataset. Accordingly, we segmented the state-space into 45 distinct subspaces to enhance the resolution of dynamic characteristics.

4.2.2 Fractal representation

To characterize the dynamic characteristics of state transition patterns, this study leverages the fractal topological structure to capture the nuanced features. Fractals are mathematical structures portrayed by self-similarity, meaning each part of the fractal replicates the whole on a smaller scale. This intrinsic property makes fractals particularly suited for modeling heterogeneous recurrences, as their recursive nature can effectively mirror the irregular and complex patterns observed in such phenomena. By employing fractals, one can capture the nuanced nonlinear and nonstationary variations inherent in heterogeneous recurrences, providing a more accurate and comprehensive understanding of their dynamics (Yang and Chen, 2014; Cheng et al., 2016; Kan et al., 2016; Yang et al., 2020).

Therefore, after the embedded EEG signals are converted into a trajectory in the heterogeneous state space, revealing the system's evolution, the trajectory is then projected into a fractal space using Iterated Function System (IFS). Notably, this IFS projection is a one-to-one mapping where each trajectory forms its own fractal structure that reveals the nuanced recurrence dynamics (as shown in Figure 5). Each transformed point strategically captures its transition order prior to its corresponding point in the state sequence.

The IFS iteratively maps each element of categorical sequence, k, which reflects the category of subspace of the corresponding embedded EEG signals, $\mathcal{L}(\boldsymbol{s}_t) = k \in \mathcal{K}$, to a unique IFS address in the fractal circle through the following function (Eq. 4):

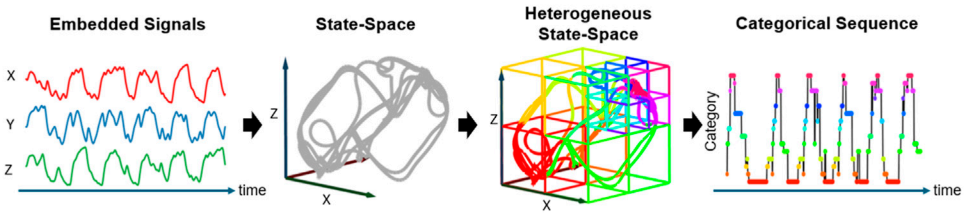


FIGURE 3
Heterogeneous State-Space Representation. This flowchart illustrates how EEG signals are transformed into a trajectory within the heterogeneous state space, and how these transitions are categorized into a dynamic sequence. The EEG signals are first transformed into a trajectory within the state space, followed by the application of Voronoi tessellation to segment the space into distinct subspaces. Each subspace, represented as a Voronoi cell, is assigned a specific category, illustrating the formation of a categorical sequence that captures the dynamic evolution of brain activity.

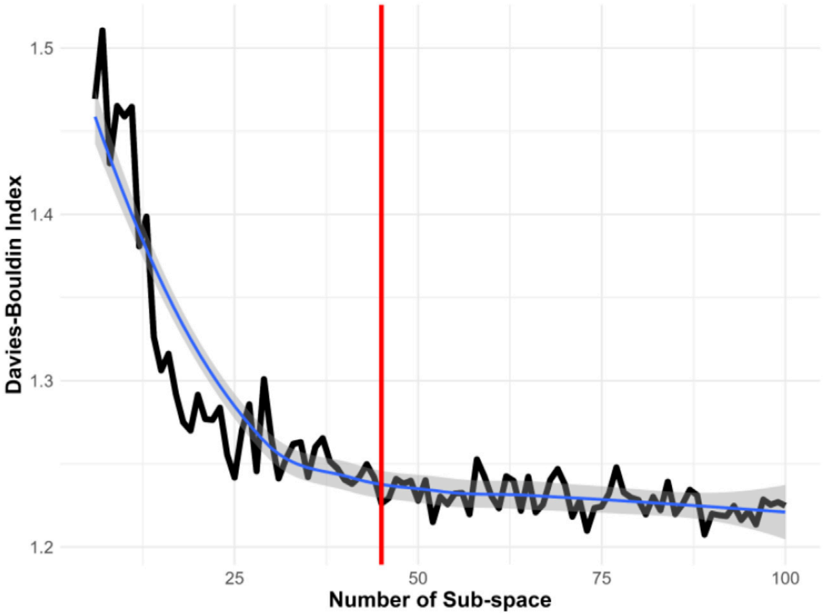


FIGURE 4
Determining the Optimal Number of Subspaces Using the Davies-Bouldin Index. This index assesses the effectiveness of different subspace configurations, with a lower score indicating better clustering quality. The analysis suggests that 45 subspaces provide the most informative clustering in this study.

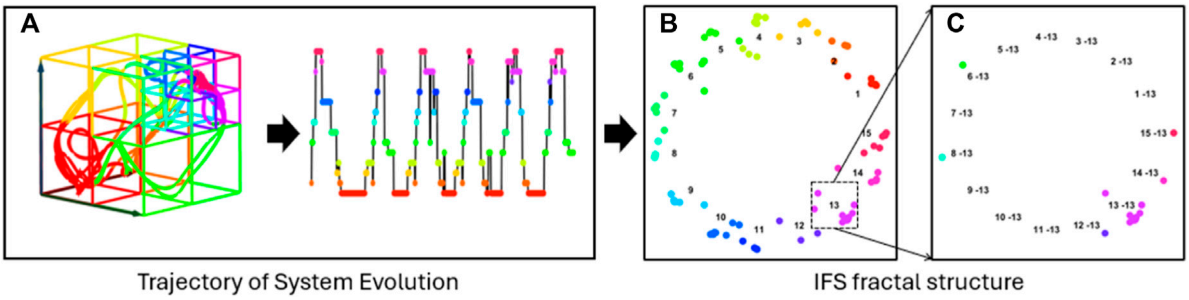


FIGURE 5
Fractal Structure Construction. Panel **(A)** displays the trajectory of system evolution as a categorical sequence, and **(B)** illustrates the projection of this trajectory into a unique fractal structure using an Iterated Function System (IFS). The self-similar nature of the fractal enables the investigation of dynamic patterns across multiple scales. **(C)** Depicts a second-level fractal derived from **(B)**, revealing dynamic characteristics on a different scale.

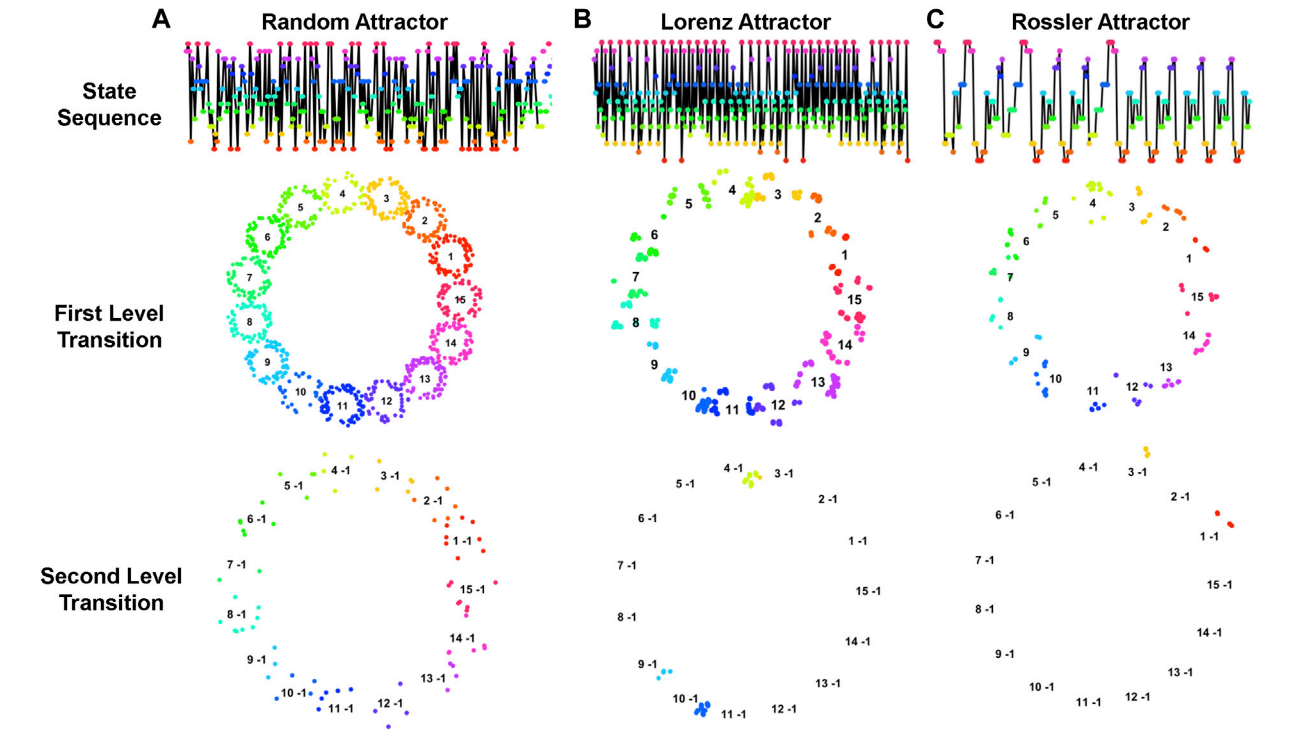


FIGURE 6
Trajectories of Dynamic Systems with Corresponding Fractal Structures. This figure illustrates the trajectories and fractal patterns of three dynamic systems: **(A)** random attractor, **(B)** Lorenz attractor, and **(C)** Rossler attractor. The top layer figures indicate the trajectories of the systems, the second- and third-layer figures illustrate the corresponding fractal structures of first- and second-level transitions. The topological structures of fractals characterize the dynamic properties of the systems.

$$I(t) = \begin{pmatrix} I_x(t) \\ I_y(t) \end{pmatrix} = \Phi\left(k, \begin{pmatrix} I_x(t-1) \\ I_y(t-1) \end{pmatrix}\right)$$

$$= \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix} \cdot \begin{pmatrix} I_x(t-1) \\ I_y(t-1) \end{pmatrix} + \begin{pmatrix} \cos\left(\frac{2\pi k}{K}\right) \\ \sin\left(\frac{2\pi k}{K}\right) \end{pmatrix}, \text{with } I(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{4}$$

where $\Phi(k, I(t-1))$ maps an IFS address $I(t)$ based on the subspace category k at time t and incorporates the influence of all previous states provided by $I(t-1)$. The circular address is determined by two components: (1) current state and its assigned category variable k, via the transformation $(\cos(2\pi k/K), \sin(2\pi k/K))^T$; (2) all the previous states, adjusted by a scaling factor α, through the iterative function. Note that α is defined as $\alpha = \tau \cdot \sin(\pi/K)/(1 + \sin(\pi/K))$ to ensure address remains distinct, where $0 < \tau < 1$ (in this study, $\tau = .99$).

This IFS is designed to provide a self-similar fractal structure that embeds the information from all previous states, thereby enabling the formation of fractal patterns of spatial transitions at multiple scales. Note that this fractal structure allows us to investigate dynamic characteristics of transitions at multiple levels. For instance, as shown in Figure 5B, the distribution of 15 individual subspaces, $\{1, 2, \ldots, 15\}$, shows the recurrence variations in different subspaces, named first-level transition; Figure 5C reveals the recurrence variations of two-state transitions, $\{\{1 \rightarrow 13\}, \{2 \rightarrow 13\}, \ldots, \{13 \rightarrow 13\}\}$, named second-level transition, in a zoomed-in fractal of Figure 5B. This fractal representation precisely captures the nuanced characteristics of system dynamics.

Notably, different trajectory patterns form various fractal structures that reveal diverse dynamic characteristics of the corresponding systems. As demonstrated in Figure 6, trajectories of three different dynamic systems, including random, Lorenz, and Rossler attractors, along with their corresponding fractal structures in the first- and second-level transitions are quite different. It is noteworthy that systems with more randomness typically yield a less informative fractal structure, whereas systems with specific patterns yield a more distinctive fractal structure that is characteristically unique. Thus, analyzing the topological structure of multilevel fractals increases the resolution of dynamic system properties.

However, fractal representation is sensitive to the categorical labels, which are presented as a sequence of consecutive positive integers from 1 to K, each indicating a specific subspace within the state-space. As Figure 7 illustrates, even when the same trajectory underlies the same heterogeneous state-space structure, various fractal structures can emerge due to different subspace label assignments. This variability significantly influences the effectiveness of dynamic characterization. Therefore, since the dynamic characteristics of the system are derived by analyzing the fractal topology and complexity, optimizing subspace label assignments is crucial for achieving the most accurate fractal representation.

However, determining the optimal subspace label assignment is a challenging task. For example, to evaluate all possible 45 subspace assignments would be 45! (approximately 1.1962e+56) scenarios, making it impractical to exhaustively test all permutations to find the best assignment. To address this challenge, we propose a novel Genetic Algorithm (GA) to achieve a heuristic solution for optimizing subspace label assignment, as illustrated in Algorithm 1. GA is a type of evolutionary algorithm which generates solutions to problems inspired by natural selection (Holland, 1992).

```
INPUT:
  S: Initial sequence from which to create the trajectory
  K: Number of distinct labels (derived from S if not provided)
  l.size: Size of the sample Pool
  slt: Number of instances to select for reproduction
  pm: Mutation probability
  ga.iter: Number of iterations for the genetic algorithm
BEGIN:
  //Initialize GA parameters
  GAobj = GNW(S,K)//Create network structure representing the trajectory of S
  //Generate initial population
  SamplePool = GAinit(K,l.size)//Create a sample pool of sequence for GA
  //Genetic algorithm main loop
  FOR iter = 1 TO ga.iter DO
    //Evaluate fractal dimension of each instance in the sample pool
    FOR EACH instance IN SamplePool DO
       fitness[instance] = Fitness(instance,GAobj)
     END FOR
     //Select top individuals for reproduction
     Selected = select_top(fitness,slt)
     //Update sampling pool through reproduction and mutation
     SamplePool = reproduce_and_mutate(Selected,pm)
     //Optional: Convergence check to break loop early
     IF check_convergence(fitness) DO
       BREAK
     END IF
   END FOR
   //Determine the best solution
   BestSolution = find_best(SamplePool)
   RETURN BestSolution
```

Algorithm 1. Genetic Algorithm for Optimizing Label Arrangements.

*Fitness function returns the fractal dimension of the fractal structure generated by the input instance.

In this study, we modified the GA as follows:

- Initial Population: Started with 50,000 random subspace label assignments, each offering a unique labeling approach within the EEG state-space.
- Evaluation: Each assignment is assessed for fractal complexity to gauge effectiveness in describing the underlying trajectory structure.
- Selection and Generation: Post-evaluation, another 50,000 assignments are generated using genetic crossover and mutation techniques to explore new solutions.
- Optimization: Assignments with the highest fractal complexity, indicative of effective system dynamics capture, are selected.
- Iteration: This cycle of generation, evaluation, and optimization continues until a fractal complexity threshold is reached or no further improvements are observed.

Note that fractal complexity in this study is measured using the Minkowski fractal dimension, which involves covering the fractal with boxes of a specific size and counting the number needed to completely cover the fractal. This process is repeated with progressively smaller boxes (Hunt et al., 1939). The Minkowski fractal dimension for a fractal $\mathcal{F}$ can be mathematically expressed as Eq. 5:

$$\dim_{box}(\mathcal{F}) = \lim_{\varepsilon\to 0} \frac{\log \xi(\varepsilon)}{\log \frac{1}{\varepsilon}} \tag{5}$$

where ξ denotes the number of boxes with a side length of ε. A higher Minkowski dimension suggests a more complex fractal, implying that it retains richer information.

4.2.3 Generalized heterogeneous recurrence quantification analysis

The fractal representation clusters the system's trajectory at multiple scales with fractal structures, which demonstrate the heterogeneous recurrence dynamics of a system on the two-dimensional coordinates. To effectively capture this heterogeneity in system recurrences, a new measurement approach has been developed that employs the fractal structure for quantifying these heterogeneous recurrences (Yang and Chen, 2014; Chen and Yang, 2015; Chen and Yang, 2016). Rather than treating all recurrences uniformly, this method, known as HRQA, specifically characterizes recurrent patterns based on the diverse states or transitions that are mapped onto the fractal structure, thereby enhancing the analytical capabilities of recurrence quantifiers. Chen and Yang derived a series of HRQA methodologies based on this fractal representation (Yang et al., 2020). However, traditional HRQA methods encounter scalability issues when attempting to quantify transitions at different levels. In response to this challenge, this research introduces a generalized HRQA system that addresses scalability issues to assess system recurrences. This advanced system allows for a more nuanced analysis of the dynamics inherent within different level transitions.

To quantify the fractal representation, the first step is to identify the sets of states falling into different level transitions in the fractal representation. Since the IFS assigns unique addresses in the circles to clusters of state sets, we define these heterogeneous recurrence sets $\mathcal{C}_{k_1,k_2,\ldots,k_N}$ as Eq. 6:

$$\begin{aligned}\mathcal{C}_{k_1,k_2,\ldots,k_N} = \{ f(k_1|k_2,\ldots,k_N)\colon \mathcal{L}(\boldsymbol{s}_t) = k_1, \mathcal{L}(\boldsymbol{s}_{t-1}) = k_2, \ldots, \\ \mathcal{L}(\boldsymbol{s}_{t-N+1}) = k_N, \forall k_t \in \mathcal{K}\}\end{aligned} \tag{6}$$

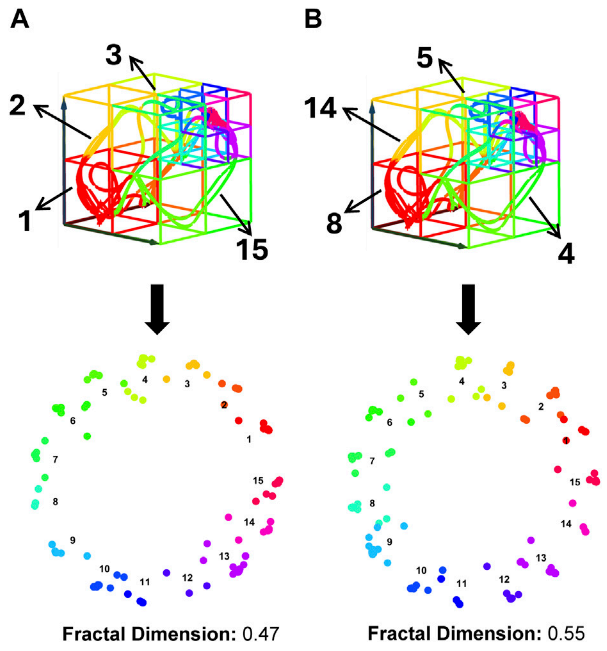


FIGURE 7
Impact of Subspace Label Assignment on Dynamic Feature Characterization. Both panels **(A, B)** display identical trajectories within the same heterogeneous state-space structure, yet they have different subspace label assignments. These differences lead to the distinct fractal structures shown in the lower layers of each panel, with varying fractal dimension values. Fractal dimension is used here to quantify the complexity of fractal structures, where higher values indicate increased complexity and greater detail retention across scales.

TABLE 1 Number of LASSO selected HRQA metrics for each emotion.

Emotion	Number of selected metrics
Neutral	216
Sad	270
Fear	89
Happy	108

Here, the subscript $k_1, k_2, .., k_N$ represents an N^{th}-level transition sequence. For instance, $\mathcal{C}_{k_1} = \{\mathcal{L}(\boldsymbol{s}_t) = k_1\}$ represents the recurrence set of first-level transition, and $\mathcal{C}_{k_1,k_2} = \{\mathcal{L}(\boldsymbol{s}_t) = k_1, \mathcal{L}(\boldsymbol{s}_{t-1}) = k_2\}$ represents the recurrence set of the second-level transition. Notably, we also define $\mathcal{C}_\phi$ as zero-level transition to represent overall transitions without specifying any transition pattern. This allows for the investigation and quantification of the system dynamics from a comprehensive system perspective. To simplify, we will use $\mathcal{N}$ to indicate the $k_1, k_2, .., k_N$ in the subsequent discussion. The generalized HRQA metrics are depicted in the following section.

4.2.3.1 Heterogeneous recurrence rate (HRR)

$$HRR(\mathcal{N}) = \left(\overline{\overline{\mathcal{C}}}/L\right)^2 \quad (7)$$

HRR quantifies the proportion of a specific N^{th}-level transition $\mathcal{N}$ occurred in an observed sequence. Note that $\overline{\overline{\mathcal{C}}}$ represents the cardinality of $\mathcal{C}_{k_1,k_2...,k_N}$ and L indicates the length of the observed sequence.

4.2.3.2 Heterogeneous recurrence mean (Hmean)

To scale the HRQA for different N^{th}-level transition, we define an adjusted distance $d_{i,j}^N$ for two addresses i and j for each $\mathcal{C}_{k_1,k_2...,k_N}$ as $d_{i,j}^N = d_{i,j}/\alpha^N$, where $d_{i,j}$ is the original distance, α is the scaling factor in Eq. 4, and N indicates the transition level. Then the generalized central tendency, variance tendency, skewness, and kurtosis of one local fractal cluster for N^{th}-level transition are quantified as in Eqs 8–13 shown below, respectively.

$$HMean(\mathcal{N}) = \frac{\sum_{i=1}^{\overline{\overline{\mathcal{C}}}} \sum_{j=i+1}^{\overline{\overline{\mathcal{C}}}} d_{i,j}^N}{\overline{\overline{\mathcal{C}}}\left(\overline{\overline{\mathcal{C}}} - 1\right)/2} \quad (8)$$

4.2.3.3 Heterogeneous recurrence variance (HVar)

$$HVar(\mathcal{N}) = \sum_{i=1}^{\overline{\overline{\mathcal{C}}}} \frac{\sum_{j=i+1}^{\overline{\overline{\mathcal{C}}}} \left(d_{i,j}^N - HMean(\mathcal{N})\right)^2}{\overline{\overline{\mathcal{C}}}\left(\overline{\overline{\mathcal{C}}} - 1\right)/2} \quad (9)$$

4.2.3.4 Heterogeneous recurrence skewness (HSkew)

$$HSkew(\mathcal{N}) = \frac{\sum_{i=1}^{\overline{\overline{\mathcal{C}}}} \frac{\sum_{j=i+1}^{\overline{\overline{\mathcal{C}}}} \left(d_{i,j}^N - HMean(\mathcal{N})\right)^3}{\overline{\overline{\mathcal{C}}}\left(\overline{\overline{\mathcal{C}}} - 1\right)/2}}{HVar(\mathcal{N})^{\frac{3}{2}}} \quad (10)$$

4.2.3.5 Heterogeneous recurrence kurtosis (HKurtosis)

$$HKurtosis\left(\mathcal{N}\right) = \frac{\sum_{i=1}^{\bar{\bar{C}}} \frac{\sum_{j=i+1}^{\bar{\bar{C}}}\left(d_{i,j}^{N} - HMean\left(\mathcal{N}\right)\right)^{4}}{\bar{\bar{C}}\left(\bar{\bar{C}}-1\right)/2}}{HVar\left(\mathcal{N}\right)^{2}} \tag{11}$$

4.2.3.6 Heterogeneous recurrence entropy (HENT)

$$HENT\left(\mathcal{N}\right) = -\sum_{b=1}^{B} \Pr\left(b\right) \ln\left(\Pr\left(b\right)\right) \tag{12}$$

4.2.3.7 Heterogeneous recurrence gini index (HGini)

$$HGini\left(\mathcal{N}\right) = 1 - \sum_{b=1}^{B} \Pr\left(b\right)^{2} \tag{13}$$

Note that the calculation of $HENT\left(\mathcal{N}\right)$ utilizes Shannon entropy, based on the probability distribution derived from the distance matrix $d_{i,j}^{N}$. The histogram of distance matrix d^{N} is segmented into B qual bins, ranging from 0 to $\max\left(d^{N}\right)$. Consequently, for every bin b up to B, the probability of b is defined as Eq. 14:

$$\Pr\left(b\right) = \frac{1}{\bar{\bar{C}}\left(\bar{\bar{C}}-1\right)} \#\left\{\frac{b-1}{B}\max\left(d^{N}\right) < d_{i,j}^{N} \leq \frac{b}{B}\max\left(d^{N}\right)\right\} \tag{14}$$

We deployed the proposed generalized HRQA to quantify the fractal representations derived from the embedded EEG. In this research, we addressed different resolutions of dynamic to the second-level transitions. A total $7 + 45 \times 7 + 45^{2} \times 7 = 14497$ HRQA metrics that delineate complex dynamic brain activity were then extracted for emotion recognition.

4.3 Phase 3: supervised ensemble learning

The final phase of our methodology is to develop a supervised machine learning model that classifies the outcome using HRQA metrics as the input. We chose ensemble learning for its ability to handle complex, nonlinear patterns and relationships within the data while achieving high accuracy in classifying the outcome. We evaluated three decision-tree-based ensemble machine learning algorithms, the adaptive boosting method (Adaboost), random forest classification (Random Forest), and extreme gradient boosting (XGBoost), for accurately identifying the four emotions.

Decision-tree-based ensemble machine learning methods effectively handle complex nonlinear relationships by integrating multiple decision trees. These methods continuously refine the model by adding new trees specifically designed to correct errors identified in existing trees. The methods evaluated in our methodology differ primarily in their training approaches: XGboost and Adaboost use boosting to focus on correcting mispredictions by adjusting data weights, while Random Forest employs bagging, sampling equally across data points. These ensemble strategies surpass single tree models by leveraging a majority vote from various trees, thus expanding the solution space and reducing overfitting through averaged outcomes.

Although tree-based models are effective at capturing complex relationships in data, their efficiency and performance can be significantly influenced by the number of predictors. These models are particularly sensitive to the inclusion of irrelevant or noisy predictors, which can increase model complexity and lead to a higher risk of overfitting, where the model learns the noise in the training data rather than the underlying patterns (Hu and Li, 2022). To overcome this issue, we employed the Least Absolute Shrinkage and Selection Operator (LASSO) for variable selection to reduce the number of HRQA metrics used in developing our emotion recognition models.

LASSO is particularly effective for models burdened by high-dimensional data, as it helps in reducing the risk of overfitting by imposing a constraint on the sum of the absolute values of the model parameters. This regularization process not only shrinks less important feature coefficients to zero but also simplifies the model by retaining only those variables that significantly contribute to the predictive power (Roth, 2004).

We executed the LASSO algorithm 30 times and selected metrics that consistently had non-zero coefficients across these runs. Table 1 illustrates the final number of HRQA metrics selected for each emotion. Our results indicate that the emotions ‘Neutral’ and ‘Sad’ are associated with a broader range of dynamic characteristics of brain activity, while ‘Fear’ and ‘Happy’ are linked to relatively fewer features.

To identify the four emotions based on their dynamic characteristics extracted from LASSO selected HRQA metrics, we tailored a classification model for each specific emotion. We evaluated three supervised ensemble learning methods, AdaBoost, XGBoost, and Random Forest, for emotion recognition. For each method we used the One-vs-All (OvA) strategy, where each emotion was classified independently as the positive class against all others grouped as the negative class. To ensure the robustness and reliability of our models, we adopted a rigorous testing protocol. The data was randomly split into a training dataset (90% of the total dataset) and a testing dataset (remaining 10% of the total dataset) to prevent any potential bias in model training. Then the training dataset was used to develop three different models (AdaBoost, XGBoost, and Random Forest) for each emotion (neutral, sad, fear, happy); this process was repeated 30 times with each model to ensure stability and consistency in the results. After training the model, the testing dataset was used to validate the performance of each model. Performance was quantitatively assessed by comparing the predicted labels against the actual labels from the testing set, calculating both the average and the standard deviation. In addition, we conducted sensitivity analyses on the emotion recognition models to investigate which dynamic characteristics are strongly associated with specific emotions. This analysis helped identify key features that significantly influence the models’ ability to accurately classify different emotional states.

We assessed the effectiveness of ensemble learning models for emotion recognition using two performance metrics: accuracy and AUC. Accuracy is defined as $(TP + TN)/(TP + TN + FP + FN)$, where True Positives (TP) represent actual positives correctly predicted as positive, True Negatives (TN) represent actual negatives correctly predicted as negative, False Positives (FP) indicate actual negatives incorrectly predicted as positive, and False Negatives (FN) refer to actual positives incorrectly predicted as negative. The ROC curve is plotted with false positive rate (1-specificity) on the x-axis against the true positive rate (sensitivity) on the y-axis at various threshold settings. Specifically, sensitivity = $TN/(TN + FP)$ and specificity = $TP/(TP + FN)$. AUC represents the area under the ROC curve, providing a single measure of overall model performance across all classification thresholds. It is particularly valuable in the presence of biased datasets, as it evaluates the model's ability to discriminate between classes without being influenced by class imbalance (Nahm, 2022). A higher AUC value indicates better model performance, with 1.0 representing perfect discrimination and 0.5 indicating no discriminative power beyond random chance.

To achieve optimal performance, we applied grid search combined with 10-fold cross-validation to fine-tune the hyperparameter settings for the supervised ensemble learning methods, including Adaboost, Random Forest, and XGBoost. The hyperparameters yielding the highest F1 score (calculated as 2·TP/(2·TP + FP + FN)) on the validation dataset were selected. This comprehensive tuning process involved exhaustively searching through a predefined set of hyperparameters to find the optimal combination, ensuring that each model was finely adjusted to achieve the best possible performance. For Adaboost, we created an ensemble of 500 weak learners without resampling with replacement and used the Breiman method for adjusting weights. For Random Forest, we built 800 trees, each considering 30 features at each split, and used a 0.5 threshold for classification. For XGBoost, we trained 500 deep trees to solve a binary classification problem using logistic regression.

5 Results

We developed a comprehensive methodology consisting of three phases to identify four emotions by analyzing the corresponding complex dynamic characteristics in EEG. In this section, we discussed the performance of the proposed methodology in three perspectives. We initially compared the performance of three ensemble learning models: AdaBoost, Random Forest, and XGBoost. Then, we discussed the performance of each individual emotion identification model under XGBoost. Finally, an overall performance comparison with other models using the same dataset was conducted.

5.1 Model performance of AdaBoost, random forest, and XGBoost

To evaluate which ensemble learning model had the best performance for emotion recognition, accuracy and AUC was calculated for each specific emotion then averaged for each model. Table 2 demonstrates that XGBoost and Random Forest consistently achieved high accuracy and AUC, signifying excellent stability across multiple trials, whereas AdaBoost did not. Since both Random Forest and XGBoost achieved at least 0.75 in both accuracy and AUC, this suggests that dynamic transition properties of brain activity extracted from high-dimensional EEG signals using the MHRA methodology, can effectively recognize emotions. Given that accuracy was our primary performance criterion, XGBoost with an average accuracy of 0.7885 and an AUC of 0.7552 was selected as the best model for emotion recognition.

TABLE 2 Performance of each ensemble model of all four emotions.

Method	Accuracy	AUC
Adaboost	0.7498 (0.0118)	0.5444 (0.0631)
Random Forest	0.7518 (0.0140)	0.7666 (0.0177)
XGBoost	0.7885 (0.0116)	0.7552 (0.0207)

*Mean (Standard Deviation).

5.2 Performance of XGBoost for each emotion

Figure 8 demonstrates the AUC curves for the XGBoost model's performance in recognizing four distinct emotions. The curves reflect the varying levels of the model's discriminatory ability for each emotion. The AUC for 'Sad' shows the highest value at 0.7931, indicating that the model is most effective at distinguishing 'Sad' from non-sad emotional states. 'Neutral' also demonstrates a robust performance with an AUC of 0.7814. However, the AUCs for 'Fear' and 'Happy' are lower, at 0.7165 and 0.7299 respectively, suggesting challenges in the model's ability to consistently differentiate these emotions from others. The lower AUC for 'Fear' indicates a particular difficulty in discrimination, which could be due to the nuanced nature of fear as an emotion. Conversely, despite 'Happy' having the highest accuracy, its AUC indicates less consistency in distinguishing happiness, likely due to overlapping features with other emotions.

In this section, we demonstrated the performance of XGBoost into each emotion model, as shown in Table 3. The results indicate that all the emotion models can achieve at least 0.77 for accuracy and at least 0.71 for the AUC. The model excels in recognizing 'Happy' emotions, achieving the highest accuracy of 0.8127. The accuracies and AUCs for 'Neutral' and 'Sad' are relatively higher and more consistent, suggesting more reliable performance for these emotions. Conversely, the AUCs for 'Fear' and 'Happy' are lower and show greater variability, reflecting differences in the model's ability to consistently distinguish these emotions from others. The small standard deviations associated with these metrics across all emotions underscore the model's stability and reliability in performance across multiple iterations or subsets of the dataset.

5.3 Performance comparison to other methodologies

To evaluate the performance of our methodology relative to other methodologies, Table 4 compares our performance to other methodologies using the same dataset: EmotionMeter, (Zheng et al., 2019b), BiHDM, (Li et al., 2019b), RGNN, (Zhong et al., 2019),

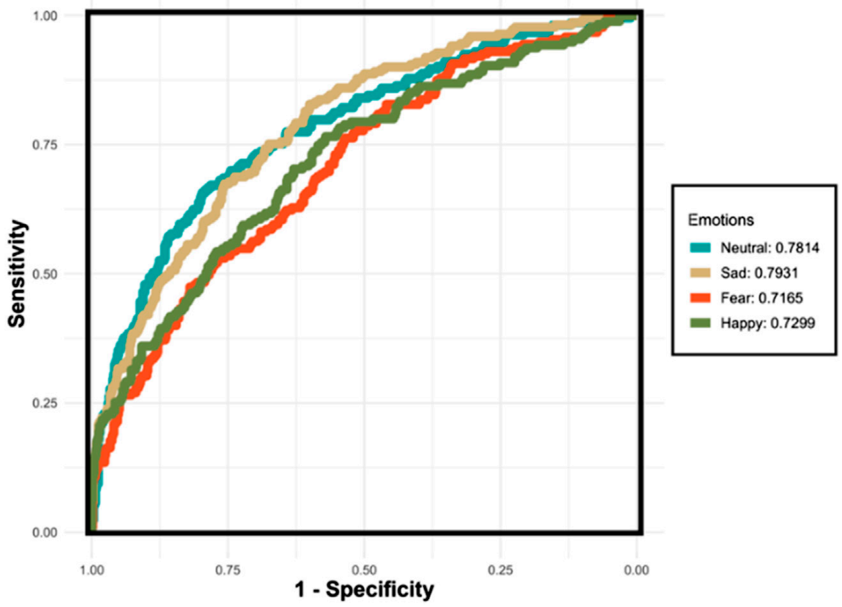


FIGURE 8
The ROC curves for the XGBoost classifier applied to the testing set using the One-vs-All (OvA) strategy for four separate emotions. The emotions "Neutral" and "Sad" exhibit relatively higher AUC values, indicating more reliable performance in distinguishing these emotions. Conversely, "Fear" and "Happy" demonstrate lower AUC values, reflecting the model's reduced consistency in differentiating these emotions from others.

TABLE 3 Performance of XGBoost method for each emotion.

Emotion	Proportion (%)	Accuracy	AUC
Neutral	27.09	0.7790 (0.0009)	0.7814 (0.0164)
Sad	27.27	0.7868 (0.0102)	0.7931 (0.0196)
Fear	24.49	0.7757 (0.0137)	0.7165 (0.0251)
Happy	21.15	0.8127 (0.0130)	0.7299 (0.0215)
Average	**25.00**	**0.7885 (0.0116)**	**0.7552 (0.0207)**

*Mean (Standard Deviation).
The bold values indicate the average performance of the four emotion models.

Fractal-SNN, (Li et al., 2024), Saliency-based CNN, (Delvigne et al., 2022), MetaEmotionNet, (Ning et al., 2024), ST-SCGNN, (Pan et al., 2024), and MISNet (Gong et al., 2024). Our methodology not only outperformed all of these models in overall accuracy (0.7885) but also demonstrated the most stable performance among the repeated experiments, as indicated by the lowest standard deviation (0.0207).

Notably, our methodology provided the most consistent recognition performance across different emotions, with average accuracies ranging from 0.7757 to 0.8127. This consistency highlights the robustness and effectiveness of our approach in capturing the subtle dynamics of brain activity. In contrast, other methods showed varying strengths across specific emotions. For example, EmotionMeter is more effective in identifying 'Happy' and 'Neutral', BiHDM is more accurate in recognizing 'Neutral' and 'Sad', RGNN and MetaEmotionNet are specifically sensitive to 'Sad' and 'Happy', respectively, and MISNet performs better in 'Sad' and 'Happy'. This implies that previous models struggle to grasp the nuanced activities in the brain, likely due to their inability to fully capture the complex characteristics of EEG signals. Collectively, this indicates that complex brain activity can be effectively characterized using dynamic recurrence properties with our novel MHRA methodology.

These results highlight the robustness and effectiveness of our approach in handling the complex, nonlinear, and nonstationary characteristics of EEG signals. Our methodology's ability to maintain high accuracy across all emotions and its stable performance in repeated experiments underscore its reliability and potential for real-world applications. By comparing our findings with the relevant literature, it is evident that MHRA not only advances the state of the art in emotion recognition but also provides a versatile method for analyzing complex brain dynamics. This comprehensive analysis reinforces the value of our contributions to the field and demonstrates the superiority of our approach over existing methods.

In addition to achieving the highest accuracy in emotion recognition, our methodology offers profound insights into the specific dynamic features that drive emotional responses, thereby enhancing our understanding of complex brain activity. We demonstrate that variations in the distribution of MHRA metrics are key indicators for emotion recognition, providing robust evidence of our model's superiority over traditional 'black box' methods. For example, Figure 9 presents a sensitivity analysis of how specific HRQA metrics vary in value across each emotion. Specifically, each panel is one unique HRQA that corresponds to a dynamic property that characterizes a specific transition between different subspaces within the constructed heterogeneous state-space.

TABLE 4 Accuracy of MHRA in emotion recognition (individual and overall) vs. other methods.

Authors	Methodology	Neutral	Sad	Fear	Happy	All (mean/s.d.)
Zheng et al. (2019)	EmotionMeter	0.7800	0.6300	0.6500	0.8000	0.7058/0.1701
Li et al. (2019)	BiHDM	0.7443	0.7273	0.5813	0.6350	0.6903/0.0866
Zhong et al. (2020)	RGNN	0.7516	0.9192	0.7185	0.7435	0.7384/0.0802
Li et al. (2023)	Fractal-SNN	-	-	-	-	0.6833/--------
Delvigne et al. (2023)	Saliency based CNN	-	-	-	-	0.7442/0.0476
Ning et al. (2024)	MetaEmotionNet	0.5393	0.6312	0.5052	0.7415	0.6120/0.0830
Pan et al. (2024)	ST-SCGNN	-	-	-	-	0.7637/0.5777
Gong et al. (2024)	MISNet	0.7071	0.8300	0.6319	0.8169	0.7460/0.0930
Wang et al. (2024)	**MHRA**	**0.7790**	**0.7868**	**0.7757**	**0.8127**	**0.7885/0.0207**

The bold values shows the results of this research.

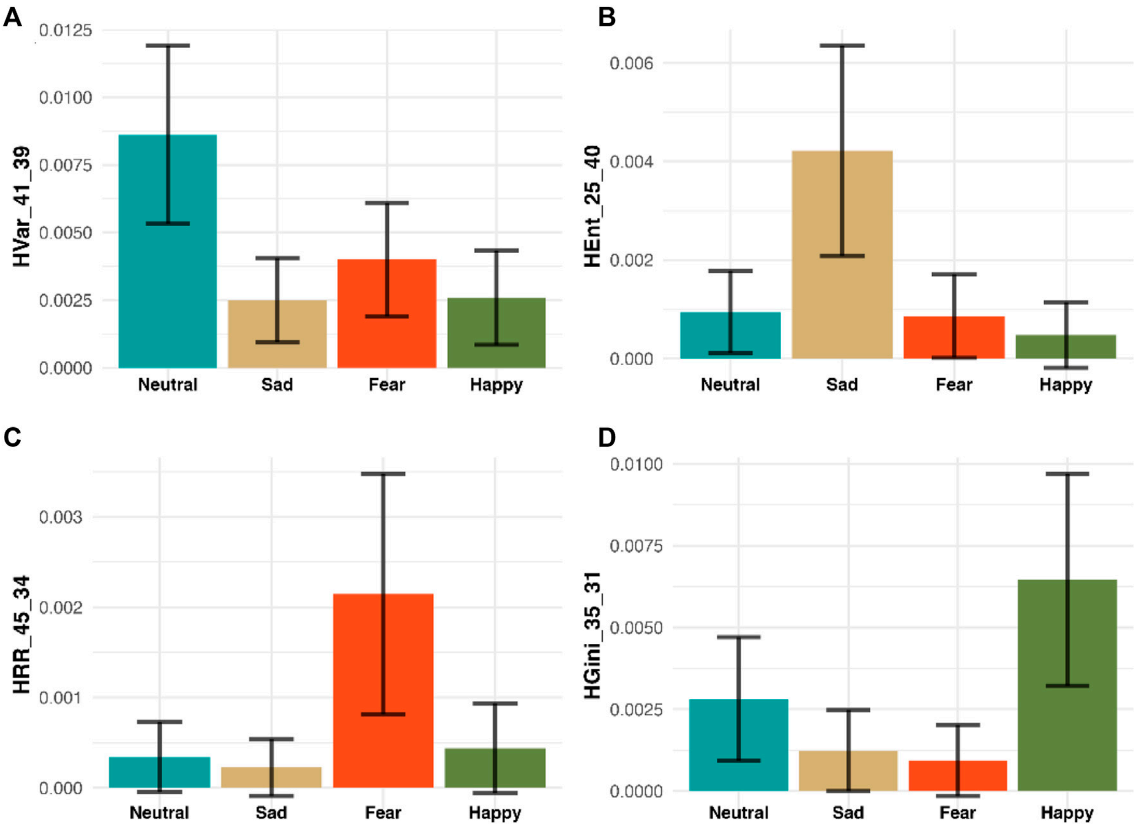


FIGURE 9
Sensitivity Analysis of Four selected HRQA Metrics. The four panels **(A–D)** display four selected HRA metrics for four emotions, respectively. Specifically, Panel A shows HVar_41_39 has the highest value in "Neutral," Panel B demonstrates HEnt_25_40 has the highest value in "Sad," Panel C illustrates HRR_45_34 has the highest value in "Fear," and Panel D presents HGini_35_31 has the highest value in "Happy." Each panel highlights a metric where one emotion scores significantly higher on average than the others, demonstrating the metric's potential to distinctly identify that emotion from the rest.

Panel A displays the HVar in the transition from subspace #21 to subspace #39. Here, the 'Neutral' emotion exhibits the highest average, suggesting significant variability during these transitions. Panel illustrates the HEnt during transitions from subspace #25 to subspace #40, with 'Sad' recording the highest average, indicating pronounced entropy in these transitions. Panel C depicts the HRR between subspaces #45 and #34. Here, 'Fear' stands out with the highest average, reflecting a notable recurrence rate. Finally, Panel D tracks inequality HGini in the transitions from subspace #35 to subspace #31, where 'Happy' demonstrates the highest average, highlighting significant inequality in these transitions. Each bar chart is accompanied by a 95% confidence interval, providing a clear visual representation of how distinct MHRA metrics correlate with each emotional state.

These findings not only confirm the efficacy of our model in identifying and interpreting emotions but also provide a methodology for investigating the subtle spatiotemporal dynamics underlying brain activity related to various emotions. By analyzing these HRQA metrics, we may infer the neural mechanisms involved in emotion recognition. For instance, the high value of entropy (HEnt), referring to a high level of uncertainty, in 'Sad' could signify chaotic neural activity patterns associated with emotional distress or cognitive load. The high value of recurrence rate (HRR), referring to a high tendency to revisit similar patterns, in 'Fear' suggests a specific pattern of repetitive neural activations, possibly related to the brain's heightened state of alertness and threat detection.

By correlating these dynamic features with known neural processes, our approach offers deeper insights into how different emotional states manifest in the brain's activity. This enhanced understanding can contribute to developing more effective interventions and therapeutic strategies for emotional and mental health disorders. Thus, our methodology not only advances the field of emotion recognition but also provides a valuable tool for exploring the neural underpinnings of emotions.

6 Discussion

Understanding how emotions are processed and represented in the brain enhances our basic scientific knowledge of neurological functions. By studying EEG patterns associated with different emotions, researchers can uncover the underlying neural mechanisms that govern emotional responses and how these might differ among individuals or across different contexts. However, the complex, nonlinear, and nonstationary characteristics of EEG signals pose significant challenges for many traditional methods in this field. Numerous studies on EEG-based emotion recognition rely on deep learning techniques, as these state-of-the-art neural network-based methods are adept at detecting subtle patterns within complex EEG signals (Jafari et al., 2023). Nonetheless, the lack of transparency in deep learning algorithms represents a substantial barrier, as physicians tend to be cautious by nature, and patients are hesitant to entrust their health to a 'black box' algorithm. In this study, we introduced a three-phase methodology, including manifold embedding, MHRA, and supervised ensemble learning, designed to address these concerns by characterizing the dynamic features of brain activity for emotion recognition while also preserving a degree of explainability.

We employed the proposed MHRA methodology to the SJTU-SEED IV database, in Phase 1, we utilized UMAP for data embedding to address the challenge of high dimensional data. The 62-lead EEG signals were transformed into four-dimensional embedded signals that retain dynamic spatiotemporal characteristics but significantly reduced computational demands to a manageable level for further analyses. In Phase 2, the embedded EEG data underwent our novel MHRA to capture the recurrence dynamics of brain activity at high resolution. This approach not only provides a more nuanced understanding of the complex nonlinear and nonstationary EEG patterns, but also extracts robust dynamic features for emotion recognition. Importantly, our generalized HRQA metrics systematically quantify recurrences across different transition levels, offering a scalable framework for analyzing dynamic EEG properties. Finally, in Phase 3 we employed advanced ensemble learning methods and demonstrated their effectiveness in classifying emotions using LASSO selected HRQA metrics. The superior performance of our models, especially XGBoost, suggests that dynamic transition characteristics are powerful predictors for emotion recognition. Our models achieved accuracy and AUC values of 0.7885 and 0.7552, respectively, both outperforming previous studies using the same dataset. Additionally, our sensitivity analysis identified specific HRQA metrics strongly associated with each emotion, providing valuable insights into the neural dynamics underlying emotional processing that cannot be obtained using "black box" algorithms alone.

The major contribution of this research is the development of MHRA, a novel technique leveraging the recurrence theorem to characterize dynamic brain activity across multiple granularities. Unlike traditional methods, MHRA captures the complex, nonlinear, and nonstationary properties of EEG signals, providing a detailed framework for analyzing intricate brain activity patterns. By utilizing HRQA metrics, MHRA offers an interpretable analysis of EEG data, aiding researchers in understanding the neural mechanisms of emotions. This transparency is crucial for building trust and facilitating the adoption of our methodology in clinical and research settings. The insights from our MHRA approach have significant implications for advancing studies in cognitive neuroscience, affective computing, neurofeedback therapy, human-computer interaction, and educational neuroscience. Traditional approaches often struggle with the nonlinear and nonstationary nature of EEG signals, while deep learning models lack explainability. Our methodology overcomes these challenges, offering both high performance and interpretability, thus advancing the field of emotion recognition and providing an effective solution for analyzing complex brain dynamics. Our methodology offers several key advantages. First, it effectively addresses the limitations of traditional linear methods by analyzing complex nonlinear nonstationary EEG signals. Second, MHRA offers interpretability by using HRQA metrics to explain features of complex systems. This transparency is crucial for building trust and facilitating adoption in clinical settings. Third, the tree-based ensemble learning methods not only achieve high accuracy to recognize emotions but also exhibit robustness in capturing nonlinear relationships of dynamic properties.

Despite these strengths, our study has some limitations that will be explored in future research. The SJTU-SEED IV database, while comprehensive, does not fully capture the diversity or unique emotional experiences across different populations. Investigating the generalizability of our methodology to other EEG datasets and real-world scenarios is an important next step. Additionally, integrating our approach with other modalities, such as facial expressions or other physiological signals such as eye movements, could further enhance the accuracy and robustness of emotion recognition. Furthermore, our research can facilitate a deeper understanding and characterization of brain activities, with potential applications in pediatric sleep studies, the development of objective metrics for PTSD, and non-invasive early detection of neurodegenerative diseases. Future research could benefit from incorporating more advanced techniques to retain the spatiotemporal characteristics of 62-lead EEG signals, such as integrating attention mechanisms with MHRA to provide more effective characterization of neural dynamics. By pursuing these directions, we aim to refine the existing methodology and broaden its applicability, thus advancing the field of emotion recognition and its practical applications in neuroscience and healthcare.

In conclusion, this study presents a novel three-phase methodology that includes manifold embedding, MHRA, and ensemble learning for EEG-based emotion recognition. Our approach not only achieves high performance but also offers interpretable insights into the dynamic properties underlying four emotions. This methodology has significant impact on the field to advance our ability to analyze nonlinear nonstationary, dynamic data of complex systems with potential applications in healthcare, human-computer interaction, and beyond.

Data availability statement

The data analyzed in this study was obtained from the Shanghai Jiao Tong University (SJTU) Emotion EEG Dataset for Four Emotions (SEED-IV), which is a specific subset of the broader SJTU Emotion EEG Dataset (available at https://bcmi.sjtu.edu.cn/~seed/). The following licenses/restrictions apply: the dataset is restricted to academic research use only and cannot be used for any commercial purposes. Distribution of the dataset or portions thereof is prohibited, except for small portions used to clarify academic publications or presentations. Access to the dataset is granted only after filling out, signing, and uploading the license agreement to the SEED website, followed by a review of the application. No warranty is provided with the dataset, and users must cite the relevant publications when using the dataset in their research. Requests to access these datasets should be directed to SEED Dataset (https://bcmi.sjtu.edu.cn/home/seed/index.html).

Ethics statement

Ethical approval was not required for the studies involving humans because this research uses data from Shanghai Jiao Tong University. Data is publicly available. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

YW: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. C-BC: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing. TI: Conceptualization, Investigation, Visualization, Writing–review and editing. IT: Conceptualization, Investigation, Visualization, Writing–review and editing. VS: Conceptualization, Investigation, Visualization, Writing–review and editing. PZ: Conceptualization, Visualization, Writing–review and editing. DL: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. We acknowledge funding by University of Miami Start-up funding PG013380 (to C-BC), the National Institutes of Health (NIH)R33HL162839, R33HL151253, R01HL151254 (to IT); (NIH)R01HL160619 (to VS); (NIH)R01HL140580, (NIH)P01AG011412, (NIH) R01HD097786, and (NIH)R01AG068579 (to PZ); (NIH) 1R01HL142981-01A1 and VA BLRD 1I01BX004872-01 (to DL).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akhand, M. A. H., Maria, M. A., Kamal, M. A. S., and Murase, K. (2023). Improved EEG-based emotion recognition through information enhancement in connectivity feature map. *Sci. Rep.* 13 (13), 13804–13817. doi:10.1038/s41598-023-40786-2

Amiri, A., Samet, H., and Ghanbari, T. (2022). Recurrence plots based method for detecting series arc faults in photovoltaic systems. *IEEE Trans. Industrial Electron.* 69, 6308–6315. doi:10.1109/tie.2021.3095819

Asghar, Q., Jalil, A., and Zaman, M. (2020). Self-organization analysed in architecture using Voronoi tessellation and particle systems. *Tech. J.* 25, 1–10.

Avdan, G., Chen, C., and Onal, S. (2024). An alternative EMG normalization method: heterogeneous recurrence quantification analysis of isometric maximum voluntary contraction movements. *Biomed. Signal Process Control* 93, 106219. doi:10.1016/j.bspc.2024.106219

Avdan, G., Chen, C. B., and Onal, S. (2023). "Investigation of an alternative EMG normalization technique: recurrence quantification analysis of maximum voluntary contractions," in *IISE annual conference and expo 2023* (IISE). doi:10.21872/2023IISE_1909

Bazgir, O., Mohammadi, Z., and Habibi, S. A. H. (2018). "Emotion recognition with machine learning using EEG signals," in *2018 25th Iranian conference on biomedical engineering and 2018 3rd international Iranian conference on biomedical engineering* (ICBME), 1–5. IEEE.

Bouabdelli, S., Meddi, M., Zeroual, A., and Alkama, R. (2020). Hydrological drought risk recurrence under climate change in the karst area of Northwestern Algeria. *J. Water Clim. Change* 11, 164–188. doi:10.2166/wcc.2020.207

Chai, X., Wang, Q., Zhao, Y.-P., Liu, X., Liu, D., and Bai, O. (2018). Multi-subject subspace alignment for non-stationary EEG-based emotion recognition. *Technol. Health Care* 26, 327–335. doi:10.3233/thc-174739

Chang, H., Zong, Y., Zheng, W., Tang, C., Zhu, J., and Li, X. (2022). Depression assessment method: an EEG emotion recognition framework based on spatiotemporal neural network. *Front. Psychiatry* 12, 837149. doi:10.3389/fpsyt.2021.837149

Chen, C., Ji, Z., Sun, Y., Bezerianos, A., Thakor, N., and Wang, H. (2023b). self-attentive channelchannel-connectivity capsule network for EEG-based driving fatigue detection. *IEEE Trans. Neural Syst. Rehabilitation Eng.* 31, 3152–3162. doi:10.1109/TNSRE.2023.3299156

Chen, C., Li, Z., Wan, F., Xu, L., Bezerianos, A., and Wang, H. (2022a). Fusing frequency-domain features and brain connectivity features for cross-subject emotion recognition. *IEEE Trans. Instrum. Meas.* 71, 1–15. doi:10.1109/TIM.2022.3168927

Chen, C., Vong, C. M., Wang, S., Wang, H., and Pang, M. (2022b). Easy Domain Adaptation for cross-subject multi-view emotion recognition. *Knowl. Based Syst.* 239, 107982. doi:10.1016/j.knosys.2021.107982

Chen, C.-B. (2019). *Recurrence analysis of high-dimensional complex systems with applications in healthcare and manufacturing*.

Chen, C. B., Wang, Y., Fu, X., and Yang, H. (2023c). Recurrence network analysis of histopathological images for the detection of invasive ductal carcinoma in breast cancer. *IEEE/ACM Trans. Comput. Biol. Bioinform* 20, 3234–3244. doi:10.1109/TCBB.2023.3282798

Chen, C. B., Yang, H., and Kumara, S. (2018). Recurrence network modeling and analysis of spatial data. *Chaos* 28, 085714. doi:10.1063/1.5024917

Chen, C.-B., Yang, H., and Kumara, S. (2019d). A novel pattern-frequency tree for multisensor signal fusion and transition analysis of nonlinear dynamics. *IEEE Sens. Lett.* 3, 1–4. doi:10.1109/lsens.2018.2884241

Chen, C.-B., Yang, H., and Kumara, S. (2017). "A novel pattern-frequency tree approach for transition analysis and anomaly detection in nonlinear and nonstationary systems," in *IIE annual conference. Proceedings*, 1264–1269.

Chen, D., Huang, H., Bao, X., Pan, J., and Li, Y. (2023a). An EEG-based attention recognition method: fusion of time domain, frequency domain, and non-linear dynamics features. *Front. Neurosci.* 17, 1194554. doi:10.3389/fnins.2023.1194554

Chen, D. W., Miao, R., Yang, W. Q., Liang, Y., Chen, H. H., Huang, L., et al. (2019b). A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition. *Sensors* 19, 1631. doi:10.3390/s19071631

Chen, J., zhang, peize, Mao, Z., Huang, Y., Jiang, D., and Zhang, Y. N. (2019a). Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks. *Ieee Access* 7, 44317–44328. doi:10.1109/access.2019.2908285

Chen, R., Imani, F., and Yang, H. (2020). Heterogeneous recurrence analysis of disease-altered spatiotemporal patterns in multi-channel cardiac signals. *IEEE J. Biomed. Health Inf.* 24, 1619–1631. doi:10.1109/JBHI.2019.2952285

Chen, R., Rao, P., Lu, Y., Reutzel, E., and Yang, H. (2019c). Recurrence network analysis of design-quality interactions in additive manufacturing. *Sci. Total Environ.*, 135907. doi:10.1016/j.addma.2021.101861

Chen, Y., and Yang, H. (2015). "Heterogeneous recurrence T-squared charts for monitoring and control of nonlinear dynamic processes," in *2015 IEEE international conference on automation science and engineering (CASE)*, 1066–1071.

Chen, Y., and Yang, H. (2016). Heterogeneous recurrence representation and quantification of dynamic transitions in continuous nonlinear processes. *Eur. Phys. J. B* 89, 155. doi:10.1140/epjb/e2016-60850-y

Cheng, C., Kan, C., and Yang, H. (2016). Heterogeneous recurrence analysis of heartbeat dynamics for the identification of sleep apnea events. *Comput. Biol. Med.* 75, 10–18. doi:10.1016/j.compbiomed.2016.05.006

Dan, Y., Tao, J., Fu, J., and Zhou, D. (2021). Possibilistic clustering-promoting semi-supervised learning for EEG-based emotion recognition. *Front. Neurosci.* 15, 690044. doi:10.3389/fnins.2021.690044

Delvigne, V., Facchini, A., Wannous, H., Dutoit, T., Ris, L., and Vandeborre, J.-P. (2022). *A saliency based feature fusion model for EEG emotion estimation*, 3170–3174.

Donner, R. V., Small, M., Donges, J. F., Marwan, N., Zou, Y., Xiang, R., et al. (2011). Recurrence-based time series analysis by means of complex network methods. *Int. J. Bifurcation Chaos* 21, 1019–1046. doi:10.1142/s0218127411029021

Donner, R. V., Zou, Y., Donges, J. F., Marwan, N., and Kurths, J. (2010). Recurrence networks-a novel paradigm for nonlinear time series analysis. *New J. Phys.* 12, 033025. doi:10.1088/1367-2630/12/3/033025

Eckmann, J.-P., Oliffson Kamphorst, S., and Ruelle, D. (1987). Recurrence plots of dynamical systems. *Europhys. Lett. (EPL)* 4, 973–977. doi:10.1209/0295-5075/4/9/004

Elgamal, T., and Hefeeda, M. (2015). *Analysis of PCA algorithms in distributed environments.* Available at: https://arxiv.org/abs/1503.05214v2 (Accessed April 23, 2024).

Eroglu, D., Marwan, N., Prasad, S., and Kurths, J. (2014). Finding recurrence networks' threshold adaptively for a specific time series. *Nonlinear Process Geophys* 21, 1085–1092. doi:10.5194/npg-21-1085-2014

Gao, Z., Cui, X., Wan, W., and Gu, Z. (2019). Recognition of emotional states using multiscale information analysis of high frequency EEG oscillations. *Entropy* 21, 609. doi:10.3390/E21060609

Gong, M., Zhong, W., Ye, L., and Zhang, Q. (2024). MISNet: multi-source information-shared EEG emotion recognition network with two-stream structure. *Front. Neurosci.* 18, 1293962. doi:10.3389/fnins.2024.1293962

Hatami, N., Gavet, Y., and Debayle, J. (2019). Bag of recurrence patterns representation for time-series classification. *Pattern Analysis Appl.* 22, 877–887. doi:10.1007/s10044-018-0703-6

Haynes, J. D., and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534. doi:10.1038/nrn1931

Holland, J. H. (1992). Genetic algorithms. *Sci. Am.* 267, 66–72. doi:10.1038/scientificamerican0792-66

Houssein, E. H., Hammad, A., and Ali, A. A. (2022). Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. *Neural Comput. Appl.* 34 (34), 12527–12557. doi:10.1007/s00521-022-07292-4

Hu, L., and Li, L. (2022). Using tree-based machine learning for health studies: literature review and case series. *Int. J. Environ. Res. Public Health* 19, 16080. doi:10.3390/ijerph192316080

Hunt, F. V., Beranek, L. L., and Maa, D. Y. (1939). Analysis of sound decay in rectangular rooms. *J. Acoust. Soc. Am.* 11, 80–94. doi:10.1121/1.1916010

Jafari, M., Shoeibi, A., Khodatars, M., Bagherzadeh, S., Shalbaf, A., García, D. L., et al. (2023). Emotion recognition in EEG signals using deep learning methods: a review. *Comput. Biol. Med.* 165, 107450. doi:10.1016/j.compbiomed.2023.107450

Jellinger, K. A. (2003). Functional magnetic resonance imaging: an introduction to methods. *Eur. J. Neurol.* 10, 751–752. doi:10.1046/j.1468-1331.2003.00657.x

Kan, C., Cheng, C., and Yang, H. (2016). Heterogeneous recurrence monitoring of dynamic transients in ultraprecision machining processes. *J. Manuf. Syst.* 41, 178–187. doi:10.1016/j.jmsy.2016.08.007

Khoo, M. C. K., Webber, C. L., and Zbilut, J. P. (1996). Assessing deterministic structures in physiological systems using recurrence plot strategies. *Bioeng. approaches Pulm. physiology Med.*, 137–148. doi:10.1007/978-0-585-34964-0_8

Li, C., Chen, B., Zhao, Z., Cummins, N., and Schuller, B. W. (2021b). *Hierarchical attention-based temporal convolutional networks for eeg-based emotion recognition*. doi:10.1109/icassp39728.2021.9413635

Li, J., Li, S., Pan, J., and Wang, F. (2021a). Cross-subject EEG emotion recognition with self-organized graph neural network. *Front. Neurosci.* 15, 611653. doi:10.3389/fnins.2021.611653

Li, W., Fang, C., Zhu, Z., Chen, C., and Song, A. (2024). Fractal spiking neural network scheme for EEG-based emotion recognition. *IEEE J. Transl. Eng. Health Med.* 12, 106–118. doi:10.1109/JTEHM.2023.3320132

Li, Y., Zheng, W., Cui, Z., Zong, Y., and Ge, S. (2019a). EEG emotion recognition based on graph regularized sparse linear regression. *Neural Process Lett.* 49, 555–571. doi:10.1007/s11063-018-9829-1

Li, Y., Zheng, W., Wang, L., Zong, Y., Qi, L., Cui, Z., et al. (2019b). *A novel Bi-hemispheric discrepancy model for EEG emotion recognition*. Available at: http://arxiv.org/abs/1906.01704.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., and Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behav. Brain Sci.* 35, 121–143. doi:10.1017/s0140525x11000446

Liu, H., Zhang, Y., Li, Y., and Kong, X. (2021). Review on emotion recognition based on electroencephalography. *Front. Comput. Neurosci.* 15, 758212–758215. doi:10.3389/fncom.2021.758212

Liu, X., Li, T., Tang, C., Xu, T., Chen, P., Bezerianos, A., et al. (2019). Emotion recognition and dynamic functional connectivity analysis based on EEG. *IEEE Access* 7, 143293–143302. doi:10.1109/access.2019.2945059

Liu, Y., Sourina, O., and Nguyen, M. K. (2010). "Real-time EEG-based human emotion recognition and visualization," in *Proceedings - 2010 international conference on cyberworlds, CW 2010*, 262–269.

Lucarini, V., Faranda, D., de Freitas, J. M. M., Holland, M., Kuna, T., Nicol, M., et al. (2016). *Extremes and recurrence in dynamical systems*. John Wiley \and Sons.

Marwan, N. (2008). A historical review of recurrence plots. *Eur. Phys. J. Special Top.* 164, 3–12. doi:10.1140/epjst/e2008-00829-1

Marwan, N., Carmen, R. M., Thiel, M., and Kurths, J. (2007a). Recurrence plots for the analysis of complex systems. *Phys. Rep.* 438, 237–329. doi:10.1016/j.physrep.2006.11.001

Marwan, N., Kurths, J., and Saparin, P. (2007b). Generalised recurrence plot analysis for spatial data. *Phys. Lett. Sect. A General, Atomic Solid State Phys.* 360, 545–551. doi:10.1016/j.physleta.2006.08.058

Mcinnes, L., Healy, J., and Melville, J. (2020). *UMAP: uniform manifold approximation and projection for dimension reduction*.

Meilă, M., and Zhang, H. (2024). Manifold learning: what, how, and why. *Annu. Rev. Stat. Appl.* 11, 393–417. doi:10.1146/annurev-statistics-040522-115238

Mosavi, A., Faghan, Y., Ghamisi, P., Duan, P., Ardabili, S. F., Salwana, E., et al. (2020). Comprehensive review of deep reinforcement learning methods and applications in economics. *Mathematics* 8, 1640. doi:10.3390/MATH8101640

Murugappan, M., and Murugappan, S. (2013) "Human emotion recognition through short time Electroencephalogram (EEG) signals using Fast Fourier Transform (FFT)," in *Proceedings - 2013 IEEE 9th international colloquium on signal processing and its applications*. IEEE, 289–294. doi:10.1109/CSPA.2013.6530058

Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean J. Anesthesiol.* 75, 25–36. doi:10.4097/kja.21209

Ning, X., Wang, J., Lin, Y., Cai, X., Chen, H., Gou, H., et al. (2024). MetaEmotionNet: spatial–spectral–temporal-based attention 3-D dense network with meta-learning for EEG emotion recognition. *IEEE Trans. Instrum. Meas.* 73, 1–13. doi:10.1109/tim.2023.3338676

Pan, J., Liang, R., He, Z., Li, J., Liang, Y., Zhou, X., et al. (2024). ST-SCGNN: a spatio-temporal self-constructing graph neural network for cross-subject EEG-based emotion recognition and consciousness detection. *IEEE J. Biomed. Health Inf.* 28, 777–788. doi:10.1109/JBHI.2023.3335854

Pang, M., Wang, H., Huang, J., Vong, C. M., Zeng, Z., and Chen, C. (2024). Multi-scale masked autoencoders for cross-session emotion recognition. *IEEE Trans. Neural Syst. Rehabilitation Eng.* 32, 1637–1646. doi:10.1109/TNSRE.2024.3389037

Peng, B., and Chen, C.-B. (2023). "Multiscale dynamic transition analysis of solar radiation prediction," in *IISE annual conference and expo 2023*. doi:10.21872/2023IISE_1787

Peng, Y., Liu, H., Li, J., Huang, J., Lu, B.-L., and Kong, W. (2023). Cross-session emotion recognition by joint label-common and label-specific EEG features exploration. *Ieee Trans. Neural Syst. Rehabilitation Eng.* 31, 759–768. doi:10.1109/tnsre.2022.3233109

Pouyet, E., Rohani, N., Katsaggelos, A. K., Cossairt, O., and Walton, M. (2018). Innovative data reduction and visualization strategy for hyperspectral imaging datasets using t-SNE approach. *Pure Appl. Chem.* 90, 493–506. doi:10.1515/pac-2017-0907

Roth, V. (2004). The generalized LASSO. *IEEE Trans. Neural Netw.* 15, 16–28. doi:10.1109/TNN.2003.809398

Roweis, S. T., and Saul, L. K. (1979). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326. doi:10.1126/science.290.5500.2323

Shu, Z. R., Chan, P. W., Li, Q. S., He, Y. C., and Yan, B. W. (2021). Investigation of chaotic features of surface wind speeds using recurrence analysis. *J. Wind Eng. Industrial Aerodynamics* 210, 104550. doi:10.1016/j.jweia.2021.104550

Si, X., Huang, D., Sun, Y., Huang, S. W., He, H., and Ming, D. (2023). Transformer-based ensemble deep learning model for EEG-based emotion recognition. *Brain Sci. Adv.* 9, 210–223. doi:10.26599/bsa.2023.9050016

Thornton, M. A., and Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proc. Natl. Acad. Sci. U. S. A.* 114, 5982–5987. doi:10.1073/pnas.1616056114

Tian, Z., Huang, D., Zhou, S., Zhao, Z.-D., and Jiang, D. (2021). Personality first in emotion: a deep neural network based on electroencephalogram channel attention for cross-subject emotion recognition. *R. Soc. Open Sci.* 8, 201976. doi:10.1098/rsos.201976

Tong, F., and Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annu. Rev. Psychol.* 63, 483–509. doi:10.1146/annurev-psych-120710-100412

Turchetti, C., and Falaschetti, L. (2019). A manifold learning approach to dimensionality reduction for modeling data. *Inf. Sci. (N Y)* 491, 16–29. doi:10.1016/j.ins.2019.04.005

Van Booven, D. J., Chen, C., Malpani, S., Mirzabeigi, Y., Mohammadi, M., Wang, Y., et al. (2024b). *Synthetic genitourinary image synthesis via generative adversarial networks: enhancing AI diagnostic precision*. doi:10.3390/jpm14070703

Van Booven, D. J., Chen, C.-B., Kryvenko, O., Punnen, S., Sandoval, V., Malpani, S., et al. (2024a). Synthetic histology images for training ai models: a novel approach to improve prostate cancer diagnosis. *bioRxiv*, 2001–2024. doi:10.1101/2024.01.25.577225

Wang, F., Wu, S., Zhang, W., Xu, Z., Zhang, Y., Wu, C., et al. (2020). Emotion recognition with convolutional neural network and EEG-based EFDMs. *Neuropsychologia* 146, 107506. doi:10.1016/j.neuropsychologia.2020.107506

Wang, H., Xu, L., Bezerianos, A., Chen, C., and Zhang, Z. (2021). Linking attention-based multiscale CNN with dynamical GCN for driving fatigue detection. *IEEE Trans. Instrum. Meas.* 70, 1–11. doi:10.1109/TIM.2020.3047502

Wang, Y., and Chen, C.-B. (2022). "Recurrence quantification analysis for spatial data," in *IIE annual conference. Proceedings*, 1–6.

Wang, Z., Chen, C., Li, J., Wan, F., Sun, Y., and Wang, H. (2023). ST-CapsNet: linking spatial and temporal attention with capsule network for P300 detection improvement. *IEEE Trans. Neural Syst. Rehabilitation Eng.* 31, 991–1000. doi:10.1109/tnsre.2023.3237319

Webber, C. L., and Marwan, N. (2015). Recurrence quantification analysis. Theory and best practices, 426.

Webber, Jr C. L., and Zbilut, J. P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. *Tutorials Contemp. nonlinear methods Behav. Sci.* 94, 26–94.

Wolpaw, J. R., and Birbaumer, N. (2006). *Brain–computer interfaces for communication and control*. doi:10.1017/cbo9780511545061.036

Xu, T., Huang, J., Pei, Z., Chen, J., Li, J., Bezerianos, A., et al. (2023b). The effect of multiple factors on working memory capacities: aging, task difficulty, and training. *IEEE Trans. Biomed. Eng.* 70, 1967–1978. doi:10.1109/TBME.2022.3232849

Xu, T., Wang, H., Lu, G., Wan, F., Deng, M., Qi, P., et al. (2023a). E-key: an EEG-based biometric authentication and driving fatigue detection system. *IEEE Trans. Affect Comput.* 14, 864–877. doi:10.1109/taffc.2021.3133443

Yang, H., Chen, C. B., and Kumara, S. (2020). Heterogeneous recurrence analysis of spatial data. *Chaos* 30, 013119. doi:10.1063/1.5129959

Yang, H., and Chen, Y. (2014). Heterogeneous recurrence monitoring and control of nonlinear stochastic processes. *Chaos* 24, 013138. doi:10.1063/1.4869306

Yang, L., Zheng, W., Wang, L., Zong, Y., and Cui, Z. (2022). From regional to global brain: a novel hierarchical spatial-temporal neural network model for EEG emotion recognition. *IEEE Trans. Affect Comput.* 13, 568–578. doi:10.1109/taffc.2019.2922912

Yuvaraj, R. ;, Thagavel, P., Thomas, J., Fogarty, J., Ali, F., Guo, Y., et al. (2023). Comprehensive analysis of feature extraction methods for emotion recognition from multichannel EEG recordings. *Sensors* 23 (23), 915. doi:10.3390/s23020915

Zhang, B., Shang, P., Mao, X., and Liu, J. (2023). Dispersion heterogeneous recurrence analysis and its use on fault detection. *Commun. Nonlinear Sci. Numer. Simul.* 117, 106902. doi:10.1016/j.cnsns.2022.106902

Zheng, W. L., Liu, W., Lu, Y., Lu, B. L., and Cichocki, A. (2019b). EmotionMeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.* 49, 1110–1122. doi:10.1109/TCYB.2018.2797176

Zheng, W. L., Zhu, J. Y., and Lu, B. L. (2019a). Identifying stable patterns over time for emotion recognition from eeg. *IEEE Trans. Affect Comput.* 10, 417–429. doi:10.1109/taffc.2017.2712143

Zhong, P., Wang, D., and Miao, C. (2019). *EEG-based emotion recognition using regularized graph neural networks*. Available at: http://arxiv.org/abs/1907.07835.

TYPE Original Research
PUBLISHED 12 August 2024
DOI 10.3389/fphys.2024.1424931

Check for updates

OPEN ACCESS

EDITED BY
Bing Yao,
The University of Tennessee, Knoxville, United States

REVIEWED BY
Juan Guillermo Diaz Ochoa,
PerMediQ GmbH, Germany
Steve McKeever,
Uppsala University, Sweden

*CORRESPONDENCE
Xiang Zhong,
✉ xiang.zhong@ise.ufl.edu

RECEIVED 28 April 2024
ACCEPTED 19 July 2024
PUBLISHED 12 August 2024

CITATION
Trevena W, Zhong X, Lal A, Rovati L, Cubro E, Dong Y, Schulte P and Gajic O (2024) Model-driven engineering for digital twins: a graph model-based patient simulation application.
Front. Physiol. 15:1424931.
doi: 10.3389/fphys.2024.1424931

Model-driven engineering for digital twins: a graph model-based patient simulation application

William Trevena[1], Xiang Zhong[1]*, Amos Lal[2], Lucrezia Rovati[2], Edin Cubro[2], Yue Dong[2], Phillip Schulte[2] and Ognjen Gajic[2]

[1]Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, United States, [2]Mayo Clinic, Rochester, MN, United States

Introduction: Digital twins of patients are virtual models that can create a digital patient replica to test clinical interventions *in silico* without exposing real patients to risk. With the increasing availability of electronic health records and sensor-derived patient data, digital twins offer significant potential for applications in the healthcare sector.

Methods: This article presents a scalable full-stack architecture for a patient simulation application driven by graph-based models. This patient simulation application enables medical practitioners and trainees to simulate the trajectory of critically ill patients with sepsis. Directed acyclic graphs are utilized to model the complex underlying causal pathways that focus on the physiological interactions and medication effects relevant to the first 6 h of critical illness. To realize the sepsis patient simulation at scale, we propose an application architecture with three core components, a cross-platform frontend application that clinicians and trainees use to run the simulation, a simulation engine hosted in the cloud on a serverless function that performs all of the computations, and a graph database that hosts the graph model utilized by the simulation engine to determine the progression of each simulation.

Results: A short case study is presented to demonstrate the viability of the proposed simulation architecture.

Discussion: The proposed patient simulation application could help train future generations of healthcare professionals and could be used to facilitate clinicians' bedside decision-making.

KEYWORDS

digital twin, virtual patient simulation, graph model, full-stack application architecture, critical care

1 Introduction

Digital twins are virtual representations of systems that interact with the physical system bi-directionally (Lal et al., 2020a). With the increasing availability of electronic health records and sensor-derived patient data, digital twins hold significant potential in the healthcare sector. In particular, digital twin technology enables the creation of computerized replicas of patients, allowing simulation of diverse clinical scenarios and testing of interventions *in silico* without subjecting real patients to avoidable risk.

A virtual patient is a digital model able to be identified from relevant bedside data and provides prediction in response to modeled inputs. Previous works have demonstrated that virtual patient simulations can be successfully utilized to train medical professionals across an array of specialties (Kononowicz et al., 2019; Lee et al., 2020; Lee and Lee, 2021; Wu et al., 2022). However, many of the previously introduced virtual patient simulation models progress only along a limited number of hand-crafted or predetermined pathways, such as looped, serious branch games, and linear text-based scenarios (Berger et al., 2018). Other examples include virtual patient simulations that progress along decision trees (Hwang et al., 2022), and another recent work (Goldsworthy et al., 2022) utilized a commercial virtual patient simulation application, First2Act, which supports only seven simulation scenarios. Although such simulation architectures have been effectively utilized to train medical professionals, they are hard to scale as each new scenario must be crafted by hand.

Recently, computational simulation models have been proposed, which seek to dynamically model the evolution of organ systems within the human body. One such simulation focused specifically on modeling how the cardiovascular system evolves based on a set of time-varying, simultaneous differential equations (Burkhoff and Dickstein, 2024). Another example is glycemic control, and there have been multiple metabolic system models based on decades of research (Chu et al., 2023). Glycemic control protocols have been optimized using these models. In addition, virtual patient models to predict lung mechanics evolution with changing ventilator settings (mechanic ventilator models) are critical to effectively managing acute respiratory symptoms for critically ill patients, but the scope of the models is very limited (Zhou et al., 2021). These models focus primarily on the one organ system and are developed based on medical, physiological, or biological knowledge, i.e., physics-based models.

In summary, digital twin applications on virtual patient modeling have gained success in modeling individual organs for drug discovery and precision medicine (Venkatesh et al., 2022; Moingeon et al., 2023), but these models rely on the full characterization of the biological and physiological functions at the cell level or the organ level. From bench to bedside, it is important to understand how the organ systems interact and orchestrate the patient's health. For critically ill patients, the capability of modeling and predicting patient trajectories under different treatment regimens would greatly support clinical decision-making, improving patient safety and health outcomes. However, our current knowledge about the human body does not allow us to accurately depict all organ system functions using physical or mechanical models (Rovati et al., 2024). There have been emerging efforts to develop patient or human digital twins based on predictive modeling using AI and machine learning (Vallée, 2023; Katsoulakis et al., 2024; Laubenbacher et al., 2024). Despite having superior predictive capacity, the interpretability of these models is typically limited. Meanwhile, graphical models of the biomarkers of each major organ system would allow us to encode essential interactions among these biomarkers and allow for good interpretability for educational purposes and practical clinical bedside use.

Alternatively, our preliminary work (Trevena et al., 2022) proposes a virtual patient simulation architecture driven by graph-based models and focuses on patient-level simulation, i.e., modeling of the evolution of the virtual patient, determined by directed acyclic graphs (DAGs) depicting the complex pathophysiological interactions that occur within the human body. This graph-based modeling provides a more accurate and transparent presentation of complex relationships between multiple variables in a complex adaptive system where the data is often characterized by intricate interdependence and association. The improved transparency and interoperability in return ensures that the underlying expert rules building upon which the DAGs are crafted can be validated using patient data. It also allows for better visualization of variable relationships and the reasoning behind the model's decision output. The modular and flexible nature of the graph-based model also provides an opportunity to independently and iterative refine different organ systems (respiratory, cardiovascular, neurological, etc.) as discrete models to improve efficiency, and to create a more streamlined approach to incorporate new knowledge in a specific organ system without overhauling the entire model.

The goal of this research is to develop a new highly scalable full-stack architecture for a cross-platform patient simulation application driven by graph-based models, and to present a proof-of-concept of the proposed architecture to illustrate its viability. To realize the graph-based virtual patient simulation at scale, we prioritize a highly reliable, fault-tolerant, and maintainable architecture. As we aim to develop the application as a bedside decision-support tool for clinicians in actual clinical settings, the application needs to adapt swiftly and efficiently to fluctuating user demand, and to accommodate a wide range of user devices including laptops, tablets, and smartphones with diverse operating systems (iOS, Android, etc.). Our proposed architectural approach addresses these needs in an integrated manner, contributing a sustainable and practical solution to the field. Specifically, the architecture comprises three core components: a cross-platform front-end application that clinicians and trainees use to run the simulation, a cloud-hosted simulation engine that performs all the necessary computations for each user's simulation, and a graph database that hosts the graph model used by the simulation engine to drive each simulation. By integrating these elements, we present a highly-scalable full-stack simulation application architecture, which effectively addresses the identified challenges and paves the way for a new paradigm in patient simulation and dynamic system simulation based on graph models. Although the application focus of this paper is on modeling a virtual patient, the architecture presented in this paper could be adapted to support other dynamic systems such as mechanical, physical, and physiological systems that are graph-based, e.g., Sanchez-Gonzalez et al. (2018); Tu et al. (2019); Yang et al. (2021).

In the following sections of this paper, we elaborate on how the components of our proposed architecture synergize to overcome practical challenges. We present a proof-of-concept case study demonstrating the architecture and graph model, discuss the overarching benefits of the architecture, and outline future research directions.

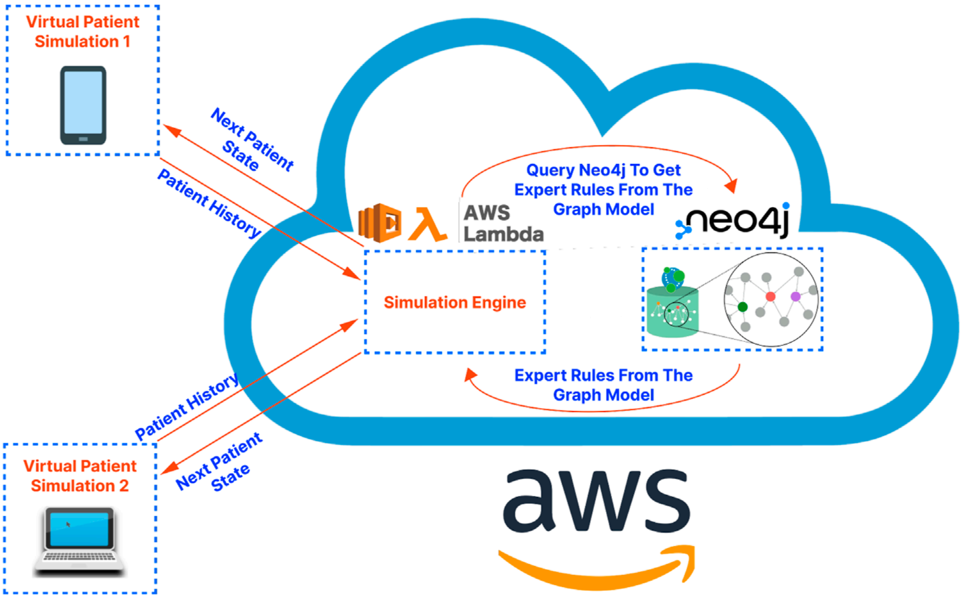

Virtual Patient Simulation 1
Next Patient State
Patient History
Query Neo4j To Get Expert Rules From The Graph Model
AWS Lambda
neo4j
Simulation Engine
Expert Rules From The Graph Model
Patient History
Next Patient State
Virtual Patient Simulation 2
aws

FIGURE 1
A high-level illustration of the proposed application architecture. The virtual patient simulations on the left-hand side of the diagram represent the front-end application. The cloud on the right-hand side of the diagram represents the cloud services serving as the "back-end" of the application. These services are hosted on Amazon Web Services (AWS) in the demo application/proof-of-concept presented in this article.

2 Materials and methods

The proposed application architecture draws upon the utility of both autoscaling serverless functions and a microservice architecture. Serverless functions are a feature offered by cloud platforms where developers write code that is executed in response to events (like a user interaction), and are automatically scaled up and down by the cloud provider. They are serverless in the sense that developers do not have to worry about server management, and their pay-as-you-go nature makes them cost-efficient for users. Microservice architecture, on the other hand, is a design pattern where an application is structured as a collection of loosely coupled services, which can be developed, deployed, and scaled independently. Anticipating usage patterns of this patient simulation application may be sporadic and synchronized, such as classroom usage leading to surges in demand, the proposed architecture is capable of scaling up and down effectively to meet these needs.

In addition, our proposed architecture considers the challenge of device heterogeneity and limited processing power, especially in the medical education setting. A cross-platform programming language is preferred, which allows developers to write a single codebase that can run on multiple platforms (like Android, iOS, and web), eliminating the need to write different versions of the application for each platform. In this case, React-Native (Masiello and Friedmann, 2017), a popular cross-platform programming language, has been employed.

For the overall architecture, the cross-platform front-end (written in React-Native) is separated from the back-end simulation engine (running on a serverless function in the cloud) and the graph database (running on a dedicated server in the cloud). This separation, characteristic of microservice-based architectures, has been shown to improve scalability, reliability, and fault tolerance while also facilitating maintenance and debugging tasks (Villamizar et al., 2015). Additionally, serverless functions, due to their autoscaling and developer-friendly nature, enable developers to focus on application logic, leaving resource provisioning and infrastructure management to cloud service providers (Chadha et al., 2022). An illustration of the proposed application architecture is shown in Figure 1. Below we present the details regarding the cross-platform front-end application, the graph database construction, and the simulation engine that drives the patient pathway simulation, respectively.

2.1 Front-end application

The cross-platform front-end application serves as the user interface for trainees and clinicians to interact with the virtual patient simulation by: (a) allowing users to set the initial state of the patient; (b) storing and showing the state of the patient over the course of a simulation; (c) allowing users to select interventions at each step of the simulation as desired; (d) sending the history of patient states to the cloud-hosted simulation engine to obtain the next state of the patient for the next step of the simulation (see Section 2.3 for more details); (e) tracking the relationships, i.e., edges in the graph-model that caused a change in the virtual patient's state at each step of the simulation; (f) allowing users to connect to the graph database to visualize the relationships defined in the graph model, which influence the trajectory of the state of the virtual patient (see Figure 2 for a sample DAG).

The microservice architecture plays a crucial role here as it does not require embedding complex simulation logic into the front-end application as would be required in a monolithic application design.

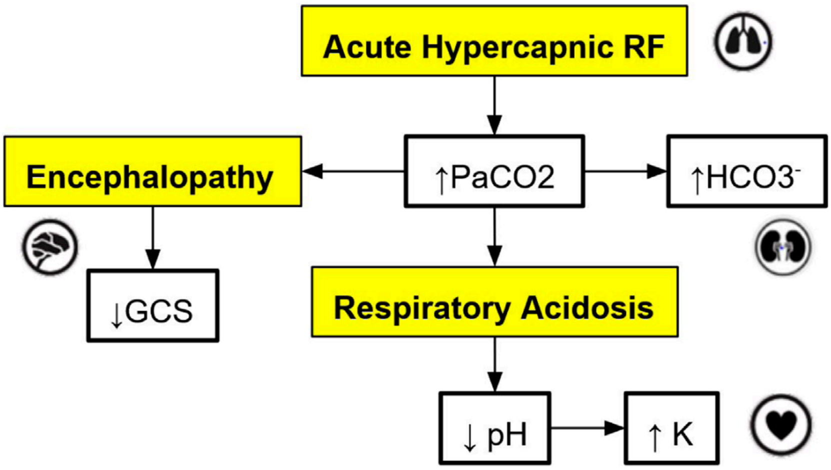


FIGURE 2
An example of a directed acyclic graph (DAG) depicting a subset of the interactions associated with respiratory acidosis. The boxes with a yellow background are medical concepts, and the boxes with a white background correspond to measurable patient vitals or clinical markers. PaCO2 = partial pressure of carbon dioxide in arterial blood, GCS = Glasgow Coma Scale, $HCO3^-$ = Bicarbonate.

This division of responsibilities keeps the front-end lightweight and modular, facilitating independent development, better error isolation, and improved overall development speed.

2.2 Graph database development

A graph database uses graph structures for semantic queries, with nodes, edges, and properties to represent and store data. This stands in contrast to a traditional SQL or noSQL database which may not natively support relationships between entities. In our study, the graph database is the heart of our simulation application, performing crucial functions like storing the graph model, enabling fast queries, providing visualization tools, and allowing developers to manage the graph model. These graph-database-powered capabilities can assist in maintaining the robustness, flexibility, and scalability of the simulation model.

For this application, the graph models are constructed based on expert rules. Our definition of expert rules takes into account the effects of clinical markers on each other and the causes (like interventions and interactions) that lead to certain effects on organ systems. Using a graph database, the expert rules (defined by clinicians and loaded into Neo4j via CSV files) that drive our simulation can be efficiently queried and updated. A very simple example DAG describing a subset of the interactions of organ systems and biomarkers associated with respiratory acidosis is shown in Figure 2. This DAG is constructed using rules presented in Table 1 (to be elaborated in this section).

Note that the simple DAG depicted in Figure 2 could be a part of a much larger DAG with many more medical concepts, measurable patient vitals, organ systems, and relationships (Lal et al., 2020b). Representing the causal pathways within the human body in an intuitive way is particularly important in a clinical setting as information overload has been correlated with an increase in medical errors (Pickering et al., 2010). Accordingly, DAGs have been utilized by clinicians in recent work to model the complex underlying causal pathways that drive the trajectory of a patient in an intuitive and visualizable way (Lal et al., 2020a). In particular, DAGs can be used to effectively model complex causal pathways within the human body as they provide a natural way to model high-dimensional directed relationships. From a simulation development perspective, instead of needing to define each new simulation scenario by hand, utilizing a graph-based simulation engine allows the number of supported scenarios to grow naturally over time as new patient vitals, clinical markers, interventions, and their associated interactions (edges) are added to the graph over the course of the iterative expert rule refinement and validation process.

The graph database utilized in this work is Neo4j (Neo4j Graph Data Platform, 2021), which has been shown to be effective at storing, querying, and analyzing graph data such as knowledge graphs (Chen, 2022). Other graph databases are also available including Amazon Neptune (Amazon Web Services, 2024) and TigerGraph (TigerGraph, 2023), among others. When developing rules for the graph model stored in the Neo4j graph database, we first define independent expert rules that have been agreed upon by the experts in the field through a formal consensus process (Gary et al., 2022). Table 1 contains sample rules expressed in the spreadsheet format to help illustrate the rule structure that is compatible with the Neo4j data structure. In the patient simulation, each rule is activated by a single triggering clinical marker or intervention (the “Cause/Input” column of the spreadsheet), and each rule causes a new incremental change or an absolute change in a single impacted clinical marker (the “Effected_Clinical_Marker” column of the spreadsheet) when all conditions for the expert rule are satisfied. Currently, states of the clinical markers are represented as integer variables (−2,-1,0,1,2) and can be color-coded in the front-end user interface. The integer values map to different value ranges of measurable biomarkers. For example, level 2 for PaCO2 corresponds to values between 71 and 120 mmHg. In the front-end application, a number randomly drawn within this range will be displayed to users, providing users with an experience closer to their regular interactions with electronic health records.

TABLE 1 The set of expert rules which define the edges in the Neo4j graph shown in Figure 4, and which represent the relationships shown in the DAG in Figure 2. These rules govern the progression of the state of the virtual patient described in the case study in Section 3.

Rule #	Cause/Input	Previous _State_ Of_Cause /Input	New _State_ Of_Cause/Input	Duration	Effected _Clinical _Marker	Impact	P	Time _Until _Effect	Simple_Conditions
1	PaCO2	2	2	30	GCS	−1	0.8	0	
2	PaCO2	1	2	30	GCS	−1	0.8	0	
3	PaCO2	0	2	30	GCS	−1	0.8	0	
4	PaCO2	−1	2	30	GCS	−1	0.8	0	
5	PaCO2	−2	2	30	GCS	−1	0.8	0	
6	PaCO2	1	2	0	pH	−1	1	15	
7	PaCO2	0	1	0	pH	−1	1	15	
8	PaCO2	−1	0	0	pH	−1	1	15	
9	PaCO2	−2	−1	0	pH	−1	1	15	
10	PaCO2	1	2	0	HCO3-	1	0.8	240	
11	PaCO2	0	1	0	HCO3-	1	0.8	240	
12	PaCO2	−1	0	0	HCO3-	1	0.8	240	
13	PaCO2	−2	−1	0	HCO3-	1	0.8	240	
14	pH	2	1	0	K	1	0.8	30	[{Given_Insulin: 0, Duration: 60} {Given_Furosemide: 0, Duration: 60}]
15	pH	1	0	0	K	1	0.8	30	[{Given_Insulin: 0, Duration: 60} {Given_Furosemide: 0, Duration: 60}]
16	pH	0	−1	0	K	1	0.8	30	[{Given_Insulin: 0, Duration: 60} {Given_Furosemide: 0, Duration: 60}]
17	pH	−1	−2	0	K	1	0.8	30	[{Given_Insulin: 0, Duration: 60} {Given_Furosemide: 0, Duration: 60}]

The first rule in Table 1 says, when the patient's PaCO2 level stays at a high level (2) for a duration of 30 min, then GCS (Glasgow Coma Scale) decreases by 1 level with a probability of 0.8. In this example, PaCO2 is the "Cause/Input" of the rule, GCS is the "Effected_Clinical_Marker", 0.8 is the "Probability", −1 is the "Impact", and 0 is the "Time_Until_Effect" (in minutes). The columns "Previous_State_Of_Cause/Input" and "New_State_Of_Cause/Input" describe what needs to happen to the value of the "Cause/Input" for the rule to be triggered. There are three possible triggers that we can account for: The "Cause/Input" increases, decreases, or stays at a particular value over the specified "Duration". In this example, the "Previous_State_Of_Cause/Input" and the "New_State_Of_Cause/Input" of PaCO2 are both high (level 2), and the "Duration" is 30 min meaning that this rule is triggered after PaCO2 has been at level 2 for 30 min. By specifying a "Duration", we can have different rules for changes that occur acutely/quickly, or which occur slowly over time. We can also model rules such as "IF PaCO2 is >70 mmHg (FOR 30 min) THEN GCS decreases" which requires that a particular "Cause/Input" (PaCO2 in this case) stays at a particular value (in this case, at a high value) for some duration. Note that, by allowing for capturing the "Duration", the simulation is no longer memoryless and the applicability of a rule is based on the historical patient trajectory.

The effect of each rule on the impacted clinical marker is stored in the "Impact" column and is represented by one of the following integers: (−2,-1,1,2). The negative (positive, resp.) integers represent a decrease (an increase, resp.) in the value or level of the impacted clinical marker. In this example (rule #1), the GCS level will be decreased by 1 level, from its current level, and the time-lapse it needs to be effective is stored in the "Time_Until_Effect" column (with zero meaning being effective immediately in this case). To handle cases where multiple rules are simultaneously applying changes to a single clinical marker during one step of the simulation, we introduce two types of rules, one causes an incremental change, meaning that its effect is additive to others that are also incremental. The other type is "absolute", which will override other rules once applied. In this simple example, all rules cause incremental changes.

For a rule to be activated, relevant conditions defined in the rule must be satisfied. The simple conditions are one or more independent conditions that all must be satisfied for a rule to take effect. Rules 14–16 in Table 1 have two simple conditions, {Given_Insulin: 0, Duration: 60} and {Given_Furosemide: 0, Duration: 60}. These conditions mean that rules 14–16 will only be applied if the patient has not been given Insulin or Furosemide during the last 60 min.

Meanwhile, complex conditions are the conditions that are satisfied if at least one of a possible set of conditions is satisfied. For example, a complex condition expressed as "[{ Brain_Swelling: 0, Duration: 0 },{ Mannitol: 1, Duration: 30 }]" requires that at least one of the following must be true: (a) the patient must have no current brain swelling (b) they must have received Mannitol 30 min ago.

If all of the conditions for a rule are satisfied, we then apply the rule with the probability listed in the "P" column. The probability characterizes the chance that a certain change in the human body will occur to maintain a level of stochasticity in the simulation model.

This precise structure for expressing expert rules allows us to capture the majority of the common rules using a systematic format that is interoperable with graph databases, and enables us to customize each expert rule based on the applicability of each property.

2.3 Cloud-hosted simulation engine

The cloud-hosted simulation engine is responsible for executing the simulation according to the graph model stored in the database and the user interactions captured by the front-end application. The engine runs on a serverless function (on a Function as a Service platform, like Amazon Web Services Lambda or Google Cloud Functions), allowing it to scale seamlessly in response to demand. These serverless computing platforms provide developers with a high degree of flexibility and scalability, as they only need to be concerned with application code and can leave infrastructure management to the service provider.

The engine is designed to take the current state of the patient, as well as any user actions (like giving a medication or performing a procedure), and calculate the resulting state of the patient. For this, it queries the graph database for relevant rules, performs calculations, and sends the new patient state back to the front-end application. As a benefit, the engine does not have to store any state itself, making it inherently scalable and resilient. Also, being decoupled from the front-end and the database, it can be independently developed, tested, and deployed, which reduces the complexity of the overall system.

All current and future rules can be processed in a uniform way using the same code (the code running in the simulation engine as shown in Figure 1). This means that rules in the graph database can be added and updated in the future independently without the need for the developers to write any new code. Specifically, to obtain the next patient state at each step in the simulation, the front-end application sends the complete patient history to the simulation engine and waits for a response which includes:

1. The next state (described by the states of all clinical markers) of the patient.
2. The rules that were applied (if any) which impacted the next state of the patient.

The upper and lower limits for the value of each clinical marker (currently some appropriate range between "very low" (−2) and "very high" (2)) and the lower and upper bound for each intervention (between "no intervention" (0) and "high dose intervention" (2)) are defined in the simulation engine and enforced at each step. Similarly, the length between each step in the simulation is defined (currently "15 min").

The procedure followed by the simulation engine at each step of the simulation is outlined in Algorithm 1 and illustrated in Figure 3. This procedure integrates several functions in a modular approach to rule application and state updates.

2.3.1 InitializeSimulation function

The InitializeSimulation procedure initializes the parameters and patient history required for the simulation. It ensures that all necessary data is correctly set up before the main simulation steps begin.

2.3.2 ApplyRules function

The ApplyRules function applies the relevant rules from the expert rules set to update the patient's state. It checks if the conditions for each rule are met and, if so, updates the patient state accordingly.

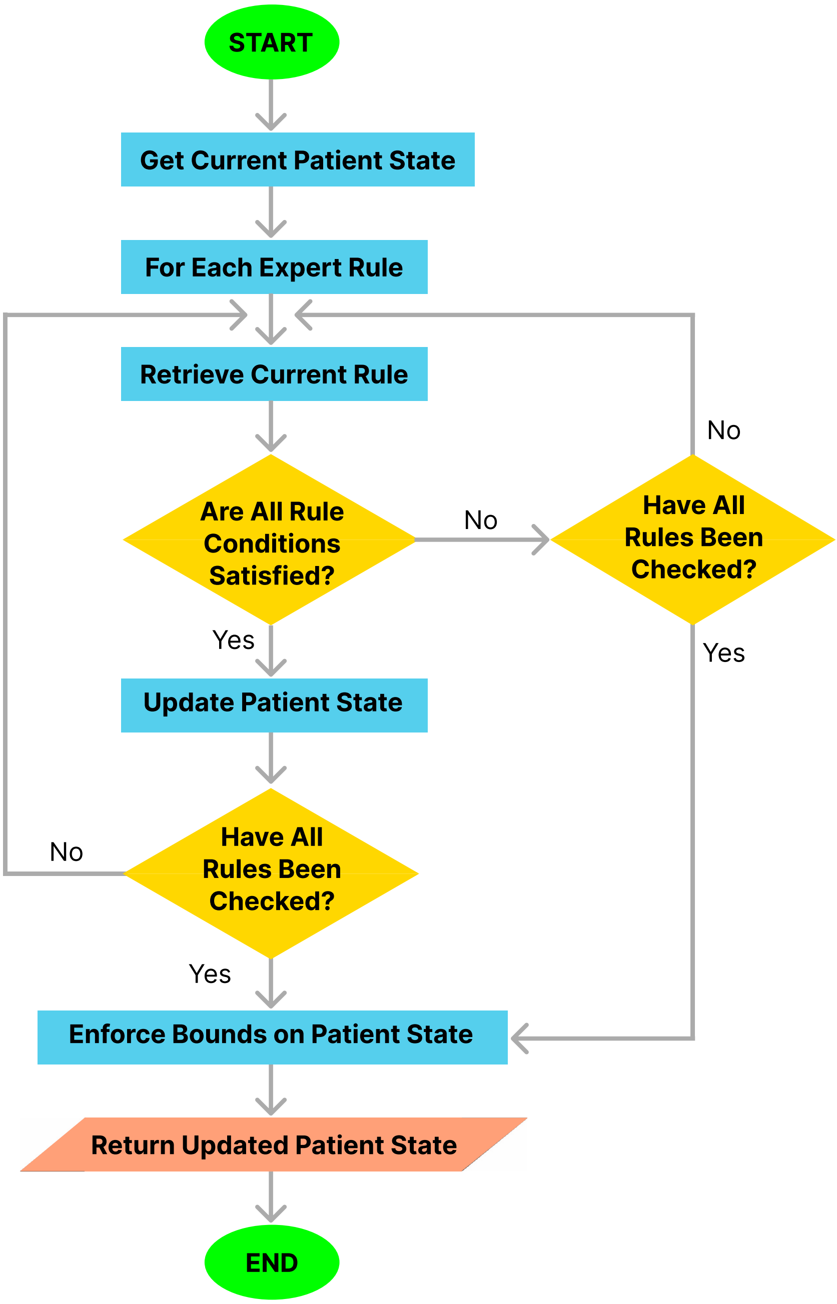

START
Get Current Patient State
For Each Expert Rule
Retrieve Current Rule
Are All Rule Conditions Satisfied?
No
Have All Rules Been Checked?
No
Yes
Yes
Update Patient State
Have All Rules Been Checked?
No
Yes
Enforce Bounds on Patient State
Return Updated Patient State
END

FIGURE 3
Flowchart of the simulation engine algorithm.

2.3.3 HandleConditions function

The HandleConditions function evaluates whether the conditions for applying a rule are satisfied based on the patient's history and the specifics of the rule. It checks whether the current rule contains a simple condition or a complex condition and whether these are satisfied over the most recent steps to be analyzed prior to moving to the next time instance. We added simple and complex conditions during the rule construction process to ensure that the expert rules are capable of fully capturing the intricate relationships between organ systems in the human body. For example, the administration of propofol to a critically ill patient should result in a drop in GCS as well as a drop in MAP. However, if phenylephrine was administrated at the same time as propofol, a drop in MAP would have not occurred. Then, administration of phenylephrine would be included in the simple condition of the rules denoted as {*Given_Phenylephrine*: 0} suggesting that phenylephrine should not be currently effective for this rule to be applicable.

The algorithm returns a Boolean variable *ConstraintsSatisfied* being "True" if all constraints are satisfied, and "False" otherwise. The condition check operation shares a similar structure as the main

```
Require: $Time\_Between\_Steps = 15$
Require: $t = 0, 1, \ldots, T$ ▷ The steps of the simulation,
         each of which is Time_Between_Steps
         minutes apart
Require: $Variable\_Names = \{Name_1, Name_2, \ldots, Name_n\}$
Require: $Lower\_Bounds = \{l_1, l_2, \ldots, l_n\}$
Require: $Upper\_Bounds = \{u_1, u_2, \ldots, u_n\}$
Require: $Patient\_History = \{h_0, h_1, h_2, \ldots, h_t\}$
Require: $Expert\_Rules \leftarrow \{Rule_1, Rule_2, \ldots, Rule_m\}$
$h_{t+1} = h_t$
InitializeSimulation (Time_Between_Steps, t,
  Variable_Names, Lower_Bounds, Upper_Bounds,
  Patient_History, Expert_Rules)
for $j = 1$ to $m$ do
    $Current\_Rule = Expert\_Rules[j]$
    ApplyRules (Current_Rule, Patient_History,
      $h_{t+1}$, Time_Between_Steps)
end for
for Varin Variable_Names do
    EnforceBounds ($h_{t+1}$, Var, Lower_Bounds,
      Upper_Bounds)
end for
    return $h_{t+1}$
```

Algorithm 1. Simulation Engine Overarching Algorithm.

```
procedure InitializeSimulation (Time_Between_Steps,
    t, Variable_Names, Lower_Bounds,
    Upper_Bounds, Patient_History, Expert_Rules)
  Initialize parameters and patient history
end procedure
```

Algorithm 2. InitializeSimulation Procedure.

algorithm, e.g., screening the states and managing the time indexes, and the details are skipped for the interest of space.

2.3.4 UpdatePatientState function

The UpdatePatientState procedure applies the impacts of a rule to the patient's state if the conditions for that rule are met.

2.3.5 EnforceBounds function

The EnforceBounds procedure ensures that the values of all clinical markers and interventions remain within their predefined bounds (e.g., when incremental rules are applied, check if the values go beyond −2 or +2). If a value exceeds its bounds, it is set to the respective limit.

The algorithmic approach modularizes the process into distinct functions, each responsible for specific aspects of the simulation, thus enhancing clarity and maintainability. The overarching algorithm (Algorithm 1) orchestrates the workflow, ensuring that all necessary steps are performed in sequence, while the individual functions handle initialization, rule application, condition checking, patient state updating, and enforcing bounds.

```
function ApplyRules (Current_Rule,
  Patient_History, $h_{t+1}$, Time_Between_Steps)
  $Duration\_Steps = \frac{Current\_Rule[Duration]}{Time\_Between\_Steps}$
  $Index\_Of\_Newest\_Measurement\_To\_Look\_At = \frac{Current\_Rule[Time\_Until\_Effect]}{Time\_Between\_Steps}$
  $Index\_Of\_Oldest\_Measurement\_To\_Look\_At = Index\_Of\_Newest\_Measurement\_To\_Look\_At + Duration\_Steps + 1$
  if $Index\_Of\_Oldest\_Measurement\_To\_Look\_At > t$
   then
   return False
  end if
  $Cause = Current\_Rule[Cause/Input]$
  if $h_{t-Index\_Of\_Oldest\_Measurement\_To\_Look\_At}[Cause] \neq$
   $Current\_Rule[Previous\_State\_Of\_Cause/Input]$
    then
    return False
   end if
   end if $h_{t-Index\_Of\_Newest\_Measurement\_To\_Look\_At}[Cause] \neq$
   $Current\_Rule[New\_State\_Of\_Cause/Input]$ then
    return False
   end if
   $MaxValue =$
   $\max(Current\_Rule[Previous\_State\_Of\_Cause/Input],$
    $Current\_Rule[New\_State\_Of\_Cause/Input])$
   $MinValue = \min(Current\_Rule$
    $[Previous\_State\_Of\_Cause/Input],$
    $Current\_Rule[New\_State\_Of\_Cause/Input])$
   for $k =$
$(t - Index\_Of\_Oldest\_Measurement\_To\_Look\_At + 1)$ to
    $(t - Index\_Of\_Newest\_Measurement\_To\_Look\_At - 1)$ do
    if $h_k[Cause] > MaxValueorh_k[Cause] < MinValue$ then
     return False
    end if
   end for
   if HandleConditions (h,Current_Rule,
    Index_Of_Newest_Measurement_To_Look_At,
      Time_Between_Steps) then
    UpdatePatientState (Current_Rule, $h_{t+1}$)
    return True
   else
    return False
   end if
  end function
```

Algorithm 3. ApplyRules Function.

To summarize, the simulation engine runs on a serverless function in the cloud and performs the following functions: (a) receives the history of a virtual patient from a user's front-end application; (b) calculates the next state of the virtual patient for the next step of the simulation by analyzing the history of past states of the virtual patient, querying the graph database to obtain

```
function HandleConditions ($h, Current\_Rule, Index\_Of\_Newest\_Measurement\_To\_Look\_At, Time\_Between\_Steps$)
    Evaluate simple and complex conditions of the rule
    return all conditions are satisfied and also $rand(Unif(0,1)) \le Current\_Rule[Probability]$
end function
```

Algorithm 4. HandleConditions Function.

```
procedure UpdatePatientState ($Current\_Rule, h_{t+1}$)
    $h_{t+1}[Effected\_Clinical\_Marker] += Current\_Rule[Impact]$
end procedure
```

Algorithm 5. UpdatePatientState Procedure.

```
procedure EnforceBounds ($h_{t+1}, Var, Lower\_Bounds, Upper\_Bounds$)
    if $h_{t+1}[Var] < Lower\_Bounds[Var]$ then
        $h_{t+1}[Var] = Lower\_Bounds[Var]$
    else if $h_{t+1}[Var] > Upper\_Bounds[Var]$ then
        $h_{t+1}[Var] = Upper\_Bounds[Var]$
    end if
end procedure
```

Algorithm 6. EnforceBounds Procedure.

the relevant relationships from the graph-model which may cause a change in the state of the patient, and applying the queried relationships as appropriate to calculate the next state of the patient; (c) returns any rules that were applied and the next state of the virtual patient for the next step of the simulation to the user's front-end application.

3 Results

To demonstrate the viability of the proposed simulation architecture, we will walk through a short case study that considers a virtual patient whose state is defined in terms of the five clinical markers shown in the DAG in Figure 2 and the corresponding nodes in the Neo4j graph in Figure 4. The trajectory of the patient will be determined by the set of edges shown in the Neo4j graph in Figure 4, each of which corresponds to an expert rule defined in Table 1. The trajectory of the patient's state throughout this case study is summarized in Table 2, and the rules from Table 1 that were applied at each step of the simulation (each step is 15 min) are described in the "Applied Rules" column of Table 2.

This case study (respiratory acidosis) is crafted to allow for a manual prospective validation to assist in a quick understanding of the simulation mechanism. In the real implementation, the user will first choose a clinical scenario (e.g., chronic obstructive pulmonary disease exacerbation, or sepsis), along with the most relevant clinical markers and the corresponding rules related to this clinical scenario will be identified. Each clinical scenario is typically associated with dozens of clinical markers and rules, e.g., 70 rules for a demonstration version for validation in a related study (Rovati et al., 2024).

3.1 Initializing the simulation

To initialize the simulation, we first need to set the lower and upper bounds for each vital/clinical marker that we have. In this case study, the simulation engine was configured to use the upper and lower bounds: *Lower_Bounds* = $\{PaCO2: -2, pH: -2, HCO3^{-}: -2, GCS: -2, K: -2\}$, *Upper_Bounds* = $\{PaCO2{:}2, pH{:}2, HCO3^{-}{:}2, GCS{:}0, K{:}2\}$.

Also, we need to define an initial *Patient_History* = $\{h_0, h_1\}$ for the patient. Let us assume that at the first step of the simulation, step $t = 0$ (row 1 of Table 2), the patient had a slightly elevated level of PaCO2 (denoted by a value of "1") and a normal level of all the other clinical markers (denoted by a value of "0"). Then, 15 min later at step $t = 1$ (row 2 of Table 2), the patient had a very elevated PaCO2 level (denoted by a value of "2"), but still had a normal level (level "0") for all the other clinical markers. In this case, the *Patient_History* described in Algorithm 1 is initialized as $h_0 = \{PaCO2{:}1, pH{:}0, HCO3^{-}{:}0, GCS{:}0, K{:}0\}$ and $h_1 = \{PaCO2{:}2, pH{:}0, HCO3^{-}{:}0, GCS{:}0, K{:}0\}$.

3.2 The patient's state trajectory during the simulation

As shown in Table 2, the first rule applied is Rule # 6 at time $t = 30$ minutes. This is expected as Rule # 6 is triggered by an increase in PaCO2 from a slightly elevated level (a value of "1") to a very elevated level (a value of "2"). Since the duration is 0 min for this rule, this rule is triggered as soon as the value of PaCO2 changes from "1" to "2". However, this rule has a delayed "Time_Until_Effect" of 15 min which means that the "Impact" of the rule is applied 15 min after the rule is triggered. Therefore, since the rule was applied at time $t = 30$ minutes, the rule was triggered 15 min earlier, at time $t = 15$ minutes. Once the rule was triggered it was guaranteed to be applied since the rule's probability, *P*, is 100%.

Next, at time $t = 60$ Rule #16 was applied. Rule #16 is triggered by a decrease in pH from a normal level (level "0") to a slightly low level (level "-1"). After the decrease occurs, this rule is delayed by a "Time_Until_Effect" of 30 min. Therefore, the change in pH must have occurred 40 min earlier, which we can see occurred in Table 2 as pH decreased from normal (level "0") at time $t = 15$ to slightly low (level "-1") at time $t = 30$. It is therefore in alignment with our expectations that Rule #16 is applied 30 min later at time $t = 60$ minutes due to the rule's "Time_Until_Effect" of 30 min.

At time $t = 75$ one rule was applied, Rule #1. Rule # 1 is triggered by PaCO2 being at level "2" for 30 min, and looking at the patient's state history in Table 2, we can see that at time $t = 75$ minutes, the patient had actually already had a PaCO2 level of "2" for 60 min. Since this rule has a "Time_Until_Effect" of 0 min, we know

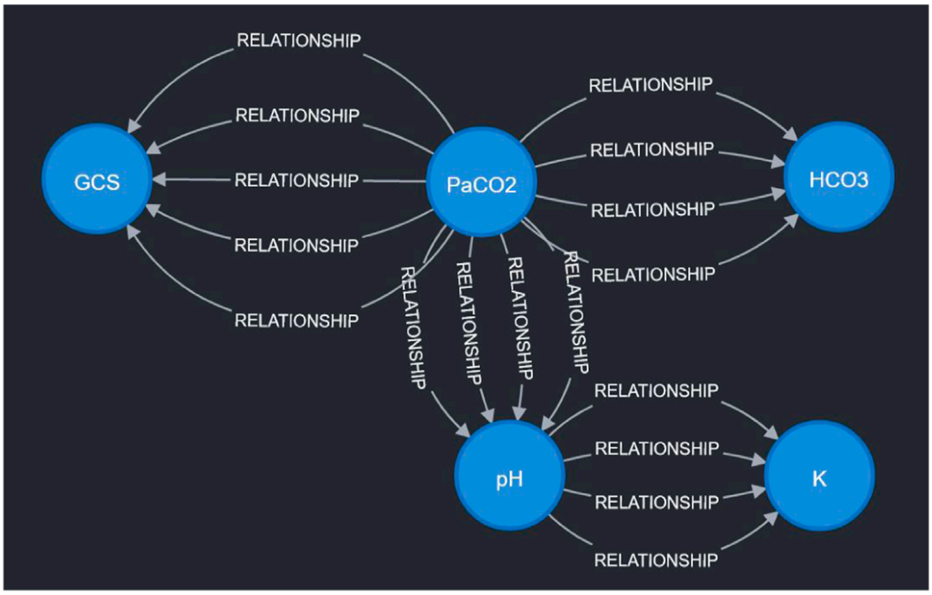


FIGURE 4
Visualization of sample expert rules stored in the Neo4j graph database. Each node in the graph corresponds to a measurable vital or clinical marker in Figure 2. Each directed edge corresponds to a specific expert rule in Table 1. The detailed cause-effect will be displayed when the specific "relationship" edge is clicked in the Neo4j workspace.

TABLE 2 The patient's state throughout Section 3 case study.

Time (min)	PaCO2	pH	HCO3-	GCS	K	Applied rules
0	1	0	0	0	0	
15	2	0	0	0	0	
30	2	−1	0	0	0	6
45	2	−1	0	0	0	
60	2	−1	0	0	1	16
75	2	−1	0	−1	1	1
90	2	−1	0	−2	1	1
105	2	−1	0	−2	1	
120	2	−1	0	−2	1	

that once this rule is triggered, its "Impact" is instantly applied. Subtracting the rule's "Duration" of 30 min from the 60 min that the patient's PaCO2 level was "2", we can see that starting at time t = 30 minutes the rule was being triggered. However, as indicated by column P of Table 1, Rule #1 only has an 80% probability of being applied each time it is triggered. This means that the rule was only applied on the third time that it was triggered (the 20% chance that the rule would not be applied hit the first two times it was triggered, at t = 45 and t = 60).

At time t = 90, Rule #1 was applied again, further decreasing GCS to its lower bound of "-2". As we can see, Rule #1 was not decreased at time t = 105 or t = 120 even though Rule #1 was still being triggered since GCS can not decrease below its lower bound (below a value of "-2").

In conclusion, we can see that the trajectory of the patient's state throughout the case study (Table 2) is in alignment with our expectations based on our expert rules (Table 1).

4 Discussion

The presented work introduces an application architecture designed to overcome various challenges inherent in the dynamic realm of healthcare simulations. Specifically, it is constructed to seamlessly scale to accommodate a growing user base with sporadic and correlated usage patterns, making it universally accessible across a multitude of platforms. It is also built to operate reliably under various conditions while ensuring fault-tolerance and easy maintainability.

A key aspect of this architecture is that it does not question the validity of expert rules, but rather focuses on the execution of these rules within the simulation. Therefore, during the validation phase, an unexpected simulation behavior due to an incorrect expert rule or its faulty implementation can be handled separately. For instance, if an erroneous simulation result is due to an incorrect expert rule, the developer only needs to update the graph database without touching the simulation engine. This will also improve the handling of the changes in the clinical management of patients in the intensive care unit where the scientific premise and the interventions change according to an evolving body of evidence.

Because of the stochastic nature of the simulation and the scale of the model, it is infeasible to validate the model based on specific values of each individual clinical marker realized in each simulation run. Rather, we focus on the clinical trajectory and examine whether the trajectory over an initial 6-h span from the time of admission is concordant with the expectation (e.g., samples from real patient trajectories or crafted virtual patients with the same clinical scenario). Our commitment to enhancing the validity and utility of this simulation application extends beyond the present study. We understand the importance of rigorous evaluation and ablation studies and are actively engaged in further research to refine and validate the expert rules that underpin the simulation. We are employing rigorous methodologies to calibrate the decision-making algorithms based on real-world patient data and physician inputs. To ascertain the application's effectiveness as an educational tool and its ability to satisfy user requirements, we have initiated a mixed-methods study involving first-year Internal Medicine residents (Gary et al., 2023; Rovati et al., 2024). These user testing sessions are specifically designed to assess the usability of the application, the workload it presents to users, the usability of the application, and the satisfaction of learners. We anticipate that the findings from these sessions will provide invaluable insights and guide iterative refinement of the application design to better cater to user needs.

Looking ahead, there are numerous avenues for enhancing the proposed architecture's scalability, reliability, efficiency, and performance. Such improvements are crucial for realizing high-fidelity graph-based simulation models capable of functioning as decision support tools for clinicians at the bedside. Our vision is to use these models as digital twins and interpretable counterparts to less transparent associative AI models, facilitating patient diagnosis and optimal treatment prediction in real-time settings (see, for example, (Komorowski et al., 2018; Chakshu and Nithiarasu, 2022; Sun et al., 2022)). The interpretability aspect is particularly crucial in healthcare, given the reluctance among clinicians to adopt "black-box" AI models (Dang et al., 2021; Lal et al., 2022).

Specifically, to utilize a data-driven approach to further validate the patient simulation application, it is necessary to extract meaningful data points from the current plethora of variables thereby improving the signal-to-noise ratio. This approach would involve the current electronic health record data being mapped to experimentally proven physiological concepts (e.g., utilizing our approach with DAGs and validated expert rules). The future iterations of this scalable patient simulation application will also include a "plug-in" feature with the current electronic health record, which will seamlessly integrate the real-time data and interoperability of the proposed virtual testing environment with the current clinical infrastructure for medical education, *in silico* research, and clinical decision support.

To realize these visions, an exciting future direction involves the utilization of graph algorithms like Graph Neural Networks for link prediction. This would improve the accuracy of the graph model that drives the virtual patient simulation. Graph Neural Networks have demonstrated state-of-the-art results in predicting synthetic lethality and drug-target interaction in biomedical networks (Long et al., 2022). Therefore, applying these algorithms to a graph model based on DAGs, illustrating causal relationships and intricate pathophysiological interactions within the human body, could potentially yield impressive results.

Another intriguing prospect is to enhance the efficiency of querying the Neo4j graph database. Currently, the simulation engine examines all rules upon querying the graph database, even those that do not meet the application conditions. Future work should aim to develop more specific queries using Neo4j's cypher query language. This could traverse only nodes or edges of a specific type or with particular properties, increasing query efficiency. However, this requires careful reconsideration of how the data is structured within the database, given the unique set of simple and complex conditions associated with each rule.

Lastly, the incorporation of parallel computing within the cloud-hosted simulation engine could significantly boost its performance. Recent research has shown that integrating parallel computing within serverless functions drastically enhances performance and reduces costs (Kiener et al., 2021). Future studies could adapt these findings to elevate the performance of our simulation engine. These initiatives, when realized, could greatly advance the capabilities of the proposed architecture, moving us closer to our ultimate goal of creating a robust and scalable tool for healthcare simulations.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The datasets used for this study are not publicly available. Requests to access these datasets should be directed to YD, dong.yue@mayo.edu.

Ethics statement

The studies involving humans were approved by Mayo Clinic Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

WT: Conceptualization, Formal Analysis, Methodology, Writing–original draft. XZ: Conceptualization, Funding acquisition, Methodology, Writing–original draft. AL: Conceptualization, Methodology, Writing–review and editing. LR: Conceptualization, Methodology, Writing–review and editing. EC: Methodology, Software, Writing–review and editing. YD: Conceptualization, Project administration, Writing–review and editing. PS: Conceptualization, Funding acquisition, Writing–review and editing. OG: Conceptualization, Methodology, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Science Foundation IIS-2123848 and IIS-2123900.

Acknowledgments

The authors would like to thank Simon Zec, Nika Zorko Garbajs, John M. Litell for their contribution to the development of expert rules.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Amazon Web Services (2024). *AWS Neptune*. Available at: https://aws.amazon.com/neptune/ (Accessed April 28, 2024).

Berger, J., Bawab, N., De Mooij, J., Sutter Widmer, D., Szilas, N., De Vriese, C., et al. (2018). An open randomized controlled study comparing an online text-based scenario and a serious game by belgian and swiss pharmacy students. *Curr. Pharm. Teach. Learn.* 10, 267–276. doi:10.1016/j.cptl.2017.11.002

Burkhoff, D., and Dickstein, M. L. (2024). *Harvi academy simulator*. Available at: https://harvi.academy/simulator/ (Accessed April 28, 2024).

Chadha, M., Pacyna, V., Jindal, A., Gu, J., and Gerndt, M. (2022). *Migrating from microservices to serverless: an iot platform case study*. New York, NY, USA: Association for Computing Machinery.

Chakshu, N. K., and Nithiarasu, P. (2022). An ai based digital-twin for prioritising pneumonia patient treatment. *Proc. Institution Mech. Eng. Part H J. Eng. Med.* 236, 1662–1674. doi:10.1177/09544119221123431

Chen, X. (2022). "Design and implementation of knowledge graph of listed companies based on Neo4j," in *International conference on high performance computing and communication (HPCCE 2021)*. Editors Y. Wang, and S. Chen (Bellingham, Washington : International Society for Optics and Photonics), 12162, 1216213. doi:10.1117/12.2628309

Chu, Y., Li, S., Tang, J., and Wu, H. (2023). The potential of the medical digital twin in diabetes management: a review. *Front. Med.* 10, 1178912. doi:10.3389/fmed.2023.1178912

Dang, J., Lal, A., Flurin, L., James, A., Gajic, O., and Rabinstein, A. A. (2021). Predictive modeling in neurocritical care using causal artificial intelligence. *World J. Crit. Care Med.* 10, 112–119. doi:10.5492/wjccm.v10.i4.112

Gary, P., Rovati, L., Dong, Y., Lal, A., Cubro, E., Wörster, M., et al. (2023). "Use of a digital twin virtual patient simulator in critical care education: a pilot study," in *A45. ICU practices, quality improvement, and medical education* (American Thoracic Society), A1681.

Gary, P. J., Lal, A., Simonetto, D., Gajic, O., and De Moraes, A. G. (2022). Results of a modified delphi approach to expert consensus for a digital twin patient model in the icu: acute on chronic liver failure. *Chest* 162, A2702. doi:10.1016/j.chest.2022.08.2198

Goldsworthy, S., Muir, N., Baron, S., Button, D., Goodhand, K., Hunter, S., et al. (2022). The impact of virtual simulation on the recognition and response to the rapidly deteriorating patient among undergraduate nursing students. *Nurse Educ. Today* 110, 105264. doi:10.1016/j.nedt.2021.105264

Hwang, G.-J., Chang, C.-Y., and Ogata, H. (2022). The effectiveness of the virtual patient-based social learning approach in undergraduate nursing education: a quasi-experimental study. *Nurse Educ. Today* 108, 105164. doi:10.1016/j.nedt.2021.105164

Katsoulakis, E., Wang, Q., Wu, H., Shahriyari, L., Fletcher, R., Liu, J., et al. (2024). Digital twins for health: a scoping review. *NPJ Digit. Med.* 7, 77. doi:10.1038/s41746-024-01073-0

Kiener, M., Chadha, M., and Gerndt, M. (2021). "Towards demystifying intra-function parallelism in serverless computing," in *Proceedings of the seventh international workshop on serverless computing (WoSC7) 2021* (New York, NY, USA: Association for Computing Machinery), 42–49. WoSC '21. doi:10.1145/3493651.3493672

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* 24, 1716–1720. doi:10.1038/s41591-018-0213-5

Kononowicz, A. A., Woodham, L. A., Edelbring, S., Stathakarou, N., Davies, D., Saxena, N., et al. (2019). Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. *J. Med. Internet Res.* 21, e14676. doi:10.2196/14676

Lal, A., Dang, J., Nabzdyk, C., Gajic, O., and Herasevich, V. (2022). Regulatory oversight and ethical concerns surrounding software as medical device (samd) and digital twin technology in healthcare. *Ann. Transl. Med.* 10, 950. doi:10.21037/atm-22-4203

Lal, A., Li, G., Cubro, E., Chalmers, S., Li, H., Herasevich, V., et al. (2020a). Development and verification of a digital twin patient model to predict specific treatment response during the first 24 hours of sepsis. *Crit. care Explor.* 2, e0249. doi:10.1097/CCE.0000000000000249

Lal, A., Pinevich, Y., Gajic, O., Herasevich, V., and Pickering, B. (2020b). Artificial intelligence and computer simulation models in critical illness. *World J. Crit. Care Med.* 9, 13–19. doi:10.5492/wjccm.v9.i2.13

Laubenbacher, R., Mehrad, B., Shmulevich, I., and Trayanova, N. (2024). Digital twins in medicine. *Nat. Comput. Sci.* 4, 184–191. doi:10.1038/s43588-024-00607-6

Lee, C. Y., and Lee, S. W. H. (2021). Review: impact of the educational technology use in undergraduate pharmacy teaching and learning – a systematic review. *Pharm. Educ.* 21, 159–168. doi:10.46542/pe.2021.211.159168

Lee, J., Kim, H., Kim, K. H., Jung, D., Jowsey, T., and Webster, C. S. (2020). Effective virtual patient simulators for medical communication training: a systematic review. *Med. Educ.* 54, 786–795. doi:10.1111/medu.14152

Long, Y., Wu, M., Liu, Y., Fang, Y., Kwoh, C. K., Chen, J., et al. (2022). Pre-training graph neural networks for link prediction in biomedical networks. *Bioinformatics* 38, 2254–2262. doi:10.1093/bioinformatics/btac100

Masiello, E., and Friedmann, J. (2017). *Mastering React native*. Birmingham, United Kingdom: Packt Publishing Ltd.

Moingeon, P., Chenel, M., Rousseau, C., Voisin, E., and Guedj, M. (2023). Virtual patients, digital twins and causal disease models: paving the ground for *in silico* clinical trials. *Drug Discov. today* 28, 103605. doi:10.1016/j.drudis.2023.103605

Neo4j Graph Data Platform (2021). *Neo4j graph data platform*. Available at: https://neo4j.com/ (Accessed April 28, 2024).

Pickering, B. W., Herasevich, V., Ahmed, A., and Gajic, O. (2010). Novel representation of clinical information in the ICU: developing user interfaces which reduce information overload. *Appl. Clin. Inf.* 1, 116–131. doi:10.4338/ACI-2009-12-CR-0027

Rovati, L., Gary, P. J., Cubro, E., Dong, Y., Kilickaya, O., Schulte, P. J., et al. (2024). Development and usability testing of a patient digital twin for critical care education: a mixed methods study. *Front. Med.* 10, 1336897. doi:10.3389/fmed.2023.1336897

Sanchez-Gonzalez, A., Heess, N., Springenberg, J. T., Merel, J., Riedmiller, M., Hadsell, R., et al. (2018). "Graph networks as learnable physics engines for inference and control," in *International Conference on machine learning (PMLR)*, 4470–4479.

Sun, T., He, X., Song, X., Shu, L., Li, Z., Lan, Q., et al. (2022). Presbyopia-correcting performance and subjective outcomes of a trifocal intraocular lens in eyes with different axial lengths: a prospective cohort study. *Front. Med.* 9, 980110. doi:10.3389/fmed.2022.980110

TigerGraph (2023). *Graph analytics platform: graph database*. Available at: https://www.tigergraph.com/ (Accessed April 28, 2024).

Trevena, W., Lal, A., Zec, S., Cubro, E., Zhong, X., Dong, Y., et al. (2022). Modeling of critically ill patient pathways to support intensive care delivery. *IEEE Robotics Automation Lett.* 7, 7287–7294. doi:10.1109/lra.2022.3183253

Tu, R., Zhang, K., Bertilson, B., Kjellstrom, H., and Zhang, C. (2019). "Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation," in *Advances in neural information processing systems*. Editors H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Red Hook, NY, United States: Curran Associates, Inc.), 32.

Vallée, A. (2023). Digital twin for healthcare systems. *Front. Digital Health* 5, 1253050. doi:10.3389/fdgth.2023.1253050

Venkatesh, K. P., Raza, M. M., and Kvedar, J. C. (2022). Health digital twins as tools for precision medicine: considerations for computation, implementation, and regulation. *NPJ Digit. Med.* 5, 150. doi:10.1038/s41746-022-00694-7

Villamizar, M., Garcés, O., Castro, H., Verano, M., Salamanca, L., Casallas, R., et al. (2015). "Evaluating the monolithic and the microservice architecture pattern to deploy web applications in the cloud," in *2015 10th computing Colombian conference (10CCC)*, 583–590. doi:10.1109/ColumbianCC.2015.7333476

Wu, Q., Wang, Y., Lu, L., Chen, Y., Long, H., and Wang, J. (2022). Virtual simulation in undergraduate medical education: a scoping review of recent practice. *Front. Med.* 9, 855403. doi:10.3389/fmed.2022.855403

Yang, C., Gao, W., Wu, D., and Wang, C. (2021). "Learning to simulate unseen physical systems with graph neural networks," in *NeurIPS 2021 AI for science workshop*.

Zhou, C., Chase, J. G., Knopp, J., Sun, Q., Tawhai, M., Möller, K., et al. (2021). Virtual patients for mechanical ventilation in the intensive care unit. *Comput. Methods Programs Biomed.* 199, 105912. doi:10.1016/j.cmpb.2020.105912

TYPE Original Research
PUBLISHED 03 October 2024
DOI 10.3389/fdgth.2024.1455477

Quantifying the impact of surgical teams on each stage of the operating room process

Adam Meyers[1]*, Mertcan Daysalilar[1], Arman Dagal[2,3], Michael Wang[3], Onur Kutlu[4] and Mehmet Akcin[1,4]

[1]Department of Industrial and Systems Engineering, University of Miami, Coral Gables, FL, United States, [2]Department of Anesthesiology, Perioperative Medicine, and Pain Management, Miller School of Medicine, University of Miami, Miami, FL, United States, [3]Department of Neurological Surgery, Miller School of Medicine, University of Miami, Miami, FL, United States, [4]DeWitt Daughtry Family Department of Surgery, Miller School of Medicine, University of Miami, Miami, FL, United States

Check for updates

OPEN ACCESS

EDITED BY
Nathan Gaw,
Air Force Institute of Technology,
United States

REVIEWED BY
Dessislava Pachamanova,
Babson College, United States
Alana Delaforce,
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia

*CORRESPONDENCE
Adam Meyers
✉ axm8336@miami.edu

RECEIVED 27 June 2024
ACCEPTED 18 September 2024
PUBLISHED 03 October 2024

CITATION
Meyers A, Daysalilar M, Dagal A, Wang M, Kutlu O and Akcin M (2024) Quantifying the impact of surgical teams on each stage of the operating room process.
Front. Digit. Health 6:1455477.
doi: 10.3389/fdgth.2024.1455477

Introduction: Operating room (OR) efficiency is a key factor in determining surgical healthcare costs. To enable targeted changes for improving OR efficiency, a comprehensive quantification of the underlying sources of variability contributing to OR efficiency is needed. Previous literature has focused on select stages of the OR process or on aggregate process times influencing efficiency. This study proposes to analyze the OR process in more fine-grained stages to better localize and quantify the impact of important factors.
Methods: Data spanning from 2019-2023 were obtained from a surgery center at a large academic hospital. Linear mixed models were developed to quantify the sources of variability in the OR process. The primary factors analyzed in this study included the primary surgeon, responsible anesthesia provider, primary circulating nurse, and procedure type. The OR process was segmented into eight stages that quantify eight process times, e.g., procedure duration and procedure start time delay. Model selection was performed to identify the key factors in each stage and to quantify variability.
Results: Procedure type accounted for the most variability in three process times and for 44.2% and 45.5% of variability, respectively, in procedure duration and OR time (defined as the total time the patient spent in the OR). Primary surgeon, however, accounted for the most variability in five of the eight process times and accounted for as much as 21.1% of variability. The primary circulating nurse was also found to be significant for all eight process times.
Discussion: The key findings of this study include the following. (1) It is crucial to segment the OR process into smaller, more homogeneous stages to more accurately assess the underlying sources of variability. (2) Variability in the aggregate quantity of OR time appears to mostly reflect the variability in procedure duration, which is a subinterval of OR time. (3) Primary surgeon has a larger effect on OR efficiency than previously reported in the literature and is an important factor throughout the entire OR process. (4) Primary circulating nurse is significant for all stages of the OR process, albeit their effect is small.

KEYWORDS

operating room, surgery, efficiency, case delay, duration, surgical team, human factors, linear mixed model

1 Introduction

Improving operating room (OR) efficiency is a key factor in controlling or reducing surgical healthcare costs (1), which are significant. Aggregate surgical healthcare expenditures comprised 29% of aggregate healthcare expenditures in the United States in 2005, as computed by Muñoz et al. (2). Moreover, aggregate surgical expenditures were forecasted to grow from 4.6% of US GDP in 2005 to 7.3% of US GDP in 2025 (2). In a more recent study by Childers and Maggard-Gibbons (3), the mean cost of ambulatory OR time across California hospitals in fiscal year 2014 was $36.14 per minute with a standard deviation of $19.53 per minute. Cerfolio et al. (4) report a significantly higher cost of $150 per minute of OR time in the main campus ORs at New York University Langone Health. Even with financial considerations aside, improving OR efficiency will likely improve patient safety, experience, and outcomes, decrease patient wait time, increase OR throughput, and improve surgical team and staff satisfaction (5, 6).

Improving OR efficiency is a multifaceted problem, and several metrics have been investigated by researchers.[1] A common approach to improving efficiency is to improve the utilization of the OR, that is, by minimizing both underutilization and overutilization (9, 10). Underutilization occurs when an OR lies unused due to cases being completed earlier than predicted, and overutilization occurs when an OR is used beyond its predicted or allotted time (5). Such inefficiencies are caused in large part by variability in OR time (11, 12), typically defined as the duration of time from when the patient is wheeled into the OR to the time the patient is wheeled out. Indeed, studies by Bokshan et al. (13) and Allen et al. (14) have shown OR time to be a significant driver of increased surgical costs. To reduce inefficiencies and associated costs, researchers have sought to identify the sources of variability in OR time. The primary conclusion in the literature is that procedure characteristics, namely, precise procedure type and type of anesthesia, are the main factors explaining the variation in OR time, followed by surgical team characteristics, primarily the surgeon (11, 12, 15, 16). Other factors such as patient characteristics (e.g., BMI) or other surgical team factors, such as the anesthesiologist, are generally found to be insignificant.

OR time, however, is an aggregate quantity that encompasses several stages of the OR process, and it does not span the entire OR process (Figure 1). As such, it has the following potential downsides. First, OR time does not include all stages of the OR process. In this study's dataset, which consists of timestamps taken from a surgery center located in a large academic hospital, OR time does not include room setup duration or room cleanup duration, nor any delays in starting the next case or beginning anesthesia induction. In addition, the dataset shows that anesthesia induction begins, on average, approximately two minutes before the patient is wheeled into the OR (i.e., two minutes before OR time begins). Thus, analyzing OR time alone will not allow for ascertaining the sources of variability in all stages of the OR process, and it may also contain some inaccuracies due to the starting and ending points of OR time not lining up with the activities in the OR process. Second, OR time itself covers several different stages of the OR process, including anesthesia induction, procedure duration, and delays in the procedure start time and in the time the patient is wheeled out after the procedure is completed. It is reasonable to hypothesize that the above four stages do not have the same sources of variability, or that shared sources of variability do not account for the same proportion of variability across all stages of the OR process. Therefore, this study's approach is to segment the OR process into more fine-grained, homogeneous stages and assess the sources of variability within each stage.

Past studies have focused on other parts of the OR process besides OR time and the more fine-grained stages that comprise OR time, including surgical procedure duration, anesthesia-related times, start time delays, and turnover time (refer to Section 2.1). However, an effort to quantify the sources of variability across all fine-grained stages of the OR process is currently lacking. Based on timing markers obtained from a surgery center located in a large academic hospital, this paper quantifies the sources of variability in several OR process stages, including first case start time delay, setup duration, anesthesia induction time, procedure start time delay, procedure duration, wheels out delay, cleanup duration, and OR time (refer to Section 3.1). The focus of this paper is on quantifying the extent to which type of procedure and members of the surgical team - primary surgeon, responsible anesthesia provider, and primary circulating nurse - and their interactions explain the variation in the fine-grained stages of the OR process. By better understanding the influence of various important factors, stakeholders and researchers can better pinpoint where interventions to improve efficiency should be targeted.

The rest of this paper is organized as follows. Section 2 provides a literature review on previous approaches to assess or improve efficiency within different stages of the OR process. Section 3 describes the dataset, process times, statistical approach, and model selection. Section 4 describe the results of the statistical analysis, primarily providing a decomposition of variability for each process time. Section 5 discusses the primary findings of this study and comments on this study's limitations and opportunities for future work. Section 6 provides concluding remarks. Additional tables and figures generated in this study are available in the Supplementary Material.

2 Research background

2.1 Related work in determining the factors driving the stages of the OR process

Numerous studies have investigated the various factors purported to cause or explain the variation in the OR process. Such work is motivated by the idea that, for OR efficiency to be improved, relevant stakeholders must first be informed about the

[1] For comprehensive reviews, refer to Lee et al. (7) and Dexter and Epstein (8).

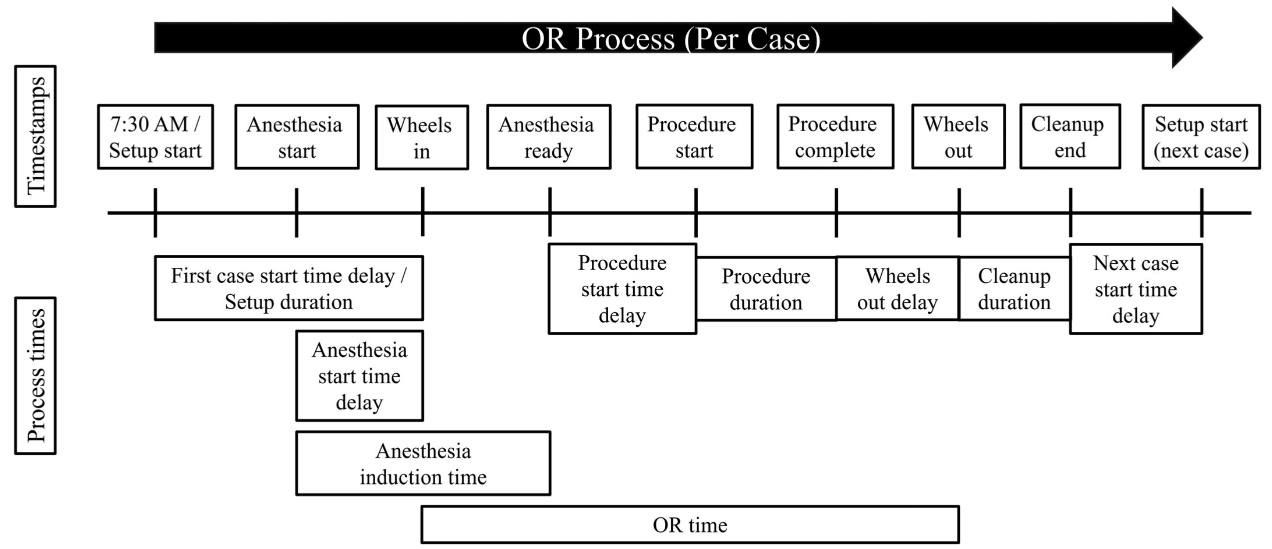


FIGURE 1
Visual depiction of the OR process, including timestamps and the span of time each OR process time covers. Formulas for each process time are given in Table 1.

primary factors driving OR inefficiencies. In addition, identifying the primary factors will allow for better predictive modeling, which in turn will allow for more accurate OR case scheduling to reduce OR underutilization and overutilization.

Variability in OR time (i.e., "wheels in" to "wheels out" time) is cited as a primary cause of inefficient OR utilization (11, 12). When a case lasts longer than planned, subsequent cases will either be delayed, potentially leading to OR overutilization, or cancelled, resulting in less OR revenue, patient dissatisfaction, and reduced quotas for surgical teams. When a case lasts shorter than expected, the OR will likely lie underutilized for some period of time, wasting resources. Exploring the factors that explain the variability in OR time, Dexter et al. (15) verified earlier findings, e.g., in Strum et al. (16), that reported the importance of three factors: precise procedure information, surgical team, and anesthetic type in predicting OR time. Eijkemans et al. (11) later identified additional factors, including the surgeon's estimate of total surgical time, operation characteristics (e.g., number of separate procedures), and team characteristics (e.g., number of surgeons). van Eijk et al. (12) found that type of procedure is the overwhelming predictor of OR time variability, with surgeon having a small but significant effect and anesthesiologist having a negligible effect. Many studies show that patient characteristics (e.g., body mass index) have little effect (11, 12).

Some studies have investigated the sources of variability in other parts of the OR process and in more fine-grained stages. The most commonly examined stage is the (surgical) procedure duration, which is typically the longest stage that comprises OR time. For instance, Strum et al. (16) found the surgeon to be the most important source of variability in procedure duration, followed by anesthesia type. Patoir et al. (17) found surgeon characteristics, center location, and surgical procedure and patient characteristics accounted for much of the variation in procedure duration. Additional factors were explored in the literature, such as surgeon factors (e.g, team composition factors, such as the presence of residents) (18), factors that increase the expected duration (e.g., communication failures) by Gillespie et al. (19), and operational (e.g., OR assignment) and temporal (e.g., whether a case was started after 5:00PM) factors by Kayis et al. (9). However, many of the studies focusing on procedure duration, e.g., Strum et al. (16), Stepaniak et al. (18), and Kayis et al. (9), perform statistical analyses separately for each surgical speciality or coarse-grained category rather than considering holistically how the specific procedure type, as indicated by a fine-grained category such as the American Medical Association's Current Procedure Terminology codes (refer to Section 3.2), accounts for the variation in procedure duration.

Other parts of the OR process explored in the literature are anesthesia-related times. For instance, Kougias et al. (20) found in their multivariate regression analysis that procedure type, anesthesia type, and BMI were statistically significant predictors of anesthesia induction time, while procedure type, anesthesia type, and operative case length were statistically significant predictors of anesthesia recovery time. van Veen-Berkx et al. (21) found that scheduling accuracy improved when looking at anesthesia-controlled time (ACT) as a proportion of total procedure time.[2] Few studies, however, have examined the

[2] ACT is defined in Dexter et al. (22) as the sum of the time from when the patient enters the OR to when the positioning or skin preparation begins, plus the time from when the surgical dressing is completed to when the patient is wheeled out of the OR.

impact of various human factors involved in the OR process on anesthesia-related times, including anesthesiologists.

Other fine-grained stages of the OR process that have been explored include start time delays, such as procedure start time delay, (any) case start time delay, and first case start time delay. Does et al. (23) employed Six Sigma techniques (24) to identify poor planning and scheduling as the primary factor causing delays in the start times of surgical procedures. The authors noted that surgical specialty and anesthesia technique also influence start time delays. A review by Halim et al. (25) identified several factors that can improve start time, including financial incentives for staff, education strategies, perioperative protocols and systems, surgical team communication, the “golden patient” initiative,[3] and the “productive operating theatre” scheme[4] A more specific approach is to look only at delays in the first case of the day, with the justification being to mitigate the cascading effect a delay in the first case has on subsequent cases in the OR. Cox Bauer et al. (27) analyzed data across three high-volume urban hospitals and found that, for cases with a documented reason for delay, the physician was the most reported reason for delay at 52%, followed in descending order by anesthesia, patient, staff, other sources, and facility. The authors did perform a regression analysis finding patient age, occurrence of late arrival, department, and facility to be significant predictors of delay. However, neither approach gives a quantification of the overall impact of a predictor on first case start time delay. Other similar work has looked at more specific events such as delays in the start of a subsequent case when the preceding case was performed by a different surgeon (28) and remaining time to exit the OR after surgical closure begins (29).

An additional stage of the OR process explored in the literature is turnover time, which is the duration of time from when a patient is wheeled out until the next patient is wheeled in. Thus, turnover time is all the remaining time in the OR process not covered by OR time (Figure 1). Bhatt et al. (30) took a systems-level approach to improve turnover time, which focused on developing a consistent “room ready” designation to reduce variability, implementing parallel processing to ensure room readiness and patient readiness occur simultaneously, and improving perioperative communication. Cerfolio et al. (4) piloted a Performance Improvement Team, called “PIT Crew,” that performed lean processing and value mapping to improve efficiency in the turnover time period. Goldhaber et al. (31) reduced turnover times significantly by collecting more granular data within the turnover time period and displaying these data to teams for regular review and accountability. The turnover time period was further divided into the followings segments: wheels out time → cleanup start time → cleanup complete time → setup start time → time room is ready for patient → wheels in time. Few studies, however, have taken the approach of quantifying the factors that explain variation in turnover time or the stages that comprise the turnover time period.

2.2 State-of-the-practice methodologies for determining important factors

There are several approaches in the related literature that seek to identify the important factors accounting for the variation in the OR process. Primary methods found in the literature include performing basic statistical analysis, fitting known probability distributions to OR process times, utilizing regression approaches for inference or prediction, utilizing systems-level approaches for improving process efficiency, and, more recently, training machine learning models for prediction.

Traditional statistical analysis, such as descriptive statistics and hypothesis testing methods, is a fundamental approach to gaining insights from gathered datasets. Such analysis dates back many decades but is still utilized today, particularly with healthcare data, as it provides insights and an overview of process efficiency. Dexter et al. (22) used two-group, one-sided t-tests to determine if eliminating ACT would allow for additional cases to be completed during a typical 8-h workday. Martin and Langell (32) used Cuzick’s test for trend to evaluate whether pre-OR timeouts and performance pay improved on-time starts, OR utilization, and OR costs. Simmons et al. (33) was interested in determining if fine-grained CPT codes, compared to coarser-grained surgical specialties, would improve accuracy in surgical scheduling. They utilized the I^2 statistic and Levine’s test to assess heterogeneity in the means and variances, respectively, of ACTs and surgical-controlled times (SCTs).[5] While traditional methods of statistical analysis can provide interpretable and meaningful summaries of data to answer questions of interest, such as determining whether differences in groups are significant following an intervention, further quantification capabilities are needed to assess the impact of factors on OR efficiency.

An early line of research involved finding distributions with a good fit to OR process time data. A main contributing paper in this approach is that of Strum et al. (34) in which the authors recommended using the lognormal distribution to model surgical procedure times. Stepaniak et al. (35) mostly corroborated the findings of Strum et al. (34), but Kayis et al. (9) found the lognormal distribution did not generally fit surgery duration well at the procedure level. Joustra et al. (36) more comprehensively fit a number of hazard models. However, as mentioned in Joustra et al. (36), such methods are less concerned with

[3] The “golden patient” initiative is a strategy where the first patient on the operating list is medically fit, thoroughly investigated, and has a clear surgical plan (25).

[4] The “productive operating theatre” scheme is a three-step intervention to increase OR efficiency (25, 26).

[5] Surgical-controlled time is defined as the duration of time from surgical incision to surgical closure.

identifying the factors contributing to OR efficiency and more concerned with prediction.

Regression models, on the other hand, do allow for evaluating sources of variability in OR process times. Strum et al. (16) employed main-effects ANOVA modeling with the logarithm of surgical time and total procedure time as separate responses and found primary surgeon and type of anesthesia to be important predictors of variability. Does et al. (23) and Stepaniak et al. (18) also utilized ANOVA models to assess the importance of select factors on start time delays and surgical procedure times. Regression modeling is similarly used to identify factors that influence OR process times. Linear regression is especially utilized for this purpose, such as in Silber et al. (37), Ying Li and Huang (38), Gillespie et al. (19), and van Veen-Berkx et al. (21). Linear regression models also have added functionalities over ANOVA models, such as regularization techniques to avoid overfitting or to perform variable selection, e.g., LASSO used in Wang et al. (39), and incorporating nonlinear terms such as in Wang et al. (40).

The literature above utilizing linear regression methods tends to treat all factors as fixed effects. However, in a fixed effects setting, when certain units, e.g., surgeons, have few observations, parameter estimates may have high sample-to-sample variability. Thus, the parameter estimates may vary substantially from dataset to dataset, implying that the model built on a given dataset may not be reliable (41). In addition, fixed effects models require dummy variables to be created for each unit (e.g., each surgeon), and a coefficient must be estimated for each unit. If a factor contains many units (this study's dataset contains over one hundred surgeons), then estimating a large number of coefficients reduces the model's degrees of freedom, diminishes the model's power, and increases the standard errors of the coefficient estimates (41). Furthermore, the present study is not concerned with estimating the effects of individual surgeons, anesthesiologists, etc., but rather the effect of these groups as a whole. For such reasons, previous papers in assessing the effects of different factors on OR process times have employed linear mixed model (LMM) approaches, which incorporate both fixed and random effects, with great success, e.g., Dexter and Ledolter (42), Eijkemans et al. (11), van Eijk et al. (12). This paper also takes an LMM approach for the above reasons.

More recently, machine learning (ML) has become a popular method for predicting quantities in the OR process. Master et al. (43) found that regression tree methods, such as gradient boosted regression trees, outperformed historical averaging, surgeon expert predictions, and other ML methods in the literature when predicting pediatric surgical durations. ML methods combined with surgeon predictions were also among the top-performing methods in Master et al. (43). Other research has used ML to improve predictions of OR process times (44–48). While ML methods may improve prediction, Wang and Dexter (49) notes that implementing ML software to increase prediction accuracy will not increase productivity unless accompanied by more allotted case time in a typical workday. More importantly, the objective of this paper is to quantify the impact that various human factors have on OR efficiency. LMMs allow for quantifying the proportion of variance explained in a response by each factor of interest. ML methods do have some options for determining similar values of impact, including variable importance in classification and regression trees (CART) (45) and Shapley additive explanations (SHAP) values (46). However, variable importance metrics may not correlate well with model variance explained by features (50), particularly when the model overfits the data on which it's trained, which is a common issue with CART methods (43). The variable importance values are also typically reported in a relative fashion (to other variables) and thus do not provide an absolute assessment of the impact a factor has on a response. SHAP values may provide a better alternative in these regards, however they are not as well-established as linear regression-based metrics and may have issues as feature importance metrics (51).

3 Materials and methods

3.1 Dataset and subjects

Data spanning from January 2, 2019 to June 30, 2023 were obtained from a surgery center in the University of Miami Hospital. The surgery center incorporates six operating rooms and a dedicated preoperative area and postoperative recovery unit. The dataset originally contained 12,375 cases, before data cleaning was performed (detailed below). The dataset included the following timestamps: setup start time, anesthesia start time, wheels in time (i.e., when the patient enters the OR), anesthesia ready time, procedure start time, procedure complete time, wheels out time (i.e., when the patient exits the OR), and cleanup end time.[6]

This study examined various critical stages of the OR process rather than focusing solely on one stage or on aggregate process times encompassing several stages. The OR process times explored in this study included first case start time delay, setup duration, anesthesia induction time, procedure start time delay, procedure duration, wheels out delay, and cleanup duration.[7] Each OR process time was defined as the elapsed time between two timestamps as described in Table 1. Figure 1 depicts the timestamps and process times.

There is a strong focus in previous literature on the aggregate quantity, OR time, defined as the elapsed time between when the patient is wheeled into the OR to when the patient is

[6] Further description of the dataset in terms of procedural categories and the range of CPT codes used are provided in Supplementary Table S1.

[7] Initially, the process times of anesthesia start time delay, computed as anesthesia start minus wheels in, and next case start time delay, computed as setup start (of next case) minus cleanup end (of previous case), were included in this study. However, after data cleaning, there were too little data to build LMMs with the desired factors; thus these process times were excluded from the analysis.

TABLE 1 Formulas for calculating each OR process time and number of cases after individual data cleaning for each process time.

Process time	Formula	Nbr. of cases
First case start time delay	Wheels in – 7:30/8:30 AM*	3,543
Setup duration	Wheels in – setup start	4,681
Anesthesia induction time	Anesthesia ready – anesthesia start	5,480
Procedure start time delay	Procedure start – anesthesia ready	11,357
Procedure duration	Procedure complete – procedure start	11,501
Wheels out delay	Wheels out – procedure complete	11,326
Cleanup duration	Cleanup end – wheels out	4,800
OR time	Wheels out - wheels in	11,467

*7:30 AM is the day's start time for Monday, Tuesday, Wednesday, and Friday, and 8:30 AM is the start time for Thursday.

wheeled out. It is hypothesized that the factors driving an aggregate quantity such as OR time, which covers various stages of the OR process (Figure 1), would not necessarily be identical across all stages comprising OR time, nor that shared sources of variability in OR time would explain the same proportion of variation in each stage comprising OR time. To evaluate these hypotheses, OR time was also included as a process time for comparison to the other seven process times.

This study's statistical analysis (refer to Section 3.2) involved building separate regression models using each OR process time as a univariate response for a total of eight models. The subset of cases containing errors corresponding to one process time were not necessarily the same as the subset of cases containing errors for a different process time. Then, because separate models were developed for each process time, the choice was made to clean the data separately for each process time, maximizing the amount of data available for each model. Data cleaning involved removing any cases with missing data, outliers, or errors. In addition, any process time labeled as a "delay" only included delay times that were positive. For instance, if the first case started on or before the day's start time, e.g., 7:30 AM, then this case was removed as there was no "delay" in the commencement of the first case. After removal of such cases for first case start time delay, the number of cases available for fitting the statistical model was 3,543 cases (Table 1). If instead the choice was made to remove the same subset of cases for all process times, then while each of the eight models would have a common pool of data, the data size would be significantly reduced and the results would not be as robust. The number of cases available after data cleaning for each process time is provided in Table 1.

All the OR process times exhibited right skewness. For instance, Figures 2(a,b) shows the original distributions of first case start time delay and procedure duration, where the right skewness is evident.[8] To address this, a common approach in the relevant literature is to use a logarithm transformation. Eijkemans et al. (11) and van Eijk et al. (12) used the log transformation on OR time, and Strum et al. (34) and Stepaniak et al. (35) showed that OR time and procedure duration follow lognormal distributions, implying that log-transforming these process times will approximately yield a normal distribution more appropriate for linear regression modeling methods. Does et al. (23) were concerned with reducing start time delays of procedures, and to address right skewness they opted for a more thorough Box-Cox transformation. However, the optimal choice for the parameter λ in the Box-Cox tranformation was found to be zero in Does et al. (23), which is simply the log transformation. This study investigated several transformations, including log, square root, Box-Cox, and more. For many of the process times, the log transformation was not "optimal" in the sense of producing a distribution that most closely fits a normal distribution relative to all other transformations, however it was near-optimal for all process times. Moreover, given that the previous literature concluded the log transformation is appropriate for several OR process times and that the log transformation has better interpretability (in contrast to, e.g., the Box-Cox transformation), the logarithm was used to transform all process times in this study.

3.2 Statistical analyses

The primary objective of this study was to quantify the extent to which the variability observed in each OR process time could be attributed to four key factors: type of procedure, primary surgeon, responsible anesthesia provider, and primary circulating nurse. These factors will henceforth be referred to as "procedure," "surgeon," "anesthesiologist," and "circulator," respectively. Such analyses can provide a more precise account and quantification of the impact each factor has on each fine-grained stage of the OR process.

To quantify sources of variability in the OR process times, a linear mixed model (LMM) approach was used. An LMM was built separately for each of the eight process times, so that the sources of variability for each stage of the OR process could be assessed and quantified. The primary factors of interest, i.e., procedure, surgeon, anesthesiologist, and circulator were treated as random effects. Table 2 shows the number of levels of each factor that occurs in each process time's corresponding dataset (after data cleaning).

The four primary factors were treated as random effects for multiple reasons. First, treating a factor as a random effect allows for estimating the factor's variance and proportion of variance explained in the response (i.e., process time). Second, Table 2 shows that each of the four primary factors has many levels, and treating each as a fixed effect would require estimating tens to hundreds of coefficients, reducing the degrees of freedom in the model. This study is also not concerned with, e.g., a particular surgeon's effect, but rather the impact of the group of surgeons as a whole. Third, the procedures, surgeons, anesthesiologists, and circulators included in the dataset do not necessarily encompass

[8] Supplementary Figures S1–S6 show the corresponding histograms for all other process times.

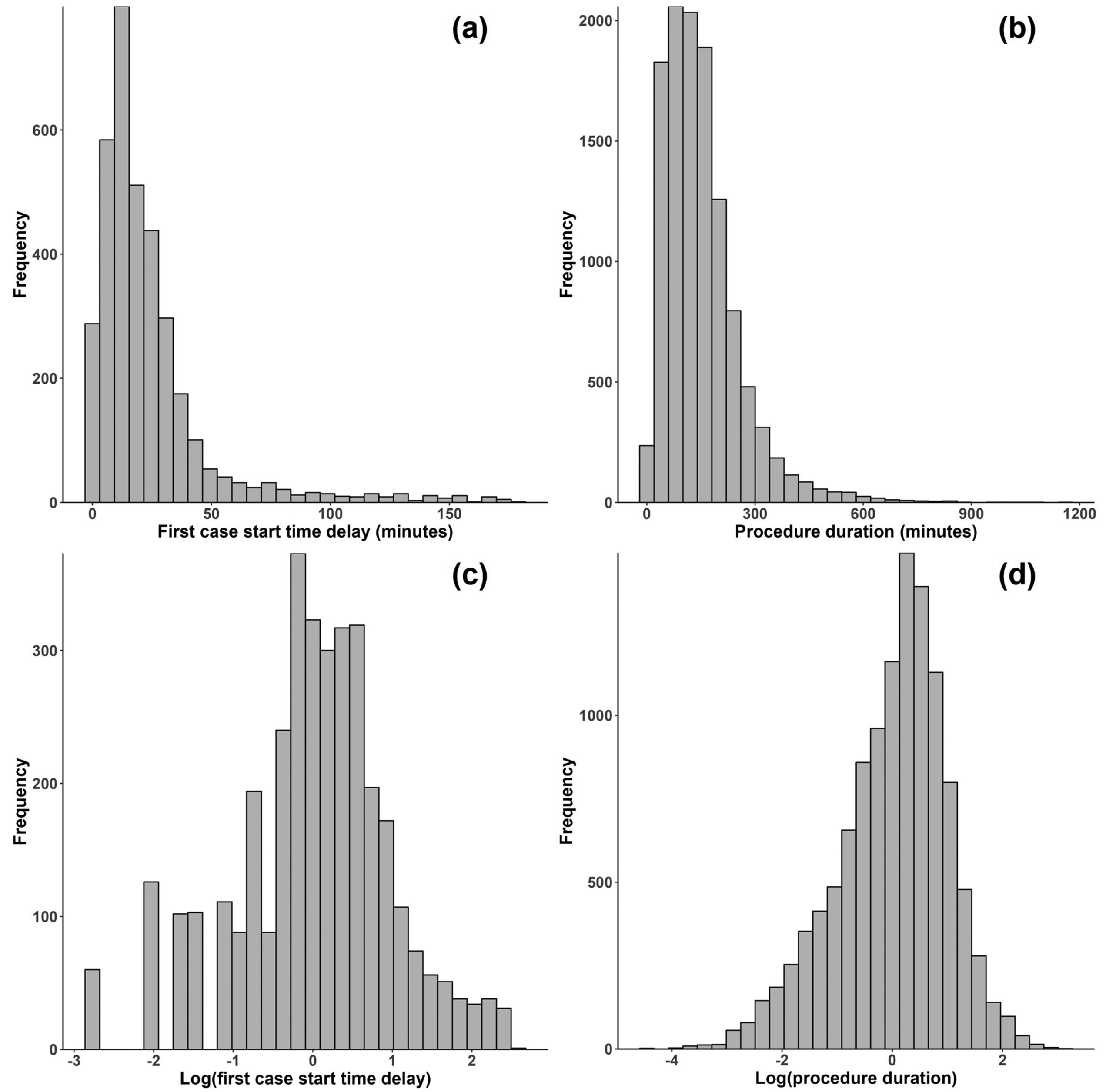


FIGURE 2
Histograms for (a) first cast start time delay and (b) procedure duration before transformation and (c) first case start time delay and (d) procedure duration after a log transformation.

TABLE 2 Number of levels of each random effect in the cleaned dataset for each OR process time.

OR process time	Random effect			
	Procedure	Surgeon	Anesthesiologist	Circulator
First case start time delay	439	106	78	112
Setup duration	568	106	76	114
Anesthesia induction time	633	118	77	117
Procedure start time delay	820	131	82	130
Procedure duration	827	132	82	131
Wheels out delay	818	131	82	130
Cleanup duration	519	107	79	121
OR time	823	132	82	131

the entire populations of these factors. Thus, treating the factors as random effects allowed for accomplishing this study's research objective and was an appropriate choice given the dataset. Note that only random intercepts were used in the LMMs.

Procedure was categorized based on the American Medical Association's Current Procedure Terminology (CPT) codes (52). Several past studies have identified the importance of categorizing procedures with high granularity, e.g., with CPT codes, rather than with low granularity, e.g., with surgical specialities such as neurosurgery, gynecology, etc. (33, 34, 38). In particular, a recent study by Simmons et al. (33) examined over 30,000 surgical cases in an academic hospital and found that both the mean and variance of ACT and SCT varied significantly between CPT codes within specialities. Their results suggest that the use of more granular categories, specifically CPT codes, will enhance the accuracy of subsequent analysis and scheduling. Accordingly, this study used the primary CPT code for each case as the procedure type.

Other factors were available in the University of Miami Hospital's database that could influence the process times. Domain expertise of this study's authors was used to select the factors believed to impact OR process efficiency. Six factors were included; they are shown in Table 3. "Position," for instance, was included as a proxy measure of the seniority and expertise of the primary surgeon. More experienced and senior surgeons were expected to be more efficient and consequently have a positive impact on OR efficiency. The six factors were treated as fixed effects for the following reasons. First, the factors were of less interest in this study and were expected to only marginally improve the model. The objective of this study was to quantify the sources of variability in the process times, focusing on procedure, surgeon, anesthesiologist, and circulator. Second, every factor had no more than five levels, with the exception of the number of procedures, which had thirty-two possible levels.[9] Third, the levels of the factors were exhaustive of the population, whereas the levels of the random effects were only a subset of their respective populations.

As stated previously, LMMs were separately built for each of the eight process times.[10] Before model selection was performed, a univariate analysis of each random effect was conducted to quantify the improvement in each model by the addition of a single random effect. Two base models were used - one consisting of a fixed intercept and the other a fixed intercept plus all six fixed effects. To each base model, a single random effect was added and the adjusted intraclass correlation coefficient (ICC) was calculated for each random effect, given by

$$\mathrm{ICC}_{(\mathrm{adj})} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}, \tag{1}$$

TABLE 3 Description of fixed effects used in the LMMs.

Factor	Levels	Description
Number of procedures	0, 1,…, 31	Number of procedures performed in a surgical case.
Number of panels	1, 2,…, 5	Number of panels in a surgical case, where a "panel" is defined as a grouping of surgical procedures performed together.
Procedure level	None, I, II, III, IV	Indicates surgical complexity of case.
Cancer/ noncancer	Cancer, noncancer	Indicates whether procedures were cancer-related or not.
Position	Assistant, associate, professor	Position of primary surgeon in academic hospital.
Patient class	Emergency, hospital ambulatory surgery, inpatient, surgery admit	Admission status of patient.

where σ_α^2 refers to the variance of the random effect. $\mathrm{ICC}_{(\mathrm{adj})}$ may be interpreted as the proportion of variance explained in the logarithm of the process time by the random effect, after controlling for the fixed effects.

After univariate analysis, multivariate analysis was performed to assess the impact of each random effect (in the presence of other significant random effects) on the process time and to control for fixed effects. Model selection proceeded as follows.[11]

First, a base model was developed, given by

$$\begin{aligned} y_i &= \beta_0 + \boldsymbol{X}_i\boldsymbol{\beta} + \alpha_{j[i]} + \epsilon_i, \\ \alpha_j &\sim N(0, \sigma_\alpha^2), \\ \epsilon_i &\sim N(0, \sigma_\epsilon^2), \end{aligned} \tag{2}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, J$ are the indices of the observations and procedure levels, respectively, y_i represents the ith observation of the logarithm of the respective process time, β_0 is the fixed intercept, $\boldsymbol{\beta}$ is the vector of fixed slopes, $\boldsymbol{X}_i$ is the vector of the ith observations of all variables associated with the fixed effects (Table 3),[12] $\alpha_{j[i]}$ is a random intercept for procedure, $j[i]$ denotes to which procedure the ith observation belongs, ϵ_i is the error term, and σ_α^2 and σ_ϵ^2 are the variances of the random effect and error, respectively.

Second, note that the base model in Equation 2 only includes a random intercept for procedure. Procedure was previously found in multiple studies to be the primary source of variability in various OR process times (11, 12, 16, 35). Thus, with procedure ostensibly explaining much of the variation in the process times, it was reasonable to begin the base model with only procedure as a random intercept. Each additional random effect was subsequently and cumulatively added to determine if the

[9] However, most cases involved only a few procedures.

[10] All LMMs were fitted using the lmer function from the R package lme4 (53). A reference on this package is provided by Bates et al. (53).

[11] A similar model selection procedure to that of van Eijk et al. (12) was used in this study.

[12] Because all fixed effects were categorical with many levels, several dummy variables were created which would be included in the vector $\boldsymbol{X}_i$.

additional random effect should be retained in the final model. A chi-squared test was used to determine the significance[13] of a model with one additional random effect compared to the (previous) model without the random effect. Akaike information criterion (AIC) was also reported as it penalizes the addition of more terms to the LMM. However, the chi-squared test was solely used for determining which random effects to keep, since AIC is more appropriate for prediction which is not the objective of this study.

Third, fixed effects were individually examined to determine whether each should be retained in the final model for each process time. For a given process time, a new base model was formed by adding all significant random effects found in step two above to Equation 2. Each of the six fixed effects were individually removed from the new base model, while retaining all other fixed effects, and chi-squared tests were performed and AIC values were computed. If the new base model (containing all fixed effects and significant random effects from step two above) was found significant over the new base model without the individual fixed effect (according to the chi-squared test), then the fixed effect was retained for the final model.

Fourth, the final model for a given process time was formed by adding all significant random effects from step two above and removing all fixed effects according to the procedure described in step three above. To assess the impact of each random effect retained in the final model, $\text{ICC}_{(\text{adj})}$ in Equation 1 was calculated for each random effect. In addition, model ICC values were calculated to give the overall proportion of variance explained in the logarithm of the process time by all random effects. Both unadjusted and adjusted model ICC values were reported. The unadjusted model ICC, denoted ICC_{LMM}, and adjusted model ICC, denoted $\text{ICC}_{\text{LMM(adj)}}$, are given by

$$\text{ICC}_{\text{LMM}} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_f^2 + \sigma_\epsilon^2} \quad \text{and} \quad \text{ICC}_{\text{LMM(adj)}} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_\epsilon^2}, \tag{3}$$

where σ_r^2 and σ_f^2 are the variances explained by all random and fixed effects, respectively.

4 Results

Figure 3 shows a summary of the data, after cleaning, for each (untransformed) process time and random effect. To calculate the values of a given box plot, the times of the corresponding process time (e.g., procedure duration) were grouped according to the levels of the corresponding random effect (e.g., surgeon) and the median was taken for each level. For all process times, the random effect of procedure displays the highest dispersion through larger interquartile ranges and wider outliers (dots). As seen in Table 2, procedure also has significantly more levels than any other random effect. However, this alone does not explain the higher dispersion observed in procedure. Rather, it is likely that procedure is an important factor, which also agrees with much of the literature concluding that procedure is the primary source of variability in various OR process times (11, 12, 16, 35). Figure 3 also shows that all process times are right-skewed, as indicated by the median line being closer to the first quartile (bottom of box) and the upper tails and outliers extending far upwards, particularly in the box plots corresponding to the random effect of procedure. Some process times only show mild skewness, such as cleanup duration.[14]

Table 4 and Supplementary Tables S2–S8 provide an assessment of the individual impact of each random effect, with and without fixed effects. In Table 4 and Supplementary Tables S2–S8, $\text{ICC}_{(\text{adj})}$ shows a small reduction when fixed effects are included; thus, the fixed effects included in this study do not explain much variation in the logarithms of the process times. Of all main random effects, procedure shows the largest $\text{ICC}_{(\text{adj})}$ for most of the process times. This observation is supported by Figure 3 in that the box plots associated with procedure show the largest variation. Exceptions include first case start time delay (Supplementary Table S2) and cleanup duration (Supplementary Table S7) in which surgeon shows the largest $\text{ICC}_{(\text{adj})}$.[15] In other cases, surgeon is not far behind procedure in terms of $\text{ICC}_{(\text{adj})}$, including setup duration (Supplementary Table S3), anesthesia induction time (Supplementary Table S4), procedure start time delay (Supplementary Table S5), and wheels out delay (Supplementary Table S6). Procedure duration (Table 4) and OR time (Supplementary Table S8) are the process times where procedure explains moderately more variation than surgeon.[16] While the fact that procedure and surgeon accounting for the most variability could in part be due to both factors having many levels, circulator also shows approximately the same number of levels as surgeon (Table 2), yet it typically accounted for very little of the variability. One final observation from Table 4 and Supplementary Tables S2–S8 is that the interaction terms typically have higher $\text{ICC}_{(\text{adj})}$ than their main effect counterparts, but the gain is marginal.

Model selection was performed for each process time as described in Section 3.2. The model selection process is

[13] The standard 0.05 level of significance was used.

[14] Skewness of the process times is further supported by the histograms displayed in Figure 2 and Supplementary Figures S1–S6.

[15] For cleanup duration (Supplementary Table S7), surgeon is only higher than procedure in $\text{ICC}_{(\text{adj})}$ when fixed effects are included.

[16] As concluded in this paper, though, the variability of OR time largely reflects that of procedure duration; thus, procedure type has a significantly higher impact on procedure duration than the primary surgeon, but this is not the case for the other stages of the OR process as examined in this study.

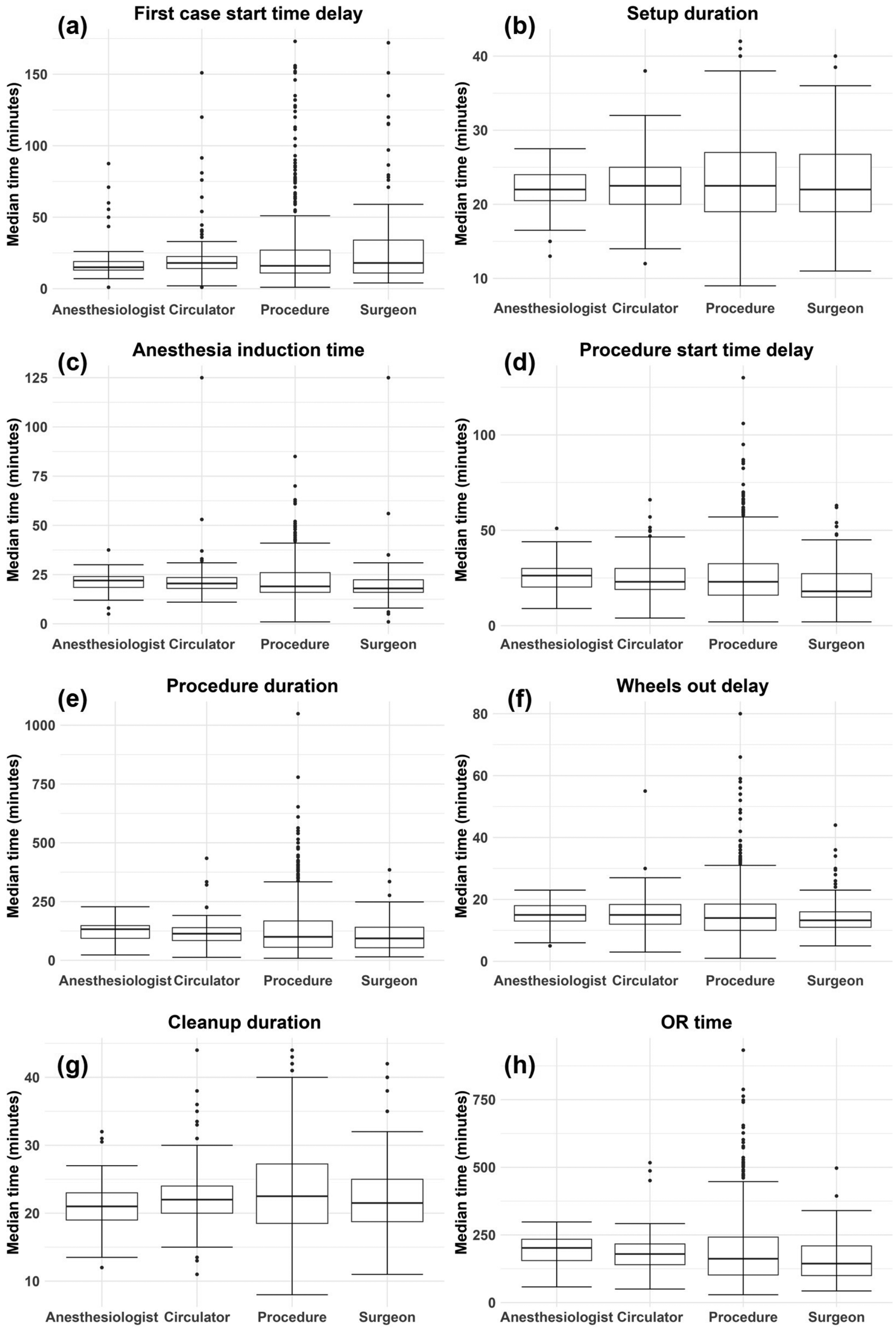


FIGURE 3
Box plots by random effect for each (untransformed) process time: (A) first case start time delay, (B) setup duration, (C) anesthesia induction time, (D) procedure start time delay, (E) procedure duration, (F) wheels out delay, (G) cleanup duration, (H) OR time. The values used to generate each box plot were the median times for each level of a given random effect and for each process time.

TABLE 4 Univariate assessment of random effects when using procedure duration as the response.

Random effect	$\text{ICC}_{(\text{adj})}$ (%), without FE	$\text{ICC}_{(\text{adj})}$ (%), with FE
Procedure	66.9	59.8
Surgeon	39.7	34.9
Procedure × Surgeon	72.0	64.3
Anesthesiologist	6.1	1.6
Procedure × Anesthesiologist	69.8	58.2
Surgeon × Anesthesiologist	46.9	35.3
Circulator	12.3	6.6
Procedure × Circulator	69.0	59.1
Surgeon × Circulator	44.3	35.7
Anesthesiologist × Circulator	22.7	12.3

Each row corresponds to an LMM with a fixed intercept and the single random effect specified in column 1. Columns 2 and 3 give $\text{ICC}_{(\text{adj})}$ values (Equation 1) for each univariate random effect model, both excluding (column 2) and including (column 3) all fixed effects in the LMM. ICC, intraclass correlation coefficient; FE, fixed effects.

TABLE 5 Model selection for choosing random effects in the LMM where procedure duration is the response.

Model	AIC	AIC gain	p-value
Base model	22516.8	–	–
+ Surgeon	21955.9	560.9	<0.001
+ Procedure × Surgeon	21750.2	766.6	<0.001
+ Anesthesiologist	21685.7	831.0	<0.001
+ Procedure × Anesthesiologist	21681.0	835.7	0.010
+ Surgeon × Anesthesiologist	21669.3	847.5	<0.001
+ Circulator	21628.8	888.0	<0.001
+ Procedure × Circulator	21597.9	918.9	<0.001
+ Surgeon × Circulator	21584.8	932.0	<0.001
+ Anesthesiologist × Circulator	21578.2	938.6	0.003

The base model is given in Equation 2 and consists of a fixed intercept, all six fixed effects, and procedure as a random intercept. Additions appearing in this table are cumulative in the sense that each subsequent random effect was added to the model in the preceding row. AIC gain is the improvement in AIC from adding additional random effects onto the base model (calculated as AIC of the base model minus AIC of the larger model). AIC, Akaike information criterion.

TABLE 6 Number of final models in which each random and fixed effect appeared.

Random effect	Frequency	Fixed effect	Frequency
Procedure	8	Number of procedures	6
Surgeon	8	Number of panels	1
Anesthesiologist	7	Procedure level	4
Circulator	8	Cancer/noncancer	1
Procedure × Surgeon	5	Position	1
Procedure × Anesthesiologist	5	Patient class	8
Procedure × Circulator	8		
Surgeon × Anesthesiologist	4		
Surgeon × Circulator	4		
Anesthesiologist × Circulator	2		

TABLE 7 Model selection for choosing fixed effects in the LMM where procedure duration is the response.

Model	AIC	AIC loss	p-value
Base model	21578.2	–	–
BM − Number of procedures	22772.2	1194.1	<0.001
BM − Number of panels	21577.0	−1.2	0.362
BM − Procedure level	21665.5	87.4	<0.001
BM − Cancer/noncancer	21576.8	−1.3	0.413
BM − Position	21574.3	−3.9	0.928
BM − Patient class	21696.0	117.8	<0.001

The base model consists of a fixed intercept, all six fixed effects, and the random effects found to be significant from Table 5. Each fixed effect was removed from the base model, and each reduced model was compared to the base model via a chi-squared test. If the base model was found significant compared to the reduced model, then the corresponding fixed effect was retained. Subtractions appearing in this table are not cumulative and denote that only the indicated fixed effect was removed from the base model and all other fixed effects were included. AIC loss is the increase in AIC from removing a fixed effect from the base model (calculated as AIC of the reduced model minus AIC of the base model). AIC, Akaike information criterion; BM, base model.

illustrated with procedure duration (Table 5).[17] The p-values were determined from performing chi-squared tests between each model and its previous (nested) model in Table 5, and they were used to determine whether to retain a particular random effect for the final LMM. In the case of procedure duration (and OR time; Supplementary Table S15), all main effects and interactions were determined to be significant and were retained for the final model. The gain in AIC exhibited by every random effect in addition to procedure indicates that including each term will likely improve prediction. The number of final models in which each random effect appeared is shown in Table 6. Procedure was by default included in every model, but surgeon, circulator, and the interaction of procedure and circulator also appeared in every model. Anesthesiologist appeared in all models except for that of first case start time delay.

A new model for each process time was formed by augmenting the base model (Equation 2) with the significant random effects shown in Table 5 and Supplementary Tables S9–S15. Then the individual impact of each fixed effect on the performance of the augmented LMM was assessed (refer to Section 3.2). Table 7 shows the performance associated with the fixed effects for procedure duration.[18] Based on the p-values, the fixed effects retained for the final model for procedure duration were the number of procedures, procedure level, and patient class. Table 6 shows the number of final models for which each fixed effect

[17] The corresponding tables for all other process times are included in the Supplementary Tables S9–S15.

[18] Supplementary Tables S16–S22 show the performance associated with the fixed effects for all other process times.

was retained. Patient class was found significant for every process time and the number of procedures was found significant for all process times except first case start time delay (Supplementary Table S16) and cleanup duration (Supplementary Table S21).

$ICC_{(adj)}$ (Equation 1) was calculated for each random effect appearing in each final model, and model ICC values, ICC_{LMM} and $ICC_{LMM(adj)}$ (Equation 3), were also calculated for each process time (Table 8). From Table 8, it is observed that surgeon is the random effect with the highest $ICC_{(adj)}$ value for five of the process times, including first case start time delay, setup duration, procedure start time delay, wheels out delay, and cleanup duration. However, the highest $ICC_{(adj)}$ value surgeon obtains is 21.1% for procedure start time delay. Procedure has the highest $ICC_{(adj)}$ value for all other process times, including anesthesia induction time, procedure duration, and OR time. Procedure accounted for 44.2% and 45.5% of variability in the logarithm of procedure duration and OR time, respectively. For all other process times, procedure accounted for approximately 11% of variation or less. While anesthesiologist was found significant for seven process times, it accounted for at most 1.1% of variation (wheels out delay). Interestingly, circulator was found significant for all models and accounted for as much as 3.4% of variation (wheels out delay). However, both anesthesiologist and circulator do not individually account for much variation.

Table 8 also shows several significant interaction terms. In particular, the interaction of procedure and circulator was significant for all models. In many cases, this interaction term accounted for more variation than circulator individually. This suggests that the effect of the primary circulating nurse is significant but their effect can depend on the type of procedure. In addition, the interaction of procedure and surgeon was significant for five models and accounted for 2.2%–8.7% of variation. This also suggests the effect of the surgeon depends on the procedure. Lastly, the interaction of surgeon and circulator accounted for a modest amount of variance in the logarithms of setup duration (6.5%) and cleanup duration (8.6%), suggesting a synergistic effect of surgical teams in some stages of the OR process.

Overall, Table 8 shows that the primary factors examined in this study - procedure type, primary surgeon, responsible anesthesia provider, and primary circulating nurse - are most impactful on procedure duration and OR time, accounting for 67.5% and 69.7% of variation (in the logarithms), respectively, after fixed effects have been accounted for. The primary factors also explained a moderate amount of variation in the logarithm of procedure start time delay (43.5%), and were mildly impactful on setup duration (32.1%), anesthesia induction time (22.3%), wheels out delay (26.5%), and cleanup duration (28.7%). The primary factors accounted for very little of the variation in the logarithm of first case start time delay (11.6%). Finally, it is noted that there were little differences between $ICC_{LMM(adj)}$ and ICC_{LMM}, further reinforcing that the fixed effects included in this study had little impact on the process times.

5 Discussion

5.1 Primary findings

The present study made several findings that both complement and add to the existing literature on OR efficiency. First, this study shows that, when investigating the impact of factors on the OR process, a fine-grained approach is necessary to pinpoint where in the process, and by how much, each factor makes an impact. In Section 1, it was hypothesized that the fine-grained stages of the OR process do not consist of the same sources of variability, nor that the common sources of variability account for the same proportion of variance in each stage. The results of this study support the above hypotheses (Table 8). Notably, OR time is an aggregate quantity consisting of the stages of the OR process in which the patient is present in the OR (i.e., "wheels in" to "wheels out"). However, the results of this study indicate that the quantification of variability in OR time mainly reflects the

TABLE 8 $ICC_{(adj)}$ values for each random effect (Equation 1) appearing in each final model.

		FCSTD	SD	AIT	PSTD	PD	WOD	CD	ORT
$ICC_{(adj)}$ (%)	Procedure	1.0	9.9	**11.4**	10.7	**44.2**	5.8	1.3	**45.5**
	Surgeon	**7.1**	**12.2**	7.3	**21.1**	10.8	**10.9**	**11.3**	13.3
	Anesthesiologist	-	0.8	0.8	0.4	0.2	1.1	0.5	0.3
	Circulator	0.2	2.3	0.7	0.7	0.5	3.4	1.5	0.4
	Proc. × Surg.	–	–	–	3.5	8.7	2.2	2.5	7.2
	Proc. × Anes.	–	–	–	0.9	0.3	1.1	2.7	0.6
	Proc. × Circ.	3.4	0.4	2.1	5.6	1.3	1.4	0.4	0.6
	Surg. × Anes.	–	–	–	0.7	0.3	0.7	-	0.7
	Surg. × Circ.	–	6.5	–	–	0.9	–	8.6	0.9
	Anes. × Circ.	–	–	–	–	0.3	–	–	0.2
$ICC_{LMM(adj)}$ (%)		11.6	32.1	22.3	43.5	67.5	26.5	28.7	69.7
ICC_{LMM} (%)		11.3	29.8	21.1	42.2	58.2	25.0	28.2	59.0

A dash (–) indicates the random effect was not selected for the final model. The random effects with the largest ICC for each process time are indicated in bold. Also provided are the model ICCs, namely ICC_{LMM} and $ICC_{LMM(adj)}$ (Equation 3). FCSTD, First case start time delay; SD, Setup duration; AIT, Anesthesia induction time; PSTD, Procedure start time delay; PD, Procedure duration; WOD, Wheels out delay; CD, Cleanup duration; ORT, OR time.

quantification of variability in procedure duration. Comparing the two process times in Table 8, their variabilities roughly decompose in the same way. For example, procedure accounted for 45.5% and 44.2% and surgeon for 13.3% and 10.8% of variability in the logarithms of OR time and procedure duration, respectively. Moreover, the random effects overall accounted for 69.7% and 67.5% of variability in the logarithms of OR time and procedure duration, respectively. In addition to procedure duration, OR time also comprises the time intervals associated with (a large proportion of) anesthesia induction time, procedure start time delay, and wheels out delay. However, the decompositions of variability for the latter three process times bear little resemblance to that of OR time. Thus, what happens in the OR during the procedure is mostly what is driving the aggregate quantity of OR time. As a result, interventions for improving efficiency in OR time should be focused on the procedure stage.

The second primary finding regards the impacts of each human factor. In particular, the primary surgeon had a larger impact in this study than what was previously reported in the literature. For instance, van Eijk et al. (12) found that the primary surgeon and second surgeon[19] only accounted for a combined 4.8% of the variability in the logarithm of OR time. In the present study, however, primary surgeon alone accounted for 13.3% of variability in the logarithm of OR time (Table 8). Surgeon also accounted for a substantial 21.1% of variability in the logarithm of procedure start time delay and for at least 7% in the logarithms of all other process times (Table 8). The above results suggest that the primary surgeon (and other surgeons in the team) have moderate impacts not only on procedure duration, but also on many stages of the OR process. The importance of the surgeon was stressed in previous literature, e.g., Strum et al. (16), however a quantification of the variability due to surgeon was usually not provided. Moreover, the impact of the surgeon depends in part on the procedure, as seen by the significant interaction term of procedure and surgeon (Table 8). Indeed, Strum et al. (16) found that variability in surgical time increased as procedure time increased, indicating an interaction effect between type of procedure and surgeon.

In agreement with previous literature, responsible anesthesia provider was often a significant factor but not impactful on OR efficiency (12, 16). Surprisingly, responsible anesthesia provider had little impact on the anesthesia-controlled times, including anesthesia induction time and wheels out delay, the latter of which includes the patient's emergence from anesthesia. Other factors not included in this study, such as patient and operation characteristics, may be important for accounting for variability in anesthesia-controlled times (54).

Lastly, this study found the primary circulating nurse, a less studied human factor in the literature regarding OR efficiency, to be a significant factor in all stages of the OR process. This is reasonable because the circulating nurse, sometimes called the "perioperative" nurse, is involved before the surgery (e.g., transporting the patient and preparing the patient for surgery), during the surgery (e.g., assisting with equipment), and after the surgery (e.g., monitoring the patient) (55). In this study, the circulating nurse had their largest effect on wheels out delay and setup duration, accounting for 3.4% and 2.3% of variability (Table 8), respectively. In addition, the interaction of procedure and circulator was also significant for every process time, and the interaction of surgeon and circulator was significant for four process times and reached an $\mathrm{ICC}_{(\mathrm{adj})}$ value as high as 8.6% (cleanup duration; Table 8). Thus, the effect of the circulating nurse depends on the procedure type and, for some stages of the OR process, also on the particular attending surgeon, indicating some team synergistic effect on OR efficiency. Indeed, studies have found that nursing staff characteristics and team effects are important components of OR efficiency (56–58). More work is needed though to investigate the role of nursing staff on OR efficiency and to design interventions with nursing staff as a central component.

5.2 Clinical implications

The primary findings of this paper have the following clinical implications. First, OR process prediction models may be improved by incorporating significant factors and interactions found in this study. Improving prediction models will improve scheduling accuracy and increase OR utilization (i.e., decrease under- and over-utilization) which directly impacts OR efficiency.[20] This paper helps to fill a gap by quantifying the effect of key members of the surgical team and procedure type on various stages of the OR process. For instance, it may be beneficial for models predicting procedure duration to not only include the procedure type and primary surgeon, but also consider their interaction (Table 8). There is likely less variability in procedure duration among surgeons for simple procedures than for more complex procedures. Thus, prediction models should take into account that a surgeon's variability itself will vary depending on the type of procedure performed. In addition, this study uniquely identifies the primary circulating nurse and various interaction terms as significant; therefore, researchers can more comprehensively consider the members of surgical teams and their synergistic effects when designing prediction models.

Second, case scheduling may also be improved by incorporating significant factors and interactions found in this study. The effect of a particular individual, e.g., the attending primary surgeon, can be considered when allocating portions of time to each stage for a

[19] "Second surgeon" is defined in van Eijk et al. (12) as the first registered assistant surgeon during a procedure.

[20] See Dexter and Epstein (8) where "OR efficiency" is defined as minimizing the "inefficiency of use of OR time," the latter of which is calculated using costs and times associated with under- and over-utilization.

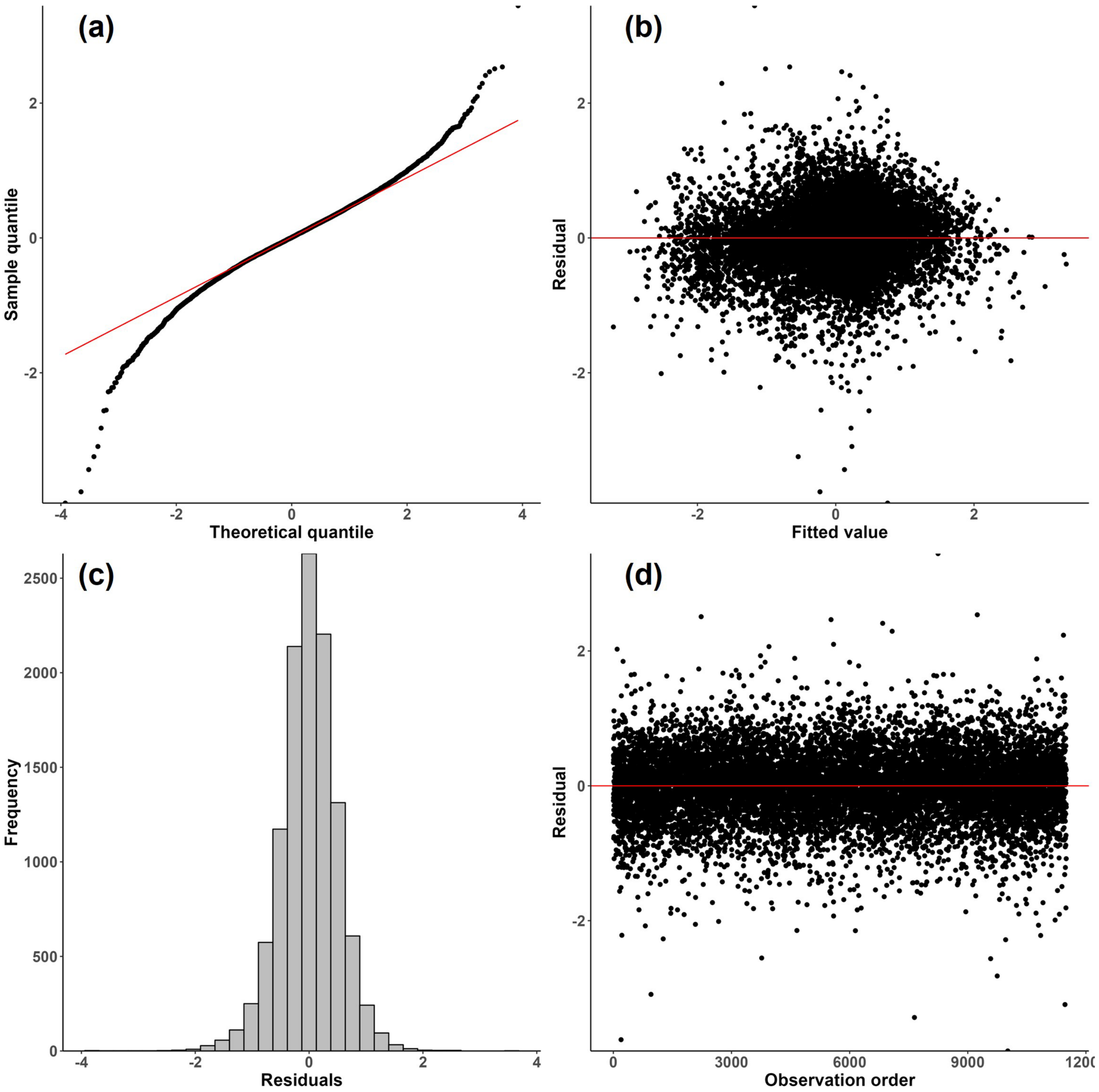


FIGURE 4
Diagnostic plots for the final LMM where procedure duration is the response. **(a)** Normal probability plot of residuals; **(b)** residuals vs. fitted values; **(c)** histogram of residuals; **(d)** residuals vs. observation order.

given case. The particular individual can be used in a more advanced prediction model as mentioned above or, more simply, the individual's historical data can be considered when allocating times. The same process can be done regarding surgical teams or combinations of surgical team members. Also, knowledge of particular surgical team members and teams themselves can inform strategies for case scheduling. For instance, if a particular surgeon or surgical team is known to have higher variability or expected completion times for a particular case, then such a case could be scheduled earlier or first in the day to allow for dynamic scheduling after the case's completion, which could allow for the completion of more cases in a day (16).

Third, OR efficiency can be improved by minimizing the variability in the stages of the OR process attributable to members of the surgical team and combinations of team members. The present study highlights areas of higher variability for surgical team members. Efforts could be made to reduce variability by identifying inefficiencies in a surgical team's or team member's practice and providing relevant training. If the area of improvement is in teamwork, for instance, training could seek to promote effective, assertive, and closed-loop communication among surgical teams to help minimize team performance variability.[21] Moreover, surgical teams can be further streamlined to match surgeons, anesthesiologists, and nurses who consistently work well together, which will in turn reduce performance variability.

5.3 Study limitations and future work

A limitation of this work was the use of a linear modeling approach and transformations to conform to model assumptions. Figure 4 shows model diagnostic plots for procedure duration.[22] Figure 4(a) shows some departure from a normal distribution for the logarithm of procedure duration; the distribution shows heavier tails as indicated by the upper right and lower left portions of the curve "peeling away" from the red line. Heavier tails indicate the presence of outliers in both directions. In addition, Figure 4(b) suggests mild heteroscedasticity in the residuals as the variation appears to decrease as fitted values increase in absolute value. Finally, the distribution of procedure duration exhibited right skewness, which was corrected by a log transformation. The above three observations were true for many of the process times. While the results provided in this paper are still relatively robust due to the large sample size of the dataset, more accurate results could possibly be obtained through the use of robust regression methods suited to handle outliers and heteroscedasticity. Moreover, generalized linear mixed models could be explored to handle the non-normality of the process times (59).

Another limitation of this work was the lack of inclusion of many potentially important fixed-effect variables. Previous literature has explored a wide range of factors that may contribute to OR efficiency (refer to Section 2). There are likely important factors missing from this analysis as they were not available in the database at the University of Miami hospital. Future work could explore a more comprehensive list of factors to maximize the potential of data to reveal OR inefficiencies. Moreover, even with a more comprehensive and retrospective assessment of influential factors, more proactive measures are needed that implement realistic interventions, in collaboration with members of surgical teams, to bring greater efficiency to the OR suite.

6 Conclusions

The primary goal of this paper was to quantify the extent to which the procedure type and key members of the surgical team accounted for variation in the fine-grained stages of the OR process. Some of the stages of the OR process and more aggregate process times have been analyzed previously in the literature (refer to Section 2.1). However, a comprehensive analysis of the impact of the primary surgical team members on the many stages comprising the OR process is lacking. This study helps to fill this gap by developing eight different linear mixed models that quantify the variability of several OR process times with respect to procedure type, primary surgeon, responsible anesthesiology provider, and primary circulating nurse.

This study found that, to more accurately account for sources of variability in the OR process, it is necessary to break up the OR process into smaller, homogeneous stages. For instance, this study found that OR time, defined as the "wheels in" to "wheels out" time of a patient in the OR, largely reflects procedure duration and is therefore not homogeneous across its entire time span. In addition, this study found that surgeon has a larger impact than previously reported in the literature and that the circulating nurse accounted for a significant, albeit small, proportion of variability in all eight process times studied. This study can serve as a foundation for quantifying the impact of important members of the surgical team on various stages of the OR process and for more targeted interventions seeking to realize more efficient and cost-effective OR suites.

Data availability statement

The datasets presented in this article are not readily available because of concerns regarding data privacy. Requests to access the datasets should be directed to Adam Meyers, axm8336@miami.edu.

Author contributions

AM: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Resources, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing. MD: Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. AD: Conceptualization, Project administration, Resources, Supervision, Writing – review & editing. MW: Conceptualization, Resources, Supervision, Writing – review & editing. OK: Resources, Supervision, Writing – review & editing. MA: Conceptualization, Data curation, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to thank the DeWitt Daughtry Family Department of Surgery and the University of Miami Hospital for generously providing the data used in this study.

[21] Granted, targeted research is needed to identify areas of inefficiency.

[22] Diagnostic plots for all other process times are provided in Supplementary Figures S7–S13.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2024.1455477/full#supplementary-material

References

1. Scott EJ. Editorial commentary: improved operating room efficiency is the best way to control orthopaedic costs. *Arthroscopy*. (2024) 40:1527–8. doi: 10.1016/j.arthro.2024.01.005

2. Muñoz E, Muñoz III W, Wise L. National and surgical health care expenditures, 2005–2025. *Ann Surg*. (2010) 251:195–200. doi: 10.1097/SLA.0b013e3181cbcc9a

3. Childers CP, Maggard-Gibbons M. Understanding costs of care in the operating room. *JAMA Surg*. (2018) 153:e176233. doi: 10.1001/jamasurg.2017.6233

4. Cerfolio RJ, Ferrari-Light D, Ren-Fielding C, Fielding G, Perry N, Rabinovich A, et al. Improving operating room turnover time in a New York city academic hospital via lean. *Ann Thorac Surg*. (2019) 107:1011–6. doi: 10.1016/j.athoracsur.2018.11.071

5. Dexter F, Traub RD, Qian F. Comparison of statistical methods to predict the time to complete a series of surgical cases. *J Clin Monit Comput*. (1999) 15:45–51. doi: 10.1023/A:1009999830753

6. Rothstein DH, Raval MV. Operating room efficiency. *Semin Pediatr Surg*. (2018) 27:79–85. doi: 10.1053/j.sempedsurg.2018.02.004

7. Lee DJ, Ding J, Guzzo TJ. Improving operating room efficiency. *Curr Urol Rep*. (2019) 20:1–8. doi: 10.1007/s11934-019-0895-3

8. Dexter F, Epstein RH. Fundamentals of operating room allocation and case scheduling to minimize the inefficiency of use of the time. *Perioper Care Oper Room Manag*. (2024) 35:100379. doi: 10.1016/j.pcorm.2024.100379

9. Kayis E, Wang H, Patel M, Gonzalez T, Jain S, Ramamurthi R, et al. Improving prediction of surgery duration using operational and temporal factors. In: *AMIA Annual Symposium Proceedings*. Vol. 2012. Washington, DC: American Medical Informatics Association (2012). p. 456. doi: 10.1016/j.surg.2021.12.032

10. Lee S-H, Dai T, Phan PH, Moran N, Stonemetz J. The association between timing of elective surgery scheduling and operating theater utilization: a cross-sectional retrospective study. *Anesth Analg*. (2022) 134:455–62. doi: 10.1213/ANE.0000000000005871

11. Eijkemans MJC, van Houdenhoven M, Nguyen T, Boersma E, Steyerberg EW, Kazemier G. Predicting the Unpredictable: A New Prediction Model for Operating Room Times Using Individual Characteristics and the Surgeon's Estimate. *Anesthesiology*. (2010) 112:41–9. doi: 10.1097/ALN.0b013e3181c294c2

12. van Eijk RP, van Veen-Berkx E, Kazemier G, Eijkemans MJ. Effect of individual surgeons and anesthesiologists on operating room time. *Anesth Analg*. (2016) 123:445–51. doi: 10.1213/ANE.0000000000001430

13. Bokshan SL, Mehta S, DeFroda SF, Owens BD. What are the primary cost drivers of anterior cruciate ligament reconstruction in the united states? a cost-minimization analysis of 14,713 patients. *Arthrosc J Arthrosc Relat Surg*. (2019) 35:1576–81. doi: 10.1016/j.arthro.2018.12.013

14. Allen AE, Sakheim ME, Mahendraraj KA, Nemec SM, Nho SJ, Mather III RC, et al. Time-driven activity-based costing analysis identifies use of consumables and operating room time as factors associated with increased cost of outpatient primary hip arthroscopic labral repair. *Arthrosc J Arthrosc Relat Surg*. (2023) 40:1517–26. doi: 10.1016/j.arthro.2023.10.050

15. Dexter F, Dexter EU, Masursky D, Nussmeier NA. Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. *Anesth Analg*. (2008) 106:1232–41. doi: 10.1213/ane.0b013e318164f0d5

16. Strum DP, Sampson AR, May JH, Vargas LG. Surgeon and Type of Anesthesia Predict Variability in Surgical Procedure Times. *Anesthesiology*. (2000) 92:1454–66. doi: 10.1097/00000542-200005000-00036

17. Patoir A, Payet C, Peix J-L, Colin C, Pascal L, Kraimps J-L, et al. Determinants of operative time in thyroid surgery: a prospective multicenter study of 3454 thyroidectomies. *PLoS One*. (2017) 12:e0181424. doi: 10.1371/journal.pone.0181424

18. Stepaniak PS, Heij C, De Vries G. Modeling and prediction of surgical procedure times. *Stat Neerl*. (2010) 64:1–18. doi: 10.1111/j.1467-9574.2009.00440.x

19. Gillespie BM, Chaboyer W, Fairweather N. Factors that influence the expected length of operation: results of a prospective study. *BMJ Qual Saf*. (2012) 21:3–12. doi: 10.1136/bmjqs-2011-000169

20. Kougias P, Tiwari V, Barshes NR, Bechara CF, Lowery B, Pisimisis G, et al. Modeling anesthetic times. predictors and implications for short-term outcomes. *J Surg Res*. (2013) 180:1–7. doi: 10.1016/j.jss.2012.10.007

21. van Veen-Berkx E, Bitter J, Elkhuizen SG, Buhre WF, Kalkmadn CJ, Gooszen HG, et al. The influence of anesthesia-controlled time on operating room scheduling in Dutch university medical centres. *Can J Anaesth*. (2014) 61:524–32. doi: 10.1007/s12630-014-0134-9

22. Dexter F, Coffin S, Tinker JH. Decreases in anesthesia-controlled time cannot permit one additional surgical operation to be reliably scheduled during the workday. *Anesth Analg*. (1995) 81:1263–8. doi: 10.1097/00000539-199512000-00024

23. Does RJ, Vermaat TM, Verver JP, Bisgaard S, Van Den Heuvel J. Reducing start time delays in operating rooms. *J Qual Technol*. (2009) 41:95–109. doi: 10.1080/00224065.2009.11917763

24. Snee RD. Six–sigma: the evolution of 100 years of business improvement methodology. *Int J Six Sigma Compet Adv*. (2004) 1:4–20. doi: 10.1504/IJSSCA.2004.005274

25. Halim UA, Khan MA, Ali AM. Strategies to improve start time in the operating theatre: a systematic review. *J Med Syst*. (2018) 42:1–11. doi: 10.1007/s10916-018-1015-5

26. Ahmed K, Khan N, Anderson D, Watkiss J, Challacombe B, Khan MS, et al. Introducing the productive operating theatre programme in urology theatre suites. *Urol Int*. (2013) 90:417–21. doi: 10.1159/000345312

27. Cox Bauer CM, Greer DM, Vander Wyst KB, Kamelle SA. First-case operating room delays: patterns across urban hospitals of a single health care system. *J Patient Cent Res Rev*. (2016) 3:125–35. doi: 10.17294/2330-0698.1265

28. Dexter F, Bayman EO, Pattillo JC, Schwenk ES, Epstein RH. Influence of parameter uncertainty on the tardiness of the start of a surgical case following a preceding surgical case performed by a different surgeon. *Perioper Care Oper Room Manag*. (2018) 13:12–7. doi: 10.1016/j.pcorm.2018.11.001

29. Epstein RH, Dexter F, Maga JM, Marian AA. Evaluation of the start of surgical closure as a milestone for forecasting the time remaining to exit the operating room: a retrospective, observational cohort study. *Perioper Care Oper Room Manag*. (2022) 29:100280. doi: 10.1016/j.pcorm.2022.100280

30. Bhatt AS, Carlson GW, Deckers PJ. Improving operating room turnover time: a systems based approach. *J Med Syst*. (2014) 38:1–8. doi: 10.1007/s10916-014-0148-4

31. Goldhaber NH, Schaefer RL, Martinez R, Graham A, Malachowski E, Rhodes LP, et al. Surgical pit crew: initiative to optimise measurement and accountability for operating room turnover time. *BMJ Health Care Inform*. (2023) 30:e100741. doi: 10.1136/bmjhci-2023-100741

32. Martin L, Langell J. Improving on-time surgical starts: the impact of implementing pre-or timeouts and performance pay. *J Surg Res*. (2017) 219:222–5. doi: 10.1016/j.jss.2017.05.092

33. Simmons CG, Alvey NJ, Kaizer AM, Williamson K, Faruki AA, Kacmar RM, et al. Benchmarking of anesthesia and surgical control times by current procedural terminology (CPT®) codes. *J Med Syst*. (2022) 46:19. doi: 10.1007/s10916-022-01798-z

34. Strum DP, May JH, Vargas LG. Modeling the Uncertainty of Surgical Procedure Times: Comparison of Log-normal and Normal Models. *Anesthesiology*. (2000) 92:1160–7. doi: 10.1097/00000542-200004000-00035

35. Stepaniak PS, Heij C, Mannaerts GH, de Quelerij M, de Vries G. Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: a multicenter study. *Anesth Analg*. (2009) 109:1232–45. doi: 10.1213/ANE.0b013e3181b5de07

36. Joustra P, Meester R, van Ophem H. Can statisticians beat surgeons at the planning of operations? *Empir Econ*. (2013) 44:1697–718. doi: 10.1007/s00181-012-0594-0

37. Silber JH, Rosenbaum PR, Zhang X, Even-Shoshan O. Influence of Patient and Hospital Characteristics on Anesthesia Time in Medicare Patients Undergoing General and Orthopedic Surgery. *Anesthesiology*. (2007) 106:356–64. doi: 10.1097/00000542-200702000-00025

38. Li Y, Zhang S, Baugh RF, Huang JZ. Predicting surgical case durations using ill-conditioned CPT code matrix. *IIE Trans*. (2009) 42:121–35. doi: 10.1080/07408170903019168

39. Wang J, Cabrera J, Tsui K-L, Guo H, Bakker M, Kostis JB. Clinical and nonclinical effects on operative duration: evidence from a database on thoracic surgery. *J Healthc Eng*. (2020) 2020:3582796. doi: 10.1155/2020/3582796

40. Wang J, Cabrera J, Tsui K-L, Guo H, Bakker M, Kostis JB. Clinical and non-clinical effects on surgery duration: statistical modeling and analysis. *arXiv* [Preprint] *arXiv:1801.04110* (2018). Available online at: https://arxiv.org/abs/1801.04110. doi: 10.48550/arXiv.1801.04110

41. Clark TS, Linzer DA. Should I use fixed or random effects? *Polit Sci Res Methods*. (2015) 3:399–408. doi: 10.1017/psrm.2014.32

42. Dexter F, Ledolter J. Bayesian Prediction Bounds and Comparisons of Operating Room Times Even for Procedures with Few or No Historic Data. *Anesthesiology*. (2005) 103:1259–167. doi: 10.1097/00000542-200512000-00023

43. Master N, Zhou Z, Miller D, Scheinker D, Bambos N, Glynn P. Improving predictions of pediatric surgical durations with supervised learning. *Int J Data Sci Anal*. (2017) 4:35–52. doi: 10.1007/s41060-017-0055-0

44. Bartek MA, Saxena RC, Solomon S, Fong CT, Behara LD, Venigandla R, et al. Improving operating room efficiency: machine learning approach to predict case-time duration. *J Am Coll Surg*. (2019) 229:346–354.e3. doi: 10.1016/j.jamcollsurg.2019.05.029

45. Fairley M, Scheinker D, Brandeau ML. Improving the efficiency of the operating room environment with an optimization and machine learning model. *Health Care Manag Sci*. (2019) 22:756–67. doi: 10.1007/s10729-018-9457-3

46. Kendale S, Bishara A, Burns M, Solomon S, Corriere M, Mathis M. Machine learning for the prediction of procedural case durations developed using a large multicenter database: algorithm development and validation study. *JMIR AI*. (2023) 2:e44909. doi: 10.2196/44909

47. Martinez O, Martinez C, Parra CA, Rugeles S, Suarez DR. Machine learning for surgical time prediction. *Comput Methods Programs Biomed*. (2021) 208:106220. doi: 10.1016/j.cmpb.2021.106220

48. Tuwatananurak JP, Zadeh S, Xu X, Vacanti JA, Fulton WR, Ehrenfeld JM, et al. Machine learning can improve estimation of surgical case duration: a pilot study. *J Med Syst*. (2019) 43:1–7. doi: 10.1007/s10916-019-1160-5

49. Wang Z, Dexter F. More accurate, unbiased predictions of operating room times increase labor productivity with the same staff scheduling provided allocated hours are increased. *Perioper Care Oper Room Manag*. (2022) 29:100286. doi: 10.1016/j.pcorm.2022.100286

50. Molnar C. *Interpretable Machine Learning*. Morrisville, NC: Lulu Press, Inc. (2020).

51. Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler S. Problems with Shapley-value-based explanations as feature importance measures. In: *Proceedings of the 37th International Conference on Machine Learning*, eds. H. D. III and A. Singh (PMLR), vol. 119 of *Proceedings of Machine Learning Research*. (2020). p. 5491–500.

52. [Dataset] American Academy of Professional Coders (2023). CPT Code Lookup. Available online at: https://www.aapc.com/codes/cpt-codes-range (accessed December 19, 2023).

53. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. (2015) 67:1–48. doi: 10.18637/jss.v067.i01

54. Brown ML, Staffa SJ, Quinonez LG, DiNardo JA, Nasr VG. Predictors of anesthesia ready time: analysis and benchmark data. *JTCVS Open*. (2023) 15:446–53. doi: 10.1016/j.xjon.2023.06.016

55. Mathenge C. The importance of the perioperative nurse. *Commun Eye Health*. (2020) 33:44.

56. Kayış E, Khaniyev TT, Suermondt J, Sylvester K. A robust estimation model for surgery durations with temporal, operational, and surgery team effects. *Health Care Manag Sci*. (2015) 18:222–33. doi: 10.1007/s10729-014-9309-8

57. Parikh N, Gargollo P, Granberg C. Improving operating room efficiency using the six sigma methodology. *Urology*. (2021) 154:141–7. doi: 10.1016/j.urology.2021.02.049

58. Xiao Y, Jones A, Zhang B, Bennett M, Mears SC, Mabrey JD, et al. Team consistency and occurrences of prolonged operative time, prolonged hospital stay, and hospital readmission: a retrospective analysis. *World J Surg*. (2015) 39:890–6. doi: 10.1007/s00268-014-2866-7

59. Nakagawa S, Johnson PC, Schielzeth H. The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J R Soc Interface*. (2017) 14:20170213. doi: 10.1098/rsif.2017.0213

frontiers | Frontiers in Physiology

TYPE Review
PUBLISHED 15 October 2024
DOI 10.3389/fphys.2024.1386760

Check for updates

OPEN ACCESS

EDITED BY
Hyo Kyung Lee,
Korea University, Republic of Korea

REVIEWED BY
Ricardo Valentim,
Federal University of Rio Grande do Norte, Brazil
Xu Huang,
Nanjing University of Science and Technology, China

*CORRESPONDENCE
Teresa Wu,
✉ Teresa.Wu@asu.edu

RECEIVED 16 February 2024
ACCEPTED 26 September 2024
PUBLISHED 15 October 2024

CITATION
Patharkar A, Cai F, Al-Hindawi F and Wu T (2024) Predictive modeling of biomedical temporal data in healthcare applications: review and future directions.
Front. Physiol. 15:1386760.
doi: 10.3389/fphys.2024.1386760

Predictive modeling of biomedical temporal data in healthcare applications: review and future directions

Abhidnya Patharkar[1,2], Fulin Cai[1,2], Firas Al-Hindawi[1,2] and Teresa Wu[1,2]*

[1]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, United States, [2]ASU-Mayo Center for Innovative Imaging, Arizona State University, Tempe, AZ, United States

Predictive modeling of clinical time series data is challenging due to various factors. One such difficulty is the existence of missing values, which leads to irregular data. Another challenge is capturing correlations across multiple dimensions in order to achieve accurate predictions. Additionally, it is essential to take into account the temporal structure, which includes both short-term and long-term recurrent patterns, to gain a comprehensive understanding of disease progression and to make accurate predictions for personalized healthcare. In critical situations, models that can make multi-step ahead predictions are essential for early detection. This review emphasizes the need for forecasting models that can effectively address the aforementioned challenges. The selection of models must also take into account the data-related constraints during the modeling process. Time series models can be divided into statistical, machine learning, and deep learning models. This review concentrates on the main models within these categories, discussing their capability to tackle the mentioned challenges. Furthermore, this paper provides a brief overview of a technique aimed at mitigating the limitations of a specific model to enhance its suitability for clinical prediction. It also explores ensemble forecasting methods designed to merge the strengths of various models while reducing their respective weaknesses, and finally discusses hierarchical models. Apart from the technical details provided in this document, there are certain aspects in predictive modeling research that have arisen as possible obstacles in implementing models using biomedical data. These obstacles are discussed leading to the future prospects of model building with artificial intelligence in healthcare domain.

KEYWORDS

biomedical temporal data, biomedical data challenges, forecasting, clinical predictive modeling, temporal data modeling problems, statistical time-series models, temporal machine learning models, deep temporal models

1 Introduction

1.1 Biomedical time series data

Clinical or biomedical data advances medical research by providing insights into patient health, disease progression, and treatment efficacy. It underpins new diagnostics, therapies, and personalized medicine, improving outcomes and understanding complex conditions. In predictive modeling, biomedical data is categorized as spatial, temporal, and spatio-temporal (Khalique et al., 2020; Veneri et al., 2012). Temporal data is key, capturing health evolution over time and offering insights into disease progression and treatment effectiveness. Time series data, collected at successive time points, shows complex patterns with short- and long-term dependencies, crucial for forecasting and analysis (Zou et al., 2019; Lai et al., 2018). Properly harnessed, this data advances personalized medicine and treatment optimization, making it essential in contemporary research.

1.2 Applications of predictive modeling in biomedical time series analysis

Predictive modeling with artificial intelligence (AI) has gained significant traction across various domains, including manufacturing (Altarazi et al., 2019), heat transfer (Al-Hindawi et al., 2023; 2024), energy systems (Huang et al., 2024), and notably, the biomedical field (Cai et al., 2023; Patharkar et al., 2024). Predictive modeling in biomedical time series data involves various approaches for specific predictions and data characteristics. Forecasting models predict future outcomes based on historical data, such as forecasting blood glucose levels for diabetic patients using past measurements, insulin doses, and dietary information (Plis et al., 2014). Classification models predict categorical outcomes, like detecting cardiac arrhythmias from ECG data by classifying segments into categories such as normal, atrial fibrillation, or other arrhythmias, aiding in early diagnosis and treatment (Daydulo et al., 2023; Zhou et al., 2019; Chuah and Fu, 2007). Anomaly detection in biomedical time series identifies outliers or abnormal patterns, signifying unusual events or conditions. For example, monitoring ICU patients' vital signs can detect early signs of sepsis (Mollura et al., 2021; Shashikumar et al., 2017; Mitra and Ashraf, 2018), enabling timely intervention.

Table 1 summarizes the example applications of these models within the context of biomedical time series.

1.3 Challenges in biomedical time series data

Regardless of the particular medical application or predictive model type used, models that manage biomedical time series data must tackle the intrinsic challenges posed by clinical and biomedical data. This includes various categories, such as electronic health records (EHRs), administrative data, claims data, patient/disease registries, health surveys, and clinical trials data. As illustrated in Table 2, each biomedical data category presents distinct challenges regarding quality, privacy, and completeness. During predictive modeling, further challenges arise. Specifically, we will investigate problems associated with missing data and imputation methods, the intricate nature of high-dimensional temporal relationships, and factors concerning the size of the dataset. Addressing these issues is crucial for developing strong and accurate predictive models in medical research.

1.3.1 Challenges in handling missing values and imputation methods in biomedical time series

Clinical data is often confronted with the issue of missing values, which can be caused by irregular data collection schedules or unexpected events (Xu et al., 2020). Medical measurements, recorded variably and at different times, may be absent, not documented, or affected by recording errors (Mulyadi et al., 2022), which makes the data irregular. Dealing with missing values in data sets usually involves either directly modeling data sets with missing values or filling in the missing values (a.k.a. imputation) to complete datasets for traditional analysis methods using data imputation techniques.

Current imputation techniques can be divided into four categories: case deletion, basic statistical imputation, machine learning-based imputation (Luo et al., 2018), and aggregating irregularly sampled data into discrete time periods (Ghaderi et al., 2023). Each of these methods comes with specific challenges in the context of handling biomedical temporal data. The deletion or omission of cases may lead to the loss of important information, particularly when the rate of missingness is high, which is critical in sensitive applications such as biomedical predictive modeling, where data is scarce and human lives are at risk. However, in certain cases, it is possible to do data omission without any potential risk to the outcome of the study. For instance, (Pinto et al., 2022), employs interrupted time series analysis to assess the impact of the "Syphilis No!" initiative in reducing congenital syphilis rates in Brazil. The results indicate significant declines in priority municipalities after the intervention. The study showcases the efficacy of public health interventions in modifying disease trends using statistical analysis of temporal data. Data collection needed to be conducted consistently over time and at evenly spaced intervals for proper analysis. To prevent bias due to the COVID-19 pandemic, December 2019 was set as the final data collection point, encompassing 20 months before the intervention (September 2016 to April 2018) and 20 months after the intervention (May 2018 to December 2019). This approach illustrates how the author addressed potential issues of irregular data or missing values in this context.

Contrary to data omission, statistical imputation techniques, such as mean or median imputation offer an alternative that reduces the effect of missing data, however, such methods do not take into account the temporal information but rather offer a summarized statistical imputation that often does not provide accurate replacement of the missing data. This could be critical in biomedical applications with scarce datasets, where the weight of a single data point could heavily affect the predictive power of the model. The use of machine learning-based imputation methods, such as Maximum Likelihood Expectation-Maximization, k-Nearest Neighbors, and Matrix Factorization, might offer a more accurate imputation that takes into account the specificity of the data point contrary to statistical aggregation methods,

TABLE 1 Overview of predictive modeling techniques for biomedical time series and their example applications across healthcare scenarios.

Model type	Description	Bio-medical application example
Forecasting	Predicts a continuous value based on historical data	Predicting blood glucose levels for diabetic patients using past glucose measurements, insulin doses, and dietary information to forecast potential hypo- or hyperglycemic events (Plis et al., 2014)
Classification	Predicts categorical outcomes based on temporal data	Detecting cardiac arrhythmias (such as normal, atrial fibrillation, or other arrhythmias) (Daydulo et al., 2023; Zhou et al., 2019; Chuah and Fu, 2007)
Anomaly Detection	Identifies outliers or abnormal patterns within time series data	Sepsis detection. (Mollura et al., 2021; Shashikumar et al., 2017; Mitra and Ashraf, 2018)

TABLE 2 Overview of clinical data types and challenges. This table lists the main types of clinical and biomedical data, their definitions, and key challenges.

Data type	Definition	Challenges
Electronic Health Records (EHRs)	Digital records of patients medical history, treatments, and outcomes	• Data Standardization: Different formats across providers • Data Quality: Missing, incomplete, or inaccurate data • Privacy and Security: Ensuring compliance with regulations like HIPAA • Interoperability: Difficulties in data exchange between systems
Administrative Data	Data related to healthcare administration, such as hospital admissions and discharge records	• Limited Clinical Detail: Lack of in-depth clinical information • Data Timeliness: Potential delays in data availability • Standardization Issues: Variability in recording and categorization • Privacy Concerns: Maintaining patient confidentiality
Claims Data	Data from insurance claims used for billing and reimbursement	• Purpose and Detail: Primarily for billing, may lack clinical details • Lag Time: Delays between care and data availability • Coding Errors: Inaccuracies in coding (e.g., ICD codes) • Complexity: Requires specialized knowledge for interpretation
Patient/Disease Registries	Databases that track patients with specific conditions or diseases	• Data Completeness: Ensuring all relevant data is captured • Data Standardization: Different definitions and methods across registries • Funding and Maintenance: Need for consistent resources • Privacy Issues: Protecting patient confidentiality
Health Surveys	Data collected from health-related surveys and questionnaires	• Response Bias: Non-response or inaccurate self-reporting • Sampling Issues: Ensuring representative samples • Data Quality: Depends on survey design and execution • Timeliness: Time-consuming design, conduct, and analysis
Clinical Trials Data	Data from controlled trials testing the efficacy of treatments or interventions	• Complexity and Cost: Expensive and logistically complex • Regulatory Hurdles: Compliance with regulatory requirements • Data Sharing: Balancing patient confidentiality and proprietary interests • Generalizability: Trial participants may not represent the broader population

however, many of them still do not consider temporal relations between observations (Luo et al., 2018; Jun et al., 2019), and they usually are computationally expensive. Furthermore, without incorporating domain knowledge, these approaches can introduce bias and lead to invalid conclusions. Both machine learning and statistical techniques may not consider data distribution or variable relationships and may fail to capture complex patterns in multivariate time-series data due to the neglect of correlated variables, potentially resulting in underestimated or overestimated imputed values (Jun et al., 2019). Additionally, in real-time clinical decision support systems, timely and accurate data is crucial, as delays or errors in imputation can lead to incorrect decisions that directly affect patient outcomes. These systems demand high-speed processing, requiring imputation algorithms to be both computationally efficient and accurate. Moreover, the dynamic nature of clinical environments, where patient conditions can change rapidly, necessitates imputation methods that can adapt quickly to evolving data.

Aggregating measurements into discrete time periods can address irregular intervals, but it may lead to a loss of granular information (Ghaderi et al., 2023). Additionally, in time series prediction, missing values and their patterns are often correlated with target labels, referred to as informative missingness (Che et al., 2018). These limitations make it ill-advised to ignore, impute, or aggregate these values when handling biomedical time series data, but rather employ a model that is capable of handling the sparsity and the irregularity of clinical time series data.

1.3.2 Complexities of high-dimensional temporal dependencies in biomedical data

Besides missing data challenges, hospitalized patients have a wide range of clinical events that are recorded in their electronic health records (EHRs). EHRs contain two different kinds of data: structured information, like diagnoses, treatments, medication prescriptions, vital signs, and laboratory tests, and unstructured data, like clinical notes and physiological signals (Xie et al., 2022; Lee and Hauskrecht, 2021), making them multivariate or high-dimensional (Niu et al., 2022).

The complexity of the relationships existing in such high-dimensional multivariate time series data can be difficult to capture and analyze. Analysts often try to predict future outcomes based on past data, and the accuracy of these predictions depends on how well the interdependencies between the various series are modeled (Shih et al., 2019). It is often beneficial to consider all relevant variables together rather than focusing on individual variables to build a prediction model, as this provides a comprehensive understanding of correlations in multivariate time series (MTS) data (Du et al., 2020). It thus becomes a requirement for predictive models employed in biomedical applications to take into account correlations among multiple dimensions and make predictions accordingly. It is equally crucial to ensure that only the features with a direct impact on the outcome are considered in the analysis. For instance, the study by Barreto et al. (2023) investigates the deployment of machine learning and deep learning models to forecast patient outcomes and allocate beds efficiently during the COVID-19 crisis in Rio Grande do Norte, Brazil. Out of 20 available features, nine were chosen based on their clinical importance and their correlation with patient outcomes, selected through discussions with clinical experts to guarantee the model's accuracy and interpretability.

In addition to the inherent high dimensionality of biomedical data sourced from diverse platforms such as EHRs, wearable devices monitoring neurophysiological functions, and intensive care units tracking disease progression through physiological measurements (Allam et al., 2021), also display a natural temporal ordering. This temporal structure demands a specialized analytical approach distinct from that applied to non-temporal datasets (Zou et al., 2019). The temporal dependency adds significant complexity to modeling due to the presence of two distinct recurring patterns: short-term and long-term. For instance, short-term patterns may repeat daily, whereas long-term patterns might span quarterly or yearly intervals within the time series (Lai et al., 2018). Biomedical data often exhibit long-term dependencies, such as those seen in biosignals like electroencephalograms (EEGs) and electrocardiograms (ECGs), which may span tens of thousands of time steps or involve specific medical conditions such as acute kidney injury (AKI) leading to subsequent dialysis (Sun et al., 2021; Lee and Hauskrecht, 2021). Concurrently, short-term dependencies can manifest in immediate physiological responses to medical interventions, such as the administration of norepinephrine and subsequent changes in blood pressure (Lee and Hauskrecht, 2021). Another instance is presented by Valentim et al. (2022), who have created a model to forecast congenital syphilis (CS) cases in Brazil based on maternal syphilis (MS) incidences. The model takes into account the probability of proper diagnosis and treatment during prenatal care. It integrates short-term dependencies by assessing the immediate effects of prenatal care on birth outcomes, and long-term dependencies by analyzing syphilis case trends over a 10-year period. This strategy aids in enhancing public health decision-making and syphilis prevention planning.

Analyzing these recurrent patterns and longitudinal structures in biomedical data is essential to facilitate the creation of time-based patient trajectory representations of health events that facilitate more precise disease progression modeling and personalized treatment predictions (Allam et al., 2021; Xie et al., 2022). By incorporating both short-term fluctuations and long-term trends, robust predictive models can uncover hidden patterns in patient health records, advancing our understanding and application of digital medicine. Failing to consider these recurrent patterns can undermine the accuracy of time series forecasting in biomedical contexts such as digital medicine, which involves continuous recording of health events over time.

Additionally, early detection of diseases is of paramount importance. This can be achieved by utilizing existing biomarkers along with advanced predictive modeling techniques, or by introducing new biomarkers or devices aimed at early disease detection. For instance, early diagnosis of osteoporosis is essential to mitigate the significant socioeconomic impacts of fractures and hospitalizations. The novel device, Osseus, as cited by Albuquerque et al. (2023), addresses this by offering a cost-effective, portable screening method that uses electromagnetic waves. Osseus measures signal attenuation through the patient's middle finger to predict changes in bone mineral density with the assistance of machine learning models. The advantages of using Osseus include enhanced accessibility to osteoporosis screening, reduced healthcare costs, and improved patient quality of life through timely intervention.

1.3.3 Dataset size considerations

The quantity of data available in a given dataset must be carefully considered, as it significantly influences model selection and overall analytical approach. For instance, when patients are admitted for brief periods, the clinical sequences generated are often fewer than 50 data points (Liu, 2016). Similarly, the number of data points for specific tests, such as mean corpuscular hemoglobin concentration (MCHC) lab results, can be limited due to the high cost of these tests, often resulting in less than 50 data points (Liu and Hauskrecht, 2015). Such limited data points pose challenges for predictive modeling, as models must be robust enough to derive meaningful insights from small samples without overfitting.

Conversely, some datasets may have a moderate sample length, ranging from 55 to 100 data points, such as the Physionet sepsis dataset (Reyna et al., 2019; 2020; Goldberger et al., 2000). These moderate-sized datasets offer a balanced scenario where the data is sufficient to train more complex models, but still requires careful handling to avoid overfitting and ensure generalizability.

In other cases, datasets can be extensive, particularly when long-span time series data is collected via sensor devices. These devices continuously monitor physiological parameters, resulting in large datasets with thousands of time steps (Liu, 2016). For example, wearable devices tracking neurophysiological functions or intensive care unit monitors can generate vast amounts of data, providing a rich source of information for predictive modeling. However, handling such large datasets demands models that are computationally efficient and capable of capturing long-term dependencies and complex patterns within the data.

The amount of data available is a major factor in choosing the appropriate model. Sparse datasets require models that can effectively handle limited information, often necessitating advanced techniques for data augmentation and imputation to make the most out of available data. Moderate datasets allow for the application of more sophisticated models, including machine learning and deep learning techniques, provided they are carefully tuned to prevent overfitting. Large datasets, on the other hand, enable the use of highly complex models, such as deep neural networks, which can leverage the extensive data to uncover intricate patterns and relationships.

1.4 Strategies in forecasting for biomedical time series data

While our discussion has generally revolved around the challenges in predictive modeling of biomedical temporal data, this review specifically emphasizes forecasting. From the earlier discourse, it is clear that a forecasting model for clinical or biomedical temporal data needs to adeptly manage missing, irregular, sparse, and multivariate data, while also considering its temporal properties and the capacity to model both short-term and long-term dependencies. The model should be able to make multi-step predictions, and the selection of a suitable model is determined by the amount of data available and the temporal length of the time series under consideration.

In this review, we initially examine three main categories of forecasting models: statistical, machine learning, and deep learning models. We look closely at the leading models within each category, assessing their ability to tackle the complexities of biomedical temporal data, including issues like data irregularity, sparsity, and the need to capture detailed temporal dependencies, alongside multi-step predictions. Since each category has its unique advantages as well as limitations in addressing the specific challenges of biomedical temporal datasets, other sets of models mentioned in the literature, known as hierarchical time series forecasting and combination or ensemble forecasting that merge the benefits of various forecasting models to produce more accurate forecasts are also covered.

The rest of the paper is structured as follows: In Section 2, statistical models are introduced. Section 3 covers machine learning models, while Section 4 focuses on deep learning models. This is followed by Section 5, which is a discussion section that summarizes the findings, discusses ensemble as well as hierarchical models, and explores future directions for the application of AI in clinical datasets. Finally, Section 6 concludes the paper.

2 Statistical models

The most popular predictive statistical models for temporal data are Auto-Regressive Integrated Moving Average (ARIMA) models, Exponential Weighted Moving Average (EWMA) models, and Regression models which are reviewed in the following sections.

2.1 Auto-Regressive Integrated Moving Average models

(Yule, 1927) proposed an autoregressive (AR) model, and (Wold, 1948) introduced the Moving Averaging (MA) model, which were later combined by Box and Jenkins into the ARMA model (Janacek, 2010) for modeling stationary time series. The ARIMA model, an extension of ARMA, incorporates differencing to make the time series stationary before forecasting, represented by ARIMA (p,d,q), where p is the number of autoregressive terms, d is the degree of differencing, and q is the number of moving average terms. ARIMA models have been applied in real-world scenarios, such as predicting COVID-19 cases. Ding et al. (2020) used an ARIMA (1,1,2) model to forecast COVID-19 in Italy. In another study, (Bayyurt and Bayyurt, 2020), utilized ARIMA models for predictions in Italy, Turkey, and Spain, achieving a Mean Absolute Percentage Error (MAPE) value below 10%. Similarly, (Tandon et al., 2022), employed an ARIMA (2,2,2) model to forecast COVID-19 cases in India, reporting a MAPE of 5%, along with corresponding mean absolute deviation (MAD) and multiple seasonal decomposition (MSD) values.

When applying ARIMA models to biomedical data, we select the appropriate model using criteria like Akaike Information Criterion (AIC) or Bayesian information criterion (BIC), estimate parameters using tools like R or Python's statsmodels, and validate the model through residual analysis. ARIMA models are effective for univariate time series with clear patterns, supported by extensive documentation and software, but they require stationarity and may be less effective for data with complex seasonality. Moreover, if a time series exhibits long-term memory, ARIMA models may produce unreliable forecasts (Al Zahrani et al., 2020), signifying that they are inadequate for capturing long-term dependencies. Additionally, ARIMA models necessitate a minimum of 50 data points in the time series to generate accurate forecasts (Montgomery et al., 2015). Therefore, ARIMA models should not be used for biomedical data that require the modeling of long-term relationships or have a small number of data points.

Several extensions such as Seasonal ARIMA (SARIMA) have been introduced for addressing seasonality. For instance, the research by Liu et al. (2023) examined 10 years of inpatient data on Acute Mountain Sickness (AMS), uncovering evident periodicity and seasonality, thereby establishing its suitability for SARIMA modeling. The SARIMA model exhibited high accuracy for short-term forecasts, assisting in comprehending AMS trends and optimizing the allocation of medical resources. An additional extension of ARIMA, proposed for long-term forecasts, is ARFIMA. In the study by Qi et al. (2020), the Seasonal Autoregressive Fractionally Integrated Moving Average (SARFIMA) model was utilized to forecast the incidence of hemorrhagic fever with renal syndrome (HFRS). The SARFIMA model showed a better fit and forecasting accuracy compared to the SARIMA model, indicating its superior capability for early warning and control of infectious diseases by capturing long-range dependencies. Additionally, it is apparent that ARIMA models cannot incorporate exogenous variables. Therefore, a variation incorporating exogenous variables, known as the ARIMAX model, has been proposed. The study by Mahmudimanesh et al. (2022) applied the ARIMAX model to forecast cardiac and respiratory mortality in Tehran by analyzing the effects of air pollution and environmental factors. The key variables encompass air pollutants (CO, NO2, SO2, PM10) and environmental data (temperature, humidity). The ARIMAX model is selected for its capacity to include exogenous variables and manage non-static time series data.

For multi-step ahead forecasting in temporal prediction models, two methods exist. The first, known as the plug-in or iterated multi-step (IMS) prediction that involves successively using the single step predictor, treating each prediction as if it were an observed value to obtain the expected future value. The second approach is to create a direct multi-step (DMS) prediction as a function of the observations, and to select the coefficients in this predictor by minimizing the sum of squares of the multi-step forecast errors. Haywood and Wilson (2009) developed a test to decide which of two approaches is more dependable based on a given lead-time. In addition to this test, there are other ways to decide which technique is most suitable for forecasting multiple steps ahead. One of these methods can be used to decide the best choice for multi-step ahead prediction either for ARIMA or other types of models depending on the amount of historical data and the lead-time.

2.2 Exponential weighted moving average models

The EWMA method, based on Roberts (2000), uses first-order exponential smoothing as a linear combination of the current observation and the previous smoothed observation. The smoothed observation $\tilde{y}_t$ at time t is given by the equation $\tilde{y}_t = \lambda y_t + (1-\lambda)y_{t-1}^{\sim}$, where λ is the weight assigned to the latest observation. This recursive equation requires an initial value $y_0^{\sim}$. Common choices for $y_0^{\sim}$ include setting it equal to the first observation y_1 or the average of available data, depending on the expected changes in the process. The smoothing parameter λ is typically chosen by minimizing metrics such as Mean Squared Error (MSE) or MAPE (Montgomery et al., 2015).

Several modifications of simple exponential smoothing exist to account for trends and seasonal variations, such as Holt's method (Holt, 2004) and Holt-Winter's method (Winters, 1960). These can be used in either additive or multiplicative forms. For modeling and forecasting biomedical temporal data, the choice of method depends on the data characteristics. Holt's method is more appropriate for data with trends. On the other hand, EWMA is suitable for stationary or relatively stable data, making it effective in scenarios without a clear trend, such as certain biomedical measurements. For instance, Rachmat and Suhartono (2020) performed a comparative analysis of the simple exponential smoothing model and Holt’s method for forecasting the number of goods required in a hospital’s inpatient service, assessing performance using error percentage and MAD. Their findings indicated that the EWMA model outperformed Holt’s method, as it produced lower forecast errors. This outcome is logical since the historical data of hospitalized patients lack any discernible trend.

EWMA models are also intended for univariate, regularly-spaced temporal data, as demonstrated in the example above (Rachmat and Suhartono, 2020), which uses a single variable (number of goods) over a period of time as input for model construction. This model is not suitable for biomedical data that involves multiple variables influencing the forecast unless its extention for multivariate data is employed. As highlighted by De Gooijer and Hyndman (2006), there has been surprisingly little progress in developing multivariate versions of exponential smoothing methods for forecasting. Poloni and Sbrana (2015) attributes this to the challenges in parameter estimation for high-dimensional systems. Conventional multivariate maximum likelihood methods are prone to numerical convergence issues and high complexity, which escalate with model dimensionality. They propose a novel strategy that simplifies the high-dimensional maximum likelihood problem into several manageable univariate problems, rendering the algorithm largely unaffected by dimensionality.

EWMA models cannot directly handle data that is not evenly spaced, and thus cannot be used to directly model biomedical data with a large number of missing values without imputation. These models are capable of multi-step ahead prediction either through DMS or IMS approach. To emphasize long-range dependencies, the parameter λ can be set to a low value, while a higher value will give more importance to recent past value (Rabyk and Schmid, 2016). The range of λ values typically used for reasonable forecasting is 0.1–0.4, depending on the amount of historical data available for modeling (Montgomery et al., 2015).

2.3 Regression models

Several regression models are available, and in this discussion, we focus on two specific types: multiple linear regression (MLR) (Galton, 1886; Pearson, 1922; Pearson, 2023) and multiple polynomial regression (MPR) (Legendre, 1806; Gauss, 1823). These models are particularly relevant for biomedical data analysis as they accommodate the use of two or more variables to forecast values. In MLR, there is one continuous dependent variable and two or more independent variables, which may be either continuous or categorical. This model operates under the assumption of a linear relationship between the variables. On the other hand, MPR shares the same structure as MLR but differs in that it assumes a polynomial or non-linear relationship between the independent and dependent variables. This review provides examination of these two regression models.

2.3.1 Multiple linear regression models

The estimated value of output y at time t, denoted as y_t with a MLR model for a certain set of predictors is given by the following Equation 1.

$$y_t = X_t\beta + \epsilon_t \tag{1}$$

where, $X_t = (1, x_{1t}, x_{2t}, \ldots, x_{kt})$ is a vector of k explanatory variables at time t, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ are regression coefficients, and ϵ_t is a random error term at time t, $t = 1, \ldots, N$ (Fang and Lahdelma, 2016). It can be solved with least squares method (Pearson, 1901) to obtain the regression coefficients.

R^2 value can be calculated to check the accuracy of model fitting. The value of R^2 that is closer to 1 indicates better model performance. Metrics such as Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Theil's inequality coefficient (TIC) are commonly utilized to assess the forecasting model's performance. While RMSE is scale-sensitive, MAPE and TIC are scale-insensitive. Lower values for these three metrics signify a well-fitting forecasting model.

Zhang et al. (2021) developed an MLR model aimed at being computationally efficient and accurate for forecasting blood glucose levels in individuals with type 1 diabetes. These MLR models can predict specific future intervals (e.g., 30 or 60 min ahead). The dataset is divided into training, validation, and testing subsets; missing values are handled using interpolation and forward filling, and the data is normalized for uniformity. The MLR model showed strong performance, especially in 60-min forward predictions, and was noted for its computational efficiency in comparison to deep learning models. It excelled in short-term time series forecasts with significant data variability, making it optimal for real-time clinical applications.

2.3.2 Multiple polynomial regression models

The estimated value of y_t with say a second-order MPR model for a certain set of predictors is given by the following Equation 2.

$$\begin{aligned} y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_n x_{nt} + \beta_{n+1} {x_{1t}}^2 + \beta_{n+2} x_{1t} x_{2t} + \cdots \\ + \beta_{2n} x_{1t} x_{nt} + \beta_{2n+1} {x_{2t}}^2 + \beta_{2n+2} x_{2t} x_{3t} + \cdots + \epsilon \end{aligned} \tag{2}$$

where, β_{1t}, β_{2t} are regression coefficients, $x_{1t}, x_{2t}, \ldots, x_{nt}$ are predictor variables, and ϵ is a random error. The ordinary least squares method (Legendre, 1806; Gauss, 1823) is applicable for solving this, similar to how it is used with MLR models. Furthermore, the evaluation metrics utilized for MLR are also suitable for MPR models.

Wu et al. (2021) utilized US COVID-19 data from January 22 to July 20 (2020), categorizing it into nationwide and state-level data sets. Positive cases were identified as Temporal Features (TF), whereas negative cases, total tests, and daily positive case increases were identified as Characteristic Features (CF). Various other features were employed in different manners, such as the daily increment of hospitalized COVID-19 patients. An MPR model was created for forecasting single-day outcomes. The model consisted of pre-processing and forecasting phases. The pre-processing phase included quantifying temporal dependency through time-window lag adjustment, selecting CFs, and performing bias correction. The forecasting phase involved developing MPR models on pre-processed data sets, tuning parameters, and employing cross-validation techniques to forecast daily positive cases based on state classification.

The various applications of multiple regression models stated above, linear or polynomial, reveal their inability to directly capture temporal patterns. Although these models can accommodate multiple input variables, their design limits them to forecasting a singular outcome with one model. One of the extentions proposed to tackle this problem is multivariate MLR (MVMLR). Suganya et al. (2020) employs MVMLR to forecast four continuous COVID-19 target variables (confirmed cases and death counts after one and 2 weeks) using cumulative confirmed cases and death counts as independent variables. The methodology includes data preprocessing, feature selection, and model evaluation using metrics like Accuracy, R^2 score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

It is clear from the design of the regression models that they are unable to process missing input data. Unless all the predictor variables are present or substituted, the value of the output variable cannot be determined. Therefore, it becomes essential to apply imputation techniques prior to employing the regression models for forecasting.

The regression models do not usually require a large amount of data; it has been demonstrated to be effective with as few as 15 data points per case (Filipow et al., 2023). Multi-step ahead prediction can be accomplished with either IMS or DMS approaches when dealing with temporal data like previous cases. Nonetheless, as mentioned previously, since these methods do not inherently capture temporal dependencies, forecasts can be generated as long as the temporal order is maintained while training, and testing the model.

3 Machine learning models

Many machine learning models are employed to construct forecasting models for temporal data sets. The most popular models for temporal data sets include Support vector regression (SVR), k-nearest neighbors regression (KNNR), Regression trees (Random forest regression [RFR]), Markov process (MP) models, Gaussian process (GP) models. We will examine these techniques in the following sections.

3.1 Support vector regression

The origin of Support Vector Machines (SVMs) can be traced back to Vapnik (1999). Initially, SVMs were designed to address the issue of classification, but they have since been extended to the realm of regression or forecasting problems (Vapnik et al., 1996). The SVR approach has the benefit of transforming a nonlinear problem into a linear one. This is done by mapping the data set x into a higher-dimensional, linear feature space. This allows linear regression to be performed on the new feature space. Various kernels are employed to convert non-linear data into linear data. The most commonly used are linear kernel, polynomial kernel, and radial basis or Gaussian kernel.

Upon transforming a nonlinear dataset x into a higher-dimensional, linear feature space, the prediction function $f(x)$ is expressed by Equation 3.

$$f(x) = w^T \phi(x) + b \tag{3}$$

The SVR algorithm solves a nonlinear regression problem by transforming the training data x_i (where i ranges from one to N, with N being the size of the training data set) into a new feature space, denoted by $\phi(x)$. This transformation allows establishing a linear relationship between input and output, using the weight matrix w and bias matrix b to further refine the model.

In SVR, selecting optimal hyperparameters (C,ϵ) is crucial for accurate forecasting. The parameter C controls the balance between minimizing training error and generalization. A higher C reduces training errors but may overfit, while a lower C results in a smoother decision function, possibly sacrificing training accuracy. The parameter ϵ sets a tolerance margin where errors are not penalized, forming an ϵ-tube around predictions. A larger ϵ simplifies the model but may underfit, whereas a smaller ϵ provides more detail, potentially leading to overfitting. Optimal values for C and ϵ may require additional methods (Liu et al., 2021).

SVR is often combined with other algorithms for parameter optimization. Evolutionary algorithms frequently determine SVR parameters. For example, Hamdi et al. (2018) used a combination of SVR and differential evolution (DE) to predict blood glucose levels with continuous glucose monitoring (CGM) data. The DE algorithm was used to determine the optimal parameters of the SVR model, which was then built based on these parameters. The model was tested using real CGM data from 12 patients, and RMSE was used to evaluate its performance for different prediction horizons. The RMSE values obtained were 9.44, 10.78, 11.82, and 12.95 mg/dL for prediction horizons (PH) of 15, 30, 45, and 60 min, respectively. It should be noted that when these evolutionary algorithms are employed for determining parameters, SVR encounters notable disadvantages, including a propensity to get stuck in local minima (premature convergence).

Moreover, SVR can occasionally lack robustness, resulting in inconsistent outcomes. To mitigate these challenges, hybrid algorithms and innovative approaches are applied. For instance, Empirical Mode Decomposition (EMD) is employed to extract non-linear or non-stationary elements from the initial dataset. EMD facilitates the decomposition of data, thereby improving the effectiveness of the kernel function Fan et al. (2017).

Essentially, SVR is an effective method for dealing with MTS data (Zhang et al., 2019). SVR, which operates on regression-based extrapolation, fits a curve to the training data and then uses this curve to predict future samples. It allows for continuous predictions rather than only at fixed intervals, making it applicable to irregularly spaced time series (Godfrey and Gashler, 2017). Nonetheless, due to its structure, SVR struggles to capture complex temporal dependencies (Weerakody et al., 2021).

It is suitable for smaller data sets as the computational complexity of the problem increases with the size of the sample Liu et al. (2021). It excels at forecasting datasets with high dimensionality Gavrishchaka and Banerjee (2006) due to the advanced mapping capabilities of kernel functions Fan et al. (2017). Additionally, multi-step ahead prediction in the context of SVR's application to temporal data can be achieved either with the DMS or IMS approach (Bao et al., 2014).

3.2 K-nearest neighbors regression

In 1951, Evelyn Fix and Joseph Hodges developed the KNN algorithm for discriminant examination analysis (Fix and Hodges, 1989). This algorithm was then extended to be used for regression or forecasting. The KNN method assumes that the current time series segment will evolve in the future in a similar way to a past time series segment (not necessarily a recent one) that has already been observed (Kantz and Schreiber, 2004). The task is thus to identify past segments of the time series that are similar to the present one according to a certain norm. Given a time series $y_N(N)$ with N samples, the segment made of the last m samples is denoted as $y_M(N)$, reflecting the current disturbance pattern. The KNN algorithm searches for k past time series intervals most comparable to $y_M(N)$ within the memory $y_N(N)$ using various distance metrics. For each nearest neighbor, a following time series of length h is generated, known as prediction contributions. Forecasting can then be done using unweighted or weighted approaches. In the unweighted approach, the prediction is the mean of the prediction contributions. In the weighted approach, the prediction is a weighted average based on the distance of each nearest neighbor from the current segment. Weights are assigned inversely proportional to the distances.

Gopakumar et al. (2016) employed the KNN algorithm to forecast the total number of discharges from an open ward in an Australian hospital, which lacked real-time clinical data. To estimate the next-day discharge, they used the median of similar discharges from the past. The quality of the forecast was evaluated using the mean forecast error (MFE), MAE, symmetric MAE (SMAPE), and RMSE. The results of these metrics were reported to be 1.09, 2.88, 34.92%, and 3.84, respectively, with an MAE error improvement of 16.3% over the naive forecast.

KNN regression is viable for multivariate temporal datasets, as illustrated by Al-Qahtani and Crone (2013). Nevertheless, its forecasting accuracy diminishes as the dimensionality of the data escalates. Consequently, it is critical to meticulously select pertinent features that impact the target variable to enhance model performance.

KNN proves effective for irregular temporal datasets (Godfrey and Gashler, 2017) due to its ability to identify previous matching patterns rather than solely depending on recent data. This distinctive characteristic renders KNN regression a favored choice for imputing missing data (Aljuaid and Sasi, 2016) prior to initiating any forecasting. Furthermore, it excels in capturing seasonal variations or local trends, such as aligning the administration of a medication that elevates blood pressure with a low blood pressure condition. Conversely, its efficacy in identifying global trends is limited, particularly in scenarios like septic shock, where multiple health parameters progressively deteriorate over time (Weerakody et al., 2021).

The KNN algorithm necessitates distance computations for k-nearest neighbors. Selecting an appropriate distance metric aligned with the dataset's attributes is essential, with Euclidean distance being prevalent, though other metrics may be more suitable for specific datasets. Ehsani and Drabløs (2020) examines the impact of various distance measures on cancer data classification, using both common and novel measures, including Sobolev and Fisher distances. The findings reveal that novel measures, especially Sobolev, perform comparably to established measures.

As the size of the training dataset increases, the computational demands of the algorithm also rise. To mitigate this issue, approximate nearest neighbor search algorithms can be employed (Jones et al., 2011). Furthermore, the algorithm requires a large amount of data to accurately detect similar patterns. Several methods have been suggested to accelerate the process; for example, (Garcia et al., 2010), presented two GPU-based implementations of the brute-force kNN search algorithm using CUDA and CUBLAS, achieving speed-ups of up to 64X and 189X over the ANN C++ library on synthetic data.

Similarly to other forecasting models, KNN is applicable for multistep ahead predictions using strategies such as IMS or DMS (Martínez et al., 2019). It is imperative to thoroughly analyze the clinical application and characteristics of the clinical data prior to employing KNN regression for forecasting, given its unique attributes. Optimizing the number of neighbors (k) and the segment length (m) through cross-validation is crucial. Employing appropriate evaluation metrics (e.g., MFE, MAE, SMAPE, RMSE) is necessary to assess the model's performance.

3.3 Random forest regression

Random Forests (RFs), introduced by Breiman (2001), are a widely-used forecasting data mining technique. According to Bou-Hamad and Jamali (2020), they are tree-based ensemble methods used for predicting either categorical (classification) or numerical (regression) responses. In the context of regression, known as Random Forest Regression (RFR), RF models strive to derive a prediction function $f(x)$ that reduces the expected value of a loss function $L(Y, f(X))$, with the output Y typically evaluated using the squared error loss. RFR builds on base learners, where each learner is a tree trained on bootstrap samples of the data. The final prediction is the average of all tree predictions as shown by Equation 4.

$$f(x) = \frac{1}{K}\sum_{k=1}^{K} l_k(x) \tag{4}$$

where K is the number of trees, and $l_k(x)$ is the k-th tree. Trees are constructed using binary recursive partitioning based on criteria such as MSE.

Zhao et al. (2019) developed a RFR model to forecast the future estimated glomerular filtration rate (eGFR) values of patients to predict the progression of Chronic Kidney Disease (CKD). The data set used was from a regional health system and included 120,495 patients from 2009 to 2017. The data was divided into three tables: eGFR, demographic, and disease information. The model was optimized through grid-search and showed good fit and accuracy in forecasting eGFR for 2015–2017 using the historical data from the past years. The forecasting accuracy decreased over time, indicating the importance of previous eGFR records. The model was successful in predicting CKD stages, with an average R^2 of 0.95, 88% Macro Recall, and 96% Macro Precision over 3 years.

The study presented in Zhao et al. (2019) indicates that RFR is effective for forecasting multivariate data. Another research by Hosseinzadeh et al. (2023) found that RFR performs better with multivariate data than with univariate data, especially when the features hold substantial information about the target. Research by Tyralis and Papacharalampous (2017) indicated that RF incorporating many predictor variables without selecting key features exhibited inferior performance relative to other methods. Conversely, optimized RF utilizing a more refined set of variables showed consistent reliability, highlighting the importance of thoughtful variable selection.

Similar to SVR, RFR is able to process non-linear information, although it does not have a specific design for capturing temporal patterns (Helmini et al., 2019). RFR is capable of handling irregular or missing data. El Mrabet et al. (2022) compared RFR for fault detection with Deep Neural Networks (DNNs), and found that RFR was more resilient to missing data than DNNs, showing its superior ability to manage missing values. To apply RFR to temporal data, it must be suitably modeled. As an example, Hosseinzadeh et al. (2023) has demonstrated one of the techniques, which involves forecasting stream flow by modeling the RFR as a supervised learning task with 24 months of input data and corresponding 24 months of output sequence. The construction of sequences involves going through the entire data set, shifting 1 month at a time. The study showed that extending the look-back window beyond a certain time frame decreases accuracy, indicating RFR's difficulty in capturing long-term dependencies when used in temporal modeling context. For a forecasting window of 24 months, the look-back window must be at least 24 months to avoid an increase in MAPE. This implies that although RFR can be used for temporal modeling, its effectiveness is more in capturing short-term dependencies rather than long-term ones. The experiments conducted by Tyralis and Papacharalampous (2017) also support this, showing that utilizing a small number of recent variables as predictors during the fitting process significantly improves the RFR's forecasting accuracy.

RFR can be used to forecast multiple steps ahead, similar to other regression models used for temporal forecasting (Alhnaity et al., 2021). Regarding data management, RFR necessitates a considerable volume of data to adjust its hyperparameters. It can swiftly handle such extensive datasets, leading to a more accurate model (Moon et al., 2018).

3.4 Markov process models

Two types of Markov Process (MP) models exist: Linear Dynamic System (LDS) and Hidden Markov Model (HMM). Both of these models are based on the same concept: a hidden state variable that changes according to Markovian dynamics can be measured. The learning and inference algorithms for both models are similar in structure. The only difference is that the HMM uses a discrete state variable with any type of dynamics and measurements, while the LDS uses a continuous state variable with linear-Gaussian dynamics and measurements. These models are discussed in more detail in the following sections.

3.4.1 Linear dynamic system

LDS, introduced by Kalman (1963), models the dynamics of sequences using hidden states and discrete time. It assumes evenly spaced time intervals within sequences, where the state transition and state-observation probabilities are given by q_i and o_i respectively. These probabilities are determined by the Equations 5, 6.

$$q_i = Aq_{i-1} + \epsilon_i \tag{5}$$

$$o_i = Bq_i + \zeta_i \tag{6}$$

The terms A and B represent the transition and emission matrices, respectively, whereas ϵ_i and ζ_i denote Gaussian noise components. Specifically, the stochastic element ϵ_i adheres to a zero-mean Gaussian distribution $\epsilon_i \sim \mathcal{N}(0, P)$, characterized by a zero-mean vector and covariance matrix P. On the other hand, the stochastic component ζ_i follows a zero-mean Gaussian distribution $\zeta_i \sim \mathcal{N}(0, R)$, which is also characterized by a zero-mean vector and covariance matrix R. The initial state distribution (q_1) is defined, with mean ξ and covariance matrix ψ, i.e., $q_1 \sim \mathcal{N}(\xi, \psi)$. The set of LDS parameters is denoted as $\lambda = (A, B, P, R, \xi, \psi)$. In applied scenarios, these parameters necessitate estimation from empirical data. Two standard approaches for learning LDS are the Expectation-Maximization (EM) (Ghahramani and Hinton, 1996) and spectral learning algorithms (Katayama, 2005; Overchee and Moor, 1996; Doretto et al., 2003). EM iteratively maximizes the likelihood of observations by cycling between expectation (E-step) and maximization (M-step). It is precise but can be slow and prone to local optima, especially with limited training data. Spectral learning algorithms provide a non-iterative, closed-form solution using singular value decomposition (SVD) to estimate LDS parameters. They are faster but may be less precise than EM.

A new data-driven state-space dynamic model was developed by Wang et al. (2014) using an extended Kalman filter to estimate time-varying coefficients based on three variate time series data corresponding to glucose, insulin, and meal intake from type 1 diabetic subjects. This model was used to forecast blood glucose levels and was evaluated against a standard model (forgetting-factor-based recursive ARX). The results showed that the proposed model was superior in terms of fit, temporal gain, and J index, making it better for early detection of glucose trends. Furthermore, the model parameters could be estimated in real time, making it suitable for adaptive control. This model was tested for various prediction horizons, demonstrating the model's suitability for multi-step ahead prediction.

The LDS is apt for modelling multivariate temporal data, yet it is confined to data sampled at regular time intervals. As a result, its application to irregularly spaced data (Shamout et al., 2021) or time series with missing values may be problematic. In such instances, modifications and extension are needed. For example, Liu et al. (2013) presented a novel probabilistic method for modeling clinical time series data that accommodates irregularly sampled observations using LDS combined with GP models. They defined the model by a series of GPs, each confined to a finite window, with dependencies between consecutive GPs represented via an LDS. Their experiments on real-world clinical time series data demonstrate that their model excels in modeling clinical time series and either outperforms or matches alternative time series prediction models.

Typically, implementing the LDS model starts with thorough data preparation, requiring uniform sampling. In cases of irregular sampling or datasets with missing values, proper management through interpolation or imputation is essential for using the model without alterations, as mentioned above. The model architecture is constructed using hidden state variables (q_i) to encapsulate the latent processes, alongside measurable observation variables (o_i) representing directly observable quantities. Parameters such as the state transition matrix (A), the emission matrix (B), and the covariance matrices for process noise (P) and observation noise (R) should be initialized based on prior knowledge or through randomization techniques. Parameter learning is facilitated through the EM algorithm or spectral learning methods, with practical considerations dictating the choice: EM being preferred for its precision with limited datasets and spectral methods for their computational expediency.

The LDS or Kalman filter remains a cornerstone for tracking and estimation due to its attributes of simplicity, optimality, tractability, and robustness. However, nonlinear system applications present complex challenges, often mitigated by the Extended Kalman Filter (EKF) (Lewis, 1986) which linearizes nonlinear models to leverage the linear Kalman filter. Also, various advancements have been proposed for LDS, particularly when addressing nonlinear or non-Gaussian dynamics. For example, approximate filtering methodologies such as the unscented Kalman filter (Julier and Uhlmann, 1997), alongside Monte Carlo-based techniques including the particle filter (Gordon et al., 1993) and the ensemble Kalman filter (Evensen, 1994), are also utilized similar to EKF. Model evaluation is conducted through cross-validation employing metrics such as MSE or RMSE. For forecasting applications, the model can be employed for one-step ahead forecasts or extended to iterative multi-step predictions.

3.4.2 Hidden markov model

Hidden Markov Models (HMMs), introduced by Baum and colleagues in the late 1960s and early 1970s (Baum and Petrie, 1966; Baum and Eagon, 1967; Baum and Sell, 1968; Baum et al., 1970; Baum, 1972), are powerful tools for linking hidden states with observed events, assuming an underlying stochastic process. An HMM consists of a set of hidden states, a transition probability matrix, a sequence of observations, observation likelihoods, and an initial state distribution. A critical assumption in HMMs is output independence, where the probability of an observation depends solely on the state that produced it.

HMMs address three fundamental problems: (1) Likelihood estimation: Using the forward or backward algorithm to compute the probability of an observed sequence given the model parameters; (2) Decoding: Employing the Viterbi algorithm to determine the optimal sequence of hidden states corresponding to a sequence of observations; and (3) Learning: Applying the Baum-Welch algorithm, a special case of the EM algorithm, to estimate HMM parameters from observation sequences.

Sotoodeh and Ho (2019) proposed a novel feature representation based on the HMM to predict the length of stay of patients admitted to the ICU. This representation was composed of a specified time resolution and a summary statistic calculated for a specific time window for each feature (e.g., average, most recent, maximum, etc.). An HMM was then trained on these features, and used to generate a series of states for each patient, with the first and last states being used as it was thought that these could better explain the variance in the length of stay. This feature matrix was then used as the input to a regression model to estimate the length of stay. Experiments were conducted to determine the optimal number of states, overlapping or non-overlapping time windows, aggregation of ICU types, summary measure for each time window, and selection of time window probabilities. The model was compared to other baseline models, and was found to have a lower RMSE than all of them.

It is evident from the application here that HMM is capable of dealing with multivariate data. Additionally, it is designed to process temporal data that is spaced at regular intervals of time (Shamout et al., 2021). Unfortunately, it is not able to process temporal data that is irregular or has missing values. Cao et al. (2015) employed both DMS and IMS strategies to forecast multiple future system states and anticipate the evolution of a fault in the Tennessee Eastman (TE) chemical process using HMM. They reported the accuracy of 1,2,3,.,20 step-ahead predictions, which were similar for both approaches, with the DMS approach being slightly more accurate than the IMS approach. This is understandable, as the IMS approach has to contend with additional complexities, such as cumulative errors, decreased precision, and increased uncertainty. This demonstrates the capability of HMM to make predictions for multiple steps in the future.

HMM can be constructed using either raw time series data or extracted features. Samaee and Kobravi (2020) introduced a forecasting model aimed at forecasting the timing of tremor bursts with a nonlinear hidden Markov model. This model was trained using the Baum-Welch algorithm, employing both raw Electromyogram (EMG) data and extracted features such as integrated EMG, mean frequency, and peak frequency. The study found that an HMM trained on raw EMG data performed better at forecasting tremor occurrences, suggesting that raw data more accurately captures tremor dynamics compared to extracted features. This is likely due to the short time window being insufficient for feature-based methods. Therefore, it is crucial to determine whether raw time series data or extracted features yield better performance in HMM construction.

In general, MP models are well-recognized for their efficacy in capturing short-term relationships (Manaris et al., 2011) between adjacent symbols or sequences with strong inter-symbol ties. However, they prove inadequate for representing long-distance dependencies between symbols that are spatially or temporally distant (Yoon and Vaidyanathan, 2006; Manaris et al., 2011). To enhance the representational scope of these models, certain methodologies must be employed. For instance, Yoon and Vaidyanathan (2006) proposed context-sensitive HMMs capable of capturing long-distance dependencies, thereby enabling robust pairwise correlations between distant symbols.

Additionally, a limitation of Markov models is that the intrinsic dimensionality of its hidden states is not known beforehand. If the dimensionality is too large, there is a risk of the model becoming overfitted. Therefore, it is often necessary to try out different training sizes and intrinsic dimensionality of the hidden states to create a model that fits (Liu, 2016).

3.5 Gaussian process models

The Gaussian process (GP), introduced by Williams and Rasmussen (Williams and Rasmussen, 2006), is a non-parametric, non-linear Bayesian model in statistical machine learning. A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution. This model extends the multivariate Gaussian to infinite-sized collections of real-valued variables, defining the distribution over random functions. A GP is represented by the mean function: $m(x) = \mathbb{E}[f(x)]$, and the covariance function: $K^G(x,x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$, where f(x) is a real-valued process and, x and x' are two input vectors.

In the context of biomedical temporal data, GP shows promises for modeling and forecasting due to their flexibility and ability to incorporate uncertainty. For example, GP can be used to model patient vital signs over time or predict disease progression (Siami-Namini et al., 2019). The key advantage of GP is their ability to provide uncertainty estimates along with predictions, which is crucial in biomedical applications where uncertainty quantification can inform clinical decisions. The GP can compute the distribution of function values for any set of inputs. This initial distribution, known as the prior, is a multivariate Gaussian represented by Equation 7.

$$f(X^*) \sim \mathcal{N}\left(m(X^*), K^G(X^*, X^*)\right) \tag{7}$$

When given observed data, the GP updates this to the posterior distribution, which also follows a multivariate Gaussian. This updated distribution incorporates the observed data, providing more accurate predictions. The posterior distribution is influenced by the observed values and accounts for noise in the data.

GPs extend the multivariate Gaussian distribution into an infinite function space, making them suitable for time series modeling. They can handle observations taken at any time, whether regularly or irregularly spaced, and can make future predictions by calculating the posterior mean for any given time index. Additionally, GPs can act as non-linear transformation operators by replacing the linear transformations used in traditional temporal models with GP, offering a flexible approach to modeling complex data.

GP parameters consist of mean and covariance function parameters. The mean function, dependent on time, represents the expectation before observations. In cases of uncertain trend directions, constant-offset mean functions are common. If prior knowledge about the long-term trend exists, it can be incorporated into GP models, optimizing mean function parameters using gradient-based methods. In clinical scenarios with diverse patient ages and circumstances, aligning time origins is challenging. A practical approach is setting mean functions to a constant ($m(t) = M$), making the GP time-invariant. The constant M is determined by averaging all patient observations. To optimize the covariance function parameters Θ, one can maximize the marginal likelihood $p(Y|X)$. The log marginal likelihood for GP is calculated where Y includes all training observations. The covariance matrix for noisy observations is represented by K^Y. It is calculated as $K^Y = K^G + \sigma^2 I$, where K^G is the covariance matrix for noise-free function values, and σ is a standard deviation of the noise, represented as, $\epsilon \sim \mathcal{N}(0, \sigma)$. The partial derivatives of the marginal likelihood with respect to each parameter in Θ are then derived. These derivatives are used in gradient-based optimization methods to maximize $p(Y|X)$, thereby optimizing the covariance function parameters.

A prevalent limitation of GP models pertains to their high computational demands. Sparse GP methodologies have been devised to mitigate this challenge (Williams and Rasmussen, 2006; Quinonero-Candela and Rasmussen, 2005), primarily by identifying a subset of pseudo inputs to alleviate computational load. Further optimization of computational efficiency can be achieved through the application of the Kronecker product (Stegle et al., 2011), synchronization of training data across identical time intervals for each dimension (Evgeniou et al., 2005), or the implementation of recursive algorithms tailored for online settings (Pillonetto et al., 2008). Applications necessitating near real-time retraining are more apt to benefit from these approaches, whereas methods that extend over more prolonged temporal frameworks exhibit reduced sensitivity to such computational constraints. Another shortcoming of GP is that it models each time series separately, disregarding the interactions between multiple variables. To tackle this problem and capture the multivariate behavior of MTS, the multi-task Gaussian process (MTGP) was proposed (Bonilla et al., 2007).

3.5.1 Multi-task Gaussian process

MTGP is an extension of GP that models multiple tasks (e.g., MTS) simultaneously by utilizing the learned covariance between related tasks. It uses K^C to model the similarities between tasks and K^G to capture the temporal dependence with respect to time stamps. The covariance function of MTGP is given by Equation 8.

$$K^M = K^C \otimes K^G + D \otimes I_T \tag{8}$$

where K^C is a positive semi-definite matrix and $K^C_{j,k}$ measures the similarity between time series j and time series k. D is an n x n diagonal matrix in which $D_{j,j}$ is the noise variance δ_j^2 for the j^{th} time series. $\otimes$ is the Kronecker product.

The parameters of GP-based models are composed of parameters that define the mean and covariance functions. Generally, the covariance function ensures that values of the function for two close times tend to have a high covariance, while values from inputs that are distant in time usually have a low covariance. These parameters can be acquired from data that includes one or multiple examples of time series. The predictions of values at future times are equivalent to the calculation of the posterior distribution for those times.

Proper data preprocessing is essential when building MTGP models for forecasting time series. This involves transformations such as detrending and applying logarithmic adjustments. Methods like spectral mixture kernels or Bayesian Nonparametric Spectral Estimation can be employed for initialization. Post-training, it is vital to visualize and interpret cross-channel correlations to better understand the inherent patterns, thereby supporting practical and accurate forecasting applications (de Wolff et al., 2021).

Shukla (2017) proposed to use MTGP to forecast blood pressure from Photoplethysmogram (PPG) signals and compared its performance to Artificial Neural Networks (ANNs). Ten features were extracted from the PPG signal, and five of them were chosen as the tasks (or targets) to construct the MTGP model. These features were systolic blood pressure, diastolic blood pressure, systolic upstroke time, diastolic time and cardiac period. Four different ANN models were built based on one or more of the above tasks. The models were evaluated on clinical data from the MIMIC Database, with the absolute error e calculated for each heart beat as the performance measure. The results showed that the performance of MTGP was either comparable to or better than the ANNs and existing methods of computing BP from non-invasive data. MTGP is thus applicable for modeling multivariate temporal data with multiple prediction targets. In a study by Dürichen et al. (2014), MTGP was employed on three diverse biomedical data sets. The experiments aimed to illustrate that forecasting all correlated variables simultaneously enhanced prediction performance, contrasting with individual variable predictions. MTGP has been demonstrated to be successful in multi-step ahead forecasting for a variety of biomedical domain applications mentioned here, as well as in other domains (Cai et al., 2020).

GP models, with an appropriate choice of covariance function, can capture rapid changes in a time series and can be applied to time series modeling problems by representing observations as a function of time. This means that there is no restriction on when the observations are made or if they are regularly or irregularly spaced in time. Liu (2016) and Cheng et al. (2020) demonstrated that, with the appropriate selection of a covariance function, it is possible to model both the short-term dependencies or long-term correlations of temporal data. GP models also work well with small amounts of data (Liu, 2016). It is possible to predict with a certain degree of certainty (confidence interval) using GP (Roberts et al., 2013), which is usually essential for temporal modeling of medical data that necessitates a certain degree of assurance to be employed by medical professionals to make their decisions. However, this approach has some limitations, the most serious being that the mean function of the GP is a function of time and must be set to a constant value in order to make the GP independent of the time origin. This significantly restricts its ability to represent changes or different modes in time series dynamics.

4 Deep learning models

The use of Deep Learning techniques for predicting time series data has gained significant attention. While there are various models available for handling time-series data, in this review, we will focus on some commonly used models for forecasting clinical data sets over time. Specifically, we will explore Recurrent Neural Networks (RNN), Long Short Term Memory Networks (LSTM), and Transformer models.

4.1 Recurrent Neural Networks

The concept of RNN was introduced by Elman (1990) for identifying patterns in sequential data. RNNs accept sequential data as input and process it recursively. In an RNN, nodes are linked

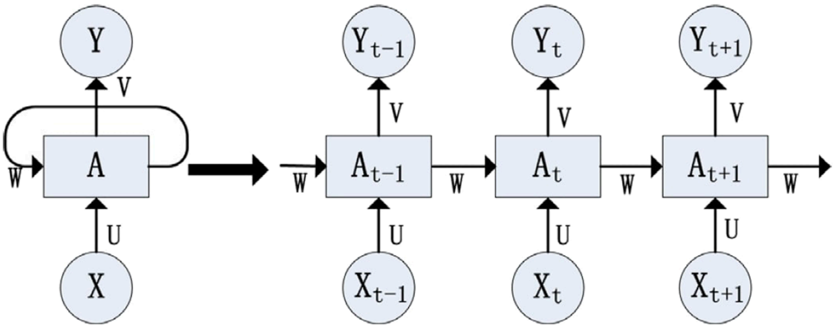


FIGURE 1
RNN structure (reproduced from Liu et al., 2021, licensed under CC BY 4.0).

sequentially, where the input at time t depends on the output at time $t-1$. The structure and functions of RNNs are depicted in Figure 1.

In this structure, the input layer (X) is weighted by U, the hidden layer (A) by W, and the output layer (Y) by V. The equations employed for calculations are Equations 9, 10.

$$Y_{t+1} = g(VA_{t+1}) \tag{9}$$

$$A_{t+1} = f(UX_{t+1} + WA_t) \tag{10}$$

The above formula is iterative in nature and can be expanded using the Equation 11 as:

$$Y_{t+1} = Vf(UX_{t+1} + Wf(UX_t + Wf(UX_{t-1} + \cdots))) \tag{11}$$

The equation above demonstrates that the RNN network's output Y_{t+1} is influenced by the current input A_{t+1}, as well as the previous inputs $A_t, A_{t-1}, ..$ RNNs effectively handle sequential and correlated data by considering historical inputs. The work of Chandra et al. (2021) demonstrates its applicability in multi-step ahead prediction. Although their demonstration focuses on univariate cases, RNN has also been successfully applied to multivariate cases. In their study, Zhu et al. (2020) utilized four data fields for each instance: sampling time, CGM values, meal intake, and insulin dose. They employed a deep learning approach using an extention of RNN, dilated RNN (DRNN), to forecast glucose levels for the next 30 min. The DRNN model exhibits superior performance compared to current models like autoregressive (ARX), SVR, and neural networks for glucose prediction (NNPG), when evaluated on the OhioT1DM dataset. The RMSE values reported are ARX: 20.1 mg/dL, SVR: 21.7 mg/dL, NNPG: 22.9 mg/dL, and DRNN: 18.9 mg/dL. RNNs are frequently used to handle missing values or irregularities in multivariate temporal datasets. There are two main approaches to achieve this: imputation and data generation, or a forecasting approach. When using the first approach, RNNs leverage temporal correlations within each series and correlations among multiple features to fill in missing values or create a time series that captures the original characteristics. On the other hand, the latter approach involves the development of more advanced RNN-based solutions that provide a deeper understanding of the missing data, as well as the patterns and relationships within the data (Weerakody et al., 2021).

Implementing RNNs for modeling and forecasting biomedical temporal data necessitates meticulous attention to data preprocessing, model structure, tuning of hyperparameters, and evaluation techniques. The recommendations for each aspect are outlined as discussed in Hewamalage et al. (2021). Deseasonalization is advised for datasets exhibiting seasonal trends unless consistent seasonal patterns exist, which RNNs can inherently manage. Data normalization enhances training convergence, while the sliding window approach divides the time series into overlapping sequences for model input. Hyperparameter tuning is crucial for achieving optimal RNN performance. Principal hyperparameters include the learning rate, batch size, and the number of layers. The learning rate must be selected judiciously; for ideal convergence, the Adagrad optimizer typically needs a higher learning rate ranging between 0.01 and 0.9, whereas the Adam optimizer performs effectively within a narrower range of 0.001–0.1. The batch size should be commensurate with the dataset size, and usually, one or two layers are sufficient, as additional layers may result in overfitting. Setting high values for the standard deviation of regularization parameters for Gaussian noise and L2 weight regularization can cause significant underfitting, reducing the neural network's efficacy in generating forecasts. One category of RNN models, stacked RNNs, which involve multiple RNN layers, are employed for forecasting and often utilize skip connections to alleviate vanishing gradient issues. Another category of RNN models, known as sequence-to-sequence (S2S) models, is typically applied in sequential data transformations and is useful for tasks like multi-step forecasting. Assessing RNN performance against traditional methods like ARIMA using standard metrics and cross-validation confirms their competitiveness. Enhancements to RNN methods, such as attention mechanisms and ensemble methods, further boost their performance. Attention mechanisms enable the model to concentrate on relevant parts of the input sequence, while ensemble methods combine several RNN models to produce robust forecasts, reducing biases and variances.

RNNs excel at capturing short-term dependencies (Helmini et al., 2019). They are more sensitive to time series data than traditional convolutional neural networks (CNNs) and can retain memory during data transmission. However, as previously mentioned, when the input sequence lengthens, the network demands more temporal references, leading to a deeper network. In longer sequences, it becomes challenging for the gradient to propagate back from later sequences to earlier ones, resulting in

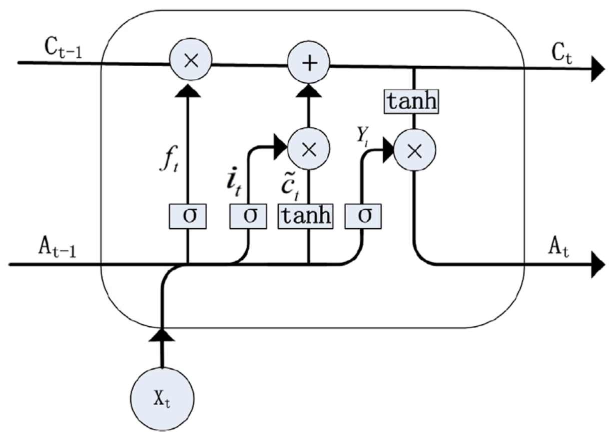


FIGURE 2
LSTM structure (reproduced from Liu et al., 2021, licensed under CC BY 4.0).

the vanishing gradient problem. Consequently, RNNs struggle with long-term dependencies. To mitigate this vanishing (or exploding) gradient issue, a modification of the RNN known as the long sshort-term memory (LSTM) model was introduced by Hochreiter and Schmidhuber (1997).

4.1.1 Long Short Term Memory Networks

To overcome the challenges of vanishing and exploding gradients in RNNs, the LSTM model was introduced. This architecture employs a cell state to maintain long-term dependencies, as discussed by Helmini et al. (2019). The model effectively manages gradient dispersion by establishing a retention mechanism between input and feedback. Figure 2 illustrates the LSTM structure (Weerakody et al., 2021). Additionally, LSTM models are proficient in capturing short-term dependencies, primarily through the use of a hidden state. LSTM units are controlled by three gates: the input gate, the output gate, and the forget gate. These gates regulate the flow of information and maintain the cell state, enabling LSTMs to retain important information over long periods. The key equations (Equations 12–17) governing LSTM operations are mentioned as follows:

$$f_t = \sigma\left(W_{fA}A_{t-1} + W_{fX}X_t + b_f\right) \tag{12}$$

$$i_t = \sigma\left(W_{iA}A_{t-1} + W_{iX}X_t + b_i\right) \tag{13}$$

$$\tilde{c}_t = \tanh\left(W_{cA}A_{t-1} + W_{cX}X_t + b_c\right) \tag{14}$$

$$c_t = \left(f_t \circ c_{t-1}\right) + \left(i_t \circ \tilde{c}_t\right) \tag{15}$$

$$Y_t = \sigma\left(W_{YA}A_{t-1} + W_{YX}X_t + b_Y\right) \tag{16}$$

$$A_t = \left(Y_t \circ \tanh\left(c_t\right)\right) \tag{17}$$

In these equations, σ represents the sigmoid function, and $\circ$ denotes element-wise multiplication. The forget gate (f_t) controls the retention of the previous cell state (c_{t-1}), the input gate (i_t) manages the incorporation of new information, and the output gate (Y_t) determines the output based on the cell state (c_t). W_{fA}, W_{fX}, W_{iA}, W_{iX}, and W_{cA} are different weights associated with the forget gate, input gate, and the current input unit state.

A deep learning neural network (NN) model based on LSTM with the addition of two fully connected layers was proposed by Idriss et al. (2019), for forecasting blood glucose levels. To determine the optimal parameters for the model, several experiments were conducted using data from 10 diabetic patients. The performance of the proposed LSTM NN, as measured by RMSE, was compared to that of a simple LSTM model and an autoregressive (AR) model. The results indicated that the LSTM NN achieved higher accuracy (mean RMSE = 12.38 mg/dL) compared to both the existing LSTM model (mean RMSE = 28.84 mg/dL) for all patients and the AR model (mean RMSE = 50.69 mg/dL) for 9 out of 10 patients. LSTM is therefore valuable in the representation of time-based information.

One popular extention of the LSTM network is a Bidirectional LSTM (BiLSTM) model which is obtained by modifying the architecture of the LSTM network to include two LSTM layers: one processing the input sequence from left to right (forward direction) and the other from right to left (backward direction). This bidirectional traversal allows the model to have information from both past and future contexts, enhancing its ability to capture complex patterns and dependencies. The outputs from both layers are concatenated at each time step, providing a richer representation of the input sequence. This approach results in improved performance for tasks like time series forecasting, as BiLSTM models can leverage additional training from both directions to better understand sequential data (Abbasimehr and Paki, 2022). For instance, in a study by Said et al. (2021), a bidirectional LSTM (Bi-LSTM) was employed to analyze multivariate data from countries grouped based on demographic, socioeconomic, and health sector indicators alongwith the information on lockdown measures, to predict the cumulative number of COVID-19 cases in Qatar from December 1st to 31 December 2020.

LSTM is also combined with multi-head attention mechanisms. This approach aims to address the non-linear patterns and complexities often found in real-world time series data, which traditional forecasting techniques struggle to predict accurately (Siami-Namini et al., 2019). When dealing with irregular temporal data that contain missing values, traditional LSTM models face challenges and may produce suboptimal analyses and predictions. This is because applying the LSTM model to irregular temporal data, either by filling in missing values or using temporal smoothing, does not enable the model to differentiate between actual observations and imputed values. Therefore, caution is advised when using an LSTM model on a dataset where multiple missing values have been imputed.

4.2 Transformer models

The Transformer model for natural language processing (NLP) was introduced by Vaswani et al. (2017). This model is composed of an encoder-decoder network, which differs from the traditional sequential structure of RNN. Transformer model utilizes the Self-Attention mechanism to enable parallel training and capture global

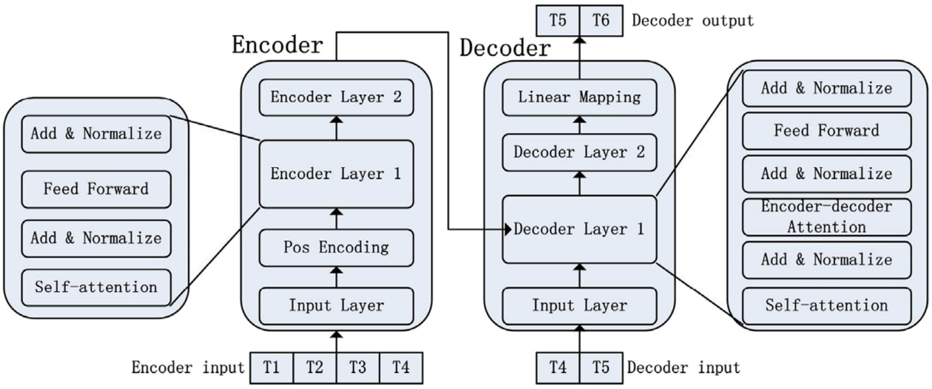


FIGURE 3
Transformer architecture (reproduced from Liu et al., 2021, licensed under CC BY 4.0).

information. The encoder takes historical time series data as input, while the decoder predicts future values using an auto-regressive approach. This means that the decoder's generated output at each step is based on previously generated outputs. To establish a connection between the encoder and decoder, an attention mechanism is employed. This allows the decoder to learn how to effectively focus ("pay attention") on relevant parts of the historical time series before making predictions. The decoder utilizes masked self-attention to prevent the network from accessing future values during training, thereby avoiding information leakage. The typical architecture of the Transformer model is depicted in the Figure 3.

Originally designed for NLP tasks, the Transformer architecture has found application in temporal forecasting as well. To model irregular temporal data, various methods have been proposed. For instance, Tipirneni and Reddy (2022) introduced the Self-supervised Transformer for Time-Series (STraTS) model, which treats each time-series as observation triplets (time, variable, value) instead of matrices as done by conventional methods. This approach eliminates the need for aggregation or imputation. STraTS utilizes a Continuous Value Embedding (CVE) scheme to retain detailed time information without discretization.

The study by Harerimana et al. (2022) utilized a Multi-Headed Transformer (MHT) model to forecast clinical time-series variables from charted vital signs, leveraging the transformer architecture's attention mechanism to capture complex temporal dependencies. The dataset is split into training and testing sets per patient, using past 24-h data for recursive future predictions. Training involves a fixed dimension of 512 for all layers, and the model is evaluated using metrics like Area under the Receiver Operating Characteristic Curve (AUC-ROC), MSE, and MAPE. The MHT model outperforms traditional models (LSTM, Temporal Convolutional Network, TimeNet) in forecasting vital signs, length of stay, and in-hospital mortality, demonstrating superior accuracy and robustness by focusing on influential past time steps, validating its efficacy in handling clinical time-series data.

The Transformer architecture is a relatively new concept, and ongoing research is being conducted to explore its capabilities. For instance, Li et al. (2019) suggest that unlike RNN-based methods, the Transformer enables the model to access any part of the time series history, disregarding the distance. This characteristic potentially makes it more adept at capturing recurring patterns with long-term dependencies. However, Zeng et al. (2023) presented an opposing viewpoint, questioning the effectiveness of Transformer-based solutions in long-term time series forecasting (LTSF). They argue that while Transformers are adept at capturing semantic correlations in sequences, their self-attention mechanism, which is invariant to permutations, may result in the loss of crucial temporal information necessary for accurate time series modeling. In support of their claim, the researchers introduced LTSF-Linear, a simple one-layer linear model, and discovered that it outperformed more complex Transformer-based LTSF models on nine real-life data sets. In addition, a temporal fusion transformer (TFT) was suggested by Zhang et al. (2022) as a method that effectively captures both short-term and long-term dependencies. Hence, when employing Transformer-based approaches for temporal forecasting, it is crucial to take into account these distinct viewpoints and conduct experiments to determine the most effective modeling technique for the specific forecasting task, considering the presence of short-term and long-term dependencies.

While DL models are capable of generating precise predictions, they are frequently perceived as black-box models that lack interpretability and transparency in their internal processes (Vellido, 2019). This presents a significant issue as medical professionals are often hesitant to trust machine recommendations without a clear understanding of the underlying rationale. In addition, significant quantities of clinical data are utilized to generate standardized inputs for training DL models. The challenge of acquiring extensive clinical data sets poses a challenge in the integration of DL clinical models into real-world clinical systems (Xiao et al., 2018).

5 Discussion

This section is comprised of two subsections. The first subsection summarizes the overview of the models and their capacities in addressing the difficulties encountered in forecasting of clinical datasets. The second subsection explores the future prospects concerning the practical obstacles in implementing AI models for biomedical data modeling.

5.1 Summary of models for biomedical temporal data forecasting

5.1.1 Summary of statistical, ML, and DL models

This review focuses on predictive models for biomedical temporal data, which face several challenges such as missing values due to irregular data collection or errors. Traditional methods use imputation or deletion, but models that handle missing values without these steps are preferable, as patterns of missing data might hold valuable information termed as "informative missingness". EHRs often feature MTS data, so models must capture these correlations. Temporal data complexity requires models to consider short-term and long-term patterns. Short-term patterns might involve events like norepinephrine administration linked to recent hypotension, while long-term patterns could involve past acute kidney injury necessitating dialysis. Models should account for these dependencies and support multi-step ahead forecasting for early disease detection. Data availability varies with clinical events, thus impacting model selection. These challenges are crucial for accurate, effective predictions in clinical settings. Table 3 summarizes the advantages and disadvantages of the discussed models, supplemented by literature insights.

Forecasting is categorized into statistical, ML, and DL methods. We focused on models frequently used in biomedical temporal modeling, evaluating their effectiveness. For statistical methods, we analyzed ARIMA, EWMA, and regression models. In ML, we assessed SVR, RFR, KNNR, MP, and GP models. For DL methods, we evaluated RNN, LSTM, and Transformer models. Our analysis found the MTGP model effective for irregularly spaced data, capturing both short-term and long-term dependencies with an appropriate covariance function. It predicts multiple steps ahead and accounts for autocorrelation within and correlation between time series, making it suitable for multivariate temporal analysis with small to moderate data. However, MTGP's computational cost can be high with large data, and a constant mean function may limit its ability to represent time series dynamics. While MTGP is suitable for biomedical temporal modeling, alternative approaches include improving current models, adopting ensemble methods, or using hierarchical approaches discussed later in this paper.

Improving existing models by incorporating new techniques can address limitations in temporal analysis of biomedical data. For instance, while RNNs struggle with long-range dependencies, they handle other temporal challenges well. To overcome this, Zhu et al. (2020) introduced a dilated RNN, enhancing neuron receptive fields to capture long-term dependencies, enabling 30-min glucose level forecasts. Similarly, HMMs lack long-range correlation modeling. Yoon and Vaidyanathan (2006) introduced context-sensitive HMM (csHMM), capturing long-range correlations by adding context-sensitivity to model states. Additionally, the interpretability in DL models is essential. Tipirneni and Reddy (2022) proposed an interpretable model with outputs as linear combinations of individual feature components. Slight modifications to the original models can address specific limitations.

Even though various modifications have been suggested to address the shortcomings of individual models, certain limitations remain insurmountable. A recently emerging solution involves combining multiple models to create a fusion model, which allows for the integration of their strengths and mitigation of their weaknesses. These fusion models, also known as combination or ensemble forecasting models, is examined in the next subsection.

5.1.2 Fusion models

A different approach to enhance forecasting precision involves merging multiple models, also known as combination or ensemble forecasting models. The paper by Wang et al. (Wang et al., 2023) provides a comprehensive overview of the evolution and effectiveness of combining multiple forecasts to enhance prediction accuracy. Combining forecasts, known as "ensemble forecasts," integrates information from various sources, avoiding the need to identify a single "best" forecast amidst model uncertainty and complex data patterns. The review covers simple combination methods, such as equally weighted averages, which surprisingly often outperform more sophisticated techniques due to their robustness and lower risk of overfitting. Linear combinations, which determine optimal weights based on historical performance, and nonlinear combinations, which account for nonlinear relationships using methods like neural networks, are also discussed. Wang et al. (2023) emphasize the potential of learning-based combination methods, such as stacking and cross-learning, which improve accuracy by training meta-models on multiple time series. In stacking, several forecasting models are trained on the original dataset, and their predictions are combined by a meta-model to provide an optimal forecast. Cross-learning builds on this by utilizing data from various time series to train the meta-model. The review also highlights the crucial role of diversity and precision in forecast combinations, pointing out that successful combinations are enhanced by diverse individual forecasts.

These techniques have been successfully applied to biomedical data forecasting. For example, Naemi et al. (2020) introduced a customizable real-time hybrid model, leveraging the Nonlinear Autoregressive Exogenous (NARX) model along with Ensemble Learning (EL) (RFR and AdaBoost), to forecast patient severity during their stay at Emergency Departments (ED). This model makes use of patient vital signs such as Pulse Rate (PR), Respiratory Rate (RR), Arterial Blood Oxygen Saturation (SpO2), and Systolic Blood Pressure (SBP), which are recorded during treatment. The model forecasts the severity of illness in hospitalized patients at ED for the upcoming hour based on their vital signs from the previous 2 hours. The effectiveness of the NARX-EL models is evaluated against other baseline models including ARIMA, a fusion of NARX and LR, SVR, and KNNR. The findings revealed that the proposed hybrid models could predict patient severity with significantly higher accuracy. Furthermore, it was noted that the NARX-RF model excels at predicting abrupt changes and unexpected adverse events in patients' vital signs, exhibiting an R^2 score of 0.978 and NRMSE of 6.16%. Kandula et al. (2018) used a super-ensemble technique to combine information from

TABLE 3 Advantages and disadvantages of models for handling biomedical temporal data.

Model type	Advantages	Disadvantages
ARIMA	- Captures linear dependencies and trends - Interpretable parameters - Works well with stationary data	- Needs data to be stationarized - Lacks ability to handle missing values - Inability to manage multivariate data - Can not capture long-range dependencies
EWMA	- Proficient in temporal modeling - Simple and computationally efficient - Adapts quickly to recent changes in data - Useful for smoothing noisy data - Effective in short-range modeling	- Not suitable for complex patterns - Unsuitable for handling multivariate data - Critical initialization and parameter selection - Capable of long-range modeling with parameter adjustment
MLR	- Interpretable coefficients - Insights into variables' relationships - Performs well with small-mid datasets - Efficiently manages multivariate data - Can be adapted for temporal modeling	- Assumes linear relationships - Sensitive to multicollinearity - Requires features to be linearly related to the target.
MPR	- Can capture higher-order relationships - More flexible than MLR. - Suitable for polynomial relationships - Manages multivariate data efficiently	- Prone to overfitting with high polynomial-degrees - Interpretation of coefficients can be complex - Unable to handle missing values
SVR	- Effective in high-dimensional spaces - Can capture nonlinear relationships - Robust to overfitting with regularization - Manages multivariate data efficiently - Although not designed for temporal modeling, but can be adapted to capture them	- Computationally complex - Needs support from other algorithms for hyperparameter tuning - Lacks robustness resulting in inconsistent outcomes - Struggles to capture complex temporal dependencies - Memory intensive for large datasets
KNNR	- Non-parametric and flexible - Can be adapted for temporal modeling - Proficient in handling missing values - Efficiently manages multivariate data - Effective in short-range modeling due to its unique structure	- Expensive for large datasets - Memory intensive - Falls short in capturing global dependencies
RFR	- Handles nonlinear relationships - Robust to overfitting - Can handle high-dimensional data - Manages multivariate data efficiently - Capable of handling irregular or missing data	- Time consuming for large datasets - Requires careful tuning of hyperparameters - Difficulty in handling long-range dependencies
LDS	- Captures temporal dependencies - Efficiently handles multivariate data - Captures short-term relationships	- Complex parameter tuning - Cannot deal with irregular data - Difficulty with nonlinear relationships
HMM	- Captures hidden influencing states - Useful for sequential data modeling - Efficiently handles multivariate data - Can capture short-term dependencies efficiently	- Training complexity - Lacks interpretability of hidden states - Prone to overfitting when intrinsic dimensionality exceeds data - Struggles with capturing long-term dependencies
MTGP	- Models multiple tasks simultaneously - Captures correlations between tasks - Provides uncertainty estimates - Can forecast efficiently with irregular data - Flexible covariance function that can capture both short-range and long-range dependencies	- Complex to implement and tune - If the GP is made time independent, it restricts the representation of changes in time series dynamics - Computationally intensive on large-scale
RNN	- Proficient in handling missing values - Can handle variable-length sequences - Effective for multivariate sequential data modeling	- Vanishing/exploding gradient problem - Training can be slow - Difficulty with very long-term dependencies

(Continued on the following page)

TABLE 3 (*Continued*) Advantages and disadvantages of models for handling biomedical temporal data.

Model type	Advantages	Disadvantages
LSTM	- Handles vanishing gradient problem - Captures long-term dependencies effectively - Robust to sequence length variations	- Training complexity - Lacks interpretability - Requires careful hyperparameters tuning - May produce suboptimal analyses and predictions when modeling imputed data
Transformer	- Highly suitable for multivariate temporal modeling - Parallel processing of sequences - Scalable to large datasets - Effective in short-range modeling	- Computationally intensive - Requires large amounts of data - Lacks interpretability - Fine-tuning can be complex - Uncertain effectiveness in managing long-term dependencies

different forecasting methods robustly. This method yielded a more accurate comprehensive forecast on average than a single model. They compared three forecasting approaches for predicting seven characteristics of seasonal influenza during the 2016–2017 USA season: a mechanistic method, a weighted average of two statistical methods, and a super-ensemble of eight statistical and mechanistic models. The study found the meta-ensemble approach to be the most accurate overall. Katari et al. (2023) employed a combination of Decision Tree (DT) and Ada Boosting algorithms for heart disease prediction. The study highlights the importance of early diagnosis due to high mortality rates. The hybrid model outperformed traditional methods in accuracy, true positive rate (TPR), and precision. Results indicate this combination approach enhances heart disease prediction and aids clinical decision-making.

It is evident that combining forecasts is a crucial component in contemporary forecasting methods for temporal biomedical datasets, providing notable benefits over using single models. Nevertheless, it is crucial to thoroughly understand the data and the aim of forecasting to create an effective ensemble model. Furthermore, it is essential to employ appropriate evaluation metrics for assessing biomedical temporal forecasts. Advancements in research on efficient combination techniques may arise from the capability to manage large and varied datasets, alongside the development of automatic selection methods that balance expertise and diversity when selecting and combining models for forecasting (Wang et al., 2023).

5.1.3 Coherent forecasting

This type of forecasting a.k.a. hierarchical time series (HTS) represents a set of data sequences organized by aggregation constraints, reflecting many real-world applications in research and industry. Forecasting in such hierarchical structures is challenging and time-consuming due to the need to ensure forecasting consistency among hierarchy levels based on their dimensional attributes, such as geography or product categories. Coherent forecasts are essential, meaning that higher-level forecasts must equal the sum of lower-level forecasts. This coherency requirement adds complexity to the original time series forecasting problem (Sagheer et al., 2021).

For biomedical data scenarios, HTS forecasting is applied in predicting instances similar to emergency medical services (EMS) requirements (Rostami-Tabar and Hyndman, 2024) and mortality rates across various U.S. states (Li and Hyndman, 2021; Li et al., 2024). Forecasting is crucial for EMS as it promotes consistency and synchronized resource allocation, enhancing decision-making processes and leading to better patient outcomes by avoiding the imbalance between demand and resources. In mortality rate predictions, forecasting addresses differences in mortality patterns across different geographic regions. Maintaining adherence between state-level and national-level mortality forecasts is vital for precise policy planning and resource management, aiding in reducing life expectancy disparities and enhancing public health results.

Different reconciliation procedures like top-down, bottom-up, and middle-out have been developed to maintain consistency across levels by generating base forecasts and then adjusting them. These procedures vary in approach: bottom-up starts from the lowest level and aggregates upwards, top-down begins at the highest level and disaggregates downwards, and middle-out combines both methods starting from an intermediate level. Each has its strengths and weaknesses, and none has proven universally superior. Hyndman et al. (2011) proposed an optimal combination approach, which independently forecasts all levels and then combines them using regression to ensure coherence. The Minimum Trace (MinT) method (Wickramasuriya et al., 2018) is another widely adopted approach for reconciliation. This technique uses the complete covariance matrix of forecast errors to generate a set of coherent forecasts. It aims to minimize the MSE of these coherent forecasts across the whole series, under the assumption of unbiasedness.

The approach detailed by Rostami-Tabar and Hyndman (2024) involves implementing forecast reconciliation for the hierarchical data of ambulance demand. It utilizes an ensemble of models: Exponential Smoothing State Space model (ETS), Poisson regression with Generalized Linear Model (GLM), and time series GLM (TSGLM). It generates base forecasts independently for each hierarchy level and reconcile them using the MinT method, minimizing forecast variances for coherence. Validation is done via time series cross-validation, with accuracy measured by mean absolute scaled error (MASE) and continuous ranked probability scores (CRPS). The methodology by Li and Hyndman (2021) ensures coherent mortality forecasts using a forecast reconciliation approach. Independent state-level forecasts are generated with the Lee-Carter model and then reconciled using the Minimum Trace (MinT) method together with the sampling approach by Jeon et al., (2019) to ensure consistency with national-level forecasts. Validation is performed using out-of-sample forecasting, with accuracy measured by MAPE and the Winkler score. The study uses U.S. mortality data from 1969 to 2017 and projects rates up to 2027. Another paper by Li et al. (2024) uses boosting with stochastic mortality models as weak learners. The authors extend gradient boosting with age-based and spatial shrinkage, iteratively fitting the Lee-Carter model to residuals and adding graph Laplacian-based penalties to align forecasts of adjacent age groups and states. Validation uses US male mortality data (1969–2019), with forecasting performance assessed using MASE.

Traditionally, methods like ARIMA and exponential smoothing generate base forecasts but fail to capture individual and grouped time series dynamics, especially with time variation or sudden changes. They also struggle with exploiting complete hierarchical information, affecting forecasting efficiency. Recently, ML algorithms like artificial neural networks, extreme gradient boosting, and SVR have been employed to improve accuracy by considering nonlinear relationships and dynamic changes. However, they often still rely on traditional methods and may overlook useful hierarchical information. Overall, HTS forecasting remains a complex problem with ongoing research aimed at finding more efficient and accurate methods to ensure coherent and reliable forecasts across all levels of the hierarchy (Sagheer et al., 2021). Note: A list and description of open source tools for forecasting is provided in the Supplementary Material of this article.

5.2 Future directions

Extensive research has been conducted to interrogate biomedical temporal data in medical and health applications. Challenges remain, and are summarized into six key areas: (1) standardizing diverse data formats; (2) managing data quality; (3) ensuring model interpretability; (4) protecting patient privacy; (5) enabling real-time monitoring; and (6) addressing bias to create fair models. To grasp the potential future developments, we present a use case to illustrate six future directions within the clinical context. Specifically, taking Mr. Smith (45 years old) as a persona who is concerned about his risk of developing Alzheimer's Disease (AD).

5.2.1 Data harmonization to standardize data format

Time series analysis plays a critical role in the early detection of AD by enabling the continuous monitoring of specific biomarkers over time. This approach is crucial for understanding the progression of the disease through its various stages, from preclinical AD to mild cognitive impairment (MCI), and ultimately to dementia. The primary biomarkers used in detecting and monitoring AD include beta-amyloid and tau proteins, which are typically measured in cerebrospinal fluid (CSF), along with imaging biomarkers such as PET scans for assessing beta-amyloid burden and MRI scans for detecting changes in brain volume. These biomarkers are indispensable for identifying the onset and progression of the disease, often before clinical symptoms become evident (Hernandez-Lorenzo et al., 2022).

During his visit to the physician, Mr. Smith is advised to undergo a series of tests, including genetic screening, neuroimaging, and cognitive assessments. These tests generate a diverse array of data types, ranging from genetic biomarkers to neuroimaging data (e.g., MRI scans) and time-series data derived from cognitive assessments. However, the data collected from Mr. Smith originate from multiple sources: a local hospital, a specialized lab for genetic testing, and a cognitive assessment app. To create a unified dataset, data harmonization is necessary, ensuring consistency across different formats, terminologies, and units. Implementing interoperable technologies can greatly facilitate seamless data exchange across disparate healthcare systems. Future research should focus on developing advanced harmonization techniques for time series data to ensure accurate and consistent integration from various sources. Additionally, integrating multi-modal data, such as clinical, genetic, and imaging information, will be crucial for creating personalized prediction models.

5.2.2 Data quality

As Mr. Smith assesses his risk of developing AD, data from various tests play a critical role in forecasting his condition. However, his data may contain missing values due to irregular monitoring, different data collection protocols, or the progression of his condition. Addressing these gaps is crucial for building a reliable predictive model. A promising approach involves filling these gaps and using the missing data as a valuable signal. Missing biomarker readings can be estimated using methods like forward-filling or zero imputation. The model can also incorporate indicators to highlight absent data points, learning from the pattern of missing data. For example, if Mr. Smith's cognitive scores are missing for several months, the model can predict these values and use the absence of scores as a feature. This allows the model to detect patterns that may reveal insights such as health changes or inconsistencies in monitoring.

Ensuring data quality is essential for reliable predictive models in clinical research. Future directions should integrate advanced ML techniques that handle missing data and leverage the temporal patterns surrounding these gaps. By combining models that analyze available data and sequences of missing data, we can improve predictive accuracy, uncover hidden trends, and identify critical periods signaling disease progression. This approach enhances timely, personalized predictions for patients like Mr. Smith.

5.2.3 Interpretability

As Mr. Smith assesses his risk of developing AD, advanced ML models analyzing the biomarkers to identify the intervention strategies become crucial. Current models offer predictive power but often function as "black boxes" making it challenging to understand risk factors and the associated impacts. To address this, interpretability methods are essential to know the factors behind risk predictions. One important future direction on interpretability is to use attention mechanisms that prioritize key biomarkers and time points, focusing on early disease prediction characteristics. For example, attention-based models can highlight critical data points, such as changes in biomarkers that signal the onset of AD.

A significant biomarker decline flagged by the model would make the risk assessment more transparent, aiding the physician's understanding and decisions (e.g., intervention). Alternatively, time-based SHAP (SHapley Additive exPlanations) techniques enhance model prediction transparency by assessing feature importance at specific times. Future work could focus on developing interpretability frameworks in personalized, real-time risk assessments for AD and other conditions, ensuring predictions are accurate and understandable for patients and clinicians.

5.2.4 Data privacy

As Mr. Smith evaluates his risk of AD, the sensitive data gathered requires rigorous privacy safeguards. Data privacy is crucial for legal compliance and maintaining trust in the healthcare system. Sharing sensitive information in research while retaining data utility is challenging. Anonymization is a technique to safeguard against reidentification while maintaining the usefulness of research data. Blockchain technology is another method, providing secure means for sharing data. Federated learning (FL) is also beneficial for collaborative studies, enabling ML models to be trained on Mr. Smith's data locally without the need for centralization, thus decreasing privacy risks. Informed consent is another essential aspect for research purposes. If consent is dynamic, it allows for real-time management, permitting alterations as new research develops. Future directions include implementing these techniques independently or as hybrid frameworks that improve privacy protection without sacrificing research utility. Establishing international standards for these methods is imperative for harmonizing global privacy practices and enhancing security and trust in collaborative research.

5.2.5 Real-time detection

Let's assume, the physician seeing Mr. Smith recommends the use of a wearable device that monitors essential physiological indicators such as sleep patterns and heart rate variability (HRV) to assess his AD risks. Note these devices have already demonstrated potential in identifying early signs of cognitive decline (Saif et al., 2020). With continuous, real-time monitoring, Mr. Smith would be empowered to take proactive actions—such as making lifestyle changes or seeking further medical evaluations—that could potentially delay the progression of the disease. We have observed an emerging trend in health domain to embed wearable devices into regular health surveillance, facilitating the early identification and treatment of AD or other disease conditions. A future direction in predictive modeling is high-fidelity model enabling real-time, or near real-time (e.g., 15 min) detection. Some related research questions include data storage (where data to be stored, cloud or locally), model calibration and fine tuning strategies (e.g., transfer learning).

5.2.6 Bias and fairness

A typical problem in AI models is the possibility of bias if they are trained on unrepresentative datasets. For example, if a model is trained mainly on data from old Asian females, it might inaccurately evaluate Mr. Smith, who is a middle-aged American male. Future directions for utilizing AI-driven models should emphasize making these models unbiased and dependable for various populations. A critical measure is the creation and validation of AI models with datasets that include a broad spectrum of demographics, such as different ages, ethnicities, and genders. Another approach to ensure fairness in AI algorithms is through regular audits and validation by independent experts. These audits can uncover and fix biases that could distort predictions. Independent audits help guarantee that AI models are equitable and effective for diverse groups thereby offering reliable health assessments. Additionally, it is essential for both healthcare providers and patients to recognize the potential biases in AI tools. By carefully reviewing AI-generated advice alongside clinical expertise and other diagnostic tools, healthcare providers can ensure that the AI model's predictions are accurate and contextual.

6 Conclusion

In summary, the review paper outlines the challenges faced in predictive modeling for biomedical temporal data, such as managing missing values, addressing correlations between variables, capturing both short-term and long-term dependencies, performing multi-step ahead predictions, and considering data availability. It assesses models in three categories—statistical, machine learning, and deep learning—to evaluate their effectiveness in forecasting data amidst these challenges. Recognizing limitations in each approach, it discusses alternative methods like model enhancements or ensemble/combination forecasting techniques to potentially improve forecasting accuracy. The review also covers hierarchical forecasting for biomedical datasets with relevant structures. Moreover, it explores issues like data quality, privacy concerns, data harmonization, interpretability, real-time detection, and bias/fairness considerations in integrating AI or ML into clinical practices. These challenges underline the necessity for thorough data evaluation, strong privacy laws, and a deep understanding of the goals of predictive modeling. Moreover, successfully implementing these models necessitates a joint effort from the different fields, along with an inclusive approach that tackles not just the technical aspects of the model but also the broader ethical and fairness issues in healthcare environments.

Author contributions

AP: Conceptualization, Formal Analysis, Investigation, Writing–original draft, Writing–review and editing. FC: Writing–original draft, Writing–review and editing. FA-H: Writing–review and editing. TW: Conceptualization, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. We extend our heartfelt appreciation to the National Science Foundation (NSF)—Partnership for Innovation—"Avoiding Kidney Injuries" (Award #2122901) and CPS: TTP Option: Medium: i-HEAR: immersive Human-On-the-Loop Environmental Adaptation for

Stress Reduction (Award #2038905) for their generous support of this work.

Acknowledgments

“Writefull” Version 2.2.0 with Overleaf is utilized for improving the quality of the writing for this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2024.1386760/full#supplementary-material

References

Abbasimehr, H., and Paki, R. (2022). Improving time series forecasting using lstm and attention models. *J. Ambient Intell. Humaniz. Comput.* 13, 673–691. doi:10.1007/s12652-020-02761-x

Albuquerque, G. A., Carvalho, D. D., Cruz, A. S., Santos, J. P., Machado, G. M., Gendriz, I. S., et al. (2023). Osteoporosis screening using machine learning and electromagnetic waves. *Sci. Rep.* 13, 12865. doi:10.1038/s41598-023-40104-w

Al-Hindawi, F., Siddiquee, M. M. R., Wu, T., Hu, H., and Sun, Y. (2024). Domain-knowledge inspired pseudo supervision (dips) for unsupervised image-to-image translation models to support cross-domain classification. *Eng. Appl. Artif. Intell.* 127, 107255. doi:10.1016/j.engappai.2023.107255

Al-Hindawi, F., Soori, T., Hu, H., Siddiquee, M. M. R., Yoon, H., Wu, T., et al. (2023). A framework for generalizing critical heat flux detection models using unsupervised image-to-image translation. *Expert Syst. Appl.* 227, 120265. doi:10.1016/j.eswa.2023.120265

Alhnaity, B., Kollias, S., Leontidis, G., Jiang, S., Schamp, B., and Pearson, S. (2021). An autoencoder wavelet based deep neural network with attention mechanism for multi-step prediction of plant growth. *Inf. Sci.* 560, 35–50. doi:10.1016/j.ins.2021.01.037

Aljuaid, T., and Sasi, S. (2016). “Proper imputation techniques for missing values in data sets,” in *2016 international conference on data science and engineering (ICDSE)*, 1–5. doi:10.1109/ICDSE.2016.7823957

Allam, A., Feuerriegel, S., Rebhan, M., and Krauthammer, M. (2021). Analyzing patient trajectories with artificial intelligence. *J. Med. internet Res.* 23, e29812. doi:10.2196/29812

Al-Qahtani, F. H., and Crone, S. F. (2013). “Multivariate k-nearest neighbour regression for time series data—a novel algorithm for forecasting UK electricity demand,” in *The 2013 international joint conference on neural networks (IJCNN)* (IEEE), 1–8. doi:10.1109/IJCNN.2013.6706742

Altarazi, S., Allaf, R., and Alhindawi, F. (2019). Machine learning models for predicting and classifying the tensile strength of polymeric films fabricated via different production processes. *Materials* 12, 1475. doi:10.3390/ma12091475

Al Zahrani, S., Al Sameeh, F. A. R., Musa, A. C., and Shokeralla, A. A. (2020). Forecasting diabetes patients attendance at al-baha hospitals using autoregressive fractional integrated moving average (arfima) models. *J. Data Analysis Inf. Process.* 8, 183–194. doi:10.4236/jdaip.2020.83011

Bao, Y., Xiong, T., and Hu, Z. (2014). Multi-step-ahead time series prediction using multiple-output support vector regression. *Neurocomputing* 129, 482–493. doi:10.1016/j.neucom.2013.09.010

Barreto, T. d. O., Veras, N. V. R., Cardoso, P. H., Fernandes, F. R. d. S., Medeiros, L. P. d. S., Bezerra, M. V., et al. (2023). Artificial intelligence applied to analyzes during the pandemic: covid-19 beds occupancy in the state of rio grande do norte, Brazil. *Front. Artif. Intell.* 6, 1290022. doi:10.3389/frai.2023.1290022

Baum, L. E., and Eagon, J. A. (1967). *An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology.*

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities* 3, 1–8.

Baum, L. E., and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. statistics* 37, 1554–1563. doi:10.1214/aoms/1177699147

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. statistics* 41, 164–171. doi:10.1214/aoms/1177697196

Baum, L. E., and Sell, G. (1968). Growth transformations for functions on manifolds. *Pac. J. Math.* 27, 211–227. doi:10.2140/pjm.1968.27.211

Bayyurt, L., and Bayyurt, B. (2020). *Forecasting of covid-19 cases and deaths using arima models*, 2020. medrxiv. doi:10.1101/2020.04.17.20069237

Bonilla, E. V., Chai, K., and Williams, C. (2007). Multi-task Gaussian process prediction. *Adv. neural Inf. Process. Syst.* 20.

Bou-Hamad, I., and Jamali, I. (2020). Forecasting financial time-series using data mining models: a simulation study. *Res. Int. Bus. Finance* 51, 101072. doi:10.1016/j.ribaf.2019.101072

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/A:1010933404324

Cai, F., Patharkar, A., Wu, T., Lure, F. Y., Chen, H., and Chen, V. C. (2023). Stride: systematic radar intelligence analysis for adrd risk evaluation with gait signature simulation and deep learning. *IEEE sensors J.* 23, 10998–11006. doi:10.1109/jsen.2023.3263071

Cai, H., Jia, X., Feng, J., Li, W., Hsu, Y.-M., and Lee, J. (2020). Gaussian process regression for numerical wind speed prediction enhancement. *Renew. energy* 146, 2112–2123. doi:10.1016/j.renene.2019.08.018

Cao, L., Fang, H., and Liu, X. (2015). “Multi-step ahead forecasting for fault prognosis using hidden markov model,” in *The 27th Chinese control and decision conference (2015 CCDC)* (IEEE), 1688–1692. doi:10.1109/CCDC.2015.7162191

Chandra, R., Goyal, S., and Gupta, R. (2021). Evaluation of deep learning models for multi-step ahead time series prediction. *IEEE Access* 9, 83105–83123. doi:10.1109/ACCESS.2021.3085085

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* 8, 6085. doi:10.1038/s41598-018-24271-9

Cheng, L.-F., Dumitrascu, B., Darnell, G., Chivers, C., Draugelis, M., Li, K., et al. (2020). Sparse multi-output Gaussian processes for online medical time series prediction. *BMC Med. Inf. Decis. Mak.* 20, 1–23. doi:10.1186/s12911-020-1069-4

Chuah, M. C., and Fu, F. (2007). “Ecg anomaly detection via time series analysis,” in *Frontiers of high performance computing and networking ISPA 2007 workshops: ISPA 2007 international workshops SSDSN, UPWN, WISH, SGC, ParDMCom, HiPCoMB, and IST-AWSN niagara falls, Canada, august 28-september 1, 2007 proceedings 5* (Springer), 123–135.

Daydulo, Y. D., Thamineni, B. L., and Dawud, A. A. (2023). Cardiac arrhythmia detection using deep learning approach and time frequency representation of ecg signals. *BMC Med. Inf. Decis. Mak.* 23, 232. doi:10.1186/s12911-023-02326-w

De Gooijer, J. G., and Hyndman, R. J. (2006). 25 years of time series forecasting. *Int. J. Forecast.* 22, 443–473. doi:10.1016/j.ijforecast.2006.01.001

de Wolff, T., Cuevas, A., and Tobar, F. (2021). Mogptk: the multi-output Gaussian process toolkit. *Neurocomputing* 424, 49–53. doi:10.1016/j.neucom.2020.09.085

Ding, G., Li, X., Jiao, F., and Shen, Y. (2020). *Brief analysis of the arima model on the covid-19 in Italy*, 2020. medRxiv. doi:10.1101/2020.04.08.20058636

Doretto, G., Chiuso, A., Wu, Y. N., and Soatto, S. (2003). Dynamic textures. *Int. J. Comput. Vis.* 51, 91–109. doi:10.1023/A:1021669406132

Du, S., Li, T., Yang, Y., and Horng, S.-J. (2020). Multivariate time series forecasting via attention-based encoder–decoder framework. *Neurocomputing* 388, 269–279. doi:10.1016/j.neucom.2019.12.118

Dürichen, R., Pimentel, M. A., Clifton, L., Schweikard, A., and Clifton, D. A. (2014). "Multi-task Gaussian process models for biomedical applications," in *IEEE-EMBS international Conference on Biomedical and health informatics (BHI) (IEEE)*, 492–495. doi:10.1109/BHI.2014.6864410

Ehsani, R., and Drabløs, F. (2020). Robust distance measures for k nn classification of cancer data. *Cancer Inf.* 19, 1176935120965542. doi:10.1177/1176935120965542

El Mrabet, Z., Sugunaraj, N., Ranganathan, P., and Abhyankar, S. (2022). Random forest regressor-based approach for detecting fault location and duration in power systems. *Sensors* 22, 458. doi:10.3390/s22020458

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res. Oceans* 99, 10143–10162. doi:10.1029/94jc00572

Evgeniou, T., Micchelli, C. A., Pontil, M., and Shawe-Taylor, J. (2005). Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.* 6.

Fan, G.-F., Peng, L.-L., Zhao, X., and Hong, W.-C. (2017). Applications of hybrid emd with pso and ga for an svr-based load forecasting model. *Energies* 10, 1713. doi:10.3390/en10111713

Fang, T., and Lahdelma, R. (2016). Evaluation of a multiple linear regression model and sarima model in forecasting heat demand for district heating system. *Appl. energy* 179, 544–552. doi:10.1016/j.apenergy.2016.06.133

Filipow, N., Main, E., Tanriver, G., Raywood, E., Davies, G., Douglas, H., et al. (2023). Exploring flexible polynomial regression as a method to align routine clinical outcomes with daily data capture through remote technologies. *BMC Med. Res. Methodol.* 23, 114. doi:10.1186/s12874-023-01942-4

Fix, E., and Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: consistency properties. *Int. Stat. Review/Revue Int. Stat.* 57, 238–247. doi:10.2307/1403797

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst. G. B. Irel.* 15, 246–263. doi:10.2307/2841583

Garcia, V., Debreuve, E., Nielsen, F., and Barlaud, M. (2010). "K-nearest neighbor search: fast gpu-based implementations and application to high-dimensional feature matching," in *2010 IEEE international Conference on image processing (IEEE)*, 3757–3760.

Gauss, C.-F. (1823). *Theoria combinations observationum erroribus minimis obnoxiae (Henricus Dieterich).*

Gavrishchaka, V. V., and Banerjee, S. (2006). Support vector machine as an efficient framework for stock market volatility forecasting. *Comput. Manag. Sci.* 3, 147–160. doi:10.1007/s10287-005-0005-5

Ghaderi, H., Foreman, B., Nayebi, A., Tipirneni, S., Reddy, C. K., and Subbian, V. (2023). A self-supervised learning-based approach to clustering multivariate time-series data with missing values (slac-time): an application to tbi phenotyping. *J. Biomed. Inf.* 143, 104401. doi:10.1016/j.jbi.2023.104401

Ghahramani, Z., and Hinton, G. E. (1996). *Parameter estimation for linear dynamical systems.*

Godfrey, L. B., and Gashler, M. S. (2017). Neural decomposition of time-series data for effective generalization. *IEEE Trans. neural Netw. Learn. Syst.* 29, 2973–2985. doi:10.1109/TNNLS.2017.2709324

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R., et al. (2000). *Physiobank, physiotoolkit, and physiome: components of a new research resource for complex physiologic signals.*

Gopakumar, S., Tran, T., Luo, W., Phung, D., and Venkatesh, S. (2016). Forecasting daily patient outflow from a ward having no real-time clinical data. *JMIR Med. Inf.* 4, e25. doi:10.2196/medinform.5650

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F (radar signal processing) (IET)* 140, 107–113. doi:10.1049/ip-f-2.1993.0015

Hamdi, T., Ali, J. B., Di Costanzo, V., Fnaiech, F., Moreau, E., and Ginoux, J.-M. (2018). Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm. *Biocybern. Biomed. Eng.* 38, 362–372. doi:10.1016/j.bbe.2018.02.005

Harerimana, G., Kim, J. W., and Jang, B. (2022). A multi-headed transformer approach for predicting the patient's clinical time-series variables from charted vital signs. *IEEE Access* 10, 105993–106004. doi:10.1109/access.2022.3211334

Haywood, J., and Wilson, G. T. (2009). A test for improved multi-step forecasting. *J. Time Ser. Analysis* 30, 682–707. doi:10.1111/j.1467-9892.2009.00634.x

Helmini, S., Jihan, N., Jayasinghe, M., and Perera, S. (2019). Sales forecasting using multivariate long short term memory network models. *PeerJ Prepr.* 7, e27712v1. doi:10.7287/peerj.preprints.27712v1

Hernandez-Lorenzo, L., Ilundain, I. S., and Rodrigo, J. L. A. (2022). "Timeseries biomarkers clustering for alzheimer's disease progression," in *2022 IEEE international conference on omni-layer intelligent systems (COINS)* (IEEE), 1–7.

Hewamalage, H., Bergmeir, C., and Bandara, K. (2021). Recurrent neural networks for time series forecasting: current status and future directions. *Int. J. Forecast.* 37, 388–427. doi:10.1016/j.ijforecast.2020.06.008

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *Int. J. Forecast.* 20, 5–10. doi:10.1016/j.ijforecast.2003.09.015

Hosseinzadeh, P., Nassar, A., Boubrahimi, S. F., and Hamdi, S. M. (2023). Ml-based streamflow prediction in the upper Colorado river basin using climate variables time series data. *Hydrology* 10, 29. doi:10.3390/hydrology10020029

Huang, J., Ghalamsiah, N., Patharkar, A., Pradhan, O., Chu, M., Wu, T., et al. (2024). An entropy-based causality framework for cross-level faults diagnosis and isolation in building hvac systems. *Energy Build.* 317, 114378. doi:10.1016/j.enbuild.2024.114378

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Comput. statistics and data analysis* 55, 2579–2589. doi:10.1016/j.csda.2011.03.006

Idriss, T. E., Idri, A., Abnane, I., and Bakkoury, Z. (2019). "Predicting blood glucose using an lsm neural network," in *2019 federated conference on computer science and information systems (FedCSIS)*, 35–41. doi:10.15439/2019F159

Janacek, G. (2010). Time series analysis forecasting and control. *J. Time Ser. Analysis* 31, 303. doi:10.1111/j.1467-9892.2009.00643.x

Jeon, J., Panagiotelis, A., and Petropoulos, F. (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. *Eur. J. Operational Res.* 279, 364–379. doi:10.1016/j.ejor.2019.05.020

Jones, P. W., Osipov, A., and Rokhlin, V. (2011). "Randomized approximate nearest neighbors algorithm," in *Proceedings of the national academy of sciences 108*, 15679–15686.

Julier, S. J., and Uhlmann, J. K. (1997). New extension of the kalman filter to nonlinear systems. *Signal Process. Sens. fusion, target Recognit. VI (Spie)* 3068, 182–193. doi:10.1117/12.280797

Jun, E., Mulyadi, A. W., and Suk, H.-I. (2019). "Stochastic imputation and uncertainty-aware attention to ehr for mortality prediction," in *2019 international joint conference on neural networks (IJCNN)*, 1–7. doi:10.1109/IJCNN.2019.8852132

Kalman, R. E. (1963). Mathematical description of linear dynamical systems. *J. Soc. Industrial Appl. Math. Ser. A Control* 1, 152–192. doi:10.1137/0301010

Kandula, S., Yamana, T., Pei, S., Yang, W., Morita, H., and Shaman, J. (2018). Evaluation of mechanistic and statistical methods in forecasting influenza-like illness. *J. R. Soc. Interface* 15, 20180174. doi:10.1098/rsif.2018.0174

Kantz, H., and Schreiber, T. (2004). *Nonlinear time series analysis*, 7. Cambridge University Press. doi:10.1017/cbo9780511755798

Katari, S., Likith, T., Sree, M. P. S., and Rachapudi, V. (2023). "Heart disease prediction using hybrid ml algorithms," in *2023 international conference on sustainable computing and data communication systems (ICSCDS)* (IEEE), 121–125.

Katayama, T. (2005). *Subspace methods for system identification*. Springer. doi:10.1007/1-84628-158-X

Khalique, F., Khan, S. A., Butt, W. H., and Matloob, I. (2020). An integrated approach for spatio-temporal cholera disease hotspot relation mining for public health management in Punjab, Pakistan. *Int. J. Environ. Res. Public Health* 17, 3763. doi:10.3390/ijerph17113763

Lai, G., Chang, W.-C., Yang, Y., and Liu, H. (2018). "Modeling long-and short-term temporal patterns with deep neural networks," in *The 41st international ACM SIGIR conference on research and development in information retrieval*, 95–104. doi:10.1145/3209978.3210006

Lee, J. M., and Hauskrecht, M. (2021). Modeling multivariate clinical event time-series with recurrent temporal mechanisms. *Artif. Intell. Med.* 112, 102021. doi:10.1016/j.artmed.2021.102021

Legendre, A. M. (1806). *Nouvelles méthodes pour la détermination des orbites des comètes: avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805 (Courcier).*

Lewis, F. L. (1986). *Optimal estimation: with an introduction to stochastic control theory. (No Title).*

Li, H., and Hyndman, R. J. (2021). Assessing mortality inequality in the us: what can be said about the future? *Insur. Math. Econ.* 99, 152–162. doi:10.1016/j.insmatheco.2021.03.014

Li, L., Li, H., and Panagiotelis, A. (2024). Boosting domain-specific models with shrinkage: an application in mortality forecasting. *Int. J. Forecast.* doi:10.1016/j.ijforecast.2024.05.001

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., et al. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Adv. neural Inf. Process. Syst.* 32.

Liu, J., Yu, F., and Song, H. (2023). Application of sarima model in forecasting and analyzing inpatient cases of acute mountain sickness. *BMC Public Health* 23, 56. doi:10.1186/s12889-023-14994-4

Liu, Z. (2016). *Time series modeling of irregularly sampled multivariate clinical data*. University of Pittsburgh. Ph.D. thesis.

Liu, Z., and Hauskrecht, M. (2015). Clinical time series prediction: toward a hierarchical dynamical system framework. *Artif. Intell. Med.* 65, 5–18. doi:10.1016/j.artmed.2014.10.005

Liu, Z., Wu, L., and Hauskrecht, M. (2013). "Modeling clinical time series using Gaussian process sequences," in *Proceedings of the 2013 SIAM international conference on data mining* (SIAM), 623–631.

Liu, Z., Zhu, Z., Gao, J., and Xu, C. (2021). Forecast methods for time series data: a survey. *Ieee Access* 9, 91896–91912. doi:10.1109/ACCESS.2021.3091162

Luo, Y., Cai, X., Zhang, Y., Xu, J., and Yuan, X. (2018). Multivariate time series imputation with generative adversarial networks. *Adv. neural Inf. Process. Syst.* 31.

Mahmudimanesh, M., Mirzaee, M., Dehghan, A., and Bahrampour, A. (2022). Forecasts of cardiac and respiratory mortality in tehran, Iran, using arimax and cnn-lstm models. *Environ. Sci. Pollut. Res.* 29, 28469–28479. doi:10.1007/s11356-021-18205-8

Manaris, B., Hughes, D., and Vassilandonakis, Y. (2011). "Monterey mirror: combining markov models, genetic algorithms, and power laws," in *Computer science department, college of charleston, SC, USA, appearred in proceedings of 1st workshop in evolutionary music, 2011 IEEE congress on evolutionary computation (CEC 2011)* (New Orleans, LA, USA (Citeseer)), 33–40.

Martínez, F., Frías, M. P., Pérez, M. D., and Rivera, A. J. (2019). A methodology for applying k-nearest neighbor to time series forecasting. *Artif. Intell. Rev.* 52, 2019–2037. doi:10.1007/s10462-017-9593-z

Mitra, A., and Ashraf, K. (2018). Sepsis prediction and vital signs ranking in intensive care unit patients. *arXiv Prepr. arXiv:1812.06686*. doi:10.48550/arXiv.1812.06686

Mollura, M., Lehman, L.-W. H., Mark, R. G., and Barbieri, R. (2021). A novel artificial intelligence based intensive care unit monitoring system: using physiological waveforms to identify sepsis. *Philosophical Trans. R. Soc. A* 379, 20200252. doi:10.1098/rsta.2020.0252

Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley and Sons.

Moon, J., Kim, Y., Son, M., and Hwang, E. (2018). Hybrid short-term load forecasting scheme using random forest and multilayer perceptron. *Energies* 11, 3283. doi:10.3390/en11123283

Mulyadi, A. W., Jun, E., and Suk, H.-I. (2022). Uncertainty-aware variational-recurrent imputation network for clinical time series. *IEEE Trans. Cybern.* 52, 9684–9694. doi:10.1109/TCYB.2021.3053599

Naemi, A., Mansourvar, M., Schmidt, T., and Wiil, U. K. (2020). "Prediction of patients severity at emergency department using narx and ensemble learning," in *2020 IEEE international Conference on Bioinformatics and biomedicine (BIBM) (IEEE)*, 2793–2799. doi:10.1109/BIBM49941.2020.9313462

Niu, K., Lu, Y., Peng, X., and Zeng, J. (2022). Fusion of sequential visits and medical ontology for mortality prediction. *J. Biomed. Inf.* 127, 104012. doi:10.1016/j.jbi.2022.104012

[Dataset] Overchee, P., and Moor, B. (1996). *Subspace identification for linear system*.

Patharkar, A., Huang, J., Wu, T., Forzani, E., Thomas, L., Lind, M., et al. (2024). Eigen-entropy based time series signatures to support multivariate time series classification. *Sci. Rep.* 14, 16076. doi:10.1038/s41598-024-66953-7

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin philosophical Mag. J. Sci.* 2, 559–572. doi:10.1080/14786440109462720

Pearson, K. (1922). Francis galton 1822-1922: a centenary appreciation, 1–28. doi:10.1037/11161-001

Pearson, K. (2023). "The life, letters and labours of francis galton," in *Scientific and medical knowledge production, 1796-1918* (London, United Kingdom: Routledge), 311–318.

Pillonetto, G., Dinuzzo, F., and De Nicolao, G. (2008). Bayesian online multitask learning of Gaussian processes. *IEEE Trans. Pattern Analysis Mach. Intell.* 32, 193–205. doi:10.1109/TPAMI.2008.297

Pinto, R., Valentim, R., da Silva, L. F., de Souza, G. F., de Moura Santos, T. G. F., de Oliveira, C. A. P., et al. (2022). Use of interrupted time series analysis in understanding the course of the congenital syphilis epidemic in Brazil. *Lancet Regional Health–Americas* 7, 100163. doi:10.1016/j.lana.2021.100163

Plis, K., Bunescu, R., Marling, C., Shubrook, J., and Schwartz, F. (2014). "A machine learning approach to predicting blood glucose levels for diabetes management," in *Workshops at the Twenty-Eighth AAAI conference on artificial intelligence*.

Poloni, F., and Sbrana, G. (2015). A note on forecasting demand using the multivariate exponential smoothing framework. *Int. J. Prod. Econ.* 162, 143–150. doi:10.1016/j.ijpe.2015.01.017

Qi, C., Zhang, D., Zhu, Y., Liu, L., Li, C., Wang, Z., et al. (2020). Sarfima model prediction for infectious diseases: application to hemorrhagic fever with renal syndrome and comparing with sarima. *BMC Med. Res. Methodol.* 20, 243–247. doi:10.1186/s12874-020-01130-8

Quinonero-Candela, J., and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.* 6, 1939–1959.

Rabyk, L., and Schmid, W. (2016). Ewma control charts for detecting changes in the mean of a long-memory process. *Metrika* 79, 267–301. doi:10.1007/s00184-015-0555-7

Rachmat, R., and Suhartono, S. (2020). Comparative analysis of single exponential smoothing and holt's method for quality of hospital services forecasting in general hospital. *Bull. Comput. Sci. Electr. Eng.* 1, 80–86. doi:10.25008/bcsee.v1i2.8

Reyna, M. A., Josef, C., Seyedi, S., Jeter, R., Shashikumar, S. P., Westover, M. B., et al. (2019). "Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019," in *2019 computing in cardiology (cinc)* (IEEE).

Reyna, M. A., Josef, C. S., Jeter, R., Shashikumar, S. P., Westover, M. B., Nemati, S., et al. (2020). Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Crit. care Med.* 48, 210–217. doi:10.1097/CCM.0000000000004145

Roberts, S. (2000). Control chart tests based on geometric moving averages. *Technometrics* 42, 97–101. doi:10.1080/00401706.2000.10485986

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Trans. R. Soc. A Math. Phys. Eng. Sci.* 371, 20110550. doi:10.1098/rsta.2011.0550

Rostami-Tabar, B., and Hyndman, R. J. (2024). Hierarchical time series forecasting in emergency medical services. *J. Serv. Res.*, 10946705241232169. doi:10.1177/10946705241232169

Sagheer, A., Hamdoun, H., and Youness, H. (2021). Deep lstm-based transfer learning approach for coherent forecasts in hierarchical time series. *Sensors* 21, 4379. doi:10.3390/s21134379

Said, A. B., Erradi, A., Aly, H. A., and Mohamed, A. (2021). Predicting covid-19 cases using bidirectional lstm on multivariate time series. *Environ. Sci. Pollut. Res.* 28, 56043–56052. doi:10.1007/s11356-021-14286-7

Saif, N., Yan, P., Niotis, K., Scheyer, O., Rahman, A., Berkowitz, M., et al. (2020). Feasibility of using a wearable biosensor device in patients at risk for alzheimer's disease dementia. *J. Prev. Alzheimer's Dis.* 7, 104–111. doi:10.14283/jpad.2019.39

Samaee, S., and Kobravi, H. R. (2020). Predicting the occurrence of wrist tremor based on electromyography using a hidden markov model and entropy based learning algorithm. *Biomed. Signal Process. Control* 57, 101739. doi:10.1016/j.bspc.2019.101739

Shamout, F., Zhu, T., and Clifton, D. A. (2021). Machine learning for clinical outcome prediction. *IEEE Rev. Biomed. Eng.* 14, 116–126. doi:10.1109/RBME.2020.3007816

Shashikumar, S. P., Stanley, M. D., Sadiq, I., Li, Q., Holder, A., Clifford, G. D., et al. (2017). Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. *J. Electrocardiol.* 50, 739–743. doi:10.1016/j.jelectrocard.2017.08.013

Shih, S.-Y., Sun, F.-K., and Lee, H.-y. (2019). Temporal pattern attention for multivariate time series forecasting. *Mach. Learn.* 108, 1421–1441. doi:10.1007/s10994-019-05815-0

Shukla, S. N. (2017). "Estimation of blood pressure from non-invasive data," in *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (IEEE), 1772–1775. doi:10.1109/EMBC.2017.8037187

Siami-Namini, S., Tavakoli, N., and Namin, A. S. (2019). "The performance of lstm and bilstm in forecasting time series," in *2019 IEEE International conference on big data (Big Data)* (IEEE), 3285–3292.

Sotoodeh, M., and Ho, J. C. (2019). Improving length of stay prediction using a hidden markov model. *AMIA Summits Transl. Sci. Proc.* 2019, 425–434.

Stegle, O., Lippert, C., Mooij, J. M., Lawrence, N., and Borgwardt, K. (2011). Efficient inference in matrix-variate Gaussian models with\iid observation noise. *Adv. neural Inf. Process. Syst.* 24.

Suganya, R., Arunadevi, R., and Buhari, S. M. (2020). *Covid-19 forecasting using multivariate linear regression*.

Sun, J., Xie, J., and Zhou, H. (2021). "Eeg classification with transformer-based models," in *2021 ieee 3rd global conference on life sciences and technologies (lifetech)* (IEEE), 92–93. doi:10.1109/LifeTech52111.2021.9391844

Tandon, H., Ranjan, P., Chakraborty, T., and Suhag, V. (2022). Coronavirus (covid-19): arima-based time-series analysis to forecast near future and the effect of school reopening in India. *J. Health Manag.* 24, 373–388. doi:10.1177/09720634221109087

Tipirneni, S., and Reddy, C. K. (2022). Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Trans. Knowl. Discov. Data (TKDD)* 16, 1–17. doi:10.1145/3516367

Tyralis, H., and Papacharalampous, G. (2017). Variable selection in time series forecasting using random forests. *Algorithms* 10, 114. doi:10.3390/a10040114

Valentim, R. A., Caldeira-Silva, G. J., Da Silva, R. D., Albuquerque, G. A., De Andrade, I. G., Sales-Moioli, A. I. L., et al. (2022). Stochastic petri net model describing the relationship between reported maternal and congenital syphilis cases in Brazil. *BMC Med. Inf. Decis. Mak.* 22, 40. doi:10.1186/s12911-022-01773-1

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science and business media.

Vapnik, V., Golowich, S., and Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. *Adv. neural Inf. Process. Syst.* 9.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. neural Inf. Process. Syst.* 30.

Vellido, A. (2019). Societal issues concerning the application of artificial intelligence in medicine. *Kidney Dis.* 5, 11–17. doi:10.1159/000492428

Veneri, G., Pretegiani, E., Rosini, F., Federighi, P., Federico, A., and Rufa, A. (2012). Evaluating the human ongoing visual search performance by eye tracking application and sequencing tests. *Comput. methods programs Biomed.* 107, 468–477. doi:10.1016/j.cmpb.2011.02.006

Wang, Q., Molenaar, P., Harsh, S., Freeman, K., Xie, J., Gold, C., et al. (2014). Personalized state-space modeling of glucose dynamics for type 1 diabetes using continuously monitored glucose, insulin dose, and meal intake: an extended kalman filter approach. *J. diabetes Sci. Technol.* 8, 331–345. doi:10.1177/1932296814524080

Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2023). Forecast combinations: an over 50-year review. *Int. J. Forecast.* 39, 1518–1547. doi:10.1016/j.ijforecast.2022.11.005

Weerakody, P. B., Wong, K. W., Wang, G., and Ela, W. (2021). A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing* 441, 161–178. doi:10.1016/j.neucom.2021.02.046

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2018). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *J. Am. Stat. Assoc.* 114, 804–819. doi:10.1080/01621459.2018.1448825

Williams, C. K., and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT press.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Manag. Sci.* 6, 324–342. doi:10.1287/mnsc.6.3.324

Wold, H. O. (1948). On prediction in stationary time series. *Ann. Math. Statistics* 19, 558–567. doi:10.1214/aoms/1177730151

Wu, J., Li, Z., and Yang, S. (2021). "Covid-19 dynamics prediction by improved multi-polynomial regression model," in *The 2nd international conference on computing and data science*, 1–7. doi:10.1145/3448734.3450847

Xiao, C., Choi, E., and Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inf. Assoc.* 25, 1419–1428. doi:10.1093/jamia/ocy068

Xie, F., Yuan, H., Ning, Y., Ong, M. E. H., Feng, M., Hsu, W., et al. (2022). Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. *J. Biomed. Inf.* 126, 103980. doi:10.1016/j.jbi.2021.103980

Xu, X., Liu, X., Kang, Y., Xu, X., Wang, J., Sun, Y., et al. (2020). A multi-directional approach for missing value estimation in multivariate time series clinical data. *J. Healthc. Inf. Res.* 4, 365–382. doi:10.1007/s41666-020-00076-2

Yoon, B.-J., and Vaidyanathan, P. (2006). Context-sensitive hidden markov models for modeling long-range dependencies in symbol sequences. *IEEE Trans. Signal Process.* 54, 4169–4184. doi:10.1109/TSP.2006.880252

Yule, G. U. (1927). Vii. on a method of investigating periodicities disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Trans. R. Soc. Lond. Ser. A, Contain. Pap. a Math. or Phys. Character* 226, 267–298. doi:10.1098/rsta.1927.0007

Zeng, A., Chen, M., Zhang, L., and Xu, Q. (2023). Are transformers effective for time series forecasting? *Proc. AAAI Conf. Artif. Intell.* 37, 11121–11128. doi:10.1609/aaai.v37i9.26317

Zhang, B., Ren, H., Huang, G., Cheng, Y., and Hu, C. (2019). Predicting blood pressure from physiological index data using the svr algorithm. *BMC Bioinforma.* 20, 109–115. doi:10.1186/s12859-019-2667-y

Zhang, H., Zou, Y., Yang, X., and Yang, H. (2022). A temporal fusion transformer for short-term freeway traffic speed multistep prediction. *Neurocomputing* 500, 329–340. doi:10.1016/j.neucom.2022.05.083

Zhang, M., Flores, K. B., and Tran, H. T. (2021). Deep learning and regression approaches to forecasting blood glucose levels for type 1 diabetes. *Biomed. Signal Process. Control* 69, 102923. doi:10.1016/j.bspc.2021.102923

Zhao, J., Gu, S., and McDermaid, A. (2019). Predicting outcomes of chronic kidney disease from emr data based on random forest regression. *Math. Biosci.* 310, 24–30. doi:10.1016/j.mbs.2019.02.001

Zhou, B., Liu, S., Hooi, B., Cheng, X., and Ye, J. (2019). Beatgan: anomalous rhythm detection using adversarially generated time series. *IJCAI* 2019, 4433–4439. doi:10.24963/ijcai.2019/616

Zhu, T., Li, K., Chen, J., Herrero, P., and Georgiou, P. (2020). Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. *J. Healthc. Inf. Res.* 4, 308–324. doi:10.1007/s41666-020-00068-2

Zou, Y., Donner, R. V., Marwan, N., Donges, J. F., and Kurths, J. (2019). Complex network approaches to nonlinear time series analysis. *Phys. Rep.* 787, 1–97. doi:10.1016/j.physrep.2018.10.005

frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 20 November 2024
DOI 10.3389/fphys.2024.1389693

Development of continuous warning system for timely prediction of septic shock

Gyumin Kim[1], Sung Woo Lee[2], Su Jin Kim[2], Kap Su Han[2], Sijin Lee[2], Juhyun Song[2]* and Hyo Kyung Lee[1]*

[1]School of Industrial Management Engineering, Korea University, Seoul, Republic of Korea, [2]Department of Emergency Medicine, Korea University Anam Hospital, Seoul, Republic of Korea

OPEN ACCESS

EDITED BY
Ahsan H. Khandoker,
Khalifa University, United Arab Emirates

REVIEWED BY
Alireza Mani,
University College London, United Kingdom
Qinghe Meng,
Upstate Medical University, United States
Jaeyoung Kim,
Korea University, Republic of Korea

*CORRESPONDENCE
Hyo Kyung Lee,
✉ hyokyunglee@korea.ac.kr
Juhyun Song,
✉ songcap97@hotmail.com

RECEIVED 22 February 2024
ACCEPTED 28 October 2024
PUBLISHED 20 November 2024

CITATION
Kim G, Lee SW, Kim SJ, Han KS, Lee S, Song J and Lee HK (2024) Development of continuous warning system for timely prediction of septic shock.
Front. Physiol. 15:1389693.
doi: 10.3389/fphys.2024.1389693

As delayed treatment of septic shock can lead to an irreversible health state, timely identification of septic shock holds immense value. While numerous approaches have been proposed to build early warning systems, these approaches primarily focus on predicting the future risk of septic shock, irrespective of its precise onset timing. Such early prediction systems without consideration of timeliness fall short in assisting clinicians in taking proactive measures. To address this limitation, we establish a timely warning system for septic shock with data-task engineering, a novel technique regarding the control of data samples and prediction targets. Leveraging machine learning techniques and the real-world electronic medical records from the MIMIC-IV (Medical Information Mart for Intensive Care) database, our system, TEW3S (Timely Early Warning System for Septic Shock), successfully predicted 94% of all shock events with one true alarm for every four false alarms and a maximum lead time of 8 hours. This approach emphasizes the often-overlooked importance of prediction timeliness and may provide a practical avenue to develop a timely warning system for acute deterioration in hospital settings, ultimately improving patient outcomes.

KEYWORDS

early warning system, machine learning, sepsis, septic shock, artificial intelligence, time-series, electronic health record

1 Introduction

Early warning of clinical deterioration can provide substantial support for clinicians by facilitating prompt identification of adverse events, allowing for proactive measures or timely interventions (Muralitharan et al., 2021). Accordingly, early warning systems hold immense potential in clinical contexts, particularly where the accurate timing of recognition or treatment is paramount. Of particular interest are sepsis and septic shock, extensively examined in early warning systems due to their elevated mortality rates and diagnostic complexity.Sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection, while septic shock is defined as a subset of sepsis in which underlying circulatory and cellular metabolism abnormalities are profound enough to substantially increase mortality (Singer et al., 2016) and is characterized by hyperlactataemia and hypotension requiring vasopressor therapy (Hotchkiss et al., 2016). While early treatment can improve patient outcomes (Evans et al., 2021), delayed intervention or recurring symptoms can lead to irreversible deterioration (Kumar et al., 2006). Thus, the development of an early warning system for septic shock can play a crucial role in timely treatment and prevention of recurrence.

Recent approaches to early warning systems for septic shock mainly employ data-driven machine learning based methodologies to generate warnings (Henry et al., 2015; Lin et al., 2018; Khoshnevisan et al., 2018; Darwiche and Mukherjee, 2018; Giannini et al., 2019; Liu et al., 2019; Fagerström et al., 2019; Yee et al., 2019; Khoshnevisan and Chi, 2020; Kim et al., 2020; Mollura et al., 2020; Misra et al., 2021; Wardi et al., 2021; Agor et al., 2022), enabling personalized early prediction with high sensitivity and specificity (Muralitharan et al., 2021). Most of these early warning systems aim to screen patients who are highly likely to show septic deterioration before onset as early as possible. These screening systems can be classified into two categories based on the timing of their alarm mechanisms. The first category, which we refer to as the 'left-aligned approach', is centered on making predictions during the initial phase of a patient's admission. In contrast, the second category, termed as the 'right-aligned approach', is designed to forecast septic shock events at a specific duration prior to their actual occurrence. Thus, the 'left-aligned approach' aligns cohort data to the start of each patient's admission, while the 'right-aligned approach' aligns data points to the onset of events or the end of a patient's admission. However, both systems may not be clinically applicable due to their inability to timely identify the risk, as they merely predict if patients would suffer from an adverse event in the future without providing sufficient information regarding the exact time of onset, making it difficult to preemptively prepare for timely actions.

We note that the development of timely early warning systems for clinical deterioration, such as septic shock, necessitates the incorporation of three components: (1) continuous calculation of future risk based on the patient's health status, (2) consideration of the timely adequacy of predictions based on their located time frame, and (3) appropriate evaluation of predictive performance achieved by the system. First, the incapability of alerting continuously restricts the system to making singular predictions, falling short in meeting the requisites of timeliness. Second, in the context of continuous warning systems, the establishment of a precise interval for timely warnings serves not only to accurately gauge the system's predictive performance but also to ensure its effective management. Lastly, given the inherent disparity between a warning system designed to capture the onset of adverse events and a screening system, standard metrics employed in previous approaches may not be able to adequately measure the performance of timely warning systems.

In Table 1, prevailing studies on early warning systems for septic shock are summarized with respect to the three essential components for timeliness. To the best of the authors' knowledge, no current frameworks satisfy all three criteria comprehensively, as most have been developed with a focus on screening rather than continuous monitoring. Although some systems leverage machine learning models capable of generating continuous warnings, such as LSTM (Long Short-Term Memory), XGBoost (Extreme Gradient Boosting), or Cox regression, and have set time windows for true warnings, these systems are still evaluated as screening tools rather than continuous warning systems. Specifically, the time windows used to define true warnings typically fall into one of three categories: the entire duration leading up to septic shock onset, the initial period post-admission, or a distant interval before the onset. As a result, despite their ability to produce continuous alerts, these systems are not optimized for issuing timely warnings. Note the difference between screening-based systems and continuous warning systems, as depicted in Figure 1.

While prevailing research on septic shock prediction systems has not adequately addressed the importance of timeliness, other prediction systems for clinical deterioration have recognized its significance (Tomašev et al., 2019; Hyland et al., 2020). Employing various machine learning techniques, these systems were designed to trigger warnings based on real-time risk score calculations for acute kidney injury and circulatory shock. They defined prediction time windows for true warnings and optimized system performance within these windows to ensure an adequate amount of lead time before the onset of deterioration. This acknowledgment of the importance of timeliness underscores its critical role in facilitating effective disease management across various clinical contexts.

Therefore, in this study, we propose a novel approach to develop a clinically applicable early warning system that addresses all aspects of timeliness. We refer to this approach as the 'timeliness focusing approach', which we apply to the development of an early warning system for septic shock, named TEW3S (Timely Early Warning System for Septic Shock). Using the MIMIC-IV (Medical Information Mart for Intensive Care) database (Alistair et al., 2022), we designed TEW3S to generate continuous timely alarms every hour.

2 Materials and methods

2.1 Cohort extraction

In this study, we utilized version 2.0 of the MIMIC-IV database, a comprehensive open-source repository containing de-identified health-related data from patients who underwent intensive critical care at Beth Israel Deaconess Medical Center between 2008 and 2019 (Alistair et al., 2022). The database encompasses records of 53,569 adult ICU patients, comprising a total of 76,943 stays. A wide array of medical information, including demographic details, laboratory findings, vital signs, test results, prescriptions, pharmaceutical information, and diagnoses, were extracted from the database to construct the sequential patient data. Supplementary Tables S1, S2 provide a detailed list of the medical information employed in this research, along with the corresponding MIMIC-IV identifiers or extraction methods. Additionally, Supplementary Table S3 summarizes the average frequency of physiological signals, encompassing vital signs and laboratory results.

Given that septic shock constitutes a subset of sepsis and the diagnosis of sepsis necessitates cohorts with suspected infections (Singer et al., 2016), our study cohorts were defined as patients with suspected infections and sepsis prior to the onset of septic shock. Hence, before selecting the cohorts, we excluded those lacking variables necessary for defining sepsis or septic shock, such as systolic blood pressure (SBP), diastolic blood pressure (DBP), PaO2, FiO2, Glasgow Coma Scale (GCS), bilirubin, platelets, creatinine, and lactate.

Suspected infection was defined for admissions meeting three conditions: (1) received antibiotics, (2) blood culture tests had been

TABLE 1 Summary of previous early warning systems developed for septic shock.

Author (Year)	Continuous warning	Timely range	Performance
Henry et al. (2015) [a]	Capable but evaluated as screening system	Whole time range before the onset of target event	AUROC: 0.83, sensitivity: 0.85, specificity: 0.67, median lead time: 28.2 h
Lin et al. (2018)	Incapable	Within 12 h after admission	AUROC: 0.9411, F1 score: 0.8623, accuracy: 0.8658, recall: 0.8408, precision: 0.8849
Lin et al. (2018)	Incapable	Before 3 h from the onset of event	AUROC: 0.8647, F1 score: 0.7731, accuracy: 0.7747, recall: 0.7676, precision: 0.7931
Khoshnevisan et al. (2018)	Incapable	Within 8 h after admission	AUROC: 0.895, F1 score: 0.808, accuracy: 0.813, recall: 0.787, precision: 0.830, Dataset: EHR from Christiana Care Health System
Khoshnevisan et al. (2018)	Incapable	Before 4 h from the onset of event	AUROC: 0.943, F1 score: 0.868, accuracy: 0.875, recall: 0.826, precision: 0.915
Darwiche and Mukherjee (2018) [a]	Incapable	Before 20 h from the onset of event	Accuracy: 0.8312, sensitivity: 0.7812, specificity: 0.8663
Giannini et al. (2019)	Capable but evaluated as screening system	Whole time range before the onset of target event	Sensitivity: 0.26, specificity: 0.98, PPV: 0.29, NPV: 0.97, median lead time: 5h 25min
Liu et al. (2019) [a]	Capable but evaluated as screening system	Whole time range before the onset of target event	AUROC: 0.93, sensitivity: 0.88, specificity: 0.84, precision: 0.52, median early warning time: 7 h
Fagerström et al. (2019) [a]	Capable but evaluated as screening system	Whole time range before the onset of target event	AUROC: 0.93, median hours before onset: 28.2 h
Yee et al. (2019) [a]	Incapable	Before 24 h from the onset of the event	AUROC: 0.81, sensitivity: 0.79, specificity: 0.66, PPV: 0.46, NPV: 0.90
Khoshnevisan and Chi (2020)	Incapable	Before 48 h from the onset of event	AUROC: 0.793, F1 score: 0.737, accuracy: 0.741, recall: 0.732, precision: 0.737
Kim et al. (2020) [b]	Incapable	At the start of ED admission (warning based on triage information)	AUROC: 0.902, AUPRC: 0.556, sensitivity: 0.706, specificity: 0.900, PPV: 0.427, NPV: 0.967
Mollura et al. (2020) [a]	Incapable	Before 15 min from the onset of event	AUROC: 0.93, F1 score: 0.84, accuracy: 0.85, sensitivity: 0.89, specificity: 0.82, PPV: 0.80, NPV: 0.90
Misra et al. (2021)	Incapable	Within 6 h after admission	AUROC: 0.9483, sensitivity: 0.8392, specificity: 0.8814
Wardi et al. (2021) [c]	Capable but evaluated as screening system	Before 8 h from the onset of the event	AUROC: 0.8, sensitivity: 0.85, specificity: 0.67
Agor et al. (2022) [d]	Incapable	Before 4 h from the onset of event	AUROC: 0.9087, accuracy: 0.8312, recall: 0.7812, precision: 0.8039, specificity: 0.8663

[a] The datasets used in these systems were from the MIMIC-II, or MIMIC-III, databases. While the MIMIC-IV, dataset may share some common cohorts with these earlier versions, the EHR, system schematics were significantly updated in MIMIC-IV, making direct comparisons between the methods of each study and our method challenging.
[b] All performances are those from ensemble (averaging) with baseline predictors only where the target event was the onset of septic shock within 20 h after admission.
[c] Some performances are reported just with lower bound, and specificity is reported only with a graphic, necessitating approximation.
[d] All performances are those from logistic regression with E1 experiment result.

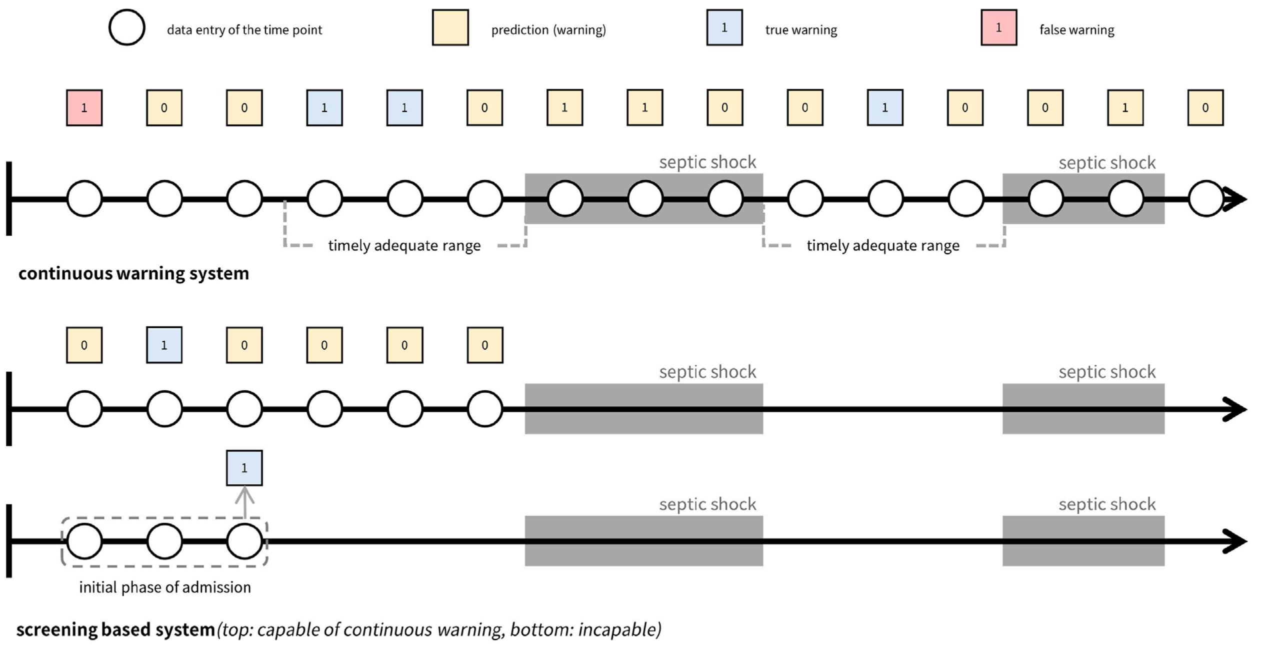


FIGURE 1
Comparison of continuous warning system versus screening system. The top figure illustrates a continuous warning system that can be managed to generate timely adequate warnings, while the figures below show two types of screening systems. Both screening systems make predictions before the event occurs and concentrates on screening whether a patient will suffer from the event in the future or not. The difference between the systems is their capability of generating continuous warnings where the first screening system is capable of generating continuous warnings while the second type is not capable.

taken, and (3) infection-related ICD-9 or 10 codes had been issued. Sepsis was only defined for cohorts with suspected infections, with its onset marked when the Sequential Organ Failure Assessment (SOFA) score reached or exceeded two points. Septic shock was only defined after the onset of sepsis, resulting in the exclusion of cohorts where septic shock occurred prior to sepsis. This decision is based on the assumption that timely prediction is more effective in cases where sepsis precedes septic shock, compared to cases where septic shock occurs prior to sepsis, as early intervention is more likely to have already taken place in the latter instances. Note that this approach can lead to the restriction of our research cohort to patients with nosocomial septic shock, and as such, our predictive model may not be applicable to cases of non-nosocomial septic shock. The onset of septic shock was determined when the lactate level equaled or exceeded 2 mmol/L and vasopressor therapy was administered, given that the definition of septic shock includes hyperlactataemia and vasopressor therapy (Hotchkiss et al., 2016).

2.2 Data refinement

We refined the data through several steps, including unit unification, outlier removal, adjustment of time errors, and correction of variable-specific errors. Initially, unit unification was applied to variables with measurements in different units, such as height, weight, temperature, vasopressors, and fluids. We consulted with professional clinicians to establish outlier criteria, drawing on the guidelines from (Hyland et al., 2020), as detailed in Supplementary Table S4. This ensured alignment with both theoretical considerations and practical feasibility in clinical settings. For instance, heart rate values were accepted within the range of 0–300 beats per minute, as values below 0 are theoretically impossible, and values above 300 are extremely rare in clinical practice. Entries falling outside these criteria were identified as errors and subsequently removed.

Time errors, defined as data entries assigned to a patient sequence with timestamps incongruent with the sequence, were adjusted. Entries recorded more than 2 days before admission or 2 days after discharge were deleted. For variables with specific timestamps, such as lab values, entries recorded outside the interval between ICU admission and discharge were excluded. Conversely, for variables recorded continuously, such as pharmaceutical variables and ventilator data, entries with start and end times outside the admission-to-discharge interval were omitted.

Additionally, errors specific to GCS and urine output were addressed. GCS comprises three component variables (eye, motor, and verbal), and its calculation relies on the summation of these components, necessitating consistency in the recorded timing of each variable. For every timestamp of the GCS components, we assumed all GCS information were recorded but some random missing entries could occur. To handle missing values, we employed a forward-and-backward imputation strategy. Urine output calculations involve unique variables, including irrigant in and out values. To accurately measure urine output at specific time points, we subtracted cumulative irrigant in amounts from cumulative irrigant out amounts. For irrigant out values immediately followed by irrigant in values, we recorded the cumulative sum of irrigant out values minus the cumulative sum of irrigant in values, retaining these records for further preprocessing. In cases where no irrigant in values preceded irrigant out values, we assigned the irrigant in value as 0.

2.3 Sequential merging and resampling

As some data entries were distributed across distinct datasets using different identifiers despite representing the same variable, merging the data into a sequential representation was necessary. We consolidated data entries in the MIMIC-IV datasets according to their respective variables. For entries categorized as vital signs, lab results, height, and weight, all values and corresponding timestamps were collated into a unified sequential timeline for each variable. Pharmaceutical instances were aggregated based on shared timestamps, while administration rates were listed individually. Age at admission and gender were also incorporated into the sequential data. Variables used to define sepsis and septic shock shared identical timestamps and were imputed accordingly based on variable-specific schemes. We adhered to predefined definitions for sepsis and septic shock, excluding cohorts without sepsis and those where sepsis occurred after the onset of septic shock. Subsequently, we performed data resampling, discretizing the concatenated data into predefined time intervals by aggregating or averaging variable values. A 1-h interval was chosen, considering both the dynamic nature of septic shock and the practical frequency of warnings in clinical settings.

For feature engineering, summary statistics were computed within each time interval, including the mean, median, maximum, and minimum values, with the mean serving as the representative value. Additionally, slope features were generated by calculating the difference between values at current and past time points (one, three, and 5 hours prior). These features capture temporal dynamics within and across intervals, facilitating the predictive model's learning process. Features were not derived for unsuitable variables such as age, drug-related items, and ventilator data. In cases where no feature values were available for a given interval, different imputation methods were applied based on the nature of the data. Lab-related features were imputed using backward and forward filling, while vital signs (excluding GCS), height, and weight were linearly interpolated. This choice of imputation methods reflects the typical frequency with which these variables are recorded in clinical practice (See Supplementary Table S3). Lab measurements are usually taken less frequently and sporadically, so forward and backward filling ensures that the last known value is carried forward until a new measurement is available, preserving temporal continuity. In contrast, vital signs are monitored more frequently, allowing for the use of linear interpolation to estimate values between measurements, which assumes a more gradual and consistent change over time. If a variable was not recorded at all across the cohort, all values for that variable were imputed as 0. To distinguish true zero feature values from imputed ones, we appended presence features indicating whether values were filled by imputation (0) or not (1). This approach accounts for the uncertainty of feature values during prediction generation, as proposed in warning systems for acute kidney injury (Tomašev et al., 2019). Finally, we defined sepsis and septic shock and excluded cohorts using the same procedures as in the sequential merging process. The resulting resampled dataset comprised 11,780 stays, of which 4,369 exhibited septic shock. We partitioned the dataset into training (70%), validation (10%), calibration (10%), and test (10%) sets.

2.4 Timeliness focusing approach via data-task engineering

The ultimate aim of our approach was to demonstrate a clinically applicable early warning system via successful integration of timeliness within the development course. In pursuit of such a goal, we introduce a timeliness focusing approach which encompasses three main considerations.

First, to ensure the clinical relevance of our system, we evaluated predictive performance from multiple perspectives. We assessed performance not only on all instances of shock onset but also specifically on the first occurrences of shock, which may hold greater clinical significance. Compared to recurring septic shocks, the first onset of septic shock may be of more clinical value as clinicians may not have been aware of the patient's deteriorating health status. Furthermore, we analyzed performance variations by adjusting the definition of timely warnings through what we termed the 'evaluation window', exploring different time points relative to shock onset to accommodate varying clinical needs. Relative to the septic shock onset time denoted as t = 0, the earliest time point of the evaluation window was defined as t = −8 and the latest time point as t = 0. Varying time points between t = −8 and t = 0 were employed to assess the robustness of our system against the varying needs of specific clinical application contexts.

Second, in addition to standard metrics used in screening systems or machine learning models, we introduced two metrics to measure timeliness: Target Event Recall (TER) and True Alarm Rate (TAR). TER measures the proportion of events warned by timely alarms, while TAR quantifies the fraction of timely warnings among both false and timely warnings. Timely warnings are defined as those occurring within the evaluation window, while false warnings exclude those generated during prolonged septic shock events. Although alarms occurring during prolonged shock events fall outside the evaluation window, they remain critical indicators of ongoing elevated risk and should not be classified as false alarms. Furthermore, we utilized modified versions of TER and TAR, termed 'TER stay' and 'TAR stay', respectively. These metrics provide an average assessment of TER and TAR specifically for stays with septic shock.

As defined, TER and TAR are calculated for individual septic shock events, which may differ from evaluation metrics commonly used in conventional machine learning classification tasks (e.g., True Positive Rate) or those in prevailing screening systems for septic shock. To distinguish these metrics, we term the metrics introduced in this study (TER and TAR) as 'event-based metrics'. In contrast, standard machine learning task metrics evaluate predictions at each time point, while screening system evaluation metrics are computed for each cohort (e.g., the proportion of cohorts adequately predicted). Hence, we classify the conventional evaluation metrics from machine learning

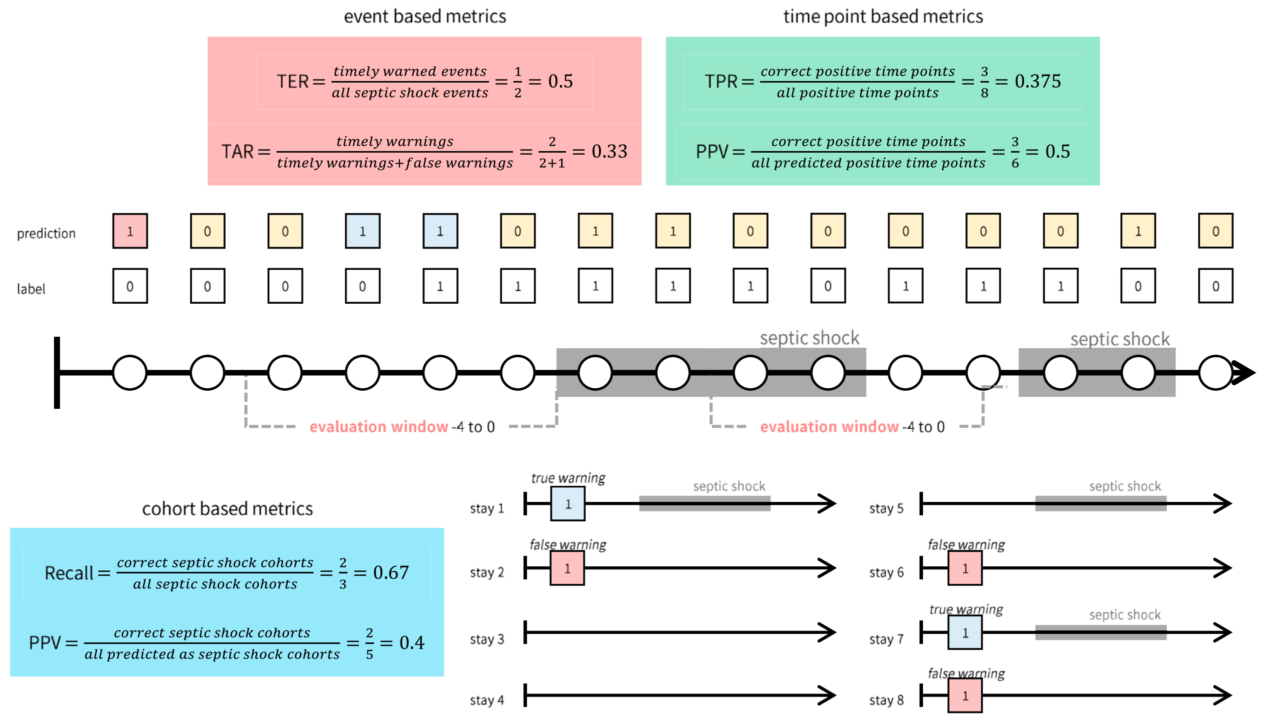

event based metrics
$TER = \frac{timely\ warned\ events}{all\ septic\ shock\ events} = \frac{1}{2} = 0.5$
$TAR = \frac{timely\ warnings}{timely\ warnings + false\ warnings} = \frac{2}{2+1} = 0.33$
time point based metrics
$TPR = \frac{correct\ positive\ time\ points}{all\ positive\ time\ points} = \frac{3}{8} = 0.375$
$PPV = \frac{correct\ positive\ time\ points}{all\ predicted\ positive\ time\ points} = \frac{3}{6} = 0.5$
prediction
1 0 0 1 1 0 1 1 0 0 0 0 0 1 0
label
0 0 0 0 1 1 1 1 1 0 1 1 1 0 0
septic shock
septic shock
evaluation window -4 to 0
evaluation window -4 to 0
cohort based metrics
$Recall = \frac{correct\ septic\ shock\ cohorts}{all\ septic\ shock\ cohorts} = \frac{2}{3} = 0.67$
$PPV = \frac{correct\ septic\ shock\ cohorts}{all\ predicted\ as\ septic\ shock\ cohorts} = \frac{2}{5} = 0.4$
stay 1
true warning
1
septic shock
stay 2
false warning
1
stay 3
stay 4
stay 5
septic shock
stay 6
false warning
1
stay 7
true warning
1
septic shock
stay 8
false warning
1

FIGURE 2
Illustrative example of how event, time point, and cohort based metrics are calculated. While event based metrics (red box) and time point based metrics (green box) are computed based on the continuous warnings generated by the system, cohort based metrics (blue box) can only be computed for each cohort.

tasks as 'time point-based metrics' and those from screening systems as 'cohort-based metrics'. Figure 2 illustrates the differences between these three types of metrics.

Third, to optimize the timeliness of our early warning system, we investigated the impact of various factors on prediction timeliness. These factors encompassed model architecture, deliberate data provision, and the utilization of calibration and oversampling techniques. Our primary focus was placed on data provision, which involves selecting data samples for training and designing prediction tasks with different time windows (prediction window). We termed this approach 'data-task engineering', akin to feature engineering, as it aims to optimize predictive performance by manipulating the relationship between input data and prediction targets. This approach distinguishes itself from traditional machine learning-based early warning systems, where timeliness is often overlooked. Even when considered, the typical methodology involves training models with fixed prediction windows and including all possible data samples in the training set. We hypothesized that each data entry possesses distinct characteristics depending on its relative timing to the onset of target events. Thus, systematic inclusion of data samples can guide the model to learn the intended relationship between input data and target events.

As depicted in Figure 3, data-task engineering encompasses three distinct schemes, each tailored to capture specific correlations between the samples and the prediction tasks. The first scheme involves manipulating the prediction window, adjusting the timeframe from 1 hour to 12 h. This variation alters the nature of the tasks learned by the model. The second scheme centers on restricting the use of data after the onset of septic shock. Specifically, we confine the training data to a window spanning from zero to 2 hours post-onset, termed the 'training window'. Additionally, we consider utilizing all training data samples post-shock onset, labeled as training window 'all'. Lastly, in the third scheme, we experiment with retaining only the data entries around the initial occurrence of septic shock, referred to as 'first shock focus'. Through data-task engineering, we aim to further refine the model's learned function, thus enhancing predictive performance beyond conventional approaches.

Focusing on the importance of timeliness, we devised a comprehensive modeling and validation process. Initially, we trained and validated the system by exploring various combinations of model architecture, data-task engineering schemes, and auxiliary techniques such as oversampling and calibration. We assessed the predictive performance of each combination, aiming to exceed a clinically applicable threshold. This threshold was meticulously determined in consultation with clinical experts and was defined as a TER of 0.9 and a TAR of 0.2 when predicting all shocks within the evaluation window of −8 to 0. Throughout the training and validation process, our primary objective was to improve TER while maintaining a TAR of 0.2. Once combinations surpassing the threshold were identified, we employed an ensemble approach to consolidate these into the final early warning system, TEW3S. This rigorous approach ensured that our system met the clinical requirements for timely detection of septic shock.

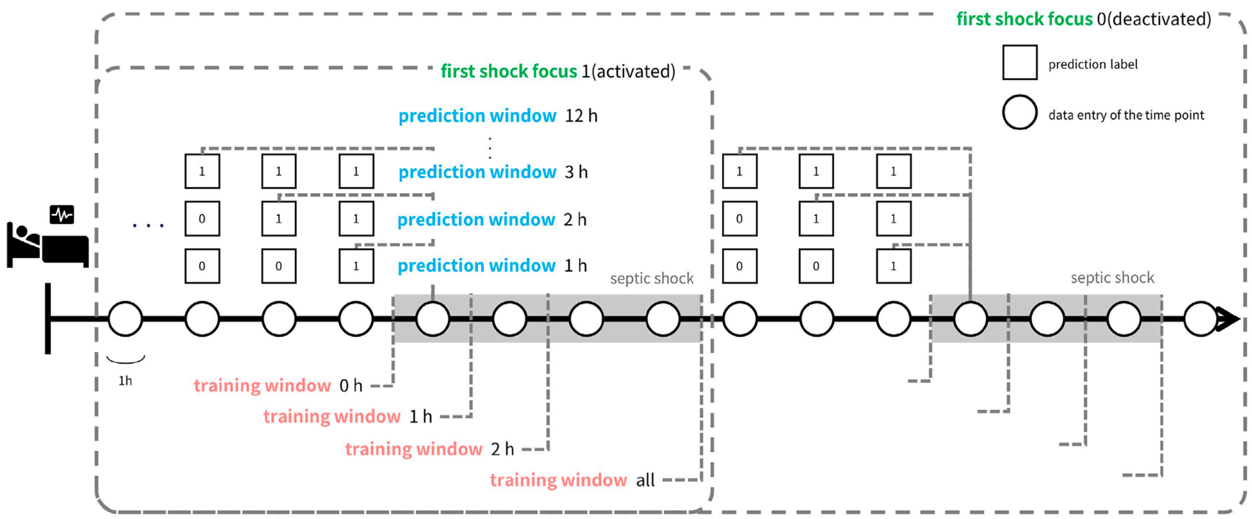


FIGURE 3
Illustration of data-task engineering approach. Data-task engineering involves three schemes: Prediction window (blue text) alters data labels by labeling the time points within prediction window prior to septic shock onset as positive labels. Training window (red text) determines how much time points after shock onset to include in the train set and first shock focus (green text) controls whether to include data points after the first shock onset in the train set. Thus, prediction window is related to the task of the learning objective while training window and first shock focus controls train set data.

2.5 Predictive modeling for TEW3S

TEW3S was developed using supervised machine learning models, including CatBoost (Hancock and Khoshgoftaar, 2020), LightGBM (Ke et al., 2017), XGBoost (Chen and Guestrin, 2016), Random Forest (Breiman, 2001), Logistic Regression (Wright, 1995), Decision Tree (CART) (Lewis, 2000), and Multinomial Naive Bayes (Webb et al., 2010). Our predictive model was designed to generate timely predictions every hour, leveraging current-hour data entries that encompassed not only the mean values but also temporal variability features within and across time steps which were derived through feature engineering, along with presence features to enhance model performance. Throughout the development process of TEW3S, auxiliary techniques such as oversampling and calibration were employed. Oversampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002) and ADASYN (Adaptive Synthetic Sampling) (He et al., 2008) were utilized to balance the class distribution, while isotonic and sigmoid regression were used for calibration of resultant risk score of prediction models. The hyperparameters of each model were set as Supplementary Table S5.

In assessing the timeliness of model predictions, we also calculated time point-based metrics such as the area under the precision-recall curve (AUPRC). AUPRC aided in selecting candidate settings during training and validation, complementing timeliness metrics by capturing the density of warnings within prediction windows. High AUPRC values, coupled with high timeliness, indicated a high true alarm rate, underscoring the importance of incorporating AUPRC in the training and validation process.

The overall training and validation process for TEW3S comprised three main steps. Initially, we trained and evaluated various supervised machine learning models with several training datasets engineered by data-task engineering scheme combinations, selecting those surpassing clinically applicable thresholds of TER and TAR. We further refined our selection based on time point-based AUPRC metric by identifying combinations with AUPRC values that exceeded the average of selected combinations. Note that these combinations consisted of which machine learning model and data-task engineering schemes to use. Subsequently, we applied auxiliary oversampling and calibration techniques to enhance either TER or TAR and diversified the pool of training settings for constructing ensemble models. This phase yielded 31 distinct training settings that met the clinically applicable criteria for an early warning system, defined as TER 0.9 and TAR 0.2. In the final phase, both soft and hard voting ensembles were constructed using the clinically applicable settings, with the hard voting ensemble outperforming the soft voting ensemble. Thus, we identified the hard voting ensemble as our ultimate choice for TEW3S, an early warning system tailored for septic shock. We implemented this predictive modeling process using Python 3.9 and relevant libraries, including Numpy, Pandas, Matplotlib, and Sklearn.

2.6 Calculation of misalignment between event versus cohort and time point-based metrics

To accurately gauge the predictive timeliness of TEW3S, we relied on event-based metrics, namely, TER and TAR. These metrics offer unique advantages over conventional types of metrics, including both cohort-based and time point-based ones, not only in terms of their intrinsic meanings but also based on the results of numerical experiments. This distinction becomes evident when

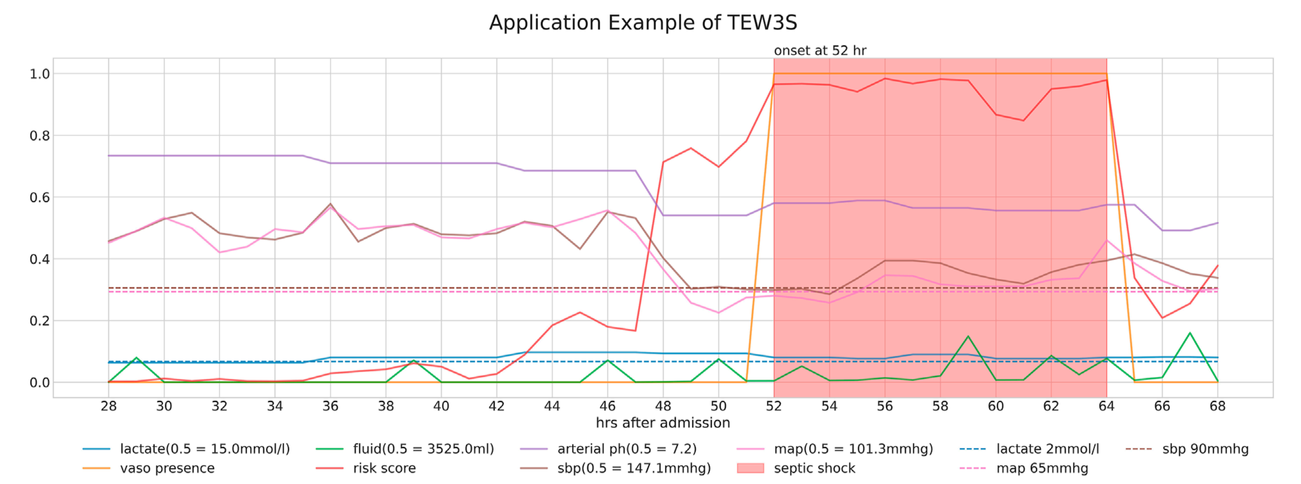


FIGURE 4
Detailed case study of a septic shock patient from our study cohorts, illustrating the sequential information relevant to their condition. The figure displays the patient's risk score, derived through our timeliness-focused approach, alongside various physiological values and intervention-related variables.

examining the discrepancy between event-based metrics and other metrics, which is calculated as follows:

1. Initially, we determined the proportion of training settings, comprising various combinations of models and three data-task engineering schemes, that exhibited clinically applicable timeliness (i.e., TER 0.9 and TAR 0.2 within the evaluation window of −8 to 0). We denote this proportion as 'p' and the set of settings meeting these criteria as 'C'. Note that in this context, settings involving oversampling and calibration techniques were excluded, as these auxiliary techniques were only applied to a subset of the training settings.

2. For the cohort and time point-based metrics which include measures related to both event sensitivity and alarm precision, we calculated the one-p percentile of these metrics.

3. Subsequently, for each cohort- and time point-based metrics, we computed the proportion of settings within set 'C' that failed to achieve the one-p percentile of the respective metric. This proportion represents the level of discrepancy observed in settings that demonstrated high timeliness but attained lower-ranked performance in other metrics.

This calculated discrepancy proportion underscores the indispensable role of TER and TAR in evaluating predictive timeliness effectively.

3 Results

3.1 Predictive performance of TEW3S

Figure 4 provides a detailed case study of a septic shock patient from our study cohorts, illustrating the sequential information pertinent to their condition. It presents the patient's risk score, derived through our timeliness focusing approach, alongside various physiological values and intervention-related variables. Note that the risk score is generated from a model utilized within the TEW3S ensemble.

The temporal changes in these variables and their correlation with the risk score yield insightful observations. During periods of low risk scores, most physiological variables remain stable, except for lactate levels. However, around the 44-h mark, a slight increase in the risk score precedes a subsequent drop in both systolic and diastolic blood pressure (SBP and DBP) approximately 5 hours later. This temporal relationship suggests our approach's potential to predict future events. Subsequently, around the 48-h mark, a significant spike in the risk score coincides with a sharp decline in both SBP and MAP, falling below critical clinical thresholds of 90 mmHg and 65 mmHg, respectively, further validating the physiological plausibility of the risk score's increase. Interestingly, despite poor physiological signs after the 44-h mark, there are instances where the risk score decreases, typically following fluid administration. However, the risk score remains significantly elevated after the onset of septic shock at the 52-h mark, persisting until the end of the observation period, even as other physiological variables return to pre-septic shock levels. This persistence suggests the model's ability to recognize the heightened risk associated with the septic shock state.

Ultimately, our predictive model aims to generate alerts based on the risk score, starting 8 h before the onset of septic shock. The risk score notably begins to rise distinctly from the 44-h mark, precisely 8 h prior to the event, underscoring its predictive capability. Therefore, the TEW3S, resulting from an ensemble of such risk score-based models, demonstrates excellent performance in reflecting both current and future physiological states and the risk of septic shock, offering valuable insights for clinical practitioners. Given the exemplary predictive performance demonstrated in this single case, we further evaluate the predictive accuracy of the hard ensemble-based model TEW3S across all patient stays. By utilizing a hard voting ensemble of models that surpass the predefined threshold, TEW3S achieved a strong performance of TER of 0.9403 and TAR of 0.2018 when predicting all shocks within the evaluation window of −8 to 0. The detailed predictive performance of the early warning system is presented in Table 2. In summary, TEW3S accurately identified 94.0% of septic shock onsets and 93.1% of first septic shock onsets within an 8-h window, with an average of one true alarm for every four false alarms. Additionally, TAR stay, representing

TABLE 2 Predictive performances of TEW3S in evaluation window −8 to 0.

Evaluation metric	All shock	First shock
TER	0.9403	0.9314
TAR	0.2018	0.1784
TER Stay	0.9314[a]	0.9347
TAR Stay	0.4305	0.7717

[a]TER, stay of all shock prediction always equals to TER, of the first shock prediction.

TABLE 3 TER variation in various evaluation windows.

Evaluation Window	−8 to 0	−8 to −1	−8 to −2
TER	0.9403	0.8230	0.7537
Evaluation Window	−7 to 0	−7 to −1	−7 to −2
TER	0.9382	0.8166	0.7452
Evaluation Window	−6 to 0	−6 to −1	−6 to −2
TER	0.9307	0.8049	0.7324
Evaluation Window	−5 to 0	−5 to −1	−5 to −2
TER	0.9254	0.7953	0.7175
Evaluation Window	−4 to 0	−4 to −1	−4 to −2
TER	0.9168	0.7836	0.6962

the average true alarm rate among septic shock cohorts, reached 0.43 for predicting all shocks and 0.77 for predicting the first shock in the evaluation window of −8 to 0.

To assess TEW3S's effectiveness in clinical settings, we analyzed the number of septic shock events predicted by timely alarms before clinicians initiated interventions (i.e., treatments for septic shock). We defined the initiation of septic shock intervention as the time point of vasopressor and fluid co-administration. Alarms triggered prior to this intervention start time were considered timely. Remarkably, 49% of septic shock events were anticipated by these timely alarms, implying almost half of septic shock events were identified through timely alarms preceding clinicians' interventions, demonstrating the practical utility of TEW3S in clinical practice.

Additionally, we analyzed TER within different evaluation windows, ranging from −8 to −4 as the earliest time point and −2 to 0 as the latest. The sensitivity analysis results presented in Table 3 illustrate that TEW3S successfully identified over 75% of septic shock events 2 hours prior to onset in the evaluation window starting from −8. Even when considering alarms only within 4 h prior to onset, 91.7% of all septic shock events were accurately predicted in advance. Notably, nearly 70% of septic shock events were timely warned by TEW3S even in the most restrictive evaluation window of −4 to −2, highlighting its robustness across various scenarios.

TABLE 4 Clinical variable level comparison between false negative cases and false positive cases.

Variables	False negative	False positive
MAP (Mean Arterial Pressure, mmhg)	76.75	76.41
Lactate (mmol/l)	1.36	2.42
Arterial pH	7.40	7.37
GCS (Glasgow Coma Scale)	9.98	9.33
Creatinine (mg/dL)	1.56	1.80
Bilirubin (mg/dL)	2.28	3.54
Platelets (K/uL)	217.67	194.41
SOFA (Sequential Organ Failure Assessment)	7.29	8.01

In comparison to existing literature, we further assessed TEW3S's predictive performance using cohort-based metrics, including sensitivity, specificity, precision, accuracy, and the F1 score. Note that as TEW3S was constructed using a hard-voting ensemble, AUROC could not be utilized. Given the focus on timeliness, our main evaluation metrics are event-based metrics (TER and TAR) as they are most appropriate for continuous warning systems. Conventional metrics were utilized for comparison purposes only, as they cannot fully capture the performance of continuous warning systems. In evaluating these metrics, we considered only the initial warning when labeling the entire cohort as positive. Consequently, TEW3S demonstrated a sensitivity of 0.9634, specificity of 0.4818, precision of 0.5230, accuracy of 0.6604, and an F1 score of 0.6779. When compared to previous research (Henry et al., 2015; Liu et al., 2019), which reported sensitivities of 0.85 and 0.88, and specificities of 0.67 and 0.84, respectively, TEW3S's predictive performance aligned closely with these prior approaches. It is worth noting that although TEW3S was primarily designed for superior timeliness rather than optimal screening performance, its effectiveness was comparable to these established models. This suggests that our proposed methodology allows for the development of an early warning system proficient not only in generating timely continuous warnings but also in effectively screening high-risk cohorts.

We further carried out failure case analysis, examining both false alarms and instances where timely alarms were absent. Several rational causes of failures were identified. First, for false alarms, we found that 95% of false alarms were associated with vasopressor, fluid administration, or mechanical ventilation within a 3-h interval. This suggests that most false alarms adequately reflected real patient risk, but subsequent septic shock onsets could have been prevented due to timely treatment by clinicians. Second, for the false negative cases, we observed that 13% of stays without timely warnings experienced a rapid onset of septic shock within 12 h after admission. This indicates that TEW3S may not have had sufficient time to generate timely predictions in these instances. Third, we compared the average values of clinical variables for each failure case: those with false negative cases versus false positive cases (refer to Table 4). The comparison revealed that false warning

TABLE 5 Misalignment proportions of conventional metrics.

Type of metric	Metric	Proportion of discrepancy	Max TER	Max TAR
Cohort Based	AUROC	1	0.92	0.21
Cohort Based	F1-Score	0.91	0.92	0.21
Time Point Based	AUPRC	0.70	0.91	0.21
Time Point Based	F1-Score	0.70	0.91	0.21

cases exhibited a worse health state on average than no warning cases. This difference was particularly notable in lactate levels, where the average lactate value for no warning cases was 1.36, whereas for false warning cases, it was 2.42. The proportion of false negative stays with lactate value below 2, 1.5, and 1.1 were 85.7%, 73.5%, and 26.5%, respectively, implying the predictive capability of TEW3S heavily relies on lactate value. Lastly, we extended the timely adequate ranges used in our evaluation criteria. Given that patient risk of septic shock may extend beyond the current 8-h window prior to onset, we evaluated warnings within 24, 48, and 96 h before septic shock onset. This adjustment led to a decline in the ratios of false negatives and false alarms from 4.2% to 42.4%–2.3% and 36.8%, 2.0% and 33.2%, and 1.6% and 30.0%, respectively. Notably, when considering all warnings prior to septic shock as true positives, consistent with the evaluation criteria of previous early warning systems, the proportions of false negatives and false positives dropped to 1.3% and 24.9%, respectively. This finding underscores the importance of incorporating timeliness metrics into the evaluation process and conducting a comprehensive review of the model's predictive capabilities. In conclusion, the majority of failure cases of TEW3S may be attributed to the mitigation of risk due to timely treatment, the intractability of temporal relationships due to insufficient time before septic shock onsets, and the evaluation criteria that accepts alarms only within 8 h window prior to septic shock onset.

3.2 Misalignment of cohort, time point, and event-based metrics

As aforementioned, disparities can exist between event-based timeliness measures and time point or cohort-based metrics due to their inherent differences, emphasizing the importance of selecting adequate evaluation metrics. To explore this incongruity across various training settings, we conducted a range of analyses.

During the initial phase of the training and validation process, we observed an intriguing pattern: training settings demonstrating clinically applicable prediction timeliness did not necessarily yield commendable performances when assessed using time point or cohort-based metrics. This observation is summarized in Table 5, which outlines the proportion of training settings exhibiting this discrepancy. The proportion was calculated as the ratio of settings in which time point or cohort-based performance ranked lower than the percentile corresponding to clinically acceptable prediction timeliness. This analysis revealed that 70% of clinically applicable settings would not be retained if time point-based metrics or cohort-based metrics were the sole criteria for evaluation. This incongruity was particularly pronounced when considering cohort-based metrics, accounting for 100% of clinically applicable settings. Moreover, the maximum event-based metric value achievable among the misaligned settings, indicated by maximum TER or TAR, underscores the potential pitfalls associated with evaluating early warning systems solely based on time point-based metrics or cohort-based metrics.

To further illustrate the misalignment between timeliness measured by event-based metrics and performance measured by cohort or time point-based metrics, we visualized the correlation of event-based metrics with the other two metrics, as depicted in Figure 5. From the variation of the mean TER in relation to the time point metric, we observed an almost flat or even declining trend in the middle bins, while the mean TAR in relation to the cohort-based metric demonstrated fluctuations at positive predictive values (PPV) lower than 0.4. Additionally, the shaded areas within the figures, indicating the minimum and maximum values of the corresponding TER or TAR, demonstrated a large dispersion of TER and TAR for each binned metric. Overall, the disparity between event-based metrics and other customary metrics was substantial and exhibited considerable variations.

3.3 Variation of prediction timeliness by data-task engineering

Based on several hypotheses regarding the impact of data-task engineering schemes and auxiliary techniques on prediction timeliness, we systematically integrated these factors into the system development process. Initially, we conjectured that factors leading to an increase in positive samples would elevate TER but decrease TAR, given their influence on augmenting the probability of positive instances within the model input distribution. Furthermore, we anticipated that data-task engineering schemes could enhance prediction timeliness while potentially introducing a trade-off between TER and TAR. For example, expanding the prediction window could broaden the model's foresight, potentially leading to heightened overall risk assessment before shock onset. Similarly, extending the training window to include samples during septic shock prolongation might enable the model to discern physiological cues indicative of critical health states, but this could also induce overreliance on these cues. Additionally, training the model using information encompassing the dynamics around every septic shock event might render the model sensitive to predicting all septic shock onsets but less so to first shock events. These scenarios could result in higher TER but reduced TAR, while the last

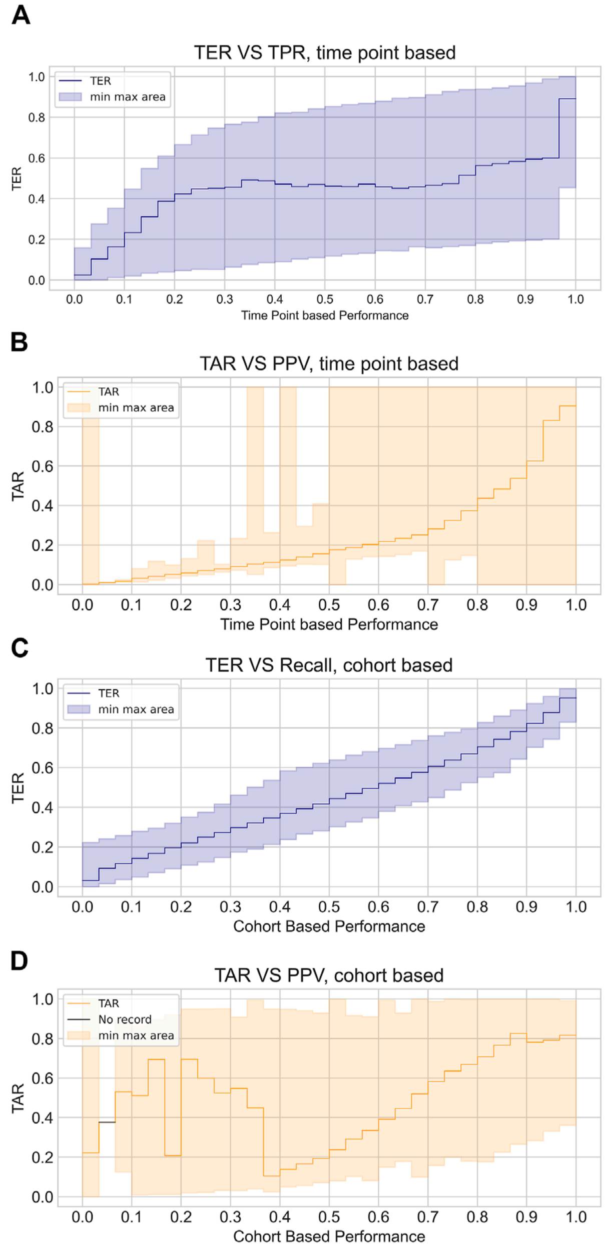


FIGURE 5
Visualized discrepancy between event based metrics and customary metrics. The plots illustrate the average event based performance corresponding to each binned intervals of time point based measures or cohort based measures, where plot **(A)** and **(B)** demonstrate TER and TAR variation with respect to time point based metrics while plot **(C)** and **(D)** show those in relation to cohort based metrics. TER is depicted in blue line and shaded area while TAR is depicted with orange line and shaded area. Line indicates average TER and TAR of each bins and shaded area shows maximum and minimum values of TER and TAR of the corresponding bins.

data engineering scheme can also provoke a trade-off of system performance on early prediction of all shock onsets versus initial onsets.

To numerically validate these hypotheses, we computed the average TER and TAR of each data-task engineering scheme, focusing on the prediction of all shocks within the evaluation window of −8 to 0. Figure 6 presents the average TER (6a, 6b, 5c) and TAR (6d, 6e, 6f) variations along the risk threshold for each setting of the prediction window, training window, and restrictive data usage around septic shock, respectively. The visualized results support our hypotheses regarding the impact of data-task engineering schemes on TER and TAR. Overall, the plots depict a tendency where wider prediction windows and training windows, as well as using all septic shock events as training samples, tend to raise TER but decrease TAR. Moreover, the TAR variation averaged by prediction window peaked at higher thresholds as the corresponding prediction window increased, aligning with the conjecture that widening prediction windows would lead to an increase in risk scores before septic shock onset. Lastly, restricting the training set to the data entries around the first shock onset only enhanced system performances. These results suggest the existence of a trade-off for each data engineering scheme, emphasizing the need for a deliberate exploration of these schemes to achieve an optimal-performing system.

Furthermore, to demonstrate the necessity of data-task engineering, we investigated training settings that achieved high timeliness (TER 0.9 and TAR 0.2 in the evaluation window −8 to 0) using conventional early warning system development approaches. In standard development approaches without considering data-task engineering, the most commonly utilized settings would involve employing the same prediction window as the evaluation window (prediction window of eight when evaluating −8 to 0), using all data samples as the training set including those during shock duration (training window ‘all’) or not using at all (training window 0), and not differentiating between first and recurring septic shock events (first shock focus 0). Notably, there were no training settings that achieved high timeliness with conventional approaches. Even when the TER criterion was reduced to 0.85, only 1.9% of the settings comprised standard schemes. These results underscore the substantial enhancement in timeliness achieved by employing data-task engineering schemes, which were scarcely employed in previous approaches.

4 Discussion

In this study, we developed TEW3S, a continuous early warning system designed to timely identify septic shock by utilizing various machine learning techniques and carefully selecting training samples from the MIMIC-IV dataset. TEW3S successfully predicted 94.1% of all septic shock onsets and 93.1% of first septic shock onsets, providing a lead time of up to 8 hours at a ratio of four false warnings for every true warning. Notably, TEW3S demonstrated a high predictive sensitivity even within highly restricted windows of early warnings, managing to predict more than 75% of septic shock events 2 hours in advance and 91% of septic shock events within a 4-h window. The strong performance of TEW3S under the constraints of a timely adequate range emphasizes the effectiveness of our development approach in constructing a clinically applicable septic shock early warning system. Furthermore, despite TEW3S not originally being designed for screening high-risk patients, it achieved comparable results to previous research studies in this regard. While our system showed significant sensitivity in anticipating septic shock events, a notable number of false warnings

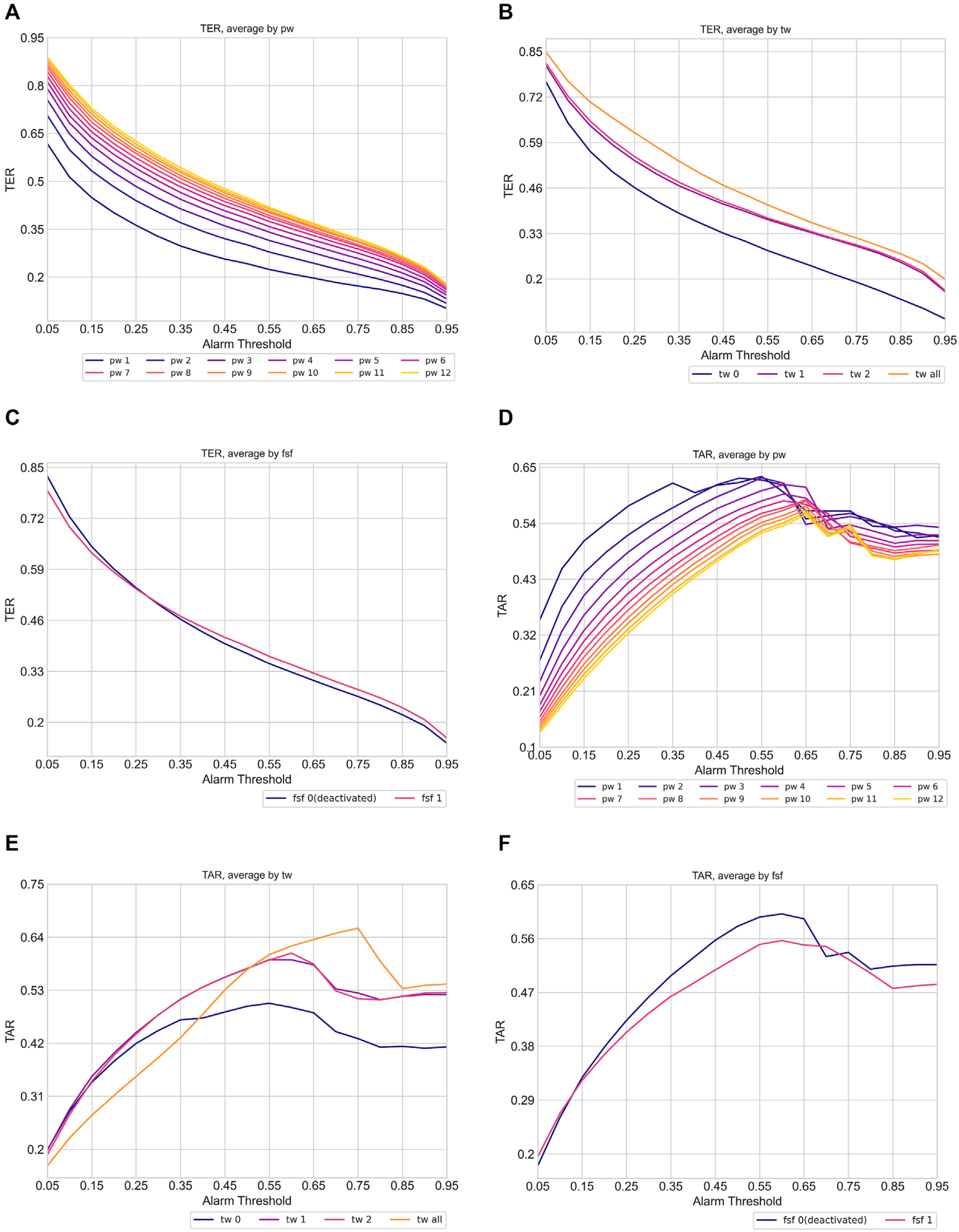


FIGURE 6
Variation of timeliness averaged by data task engineering schemes (pw: prediction window, tw: training window, fsf: first shock focus). The plot **(A)**, **(B)** and **(C)** show TER variation while plot **(D)**, **(E)**, and **(F)** indicate TAR variation. Each line in the plots indicates the average TER and TAR of corresponding data task engineering scheme configuration. The color of the line becomes brighter when corresponding scheme configuration increases.

were generated. However, many of these false alarms were associated with the initiation of interventions such as the administration of vasopressors and fluids, indicating an existing risk at the time of the alarm.

Our study introduced a novel approach focusing on timeliness in early warning system development by incorporating data-task engineering schemes and novel metrics for timeliness assessment. Analyses of timeliness metrics and the impact of data-task engineering on timeliness emphasized the importance of precise metrics for measuring the timeliness of such systems. Therefore, future efforts in developing timely early warning systems should consider data-task engineering schemes and appropriate timeliness metrics as essential components.

The primary limitations of our study stem from the architecture of the TEW3S prediction models and the absence of external validation. Although we employed a diverse array of machine learning models for prediction, we did not explore deep learning models, potentially overlooking architectures that could enhance predictive performance. Furthermore, our system was solely validated using the MIMIC-IV dataset, lacking validation on external databases which is crucial for ensuring the generalizability of our system. Additionally, as this system utilized an ensemble approach to maximize predictive performance, implementing the model in clinical practice may be burdensome. However, it is important to note that our study's primary focus was to propose an approach for constructing a timely early warning system by emphasizing the impact of data-task engineering schemes on timeliness. Future research endeavors could delve deeper into optimizing model architectures specifically geared towards maximizing timeliness, and validate such architectures on external datasets to ensure their robustness and generalizability. Additionally, our analysis of TEW3S failure cases highlighted the association of interventions with false alarms. This suggests potential areas for future research, such as mitigating false alarms by considering intervention information. For instance, one avenue could involve suppressing alarms triggered by moderate risk levels immediately following interventions, thereby refining the system's predictive accuracy.

Despite these limitations, our study remains novel as the first successful approach to building a timely early warning system by implementing prerequisites for timeliness and introducing data-task engineering methods. Our comprehensive analysis of timeliness sheds light on its unique characteristics compared to other types of performance metrics, highlighting the relationship between timeliness and data and task manipulation. Based on these promising results, we believe that our approach holds the potential to become a clinically applicable method for addressing acute deterioration in hospitals, potentially becoming routine clinical practice.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://physionet.org/content/mimiciv/2.0/.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participant's legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

GK: Conceptualization, Methodology, Writing–original draft, Writing–review and editing, Data curation, Formal Analysis, Investigation, Visualization. SwL: Conceptualization, Methodology, Writing–review and editing. SK: Conceptualization, Methodology, Writing–review and editing. KH: Conceptualization, Methodology, Writing–review and editing. SiL: Conceptualization, Methodology, Writing–review and editing. JS: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing–original draft, Writing–review and editing. HL: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was funded in part by the National Research Foundation (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00212713) (received by H. K. L) and the Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0020649, The Competency Development Program for Industry Specialist) (received by H. K. L). This work was also funded in part by the National Research Foundation of Korea through a grant from the Korean government (MSIT) (Grant No. RS-2023-00208807) (received by J. S.). This work was also supported in part by grants from the Korea University Medical Center and Korea University (Grant No. K2211881, K2211061) (received by J. S. and H. K. L.).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer JK declared a shared affiliation with the authors to the handling editor at the time of review.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2024.1389693/full#supplementary-material

References

Agor, J. K., Li, R., and Özaltın, O. Y. (2022). Septic shock prediction and knowledge discovery through temporal pattern mining. *Artif. Intell. Med.* 132, 102406. doi:10.1016/j.artmed.2022.102406

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi:10.1613/jair.953

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Darwiche, A., and Mukherjee, S. (2018). "Machine learning methods for septic shock prediction," in *Proceedings of the 2018 international conference on artificial intelligence and virtual reality*, 104–110.

[Dataset] Alistair, J., Lucas, B., Tom, P., Steven, H., Anthony, C. L., and Mark, R.(2022). Mimic-iv. doi:10.13026/7vcr-e114

Evans, L., Rhodes, A., Alhazzani, W., Antonelli, M., Coopersmith, C. M., French, C., et al. (2021). Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Crit. care Med.* 49, e1063–e1143. doi:10.1097/CCM.0000000000005337

Fagerström, J., Bång, M., Wilhelms, D., and Chew, M. S. (2019). Lisep lstm: a machine learning algorithm for early detection of septic shock. *Sci. Rep.* 9, 15132. doi:10.1038/s41598-019-51219-4

Giannini, H. M., Ginestra, J. C., Chivers, C., Draugelis, M., Hanish, A., Schweickert, W. D., et al. (2019). A machine learning algorithm to predict severe sepsis and septic shock: development, implementation and impact on clinical practice. *Crit. care Med.* 47, 1485–1492. doi:10.1097/CCM.0000000000003891

Hancock, J. T., and Khoshgoftaar, T. M. (2020). Catboost for big data: an interdisciplinary review. *J. big data* 7, 94–45. doi:10.1186/s40537-020-00369-8

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks* IEEE world congress on computational intelligence Ieee, 1322–1328.

Henry, K. E., Hager, D. N., Pronovost, P. J., and Saria, S. (2015). A targeted real-time early warning score (trewscore) for septic shock. *Sci. Transl. Med.* 7, 299ra122. doi:10.1126/scitranslmed.aab3719

Hotchkiss, R. S., Moldawer, L. L., Opal, S. M., Reinhart, K., Turnbull, I. R., and Vincent, J.-L. (2016). Sepsis and septic shock. *Nat. Rev. Dis. Prim.* 2, 1–21. doi:10.1038/nrdp.2016.45

Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., et al. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Med.* 26, 364–373. doi:10.1038/s41591-020-0789-4

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: a highly efficient gradient boosting decision tree. *Adv. neural Inf. Process. Syst.* 30.

Khoshnevisan, F., and Chi, M. (2020). "An adversarial domain separation framework for septic shock early prediction across ehr systems," in *2020 IEEE international conference on big data (big data)* (IEEE), 64–73.

Khoshnevisan, F., Ivy, J., Capan, M., Arnold, R., Huddleston, J., and Chi, M. (2018). "Recent temporal pattern mining for septic shock early prediction," in *2018 IEEE international conference on healthcare informatics (ICHI)* (IEEE), 229–240.

Kim, J., Chang, H., Kim, D., Jang, D.-H., Park, I., and Kim, K. (2020). Machine learning for prediction of septic shock at initial triage in emergency department. *J. Crit. care* 55, 163–170. doi:10.1016/j.jcrc.2019.09.024

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit. care Med.* 34, 1589–1596. doi:10.1097/01.CCM.0000217961.75225.E9

Lewis, R. J. (2000) "An introduction to classification and regression tree (cart) analysis," in *Annual meeting of the society for academic emergency medicine in San Francisco, California (Citeseer)*, 14.

Lin, C., Zhang, Y., Ivy, J., Capan, M., Arnold, R., Huddleston, J. M., et al. (2018). "Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-lstm," in *2018 IEEE international conference on healthcare informatics (ICHI) (IEEE)*, 219–228.

Liu, R., Greenstein, J. L., Granite, S. J., Fackler, J. C., Bembea, M. M., Sarma, S. V., et al. (2019). Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the icu. *Sci. Rep.* 9, 6145. doi:10.1038/s41598-019-42637-5

Misra, D., Avula, V., Wolk, D. M., Farag, H. A., Li, J., Mehta, Y. B., et al. (2021). Early detection of septic shock onset using interpretable machine learners. *J. Clin. Med.* 10, 301. doi:10.3390/jcm10020301

Mollura, M., Romano, S., Mantoan, G., Lehman, L.-w., and Barbieri, R. (2020). "Prediction of septic shock onset in icu by instantaneous monitoring of vital signs," in *2020 42nd annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (IEEE), 2768–2771.

Muralitharan, S., Nelson, W., Di, S., McGillion, M., Devereaux, P., Barr, N. G., et al. (2021). Machine learning–based early warning systems for clinical deterioration: systematic scoping review. *J. Med. Internet Res.* 23, e25187. doi:10.2196/25187

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama* 315, 801–810. doi:10.1001/jama.2016.0287

Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., et al. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 572, 116–119. doi:10.1038/s41586-019-1390-1

Wardi, G., Carlile, M., Holder, A., Shashikumar, S., Hayden, S. R., and Nemati, S. (2021). Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. *Ann. Emerg. Med.* 77, 395–406. doi:10.1016/j.annemergmed.2020.11.007

Webb, G. I., Keogh, E., and Miikkulainen, R. (2010). Naïve bayes. *Encycl. Mach. Learn.* 15, 713–714. doi:10.1007/978-0-387-30164-8_576

Wright, R. E. (1995). Logistic regression.

Yee, C. R., Narain, N. R., Akmaev, V. R., and Vemulapalli, V. (2019). A data-driven approach to predicting septic shock in the intensive care unit. *Biomed. Inf. insights* 11, 1178222619885147, doi:10.1177/1178222619885147

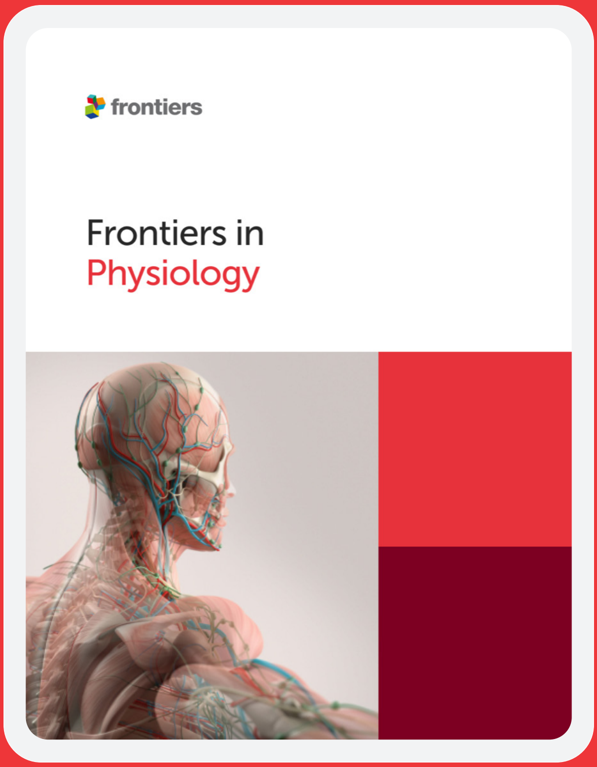
Frontiers in Physiology

Understanding how an organism's components work together to maintain a healthy state

The second most-cited physiology journal, promoting a multidisciplinary approach to the physiology of living systems - from the subcellular and molecular domains to the intact organism and its interaction with the environment.

Discover the latest Research Topics

See more →



frontiers

Frontiers in Physiology

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact