IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 June 2024, accepted 19 August 2024, date of publication 23 August 2024, date of current version 2 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3447759

RESEARCH ARTICLE

Explainable Deep Contrastive Federated Learning System for Early Prediction of Clinical Status in Intensive Care Unit

TRONG-NGHIA NGUYEN[1], HYUNG-JEONG YANG[1], (Member, IEEE), BO-GUN KHO[2], SAE-RYUNG KANG[3], (Member, IEEE), AND SOO-HYUNG KIM[1], (Member, IEEE)

[1]Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea
[2]Pulmonology and Critical Care Medicine, Chonnam National University Hospital, Gwangju 61186, South Korea
[3]Department of Nuclear Medicine, Chonnam National University Hwasun Hospital and Medical School, Hwasun-gun 58128, South Korea

Corresponding authors: Soo-Hyung Kim (shkim@jnu.ac.kr) and Bo-Gun Kho (mdrkbg@gmail.com)

This work was supported in part by the Chonnam National University Hwasun Hospital Institute for Biomedical Science under Grant HCRI23001; and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) through the Artificial Intelligence Convergence Innovation Human Resources Development, Korean Government [Ministry of Science and Information Technology (MSIT)] under Grant IITP-2023-RS-2023-00256629.

ABSTRACT Early identification of patients' clinical status plays a critical role in intensive care unit (ICU) care. The increased adoption of electronic health records (EHRs) in the ICU creates prospects for deep learning (DL) application systems in this discipline. However, monitoring and prediction systems in the ICU encounter problems with security, alarm errors, and interpretation. This research presents deep contrastive federated learning (Deep-CFL), an approach that leverages explainable AI (XAI), CFL, and imbalanced supervised learning techniques to address these problems. CFL introduces an innovative approach to minimize the difference in local and global model prediction ability while increasing the gap in prediction performance of the current local model and its previous model in a communication round. When paired with imbalanced learning, this strategy substantially mitigates error alarm problems while ensuring data security. The XAI technique, specifically integrated gradient, is employed to refine the DL-based model architecture to enhance system interpretability. Extensive experiments and in-depth analyses across three significant clinical datasets highlight the superior performance of Deep-CFL over local and centralized learning-based approaches. The results involving 25, 329 patients admitted to Chonnam National University Hospital reveal that Deep-CFL, with an area under the receiver operating characteristics curve of 0.879, an area under the precision-recall curve of 0.886, and an average precision of 0.884, surpasses systems based on centralized learning while reducing the late alarm rate by up to 10.3%.

INDEX TERMS Deep learning, federated learning, explainable artificial intelligence, intensive care unit, electronic health record.

I. INTRODUCTION

Despite the enhancements in treatment experience and support equipment, mortality rates in the intensive care unit (ICU) have increased over the past 35 years [1]. As the number of ICU admissions has significantly surged, this rate has escalated due to the strain caused by medical excess. In the era of the coronavirus disease 2019 (COVID-19), 29% of ICU mortality patients did not receive mechanical ventilation. Up to 53.2% of individuals requiring ICU care were unable to obtain it due to resource constraints [2]. After examining mortality rates upon ICU admission, studies [3], [4] have concluded that poor outcomes in ICU care are related to resource allocation in overloaded ICUs. These studies have highlighted the importance of advanced algorithmic applications for ICU decision support systems.

Two typical factors that improve ICU care are the monitoring system and the detection system. For the monitoring

The associate editor coordinating the review of this manuscript and approving it for publication was M. Venkateshkumar.

system, the integration of electronic health records (EHRs) [5] in the ICU [6] has created favorable conditions for developing algorithms to detect patients' conditions. For the detection procedure, the rapid application of early detection methods for ICU clinical status, including scoring techniques [7] or the application of artificial intelligence (AI), provides effective support for nurses during the decision process [8]. However, each technique corresponding to the two systems mentioned above is facing its own limitations.

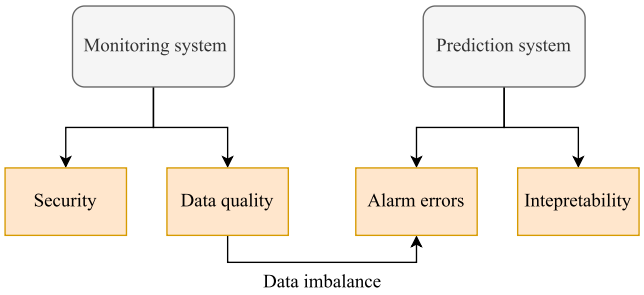


FIGURE 1. Current issues with the application systems in ICU care. While the monitoring system encounters security and data quality problems, the prediction system faces alarming error performance, and lack of interpretability. It is worth noting that data imbalance - a factor in data quality problems - is the direct cause of the alarm error problem.

A. MONITORING SYSTEM'S PROBLEM

Monitoring systems in the ICU are limited by two main issues: security and data quality. Despite the benefits to physicians and healthcare services, EHR adoption is limited by privacy and security regulations [9]. Research [10] has found threats due to violating privacy and security regulations provided by the Health Insurance Portability and Accountability Act in the United States during EHR use. The study [10] highlighted security threats and their influence on the day-to-day operations in a healthcare environment in a closed private network. The economic consequences of security threats are severe, with losses exceeding $7 million for businesses and $13,500 for individuals. These financial repercussions highlight the severe consequences of security breaches. The consequences of security breaches go beyond the people and medical institutions whose information is compromised. Medical care operations are also vulnerable to disruption from numerous risks, including insider threats and phishing attempts.

Data quality plays a significant role in contributing to the security problems surrounding EHR data [11], [12]. Besides the "trade-off" relationship with data security [13], data quality has direct effects on the ability to predict and diagnose in the ICU [14], [15], [16]. Data imbalance is considered a typical problem of data quality. This problem stems from errors in data collection or uneven distribution among patient groups who are admitted to the ICU. Data imbalance occurs when outcomes or classes are overrepresented in a dataset, leading to biased analytical models and decisions.

B. PREDICTION SYSTEM'S PROBLEM

Data imbalance negatively affects the ability of physicians and prediction systems to predict disease conditions in the ICU [17], [18], [19], [20]. Besides, the other problem with prediction systems is interpretability. When delving into the field of AI, especially prediction models applying machine learning (ML) and DL, the above problems are the main factors affecting prediction performance and accuracy [21], [22], [23].

Data imbalance might cause models to acquire biases toward more common outcomes or attributes [24]. This challenge skews the system's predictions, making them less reliable for less common conditions. The inevitable consequence of this process is a late alarm situation leading to untimely intervention in RRT. For interpretability, doctors often view AI models used in healthcare as "black boxes" [25], [26]. This lack of interpretability can limit the confidence and acceptance of AI systems in critical care because physicians must comprehend the reasoning behind AI suggestions to make informed judgments.

C. MOTIVATION AND CONTRIBUTION

From the descriptions in the previous section, this study synthesizes three main problems of effective integrated systems in the ICU: Security, alarm errors for clinical deterioration of patients, and poor interpretability of AI application systems. As mentioned, interpretability problems may limit the confidence and acceptance of AI systems in critical care. The necessity to secure the EHR's security, combined with the challenge of data quality and prediction contexts, impacts the prediction system's performance, resulting in alarm error problems. This study proposes the deep emergency contrastive federated learning (CFL) system (Deep-CFL), an intelligent prediction system that applies explainable AI (XAI) techniques integrated with federated averaging (FedAvg) [27] and contrastive learning [28] to predict patient's clinical status in the ICU.

First, to solve the security problem, the proposed approach is built on the federated learning (FL) framework, which includes multiple healthcare data centers and a central server tasked with aggregating local models into a comprehensive server model for cross-data-center predictions. This structure avoids the requirement of direct data sharing across centers, favoring the interchange of local model weights.

For the alarm error problem, the concept of integrating contrastive learning with FedAvg arose from the continuous findings in FL studies that global models outperform their local equivalents. Therefore, in the process of updating a local model (denoted as M^t at round t), we apply contrastive loss to minimize the performance disparity between it and the global model (denoted as M_g), simultaneously maximize the distinction between M^t and the local model from the previous iteration (denoted as M^{t-1}). This approach attempts to capitalize on the global model's strengths to improve the local model's ability throughout each update cycle. Besides,

imbalance learning is also applied in the system to optimize the model's ability to predict the minority class.

We addressed the interpretability problem by proposing an XAI method called integrated gradient (IG) for determining the structure of the prediction DL model. This approach is a mainstream XAI technique that leverages the concept of axiomatic attribution [29]. The main idea behind IG is to quantify the contribution of individual features to model prediction by systematically integrating gradients in the input space, which provides useful insight into feature importance while also assisting with model interpretability and optimization.

The key contributions of this investigation are:

- Security through Federated Learning: We provide federated learning (FL) infrastructure that includes numerous healthcare data centers and a central server. This topology protects data privacy by avoiding direct data sharing between centers and instead relying on the exchange of local model weights.
- Alarm Error Reduction: By leveraging contrastive learning and FedAvg, our strategy reduces the performance gap between local and global models while increasing the differentiation between subsequent local model updates. This method uses the global model's capabilities to improve local model performance while effectively reducing false and late alerts. Additionally, imbalanced learning techniques are used to improve the prediction of minority classes.
- Improved Interpretability with Explainable AI: We suggest using IG to determine the structure of a deep learning prediction model. IG measures the contribution of individual features to model predictions, revealing feature importance and aiding in model interpretation and optimization.

To our knowledge, this is the first study to simultaneously address the problems of security, alarm errors, and XAI in emergency medicine. The remainder of this study is organized as follows: Section II introduces related research. Section III describes the proposed framework in detail. Section IV describes the datasets, experimental settings, and experimental results. Section V provides discussion, benefits of the application, and limitations. Finally, section VI provides conclusions and future works for the proposed method.

II. RELATED WORKS

This section introduces and describes related methods involved in predicting early clinical deterioration or ICU clinical status. According to the development trend of prediction systems in the ICU, related methods are divided into three categories: traditional clinical status prediction, DL application, and FL application systems.

A. IN-HOSPITAL CLINICAL PREDICTION ALGORITHM

Numerous decision support systems for clinical care have emerged, resulting in substantial advances in emergency medical treatment [30], [31], [32], [33]. The rapid response system (RRS) [34] is a pioneering and representative example in this field. The operating premise of RRS is to monitor, detect, and respond to any indicators of clinical deterioration of the patient to provide timely intervention and avoid cardiac arrest or mortality in the hospital. Much related research has focused on the detection process, involving the introduction and application of numerous algorithms. Traditional approaches include scoring methodologies [35] in which the system frequently depends on basic vital signs to assess the patient's condition using a single "risk score" scale. The Acute Physiology and Chronic Health Evaluation (APACHE) II [36] and III [37] are introduced as scoring systems that assess the severity of disease in critically ill patients. Olsson et al. [38] introduced the Rapid Emergency Medicine Score (REMS) to improve the accuracy of the scoring system in nonsurgical patients regarding in-hospital mortality and length of stay (LOS). REMS is superior to the Rapid Acute Physiology Score (RAPS) [39] in predicting in-hospital mortality in both critically ill patients transferred to the ICU and in the overall sample (AUC 0.910 ± 0.015 for REMS compared to 0.872 ± 0.022 for RAPS). The MEWS is a well-known approach to detecting clinical deterioration and the possible need for higher levels of care [40]. Despite various institutions adopting diverse strategies to evaluate vital signs, MEWS evaluates patients' clinical outcomes using five fundamental vital signs: blood pressure, temperature, respiratory rate, heart rate, and the alert, voice, pain, and unresponsive (AVPU) score, as listed in Table 1. The effectiveness of MEWS was corroborated in a study by Gardner-Thorpe et al. [41] on 334 consecutively

TABLE 1. Evaluate MEWS score based on vital sign indicators.

Vital sign index	Values	MEWS
Blood Pressure	$\leq$ 70 mmHg	+3
	71-80 mmHg	+2
	81-100 mmHg	+1
	101-199 mmHg	0
	$\geq$ 200 mmHg	+2
Heart rate	<40 bpm	+2
	41-50 bpm	+1
	51-100 bpm	0
	101-110 bpm	+1
	111-129 bpm	+2
	$\geq$ 130 bpm	+3
Temperature	<35°C / 95°F	+2
	35–38.4°C / 95–101.1°F	0
	$\geq$ 38.5°C / 101.3°F	+2
Respiratory rate	<9 bpm	+2
	9-14 bpm	0
	15-20 bpm	+1
	21-29 bpm	+2
	$\geq$ 30 bpm	+3
AVPU Score	Alert	0
	Reacts to voice	+1
	Reacts to pain	+2
	Unresponsive	+3

treated patients. In this study, a MEWS score of 4 or above was found to have 75% sensitivity and 83% specificity for the requirement of intensive care. MEWS scores of 5 or higher were less sensitive (38%), but more specific (89%). NEWS [42] has been adopted for ICU prediction with a broader set of variables than the MEWS. The investigations on 440 patients revealed that NEWS reached 0.920 (95% confidence interval [0.890, 0.940]) of the area under the receiver operating characteristics curve (AUROC), with a sensitivity of 93.6% and a specificity of 82.2% to detect early clinical deterioration. Traditional prediction techniques display positive contributions to ICU assessment. However, their inherent predictability is limited by the emphasis on the patient's current survival status, and few studies have focused on the context of future prediction. With the development of AI technology in healthcare, several investigations have concentrated on applying machine learning and DL to predict clinical deterioration, aiming to overcome the limitations of traditional methods in addressing the temporal prediction context.

B. DEEP LEARNING

The deep early warning score (DEWS) was first mentioned in a study by Kwon et al. [43]. Aiming to address problems of low sensitivity and elevated false alarm rates, the authors in this study introduced a DL approach that estimates the probability of events for individual patients, moving beyond conventional risk-scoring techniques. The DEWS uses an RNN structure with a long short-term memory (LSTM) unit to manage the time-series data input. Experiments on 52, 131 patients revealed that the DEWS (AUROC: 0.850 and area under the precision-recall curve (AUPRC): 0.044) outperformed the MEWS (AUROC: 0.603 and AUPRC: 0.003), reducing the number of alarms by 82.2% at the same sensitivity. In the study [44], a more complicated DL structure was applied to the DEWS, with a temporal convolutional network (TCN) used to predict clinical events over the next 1 to 6 h. Experiments studying 4, 713 ICU admission patients from 2014 to 2018 indicate that this method achieves superior results in terms of sensitivity compared to the LSTM and feed-forward network structures. With the emergence of the attention mechanism [45], the DEWS began applying this structure to improve the performance and interpretability of DL-based healthcare systems. Shamout et al. [46] developed an innovative, deep, interpretable end-to-end system that estimates patient event probability using time-series data and Gaussian process regression [47]. Their model architecture incorporated a bidirectional-LSTM encoder and an attention mechanism that generated context vectors from the mean and variance of the input vital signs, resulting in outstanding performance (AUROC: 0.880). This approach outperformed the precision of the previous NEWS (AUROC: 0.866), as indicated by the comparison to the DEWS. The study [48] combined graph neural networks with an attention mechanism to learn the complicated dependencies between multivariate time series, achieving higher performance than DL-based approaches on two large medical datasets. Despite the notable achievements due to the variety of DL structures in the DEWS, these studies were localized to individual healthcare centers, each with unique qualities and characteristics. Thus, a DEWS that exhibits robust performance in one healthcare center might fail in another. This phenomenon is attributed to the generalization problem arising from data heterogeneity in local learning (LL). Addressing this challenge requires collaborative learning between the healthcare centers, allowing for collaborative data optimization across centers to create a global model with consistently high performance across multiple hospitals.

C. FEDERATED LEARNING

In contrast to DL, the application of collaborative learning, specifically FL, is a novel concept in ICU care. The investigation [49] introduces a novel exploration into hospital mortality prediction employing FL. The results demonstrate that FL can match the effectiveness of centralized learning (CL) without necessitating data sharing between hospitals. Furthermore, the investigation provides an empirical comparison of two widely employed model aggregation techniques in the ICU context, FedAvg [27] and FedProx [50], assessing their performance in enhancing prediction accuracy. Personalized one-time local adaptation [51] is proposed to address the problem with the task of predicting hospital mortality in a realistic multicenter ICU EHR database to preserve the original unbalanced and non-IID data distribution. The experimental results demonstrate that personalized, one-time local adaptation effectively improves the prediction performance of FL and significantly reduces the communication rounds, compared with FedAvg and personalized FL with Moreau envelopes [52]. Researchers [53] have examined the performance of various FL algorithms for mortality prediction in a realistic FL environment with FL clients with 'extreme' data distributions, employing real-world data from an integrated research dataset. Overall, FL models outperform the local models of participating hospitals and marginally underperform the 'ideal' CL model. As a prominent FL approach that integrates into ICU care, FL for ICU mortality prediction [54], is proposed as an alternative to CL and LL and is evaluated by introducing FL to the binary classification task of predicting ICU mortality. This study examined multivariate time-series data from the MIMIC III database, with an emphasis on laboratory values and vital signs. This study also examined multivariate time-series data from the MIMIC III database, with an emphasis on laboratory values and vital signs. They compared the performance of four deep sequential models (the fully connected RNN, LSTM, gated recurrent unit (GRU), and one-dimensional convolutional neural network) over varying patient history windows (8, 16, 24, and 48 h) and FL client counts (two, four, and eight). The results reveal that CL and FL produce comparable AUPRC and F1-score measures, demonstrating their efficacy in predicting clinical outcomes.

Most research listed above indicates the capacity of FL to solve data security. Because of the trade-off between security and data quality, FL in related algorithms achieves equivalent but not superior results to CL, which benefits from sharing training data across data centers. This situation motivates this investigation to propose a comprehensive FL application method that ensures confidentiality while providing optimal performance in clinical status prediction tasks.

III. PROPOSED METHODS

This section provides the problem definition and proposed framework for Deep-CFL for predicting ICU clinical status. The main purpose of this study is to develop a DL risk score to be implemented in the ICU. When this score reaches a specific threshold value, it triggers an alarm that alerts the doctors. When the alarm is triggered, doctors access the medical records or go directly to the patient to assess their condition and take the necessary actions. Through this process, the emergency team primarily aims to prevent patients from needing cardiopulmonary resuscitation (CPR) or intubation. They strive to use their skilled techniques to prevent patient mortality if such measures are unavoidable.

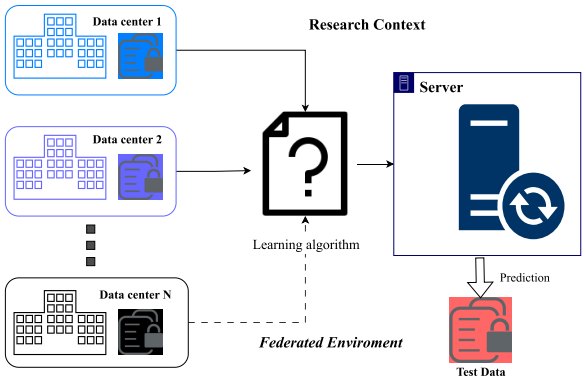


FIGURE 2. Problem definition. This research strategy indeed sets a suitable context for developing a federated learning (FL) system, which is designed to leverage decentralized datasets for model training while preserving data privacy.

A. PROBLEM DEFINITION

This study initially establishes an optimal context for developing a prediction system in a federated environment. Figure 2 illustrates the designed context to develop Deep-CFL. The FL environment comprises N local healthcare centers with their corresponding local datasets D and a central server responsible for aggregating local models (M_i) and updating the global model (M_g). The primary goal is to engineer a learning algorithm that harnesses data from these local data centers and the server to improve the performance of Mg while preserving data privacy. The objective is to solve the following:

$$\arg\min_{M_g}\left(\frac{1}{N}\sum_{i=1}^{N}L_i\left(M_g, D_i\right)\right) \tag{1}$$

where $L_i\left(M_g, D_i\right)$ is the empirical loss of D_i.

Algorithm 1 Deep Contrastive Federated Learning (Deep-CFL) for In-ICU Clinical Status Prediction

Input: N local datasets, initial local model M_i. Number of communication rounds T, temperature τ, weighting coefficient μ, learning rate η.
Output: Updated global model M_g.
Global Initialization: Server initializes the global model M_g.
Begin FL Process
for $t = 0$ to $T - 1$ **do**
  Step 1: Distribute Global Model:
  Server sends M_g^t to all participating local healthcare centers.
  Step 2: Local Model Updates:
  for $i = 1$ to N **do**
    Update local model M_i^t with local data D_i.
    For each input data point $x^{(i)} \in D_i$:
    $z_g \leftarrow M_g(x)$
    $z_t \leftarrow M_i^t(x)$
    $z_{t-1} \leftarrow M_i^{t-1}(x)$
    Contrastive loss:
    $Lc \leftarrow -\log \frac{\exp(\mathrm{sim}(z_t, z_g)/\tau)}{\exp(\mathrm{sim}(z_t, z_g)/\tau)+\exp(\mathrm{sim}(z_t, z_{t-1})/\tau)}$
    Focal loss:
    $L_f\left(p_i\right) \leftarrow -y_i\left(1-p_i\right)^{\gamma}\log\left(p_i\right) - \left(1-y_i\right)p_i^{\gamma}\log\left(1-p_i\right)$
    Imbalance loss:
    $L_i \leftarrow 1 - \frac{2\times\sum_{i=1}^{n} w_i\times y_i\times p_i+\epsilon}{\sum_{i=1}^{n} w_i\times y_i+\sum_{i=1}^{n} w_i\times p_i+\epsilon}$
    Classification loss:
    $L_{\mathrm{sup}} = L_f\left(p_i, y_i\right) + L_i\left(p_i, y_i\right)$
    The **total local loss**:
    $L = L_{\mathrm{sup}}\left(M_i^t; (x^{(i)}, y_i)\right) + \mu \cdot L_c\left(M_i^t; M_i^{t-1}; M_g; x^{(i)}\right)$
  end for
  Step 3: Collect Local Models:
  Local centers send updated models M_i^t back to the server.
  Step 4: Update Global Model using Weighted Averaging:
  $M_g^{t+1} \leftarrow \sum_{i=1}^{N}\frac{|D_i|}{\sum_{j=1}^{N}|D_j|}M_i^t$
end for
End FL Process

B. AN OVERVIEW OF PROPOSED METHOD

We proposed Deep-CFL based on the FedAvg algorithm. Figure 3 presents the overview pipeline of Deep-CFL. First, we defined the general structure of the prediction model (M) on both local healthcare centers and the server. Figure 3A illustrates the architect of the based model. The proposed network has three components: A base encoder (E), a projection head (J), and a classification head (C). For each input x_i of local data i, E extracts the representation z_i from x_i. Refer from [55], projector J is added to map z_i into a space of fixed dimension. The overall function of this step could be

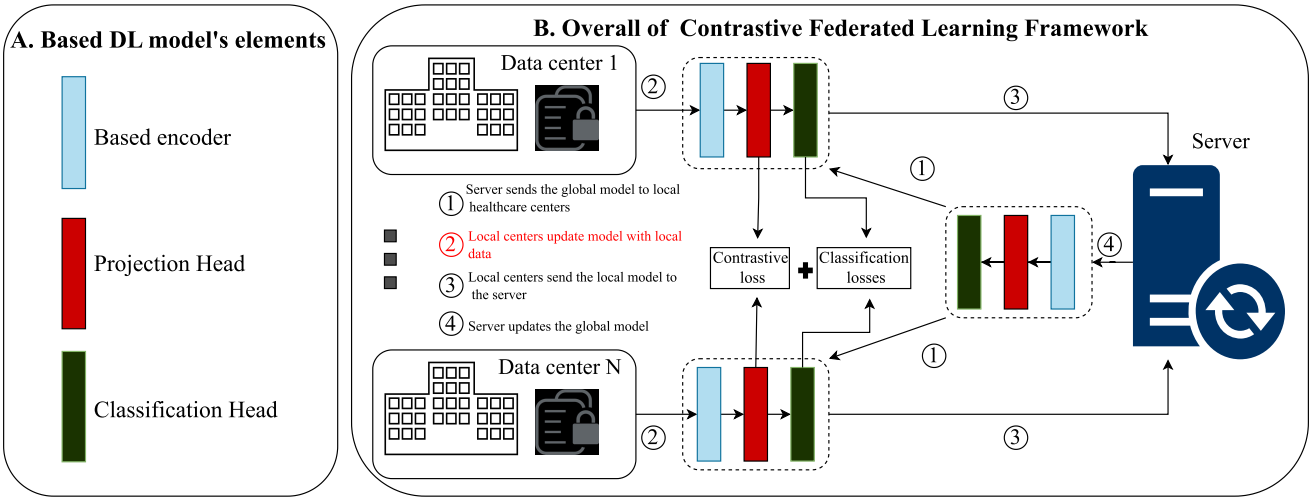


FIGURE 3. Overview of the Deep Contrastive Federated Learning (Deep-CFL) framework. Panel A presents the architecture of the foundational prediction model used within both the local healthcare centers and the central server, including three components: a base encoder for initial data processing, a projection head for dimensionality reduction, and a classification head for outcome prediction. Panel B depicts the four-step FL process: (1) Distribution of the global model from the server to local healthcare centers; (2) Local model updates at each center, incorporating contrastive learning for feature representation optimization and classification losses for accuracy enhancement; (3) Collection of locally updated models by the server; and (4) Global model updating through the averaging of local model weights.

represented as:

$$z_i = J\left(E\left(x_i\right)\right) \tag{2}$$

As regularly supervised learning algorithms, the classification head C uses extracted feature z_i to produce the predicted values for clinical outcomes (y_i).

Algorithm 1 describes the FL workflow as part of the Deep-CFL ICU clinical status prediction. This FL cycle, also illustrated in Figure 3B, unfolds across four steps: In *step (1)*, the server distributes the global model M_g to all participating local healthcare facilities, ensuring that each center uses the most recent version of the model for local training. *Step (2)*, is considered the most important stage, involves the model refining in each local dataset D_i. This refinement entails using contrastive learning on each input data point $x^{(i)}$ to optimize the feature representation $z = J(E(x^{(i)}))$, followed by applying classification loss to improve the prediction accuracy of $y = C(z)$. *Step (3)* involves aggregating the locally updated models back at the server, demonstrating the collaborative component of the Deep-CFL system. In *Step (4)*, the server averages the weights of the obtained local models to update the global model M_g. Updated M_g is then used as the foundation for the next round of FL, gradually improving the model's accuracy and generalizability with each iteration.

More detailed descriptions of the proposed framework components are introduced in the following sections.

C. DEEP LEARNING MODEL ARCHITECTURE

In our exploration of the Deep-CFL framework, we focus on two key case studies that highlight the flexibility and resilience of our approach: (1) determining present clinical status and (2) predicting future clinical status. The structure of the base prediction model M is configured differently depending on the specific needs of each prediction scenario. Figure 5 describes the specific structure of M for each case study.

We configure the structure and parameters of the base models based on two criteria: Highly interpretable and suitable for the prediction context.

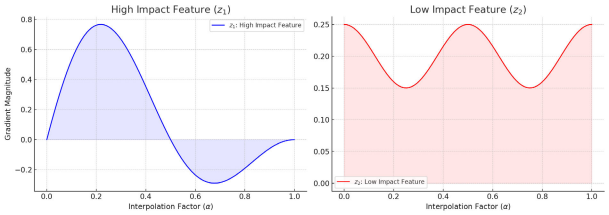


FIGURE 4. Visualization of integrated gradients (IG) for high and low impact features. The left panel shows the gradient magnitude for z_1, which shows significant volatility and a strong influence on the model's output. The usage of a sine wave adjusted by the interpolation factor (α) to decrease towards the end illustrates a genuine scenario where the feature's impact may peak at specific times and subsequently fade, demonstrating its vital role in deciding model predictions. The right panel shows z_2 with low gradient magnitude alterations, indicating a negligible effect on the model's predictions. The soft cosine wave, slightly offset to ensure it remains positive, shows instances in which a characteristic has a consistent but minor influence over its range, emphasizing its relevance in determining the outcome.

1) EXPLAINABLE ARTIFICIAL INTELLIGENCE WITH THE INTEGRATED GRADIENT METHOD

Algorithm 2 describes the base model optimization method using IG. From the initial structure of E, we evaluated the IG value calculated from the extracted representation z to quantify the contributions toward the model predictions. Based on the value of IG, we restructured E to increase the

Algorithm 2 Optimizing Model Components Using Integrated Gradients (IG) for Enhanced Interpretability and Performance

Input: $x, \bar{x}, E, J, C$
Output: Optimized E, J, C
Initialization: Initialize E, J, C
Step 1: Feature Extraction and IG Computation
for x_i in input **do**
  $z_i = J(E(x_i))$
  for z_{ij} in z_i **do**
    Define baseline $\bar{z}_{ij}$
    $\mathrm{IG}_{ij} = (z_{ij} - \bar{z}_{ij}) \times \int_{\alpha=0}^{1} \frac{\partial F(z'(\alpha))}{\partial z_{ij}} d\alpha$
    with $z'(\alpha) = \bar{z}_{ij} + \alpha(z_{ij} - \bar{z}_{ij})$
  end for
end for
Step 2: IG Analysis
for z_{ij} in z_i **do**
  $$\text{Influence}(z_{ij}) = \begin{cases} \text{positive}, & \text{if } \mathrm{IG}_{ij} > 0 \\ \text{negative}, & \text{if } \mathrm{IG}_{ij} < 0 \end{cases}$$
  Note: $|\mathrm{IG}_{ij}|$ indicates influence magnitude
end for
Step 3: Model Optimization
Optimize E, J for interpretability, focus on z_i with high $|\mathrm{IG}_{ij}|$
Optimize C for performance, prioritize z_i with significant IG_{ij}
Step 4: Model Refinement and Evaluation
Refine E, J, C integrating IG insights, evaluate on validation/test data
Return: Optimized E, J, C

interpretability and performance of z for classification tasks. The primary purpose of this process is to solve the black-box problem of DL in medicine through a simple XAI technique. Specifically, the steps in IG value analysis are performed as follows:

- The algorithm begins by defining the initial structures and parameters for the model components E, J, and C, which lay the framework for a thorough feature extraction and evaluation process.
- The model extracts features z_i from input data x_i using the encoder E and projection head J. The IG values for these representations are then computed against a predetermined baseline ($\bar{z}_i$) as a reference point. The IG value of j^{th} representation of i^{th} data point (z_{ij}) is calculated as:

$$\mathrm{IG}_{ij} = \left(z_{ij} - \bar{z}_{ij}\right) \times \int_{\alpha=0}^{1} \frac{\partial F\left(z'(\alpha)\right)}{\partial z_{ij}} d\alpha \qquad (3)$$

  where $\bar{z}_{ij}$ is the baseline value - a reference or starting point for the IG calculation and is often chosen to represent the feature's 'absence' or 'neutral' condition within the context of the model's task. The interpolated feature vector between $\bar{z}_{ij}$ and z_{ij}, denoted as $z'(\alpha) = \bar{z}_{ij} + \alpha \times \left(z_{ij} - \bar{z}_{ij}\right)$. $z'(\alpha)$ enables the approach to consider how the model's prediction changes as the feature value shifts from the baseline to its actual value. The variable α, which ranges from 0 to 1, is used to interpolate between the baseline and actual feature values. As α increases from 0 to 1, $z'(\alpha)$ transitions from $\bar{z}_{ij}$ to z_{ij}, tracing a path along which the gradient of the model's output about the feature z_{ij} is integrated. $\frac{\partial F(z'(\alpha))}{\partial z_{ij}}$ is the partial derivative of the model's output F for the feature z_{ij}, evaluated at the interpolated point $z'(\alpha)$. This gradient measures how changes in the feature value impact the model's prediction along the interpolation path. To represent this process visually, we illustrate and explain the gradient magnitude of 2 extracted representations z_1 and z_2, which are extracted from 2 different encoder structures, in Figure 4.
- We assess the generated IG_{ij} values to discover features with positive contributions (showing an increase in the model's output) and negative contributions (indicating a decrease). In particular, higher IG_{ij} values indicate a better influence of a specific feature on the model's predictions.
- Finally, the structures and parameters of M that have the highest IG value are selected and updated to the initial parameters.

Regarding the encoder-to-decoder structure, IG contributes a large portion of the system interpretability by quantifying the contribution of each extracted representation. This method assists physicians in understanding the reasons for using network structures in DL models. Besides IG, the process of shaping the structure of the base model depends on the clinical prediction context, which is described in the following two sections.

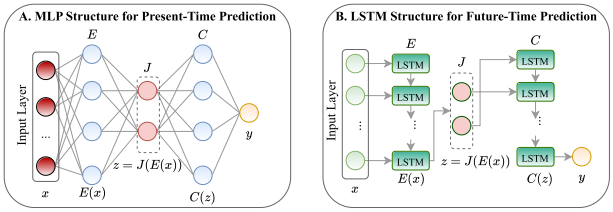


FIGURE 5. Base model architecture for both case studies within the Deep-CFL framework. Panel A depicts the architecture designed for predicting the current clinical status employing a Multilayer Perceptron (MLP) structure for the prediction. Panel B shows the model configured for prediction of clinical status at a future time point, incorporating a Recurrent Neural Network (RNN) as the core structure for handling sequential data, aiming to capture the temporal dynamics of the clinical data for prediction.

2) PRESENT-TIME PREDICTION

By including IG in the decision process for selecting a multilayer perceptron (MLP) as the principal architecture for the present-time prediction problem, we dramatically improved the interpretability and practical applicability of the proposed model. As illustrated in 5A, the present-time

prediction model, denoted as M_s uses an encoder-to-decoder structure using the MLP design for both E and C. Numerous studies [56], [57], [58], [59], [60], [61], [62], [63] have demonstrated the popularity and effectiveness of MLP in the medical field, including critical applications in ICU and mortality prediction tasks. The projection head J, for all prediction contexts, uses the dense layer with the rectified linear unit (ReLU) activation function to refine and enhance feature representations extracted by E, introducing non-linearity to enable the learning of more complex patterns. This additional processing step aims to improve the quality and interpretability of the features before they are utilized for further tasks or predictions, effectively bridging the gap between raw data and actionable insights.

3) FUTURE-TIME PREDICTION

Within the scope of this predictive task, to simulate the complexities associated with time-series classification properly, the input data must be transformed into a sequential format. In particular, the proposed approach entails using previous data to estimate outcomes at future time intervals. From a healthcare perspective, this approach is similar to how a physician examines a patient's medical history to assess and estimate health outcomes [43], such as survival prediction [64].

For medical time-series classification, the RNN architecture is widely recognized as a potent and frequently employed technique [65], [66], [67]. The LSTM network displays improvements over traditional RNNs by solving vanishing gradient problems [68]. This method exhibits high compatibility and interpretation for ICU systems [69], [70], [71] and EHR data [72], [73], [74]. Incorporating insights from IG, we have selected an RNN structure with LSTM units (as depicted in Figure 5B) to serve both E and C within the context of Future-Time Prediction model (denoted as M_d), while J still retains the Dense layer structure with ReLU activation.

The detailed parameters of the base models are described in Section IV-C3.

D. LOCAL OBJECTIVE

As presented in Figure 6, the local loss of Deep-CFL contains two parts. The initial element is the contrastive loss, denoted by L_c. The following components include losses associated with supervised learning, notably the focal loss (expressed by L_f) and imbalance loss (denoted by L_i). These losses were all designed to address significant problems encountered in the ICU prediction framework.

1) CONTRASTIVE FEDERATED LEARNING (CFL)

If local data D_i are employed in local training. For every input data point x, we extract representation z of x from global model M_g (i.e., $z_g = M_g(x)$), representation of x from local model of last round $M_i^{t-1}(x)$ (i.e., $z_{t-1} = M_i^{t-1}(x)$), and the representation of x from local model being updated

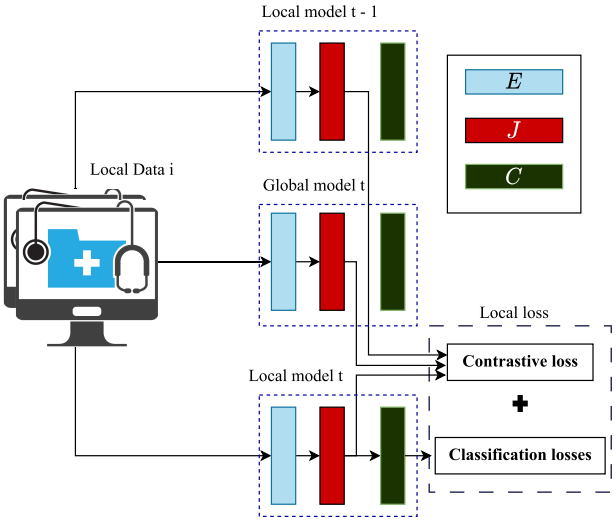


FIGURE 6. The local loss in Deep-CFL.

(i.e., $z = M_i^t(x)$). Since the M_g should be able to extract better representations, our purpose is to reduce the distance between z_i and z_g while increasing the distance between z_t and z_{t-1}. The model-contrastive loss is defined based on NT-Xent loss [75]:

$$L_c = -\log \frac{\exp\left(\text{sim}\left(z_t, z_g\right)/\tau\right)}{\exp\left(\text{sim}\left(z_t, z_g\right)/\tau\right) + \exp\left(\text{sim}\left(z_t, z_{t-1}\right)/\tau\right)} \tag{4}$$

where $\exp\left(\text{sim}\left(z_t, z_g\right)/\tau\right)$ calculates the exponential of the similarity between the current local model representation z_t and the global model representation z_g, scaled by a temperature parameter τ. The denominator sums the exponentiated similarities of z_t to both z_g and z_{t-1}, hence leveling the similarity measure. By including z_{t-1}, the model not only aligns z_t with z_g but also distinguishes z_t from the prior local model's representation, promoting representation variety and adaptability. The contrastive loss mechanism underpins the approach of Deep-CFL to fine-tune feature representations in FL settings, ensuring that each iteration brings the local models closer to extracting clinically relevant, predictive features from EHR data, improving the overall prediction accuracy.

2) SUPERVISED LEARNING

To improve the predicted accuracy of the classification head in the presence of imbalanced data, we applied the focal loss (L_f) alongside the imbalance loss (L_i). Focal loss is an alternative form of cross-entropy loss specifically designed to address problems associated with imbalanced datasets [76], [77]. The effectiveness of L_f has also been proven in the FL frameworks [78]. The function of L_f is expressed as:

$$L_f(p_i) = -y_i(1-p_i)^\gamma \log(p_i) - (1-y_i)p_i^\gamma \log(1-p_i) \tag{5}$$

where y_i indicates the ground truth for data point i^{th}. $y_i = 1$ represents an "event" occurring in ICU and vice versa.

p_i is the predicted probability for the i^{th} data point being an "event". The focusing parameter γ influences the degree of focus. A smaller γ emphasizes cases that are difficult to classify. L_f prioritizes misclassified situations, particularly complex ones. The goal is to minimize the weight of simple cases while raising the weight of more challenging ones, which are frequently underrepresented in the dataset. This method allows the model to focus on tough examples, resulting in superior representations.

To complement the ability of the model to manage imbalance classification, in addition to L_f, we used imbalance loss L_i, which is conceptually derived from the Dice coefficient loss [79]. The Dice coefficient (denoted as S_{dice}), which is commonly used in segmentation tasks, evaluates the overlap between two samples. Its efficacy in segmentation has led to its application as the Dice loss for dealing with class imbalance in classification problems. The formula of S_{dice} is:

$$S_{dice} = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{6}$$

where X and Y represent the predicted and actual values, respectively. For clinical prediction context, this translates to evaluating the overlap between the predicted positive class (predicted as "event") and the actual positive class (actually "event"), normalized by the size of each class. L_i (or Dice loss) is implemented as $1 - S_{dice}$. It concentrates on correctly identifying positive (often minority) instances and penalizes the model more strongly for misclassifications in this class. This is especially essential in datasets when positive examples are uncommon but of high interest. L_i is defined mathematically as follows:

$$L_i = 1 - \frac{2 \times \sum_{i=1}^{n} y_i \times p_i + \epsilon}{\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} p_i + \epsilon} \tag{7}$$

where y_i is the ground truth for sample i; p_i is the projected probability of sample i; and ϵ is a smoothing factor to avoid division by zero. To further mitigate the imbalance challenge, we adjust this function by incorporating class weights inversely proportional to their frequencies within the dataset:

$$L_i = 1 - \frac{2 \times \sum_{i=1}^{n} w_i \times y_i \times p_i + \epsilon}{\sum_{i=1}^{n} w_i \times y_i + \sum_{i=1}^{n} w_i \times p_i + \epsilon} \tag{8}$$

where w_i signifies the weight attributed to the i^{th} sample, dependent on its class label. The weight for samples in the "event" class is determined as $\frac{n_{neg}}{n_{pos}}$, where n_{neg} and n_{pos} are the counts of negative and positive cases, respectively. For "non-event" samples, w_i is set to 1. This weighting method increases the importance of the minority class - often the crucial "event" cases in clinical settings - encouraging the model to pay attention to these underrepresented but critical occurrences.

For t^{th} local model M_i^t, the classification loss of supervised learning progress $L_{\text{sup}}\left(M_i^t; (x, y)\right)$ for an input (x, y) is:

$$L_{\text{sup}}\left(M_i^t; (x, y)\right) = L_f\left(p_i, y_i\right) + L_i\left(p_i, y_i\right) \tag{9}$$

where p_i is the probability of the positive class predicted by M_i^t for x.

3) FINAL LOCAL TRAINING LOSS

The final local loss of an input (x, y) is calculated as the amalgamation of supervised and contrastive losses:

$$L = L_{\text{sup}}\left(M_i^t; (x, y)\right) + \mu \cdot L_c\left(M_i^t; M_i^{t-1}; M_g^t; x\right) \tag{10}$$

The weighting coefficient μ balances the contrastive loss in the total loss equation, allowing for variable modifications to prioritize the accuracy of the supervised learning and the quality of the representation learning based on unique training demands or objectives. The local objective of Deep-CFL is to minimize the following:

$$\min_{M_i^t} \mathbb{E}_{(x,y)\sim D^i}\left[L_{\text{sup}}\left(M_i^t; (x, y)\right) + \mu L_{\text{c}}\left(M_i^t; M_i^{t-1}; M_g^t; x\right)\right] \tag{11}$$

E. UPDATING THE GLOBAL MODEL THROUGH WEIGHTED AVERAGING

In the next stage, the local centers send the revised model M_i^t back to the central server. This step emphasizes the essence of the FL framework, in which localized insights are gathered independently across various healthcare environments and then combined into a single, global perspective. After collecting local models, the server combines them to create the global model, M_g^{t+1}, for the following iteration. This synthesis uses a weighted average method [80], as follows:

$$M_g^{t+1} \leftarrow \sum_{i=1}^{N} \frac{|D_i|}{\sum_{j=1}^{N} |D_j|} M_i^t \tag{12}$$

The contribution of each local model to the updated global model is correspondingly weighted by the size of the relevant dataset D_i. Weighted FedAvg guarantees that insights from larger datasets do not eclipse those from smaller ones, preserving a fair balance of learning across the network. By weighting the contributions depending on the size of D_i, M_g better matches the real-world distribution of data across local centers. Finally, this technique improves the resilience and generalizability of M_g by including a wide range of localized insights, allowing the model to navigate the intricacies of many patient populations and diseases.

IV. EXPERIMENTAL RESULTS

This section provides information on the experimental data, preprocessing steps, experimental settings, results, and interpretation.

A. DATA COLLECTION

The Chonnam National University Hospital Independent Institutional Review Board approved the study protocol. To simulate the prediction process in the ICU realistically, all data in this study are clinical EHR data. To assess the efficacy of Deep-CFL, we employed three datasets consisting

TABLE 2. Description of three experimental datasets in this work. RRT stands for 'Rapid response team' database, MIMIC III stands for 'Medical information mart for intensive Care III' database, and eICU stands for 'eICU collaborative research' database.

Dataset (Task)		RRT (Clinical deterioration)	MIMIC III (ICU mortality)	eICU (ICU mortality)
Population		25,329	18,281	27,865
Input variables	Categorical	Diagnosis, age, gender, hospitalization department, hospitalization route, inpatient ward.	Gender, age, ethnicity.	Ethnicity, gender height, weight.
	Numerical	Heart rate, Systolic/Diastolic Blood Pressure, Respiration Rate, Temperature, pO2, Potassium(K+), Alkaline phosphatase, pCO2, CRP, CRP2, pH, Albumin, Total protein, PT, Sodium, WBC, AST, Hgb, BUN, aPTT, Chloride, Plt, Total calcium, SaO2, ALT, Creatinin, Total bilirubin, Glucose, HCO3, Lactate.	Heart rate, Systolic/Diastolic Blood Pressure, Temperature, Respiration rate, spo2, albumin, bun, bilirubin, lactate, bicarbonate, bands, chloride, creatinine, glucose, hemoglobin, hematocrit, platelet, potassium, ptt, sodium, wbc.	Heart Rate, MAP (mmHg), Systolic/Diastolic Blood Pressure, O2 Saturation, Respiratory Rate, Temperature, glucose, FiO2, pH.
Output variables		0 for non-event and 1 for event	0 for alive and 1 for dead in ICU	0 for alive and 1 for dead in ICU
Class imbalance rate (minority/ majority)		0.64%	4.60%	12.97%

of one private and two public datasets. For the in-house data, we presented the Chonnam Hospital Rapid Response Team (RRT) dataset. The public datasets used in the experiment are Medical Information Mart for Intensive Care (MIMIC III) [81], and the eICU Collaborative Research Database (eICU) [82]. The description of three experimental datasets is shown in Table 2.

1) CHONNAM HOSPITAL RAPID RESPONSE TEAM DATASETS

The RRT dataset is a database that includes over 39 kinds of clinical information from 25, 329 patients admitted to Hakdong Chonnam National University Hospital in South Korea between February 1, 2021, and November 30, 2021. Regarding data types, the input variables from the RRT dataset are divided into two main types: categorical and numerical. Categorical groups include patient demographic data (e.g., age, gender, and hospitalization) and have constant values over time. Numerical data are patient clinical characteristics measured hourly (e.g., vital signs, such as heart rate and blood pressure, or laboratory tests, such as pCO2 and albumin). The RRT dataset used 39 variables (including seven demographic data, five vital signs, and 27 laboratory tests) as the input features for the system. The target variable is binary, with 1 denoting an event and 0 representing a non-event. An event is defined as a situation in which medical professionals detect clinical deterioration in ICU patients, necessitating interventions, such as cardiopulmonary resuscitation [83] or tracheal intubation [84].

The class distribution of the RRT dataset displays a significant imbalance between these two classes. With 162 event samples compared to 25, 167 non-event samples (a 0.64% imbalance rate), this dataset represents a significant challenge for the imbalance classification task. The experimental process maintains this imbalance rate without using any class balancing method in the preprocessing step to highlight the classification ability of Deep-CFL for challenging datasets.

2) MEDICAL INFORMATION MART FOR INTENSIVE CARE III DATASET

The large, single-center MIMIC III dataset contains information on 38, 597 patients admitted to critical care units at a large tertiary care facility. We followed the tutorial in [85] to streamline data gathering and eliminate unnecessary data groups due to the enormous size of this dataset, which exceeded the research resource capabilities. We concentrated solely on the subset of patients admitted to ICUs who had the complete baseline data necessary for predicting clinical events, such as vital signs and laboratory tests. Consequently, this approach resulted in selecting 18, 281 samples for inclusion in this study from the raw dataset. The input variables from MIMIC III include 24 features (three categorical and 21 numerical features). Regarding the medical context, input features also include demographics, vital signs, and laboratory test data but are different in quantity compared to those in the RRT dataset. The output variable for MIMIC III is also binary, with 0 for alive and 1 for ICU mortality cases. Similar to RRT, MIMIC III exhibits a class imbalance with a 4.6% imbalance rate. Specifically, the minority group (ICU mortality) comprises 804 samples compared to 17, 477 survival samples.

3) EICU COLLABORATIVE RESEARCH DATABASE

The multicenter eICU database contains high-granularity data on over 200, 000 ICU admissions monitored by eICU programs in the United States. With the same processing method as in MIMIC III, following the tutorials [86], we collected data on 27, 865 patients admitted to the ICU from the raw data. This dataset has 14 input features, comprising four categorical and ten numerical features. Similar to MIMIC III, the binary outcome of eICU includes 1 for ICU mortality and 0 for alive.

Among the three experimental datasets, the eICU dataset exhibits the least severe class imbalance, with the minority

class constituting 12.97% (3, 523 mortality cases compared to 27, 160 survival cases). This study purposefully selects three datasets—RRT, MIMIC III, and eICU—each sharing class imbalance challenges but differing in the variation of their imbalance ratios, arranged in descending order of imbalance. This strategic selection enables us to evaluate the performance of the proposed model thoroughly under various scenarios of class imbalance. We aim to determine the stability and broad applicability of the model by evaluating its performance on datasets with varying imbalance rates. This technique reduces the possibility of biased evaluations, which could occur if models were only evaluated on datasets with a minimal imbalance, where improved performance could be attributed to a thorough representation of the minority class. Conversely, the underperformance of the model on datasets with a severe imbalance may limit its utility in broader contexts. Thus, this methodology ensures a fair and complete analysis of the capabilities of the model, highlighting the need for adaptability in varying class distributions.

B. DATA PROCESSING

The previous section discussed the cohort and variable selection steps along with the experimental datasets. This section outlines the variable processing, data transformation into suitable input for DL models specific to each case study, and client data splitting within the FL environment.

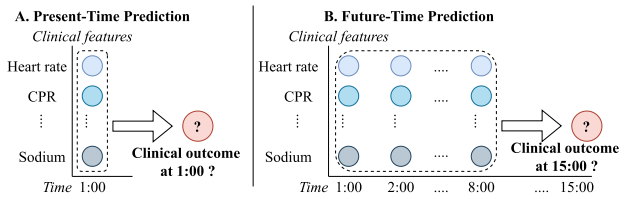


FIGURE 7. Prediction context. Panel A: Example of the present-time prediction context, determining the clinical status of a patient at the current time point. Panel B: Example of the future-time prediction context, predicting the clinical outcome of a patient in the future time point.

1) VARIABLES PROCESSING

As mentioned in the section I, we assessed Deep-CFL in two case studies: determining the *clinical status at present* and predicting the *clinical status at a future time*. Figure 7 6 presents an example and explanation of the two prediction cases.

In the present-time prediction case, we focus on the capacity of the system for instant activation, simulating a rapid response process in the RRS for each piece of clinical information at a given time. Specifically, for the collected clinical data (X_t) at a time point t, the system aims to predict the immediate clinical status (Y_t) at t. The formula for this problem is expressed as follows:

$$Y_t = M_s\left(X_t\right) \tag{13}$$

where M_s represents the optimal static prediction model for this case study. The input data of experimental datasets has two main types: categorical (denoted as x_{cat}) and numerical (denoted as x_{num}). For x_{num}, we apply standard scaler to scale the numerical data to have zero mean and unit variance. The mathematical function for this step is:

$$\psi\left(x_{num_i}\right) = \frac{x_{num_i} - \mu_i}{\sigma_i} \tag{14}$$

where $\psi\left(x_{num_i}\right)$, μ_i, and σ_i are the standardized value, mean, and standard deviation of the $i-th$ numerical feature across all observations in the dataset, respectively. For x_{cat}, we apply one-hot encoding to transform categorical variables into a binary matrix, ensuring compatibility with ML models without imposing ordinality. The function of this step could be expressed as:

$$\phi\left(x_{cat}\right)_j = \delta_{c_k,c_j} = \begin{cases} 1 & \text{if } c_k = c_j \\ 0 & \text{if } c_k \neq c_j \end{cases} \tag{15}$$

where $\phi\left(x_{cat}\right)_j$ is $j-th$ element of the one-hot encoded vector x_{cat}. δ_{c_k,c_j} is the Kronecker delta function [87], comparing the actual category c_k of x_{cat} to each possible category c_j in the dataset. These processing methods are implemented using the sci-kit learn library [88].

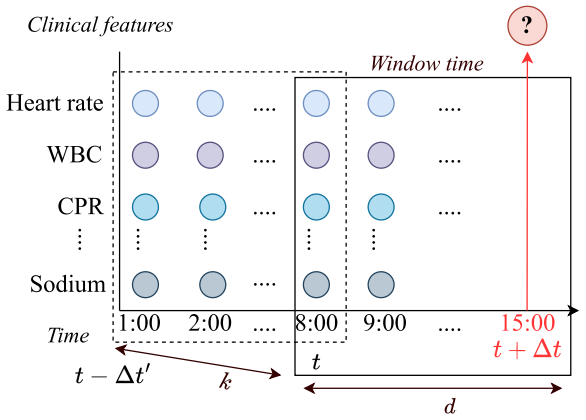


FIGURE 8. Example of a window interval sliding (WIS) mechanism, where d denotes the window size, and k denotes the window sliding size. The input in this example is the clinical information from 1:00 to 8:00, the prediction target is the clinical outcome at 15:00. Setting d = 8, k = 7, the first window (query for prediction window) $w_1 = \{(1-1)*7+1, \ldots, (1-1)*7+8\} = \{1, \ldots, 8\}$. The 2^{nd} window $w_2 = \{(2-1)*7+1, \ldots, (2-1)*7+8\} = \{8, \ldots, 15\}$ contain the target time (15:00) is the prediction window. In this example, the WIS must perform one sliding step, with a window size of $d = 8$ and a sliding step size of $k = 7$, to facilitate the prediction at the desired future time point. For extended prediction periods or time points, the system might execute additional sliding steps, the quantity of which is dictated by the values of d and k.

The future-time prediction context focuses on predicting the clinical outcomes at a future point $t + \Delta t$ based on historical and current clinical data observed from $t - \Delta t'$ to t, where $\Delta t'$ and Δt represent the lengths of the observation window and the prediction interval, respectively. The function of the dynamic predictive model M_d of this case study is expressed as follows:

$$Y_{t+\Delta t} = M_d\left(X_{t-\Delta t':t}\right) \tag{16}$$

In this case study, both standard scaling and one-hot encoding were used to preprocess the data. Additionally, the input data were transformed into sequences that correspond to the temporal nature of the model structure to meet the requirements of the future-time prediction context [89]. To represent the time-series input in the temporal dimension, we transformed the scaled data into a set of window samples $W = \{w_1, \ldots, w_T\}$, applying the Window Interval Sliding (WIS) mechanism [90], a technique widely employed in medical time-series analysis. Each t^{th} window is denoted as $w_t = \{x_{(t-1)*k+1}, \ldots, x_{(t-1)*k+d}\}$, where $t \in \{1, .., T\}$; T is the number of sliding steps that WIS needs to perform for a specific prediction case. w_t comprises d consecutive time steps and a sliding size of k, as shown and explained in Figure 8.

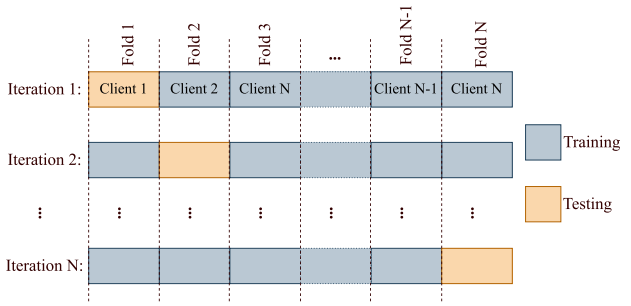


FIGURE 9. Client data splitting strategy for FL environment.

2) DATA SPLITTING

For client data splitting, we applied a strategy based on stratified k-fold cross-validation (CV), as depicted in Figure 9. This method is particularly effective in scenarios where data are inherently decentralized, such as in FL environments. Initially, the experimental dataset is partitioned into N distinct subsets to represent the distribution of data across N clients in an FL setup. Each subset is designed to imitate the client's local dataset, ensuring that each component of the data is indicative of the entire distribution. This technique iteratively uses one subset as independent local data for testing and the remaining N – 1 subsets for training/performing the FL process. This method is performed N times, with each of the N subsets serving as the testing set just once. This arrangement resembles k-fold CV, adapted for the FL context, where N represents the number of folds.

In each iteration of the FL context, a unique subset is designated as the independent local dataset for testing, whereas the remaining N – 1 subsets are used collectively for FL. This collaborative method promotes the construction of an optimal global model, which is applied to the independent local dataset designated for validation. This iterative process ensures that each subset serves as an independent testing set at least once during the validation cycle while contributing to the training of the global model in subsequent iterations. During the training phase, insights from N – 1 subsets are aggregated to develop the global model, allowing the distributed data to be used without direct sharing. Evaluating the optimized global model on untouched local data assesses its efficacy on previously unseen information, offering insight into the ability of the model to generalize across varied data landscapes and adapt to unique data distributions encountered in the FL framework. This cyclic validation approach highlights the broad application of the model and underlines the fundamental privacy and security principles of the FL paradigm.

C. EXPERIMENTAL SETTING

The proposed method and all experiments were implemented using Python and TensorFlow (v. 2.11.0) on an Nvidia GeForce RTX 3090 graphics processing unit (driver v. 530.30.02) and CUDA (v. 12.1) with 64 GB memory.

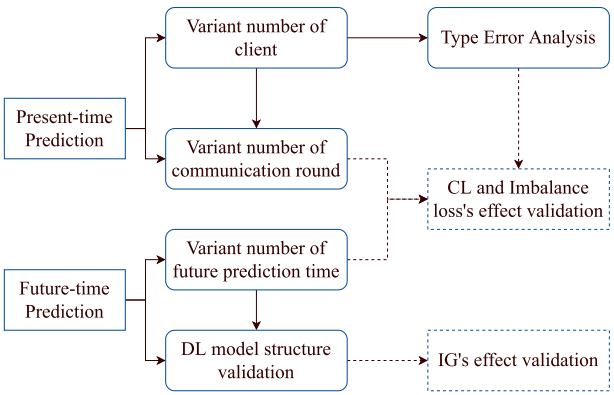


FIGURE 10. Overall validation strategy.

1) VALIDATION STRATEGY

Figure 10 presents the overall validation strategy. This investigation divides the experiment into two parts corresponding to each prediction context.

For the present-time prediction context, we experimented with the comparison methods with variations in the number of clients. Then, we extracted sample results from this main experiment to perform two additional experiments that included experimental results over a certain number of communication rounds and an error analysis. The experimental results over a certain number of communication rounds evaluate the stability of FL-based methods. With the ablation experiment setting for comparison methods, this strategy aims to evaluate the contribution of CFL and imbalance learning in the system. Through the error pattern analysis, this study provides an intuitive view of the comparison results and evaluates the effects of imbalance learning on the balance between sensitivity and specificity.

For the future-time prediction context, we experimented with the comparison methods by varying the future prediction interval while maintaining a certain number of clients. This analysis explores the predicted accuracy and reliability of each method across long time horizons, which are critical for successful ICU decision-making in dynamic contexts.

By comparing the experimental methods and representing the probabilistic result patterns for event time prediction, we verified the contribution of CFL and imbalance learning. An extensive experiment was also performed in this case. Specifically, to evaluate the effectiveness of IG, we conducted Deep-CFL experiments with various base model structures.

TABLE 3. Repetition of Deep-CFL components in comparison methods for ablation study.

Deep-CFL's component	IG	CFL (L_c)	L_i	L_f
Deep-LL	Yes	No	Yes	Yes
Deep-CL	Yes	No	Yes	Yes
Deep-FedAvg	Yes	No	No	Yes
MEWS	No	No	No	No

2) COMPARISON METHODS

Based on related studies described in Section II, we used four comparison methods corresponding to the three main learning algorithms in FL studies, including LL-, CL-, and FL-based methods. For LL-based approaches, where each local healthcare center trains a model independently using its dataset, the objective function is defined as minimizing the empirical loss $L(M, D)$ for each local dataset D individually:

$$\arg\min_{M_i} L_i\left(M_i, D_i\right) \tag{17}$$

Two comparable methods in this type of learning are Deep-LL and MEWS. With Deep-LL, we designed a DL-based model structure using IG with each prediction context similar to the proposed method and then trained them independently on each local data point. The MEWS is a special method because it directly evaluates outcomes through basic vital signs on the testing sample. This traditional method also operates independently in each healthcare center; thus, we classified it as LL. When evaluating the MEWS, to balance the sensitivity and specificity based on [41], all samples with MEWS ≥ 5 are considered as events.

For CL-based approaches, where all data from the local healthcare centers are aggregated and used to train a centralized model (M_c), the objective function aims to minimize the empirical loss $L\left(M_c, D_i\right)$ across the entire aggregated dataset. This loss is represented as follows:

$$\arg\min_{M_c}\left(\sum_{i=1}^{N} L\left(M_c, D_i\right)\right) \tag{18}$$

The experimental method corresponding to this principle is Deep Centralized Learning (Deep-CL). We design a DL-based model structure using IG with each prediction context and then train them on the aggregation data.

The experimental method corresponding to this principle is Deep-CL. We designed a DL-based model structure using IG with each prediction context and then trained them on the aggregation data. The objection function for FL was described in 1. To evaluate the effectiveness of CFL, we used Deep-FedAvg, a method that uses the same FL processes as Deep-CFL but removes the contrastive learning part as an FL-based method for comparison.

During the training procedure, we set up the structure and loss function for the ablation study comparison methodologies. Specifically, except for the MEWS, which is a method that directly evaluates the testing sample, other comparison methods all repeat the components or loss functions in the proposed method. As listed in Table 5, Deep-LL, Deep-CL, and Deep-FedAvg all use the same base model structure as Deep-CFL, which was configured using IG. However, we only used Li during Deep-CL and Deep-LL training, whereas Lf was applied for all three methods. Contrastive learning is a special structure that represents the proposed method; thus, CFL is not repeated on any of the mentioned comparison methods. The training and validation processes discussed in Section IV-B2 apply the stratified k-fold CV strategy. The training and validation process has been mentioned in Section IV-B2 - applying the Stratified K-Fold CV strategy.

TABLE 4. Parameter space settings for the base models.

Model structure		MLP	LSTM
Parameter Name	Type	Searching space	
Number of hidden layers	int	[1, 2, 3]	[1, 2, 3]
Units per layer	int	[16, 32, 64, 128, 256]	[16, 32, 64]
Activation function	list	['tanh', 'relu']	['tanh', 'relu']
Learning rate	float	[0.1, 0.01, 0.001, 0.0001]	[0.1, 0.01, 0.001, 0.0001]
Dropout rate	float	[0.2, 0.3, 0.4, 0.5]	[0.2, 0.3, 0.4]
Weight delay	float	[1.0e-03, 1.0e-04, 1.0e-05]	–
Regularization	list	['none', 'l1', 'l2']	_
Recurrent activation	list	–	['none', 'sigmoid', 'linear']
Recurrent dropout rate	float	–	[0.1, 0.2, 0.3, 0.4]

3) PARAMETERS SETTINGS

We set the parameters for the base model corresponding to each prediction context. Table 4 displays the search space for each parameter corresponding to each model structure. Rather than conducting expensive training and evaluations for each parameter in the search space, calculating the IG values allows efficiently selecting the most effective parameter. This method saves substantial resources and time while streamlining the hyperparameter tuning procedure.

Based on this optimization process, we applied the MLP encoder-to-decoder architecture as the main structure for the present-time prediction model M_s, which has three hidden layers with 64, 128, and 256 units each, with dropout rates of 0.2, 0.3, and 0.2 for each layer in the sequence. The ReLU activation function was used across all layers, with a learning rate of 0.001, L2 regularization, and a weight decay

of 1.0e-3. A dense layer with two units and sigmoid activation was included at the end of the decoder to obtain the final prediction output.

For the future-time prediction model M_d, we employed the LSTM encoder-to-decoder framework as the foundational architecture. This model was built with three LSTM layers, with 16, 32, and 64 units each, and dropout rates of 0.2, 0.2, and 0.1 to improve generalization. The activation function for all layers was ReLU, with a learning rate of 0.0001 to prevent overfitting. Next, M_d includes a dense layer with two units and a sigmoid activation function at the end of the decoder to facilitate the final predictive output, allowing for exact future-time prediction results.

The global model M_g was set up using the server to start FL training. Training consists of N local datasets identified via k-fold CV. For the experiment varying the number of clients, we conducted five-, seven-, and nine-fold CV. Each local healthcare center receives M_g^t and updates the local model M_i^t over 50 communication rounds. We set learning rate $\eta = 0.01$, temperature $\tau = 0.07$ (only for Deep-CFL), and weighting factor $\mu = 1.0$. The local models are trained for 100 epochs on their respective dataset D_i. The updated local models M_i^t were aggregated via weighted averaging to form the updated global model M_g^{t+1}. For the CL- and LL-based methods, the same settings regarding the base model structures, data division, number of epochs, and learning rate compared to FL-based methods were performed. The distinction is that these methods do not undergo evaluation across multiple communication rounds or incorporate τ and μ as FL-based methods.

4) EVALUATION METRICS

The main evaluation metrics in this study were calculated in all included prediction cases: the AUROC, AUPRC, and the average precision (AP). The three metrics are used in the context of imbalance classification [91]. The AUROC metric assesses the capacity of the model to discriminate across classes, emphasizing the binary classification of event and nonevent cases in the ICU. For the unequal distribution of the experimental datasets, where event cases are much less prevalent than non-vent cases, AUPRC is an important metric. It assesses the model precision-recall balance, which is especially relevant in cases where the positive class is scarce, and the cost of false negatives (FNs) is substantial. A higher AUPRC score indicates a better balance of precision and recall, indicating that the model is good at recognizing the true positive (TP) cases without excessive false positives (FPs). The AP computes the average value of precision across multiple recall levels. Similar to AUPRC, AP is beneficial for dealing with imbalanced datasets because it provides information on model performance throughout the range of classification criteria. A higher AP value indicates that the model can maintain high precision at increased levels of recall. For all metrics, we reported a 95% confidence interval over k-fold CV.

Extended experiments include additional comparison criteria based on confusion matrices, providing a complete perspective of TPs, true negatives (TNs), FPs, and FNs. This approach enables a detailed study of Type I and II errors and evaluates the ability of comparative models to solve the error alarm problem. Type I errors (FP rate) arise when the model predicts an event that did not occur. Minimizing Type I errors in clinical settings is critical for avoiding unnecessary interventions. Type II errors (FN rate) refer to the failure to predict an occurrence when it occurs. Reducing Type II errors is critical to provide timely medical treatment for patients.

Furthermore, we present some individual probability testing samples in terms of the future-time prediction case to comprehend the model decision-making process, improving the interpretability in the practical clinical setting. Through these varied evaluation measures and analyses, we aim to provide a comprehensive perspective of the performance of the proposed method and its implications in clinical settings.

D. PRESENT-TIME PREDICTION RESULTS

1) EXPERIMENT RESULTS BY NUMBER OF CLIENTS

Tables 5, 6, and 7 list the present-time prediction performance of the comparison methods with a variation of the number of clients on the RRT, MIMIC III, and eICU datasets, respectively. In the RRT dataset, with five clients, Deep-CFL achieves a 0.906 AUROC and 0.879 for both AUPRC and AP, outperforming all comparison methods on three metrics. As the number of clients increases to seven, fragmenting the training data into smaller segments, performance declines across all comparison approaches. For instance, Deep-LL reduces AUROC by up to 4.1%, Deep-FedAvg decreases AUPRC by 1.8%, and Deep-CL decreases AP by 8.7%. In contrast, Deep-CFL retains stable performance, with only a 1.6% loss in AUROC and 0.2% loss in AP, whereas its AUPRC metric is unaffected. The performance of Deep-CFL remains stable when assessed with nine clients, exhibiting improvement in AUPRC and AP (both increasing by 0.7%).

The observed decrease in performance for Deep-LL, Deep-FedAvg, and Deep-CL as the number of clients grows is most likely owing to data fragmentation, in which smaller training subsets result in less representative models. Deep-CFL's consistent performance, even with additional clients, demonstrates its durability, which is ascribed to the contrastive learning mechanism that aligns local models with the global model, reducing performance disparities. This alignment contributes to great performance despite fragmented data. Furthermore, the improvement in AUPRC and AP with nine customers indicates that Deep-CFL efficiently uses data diversity, hence improving its generalization potential. The stability of the AUPRC metric suggests that Deep-CFL maintains a balance of precision and recall, which is critical for finding real positives while avoiding false positives. These results highlight the superior robustness and adaptability of Deep-CFL in managing data fragmentation

TABLE 5. Results of present-time prediction of clinical deterioration by the number of clients on RRT dataset. CI: Confidence interval.

Number of clients	Method	Metrics (95% CI)		
		AUROC	AUPRC	AP
5	Deep-LL	0.904 (0.852-0.905)	0.874 (0.872-0.875)	0.870 (0.863-0.879)
	Deep-CL	0.899 (0.891-0.902)	0.872 (0.869-0.873)	0.862 (0.860-0.863)
	Deep-FedAvg	0.886 (0.884-0.888)	0.875 (0.873-0.878)	0.873 (0.870-0.875)
	MEWS	0.876 (0.871-0.877)	0.862 (0.860-0.865)	0.871 (0.869-0.873)
	Deep-CFL (ours)	**0.906 (0.904-0.908)**	**0.879 (0.877-0.881)**	**0.879 (0.877-0.882)**
7	Deep-LL	0.863 (0.841-0.868)	0.857 (0.851-0.859)	0.776 (0.772-0.778)
	Deep-CL	0.895 (0.892-0.897)	0.854 (0.851-0.858)	0.775 (0.773-0.778)
	Deep-FedAvg	0.859 (0.852-0.860)	0.857 (0.854-0.859)	0.855 (0.854-0.858)
	MEWS	0.872 (0.870-0.875)	0.871 (0.868-0.873)	0.870 (0.867-0.871)
	Deep-CFL (ours)	**0.890 (0.887-0.893)**	**0.879 (0.877-0.881)**	**0.877 (0.874-0.878)**
9	Deep-LL	0.854 (0.851-0.855)	0.818 (0.815-0.819)	0.742 (0.739-0.745)
	Deep-CL	0.869 (0.864-0.871)	0.849 (0.845-0.851)	0.771 (0.764-0.873)
	Deep-FedAvg	0.851 (0.849-0.853)	0.856 (0.854-0.858)	0.854 (0.851-0.855)
	MEWS	0.862 (0.861-0.864)	0.851 (0.848-0.853)	0.852 (0.848-0.853)
	Deep-CFL (ours)	**0.879 (0.875-0.881)**	**0.886 (0.884-0.888)**	**0.884 (0.882-0.887)**

TABLE 6. Results of present-time prediction of ICU mortality by number of clients on MIMIC III.

Number of clients	Method	Metrics (95% CI)		
		AUROC	AUPRC	AP
5	Deep-LL	0.816 (0.812-0.817)	0.663 (0.661-0.665)	0.627 (0.625-0.628)
	Deep-CL	0.819 (0.815-0.820)	0.671 (0.668-0.674)	0.637 (0.630-0.639)
	Deep-FedAvg	0.817 (0.812-0.819)	0.665 (0.663-0.668)	0.631 (0.628-0.634)
	MEWS	0.821 (0.819-0.823)	0.667 (0.665-0.670)	0.668 (0.664-0.670)
	Deep-CFL (ours)	**0.875 (0.871-0.878)**	**0.711 (0.704-0.714)**	**0.714 (0.711-0.715)**
7	Deep-LL	0.819 (0.817-0.821)	0.637 (0.632-0.638)	0.638 (0.635-0.641)
	Deep-CL	0.820 (0.818-0.823)	0.634 (0.631-0.635)	0.633 (0.632-0.635)
	Deep-FedAvg	0.817 (0.816-0.819)	0.671 (0.668-0.673)	0.638 (0.636-0.639)
	MEWS	0.820 (0.817-0.821)	0.675 (0.670-0.677)	0.674 (0.671-0.675)
	Deep-CFL (ours)	**0.863 (0.861-0.865)**	**0.705 (0.698-0.706)**	**0.702 (0.695-0.704)**
9	Deep-LL	0.820 (0.818-0.823)	0.622 (0.618-0.224)	0.636 (0.632-0.638)
	Deep-CL	0.817 (0.815-0.819)	0.662 (0.661-0.665)	0.631 (0.627-0.634)
	Deep-FedAvg	0.821 (0.819-0.824)	0.673 (0.670-0.676)	0.643 (0.642-0.645)
	MEWS	0.819 (0.815-0.821)	0.671 (0.695-0.674)	0.635 (0.632-0.636)
	Deep-CFL (ours)	**0.870 (0.868-0.872)**	**0.714 (0.711-0.715)**	**0.705 (0.702-0.708)**

TABLE 7. Results of present-time prediction of ICU mortality by number of clients on eICU.

Number of clients	Method	Metrics (95% CI)		
		AUROC	AUPRC	AP
5	Deep-LL	0.806 (0.805-0.808)	0.507 (0.503-0.509)	0.515 (0.513-0.521)
	Deep-CL	0.852 (0.849-0.855)	0.502 (0.495-0.504)	0.407 (0.405-0.408)
	Deep-FedAvg	0.845 (0.841-0.846)	0.309 (0.305-0.311)	0.404 (0.402-0.405)
	MEWS	0.832 (0.829-0.835)	0.305 (0.302-0.306)	0.401 (0.398-0.403)
	Deep-CFL (ours)	**0.908 (0.905-0.911)**	**0.658 (0.654-0.661)**	**0.658 (0.654-0.662)**
7	Deep-LL	0.857 (0.853-0.858)	0.502 (0.498-0.505)	0.409 (0.405-0.411)
	Deep-CL	0.866 (0.865-0.868)	0.492 (0.490-0.495)	0.453 (0.451-0.458)
	Deep-FedAvg	0.851 (0.847-0.852)	0.515 (0.512-0.517)	0.392 (0.389-0.393)
	MEWS	0.858 (0.853-0.861)	0.511 (0.508-0.514)	0.395 (0.384-0.396)
	Deep-CFL (ours)	**0.908 (0.903-0.909)**	**0.630 (0.627-0.635)**	**0.626 (0.622-0.628)**
9	Deep-LL	0.852 (0.850-0.854)	0.556 (0.554-0.559)	0.401 (0.398-0.402)
	Deep-CL	0.869 (0.862-0.872)	0.491 (0.487-0.495)	0.453 (0.451-0.456)
	Deep-FedAvg	0.848 (0.844-0.851)	0.489 (0.488-0.491)	0.406 (0.402-0.409)
	MEWS	0.845 (0.840-0.846)	0.494 (0.492-0.495)	0.452 (0.451-0.453)
	Deep-CFL (ours)	**0.899 (0.895-0.901)**	**0.616 (0.615-0.618)**	**0.614 (0.611-0.618)**

across an increasing number of local data centers, solidifying its effectiveness in FL environments, particularly in class imbalance challenges in the RRT dataset.

In MIMIC III and eICU, Deep-CFL demonstrates outstanding and stable performance across varying client numbers in ICU mortality prediction tasks. Although performance

typically declines for comparative approaches as the client count increases, the performance of the proposed method significantly improves with an increase to nine clients. During the MIMIC III dataset shift from seven to nine clients, the MEWS loses 0.2% in AUROC, 0.4% in AUPRC, and 3.9% in AP, whereas Deep-CFL gains 0.7%, 0.9%, and 0.3% in AUROC, AUPRC, and AP, respectively. The improved communication efficiency between clients in Deep-CFL compared to other learning methods is the reason for its superior performance as the number of clients increases [92], [93]. The contrastive learning component allows the local models to benefit from the global model's knowledge, maintaining high predictive accuracy even with increased client fragmentation. These findings indicate that Deep-CFL is particularly adept at handling data diversity and fragmentation, ensuring robust and reliable performance across different ICU settings. The ability to improve rather than degrade with more clients highlights Deep-CFL's scalability and effectiveness in federated learning environments, making it a valuable tool for ICU mortality prediction tasks. In eICU, although comparison approaches achieve high AUROC scores (more than 0.80), the majority perform poorly in AUPRC and AP metrics, with scores typically dropping below 0.50. This finding demonstrates the effect of the class imbalance on model evaluation measures. In addition, AUROC can still be relatively high even if the model performs poorly in the ICU mortality (minority class) because it assesses the ability to distinguish classes without considering the class distribution. In contrast, AUPRC and AP are more sensitive to an interclass imbalance because they are explicitly concerned with the minority class prediction performance.

Moreover, the results indicate a considerable precision-recall trade-off in these models [94]. They may accomplish better recall (hence the decent AUROC scores) but at the expense of extremely low precision, as indicated by the low AUPRC and AP scores. This trade-off is apparent in imbalanced datasets, where accurately identifying the minority class without producing too many FPs is difficult. Despite these results, Deep-CFL still achieves stable AUPRC and AP scores (e.g., 0.908 for AUROC, 0.658 for AUPRC, and AP for five clients). In Deep-CFL, CFL enables the local model to acquire the robust prediction abilities of the global model during each communication round, resulting in a final global model with optimal prediction efficiency. This observation is corroborated by the ablation study experimental results, revealing that Deep-CFL outperforms comparable approaches that skip CFL during implementation. The inclusion of L_i serves as a corrective solution, directly addressing the skewed distribution of classes. In addition, L_i penalizes misclassifications of the minority class more harshly, ensuring that the model does not neglect these critical cases, enhancing recall. This enhancement is accomplished without a considerable increase in FPs, ensuring high precision. To evaluate the performance of the models in solving alarm error problems, we performed an error analysis on the experimental results of a random local dataset, presented in the next section.

2) TYPE ERROR ANALYSIS

To evaluate the level of late and false alarms for prediction models on the likelihood of events occurring in the ICU, Figures 11 and 13 present the confusion matrix of some experimental methods in the RRT dataset (for seven clients) and MIMIC III dataset (for five clients), respectively. Figures 12 and 14 provide detailed analyses of each error type in the same cases for the comparison methods for the RRT and MIMIC III datasets, respectively.

In a random local data subset from the RRT dataset comprising 3, 619 samples dispersed across seven clients, Deep-CFL reliably detected 24 TPs from 28 event occurrences, the highest among the comparison methods. Without the implementation of L_i during its training phase, Deep-FedAvg demonstrated less effectiveness, correctly predicting the occurrence of clinical events in 18 samples. In addition, Deep-CL is ineffective despite aggregating data from local data during training when the event class prediction performance is lowest (12 samples). However, this method is effective in reducing false alarms with six FP samples.

The two main error types in the targeted clinical setting context are Type I (false alarm) and II (late alarm) errors. Type I errors lead to unnecessary anxiety, additional testing, and therapies that are not required, straining resources and harming patient well-being. Type II errors are especially concerning in ICU settings because they reflect missed detections of critical events, potentially leading to delays in necessary treatment and affecting patient outcomes. Minimizing Type II errors is critical because they significantly contribute to unsatisfactory outcomes in ICU settings [95]. Figure 12 depicts the proportion of error prediction samples of the comparison methods for each error type. For Type I errors, Deep-CL has the highest efficiency, at 9.4% of the total for this error type (Figure 12(A)).

However, as an inevitable consequence of each trade-off relationship between Type I and II errors [96], [97], this method suffers from high type II errors, with 16 out of 59 predicted samples encountering type II errors (27.1% - (Figure 12(B)). While maintaining a high degree of sensitivity that may result in several false alarms, Deep-CFL reduces the occurrence of Type I mistakes to the greatest extent, accounting for just 17% of Type I errors—a figure only slightly exceeded by Deep-CL. Critically, in addressing the problem of delayed alerts, the proposed method restricts the number of instances to four cases, accounting for only 6.7% of all Type II failures.

In Figure 13, for 3, 657 samples in a random local data of MIMIC III, with splitting for five clients, Deep-CFL exhibits the most TPs (147 samples) and TNs (3, 370 samples), displaying a superior capacity to detect event and nonevent instances accurately. This result reveals that Deep-CFL

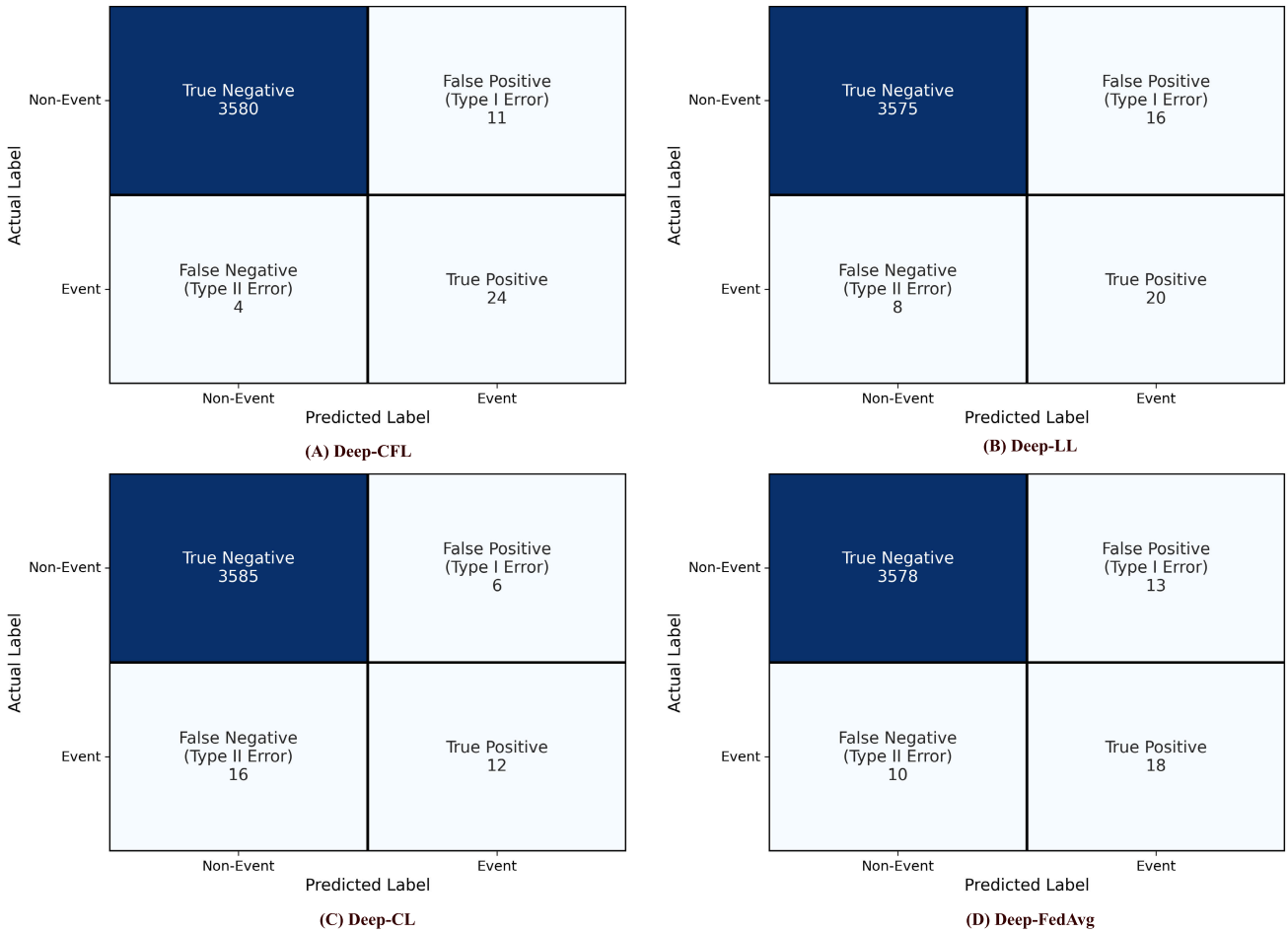


FIGURE 11. Confusion matrices of the comparison methods on the random testing local data from the RRT dataset for seven client cases. Panels (a), (b), (c), and (d) present the confusion matrices for Deep-CFL, Deep-LL, Deep-CL, and Deep-FedAvg, respectively.

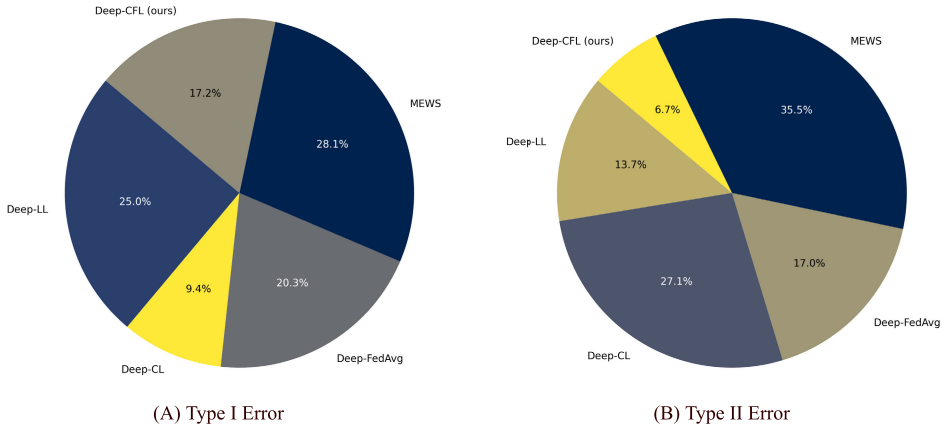


FIGURE 12. Error distribution of the comparison methods (for each error type) on the RRT dataset for splitting for seven clients.

effectively balances sensitivity and specificity, achieving high performance across both classes. Moreover, the MEWS, Deep-CL, and Deep-FedAvg perform slightly worse in TPs and TNs than Deep-CFL, indicating that these approaches may struggle more to categorize positive and negative cases reliably.

Out of 214 testing samples with Type I errors, Deep-CFL accounts for the lowest percentage, with 16.4% (35 error

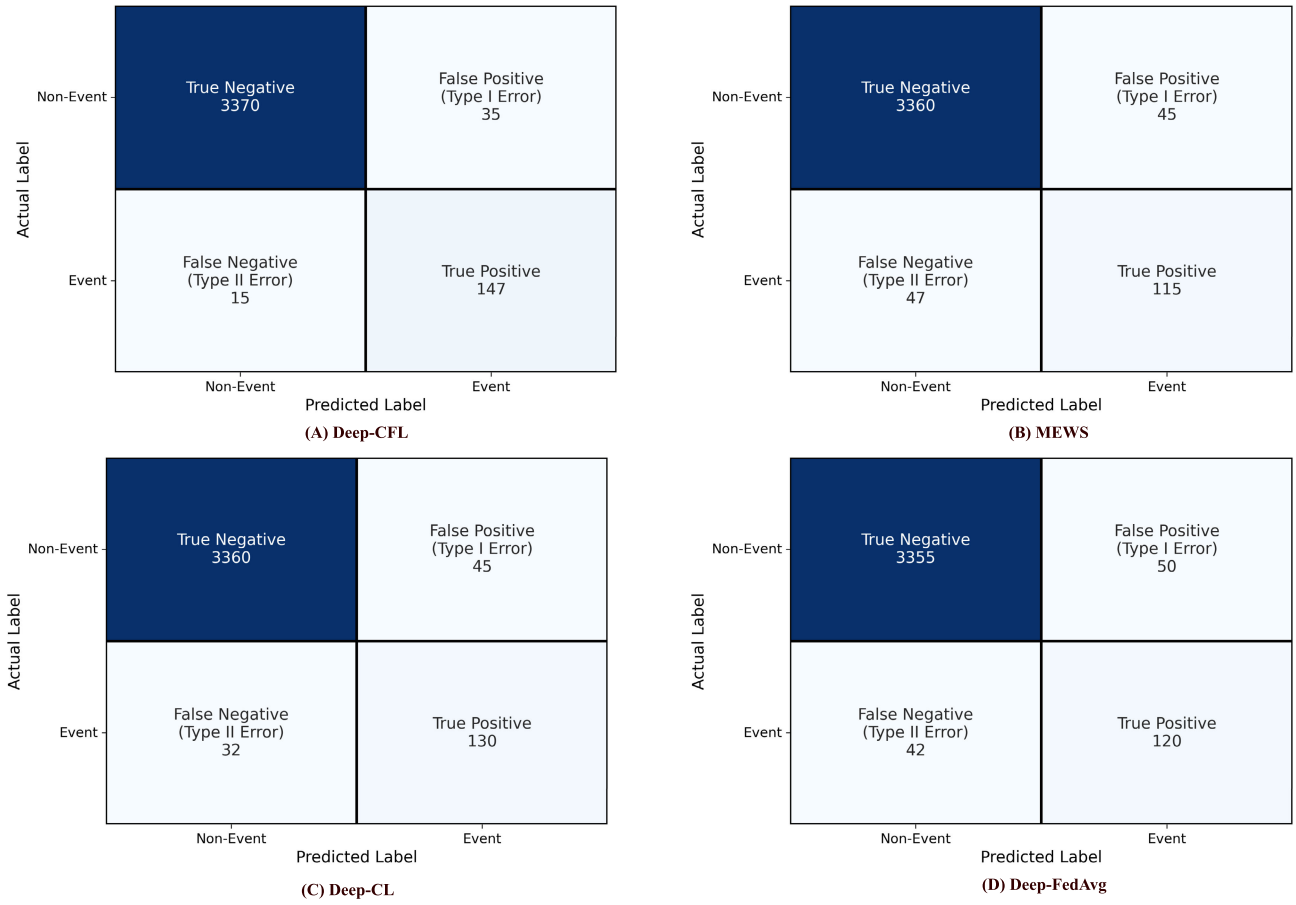


FIGURE 13. Confusion matrices of the comparison approaches on the random testing local data from the MIMIC III dataset for five client cases. Panels (a), (b), (c), and (d) present the confusion matrices of Deep-CFL, the MEWS, Deep-CL, and Deep-FedAvg, respectively.

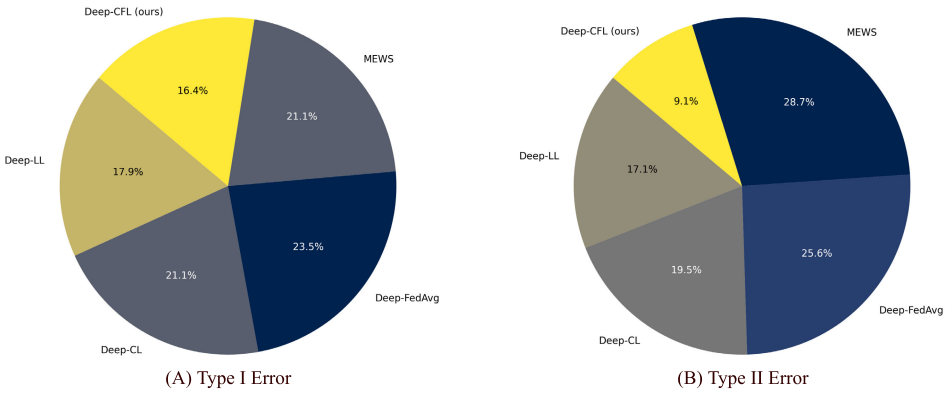


FIGURE 14. Error distribution of comparison methods (for each error type) on MIMIC III for splitting for five clients.

samples), whereas Deep-FedAvg accounts for the highest percentage, with 23.5% (Figure 14). In contrast, Deep-LL, which is expected to perform better due to direct data sharing, accounts for 21.1% of the total samples with Type I errors, on par with traditional methods, such as the MEWS. This outcome demonstrates that data sharing violates security principles and contributes less to prediction performance in this context. The increased precision of Deep-CFL in avoiding false alarms indicates a considerable improvement in ICU performance measures. In a study of 165 samples that resulted in Type II mistakes, the suggested Deep-CFL approach demonstrated superior performance, accounting for

just 9.1% (15 error samples). Further, Deep-CL, the MEWS, and Deep-FedAvg accounted for a higher proportion of these mistakes, with percentages of 19.5%, 28.7%, and 25.6%, respectively. Notably, Deep-FedAvg, which excludes L_i throughout the training phase, has some of the poorest results, emphasizing the importance of using imbalance learning techniques to address the problem of delayed detections in ICU settings successfully.

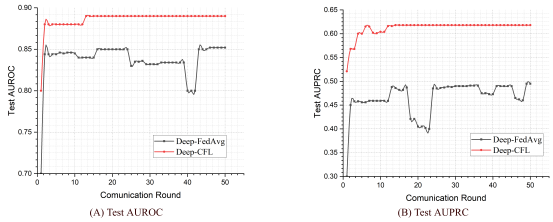


FIGURE 15. Prediction results of federated learning-based methods over 50 communication rounds on eICU (nine-client splitting case).

3) EXPERIMENTAL RESULTS OVER COMMUNICATION ROUNDS

Figure 15 illustrates the average prediction AUROC and AUPRC for FL-based methods (Deep-FedAvg and Deep-CFL) in eICU for the splitting case with nine clients. For AUROC (Figure 15(A)), the performance results of Deep-CFL and Deep-FedAvg rise rapidly by the second communication round, with Deep-CFL maintaining its advantage. Notably, following this first surge, both approaches substantially stabilized, although Deep-CFL regularly reports a higher AUROC in subsequent rounds (over 0.800 of AUROC).

The AUROC for Deep-CFL remains consistent at 0.860 from the fifth round onward, eventually rising to 0.890 in Round 13 and maintaining stability. This constant result demonstrates that Deep-CFL can preserve predictions throughout numerous communication rounds, demonstrating a stable and robust learning process. The AUROC performance for Deep-FedAvg varies, with a notable decrease to 0.830 in Round 25 and oscillations around 0.850 in later rounds. This method ends on a slight increase, reaching and sustaining an AUROC of 0.852 from Round 46 onward. In Figure 15(B), Deep-CFL displays superior performance compared to the traditional FL method for the AUPRC. In particular, Deep-FedAvg experiences a rapid increase in AUPRC by the second communication round and gradually improves until reaching a peak of 0.495 by Round 50. A reduction in the AUPRC occurs between Rounds 18 and 23, exhibiting performance volatility. A result below 0.500 for AUPRC is unconvincing in the problem of class-imbalanced binary classification, especially in clinical settings, where a high degree of event recognition is required.

Moreover, Deep-CFL presents acceptable results with a stable value of 0.618 for AUPRC from the 13^{th} round onwards. From the same FedAvg structure, when adding CFL and L_i, the proposed method demonstrates remarkable improvement in performance. This result suggests that CFL and L_i are better suited to managing the precision-recall balance in FL contexts, particularly in instances including class imbalance. In contrast, Deep-FedAvg, which omits CFL and L_i, fails to achieve the early and consistent performance improvements achieved in Deep-CFL. This result emphasizes the significance of personalized techniques for managing class imbalances in a FL environment. Despite the superior results compared with the comparison methods, the current performance of Deep-CFL must be improved in future studies for practical applications.

E. FUTURE-TIME PREDICTION RESULTS

1) EXPERIMENT WITH FUTURE PREDICTION TIMES

Tables 8, 9, and 10 exhibit the experimental findings of the future-time prediction performance of comparison methods on the RRT, MIMIC III, and eICU datasets, respectively, with variations in the prediction interval length ($k = 8$, 16, and 24) and the observation window size ($d = 8$ h). Table 8 demonstrates that Deep-CFL surpasses the comparative approaches across all three ranges of future prediction intervals. When employing an observation window size of 8 h and forecasting the clinical state in the following 8 h, with a sample size of five clients, Deep-CFL achieves an AUROC of 0.911, an AUPRC of 0.855, and an AP of 0.853. These results are an increase of 2.6% for AUPRC and 6.9% for both AUPRC and AP compared to Deep-CL, which has superior results compared to the rest. When the range of future predictions is expanded while maintaining a fixed value of the observation window, the forecasting process becomes more challenging due to the uncertainty regarding the adequacy of previous information in determining outcomes over an extended future time. The assertion is substantiated by the observation that all comparative methodologies decline in efficacy when the future prediction range is expanded to 16 and 24 h, correspondingly. Nevertheless, Deep-CFL ensures low degradation with a decrease of 1.9% in AUROC, 9.6% in both AUPRC and AP for 16 h and 5.0% in AUROC, 5.1% in AUPRC, and 9.9% in AP for 24 h when the future prediction interval is increased, compared to the scenario in which k is set to 8 h. In contrast, Deep-FedAvg declines by up to 5.0%, 11.6%, and 8.4% in AUROC, AUPRC, and AP values, respectively, when scaling k to 24 h, demonstrating how model-contrastive and imbalance learning contribute to the predictive flexibility of the system. For $k = 24$ h, except Deep-CL and Deep-CFL, the performance of most experimental methods degrades below 0.800 in the AUROC and 0.700 for both AUPRC and AP. Known for its sharing of local data inside a CL framework, Deep-CL has exhibited exceptional performance compared to traditional LL- and FL-based methodologies. However, Deep-CFL demonstrates superior performance compared to Deep-CL, highlighting its capacity to offer accurate predictions while conforming to the rigorous data security protocols necessary for EHR data in the proposed approach.

TABLE 8. Results of future-time prediction of clinical deterioration by number of future prediction intervals on RRT dataset (the observation window size d = 8h for five clients).

Future prediction interval (k)	Method	Metrics (95% CI)		
		AUROC	AUPRC	AP
Next 8 h	Deep-LL	0.847 (0.839 - 0.848)	0.754 (0.751 - 0.759)	0.757 (0.755 - 0.758)
	Deep-CL	0.885 (0.892 - 0.898)	0.786 (0.781 - 0.791)	0.784 (0.781 - 0.794)
	Deep-FedAvg	0.835 (0.831 - 0.837)	0.744 (0.740 - 0.748)	0.745 (0.742 - 0.746)
	MEWS	0.791 (0.782 - 0.794)	0.725 (0.718 - 0.726)	0.728 (0.720 - 0.731)
	Deep-CFL (ours)	**0.911 (0.906 - 0.915)**	**0.855 (0.852 - 0.858)**	**0.853 (0.850 - 0.856)**
Next 16 h	Deep-LL	0.872 (0.868 - 0.875)	0.731 (0.724 - 0.733)	0.733 (0.728 - 0.736)
	Deep-CL	0.888 (0.882 - 0.890)	0.756 (0.751 - 0.759)	0.755 (0.752 - 0.758)
	Deep-FedAvg	0.806 (0.804 - 0.811)	0.737 (0.732 - 0.801)	0.734 (0.732 - 0.736)
	MEWS	0.721 (0.699 - 0.725)	0.659 (0.654 - 0.662)	0.659 (0.652 - 0.663)
	Deep-CFL (ours)	**0.892 (0.886 - 0.893)**	**0.759 (0.752 - 0.764)**	**0.757 (0.754 - 0.759)**
Next 24 h	Deep-LL	0.781 (0.775 - 0.788)	0.695 (0.690 - 0.671)	0.693 (0.690 - 0.698)
	Deep-CL	0.843 (0.841 - 0.845)	0.736 (0.734 - 0.737)	0.732 (0.730 - 0.735)
	Deep-FedAvg	0.785 (0.783 - 0.801)	0.656 (0.652 - 0.670)	0.661 (0.659 - 0.667)
	MEWS	0.714 (0.708 - 0.716)	0.628 (0.622 - 0.629)	0.631 (0.625 - 0.634)
	Deep-CFL (ours)	**0.861 (0.855 - 0.864)**	**0.753 (0.750 - 0.757)**	**0.754 (0.752 - 0.755)**

TABLE 9. Results of future-time prediction of in-ICU mortality by number of future prediction intervals on MIMIC III (the observation window size d = 8h for clients).

Future prediction interval (k)	Method	Metrics (95% CI)		
		AUROC	AUPRC	AP
Next 8 h	Deep-LL	0.862 (0.858 - 0.865)	0.701 (0.697 - 0.702)	0.704 (0.702 - 0.709)
	Deep-CL	0.911 (0.908 - 0.914)	0.731 (0.729 - 0.735)	0.736 (0.733 - 0.738)
	Deep-FedAvg	0.905 (0.903 - 0.910)	0.720 (0.716 - 0.728)	0.724 (0.719 - 0.725)
	MEWS	0.861 (0.854 - 0.862)	0.701 (0.695 - 0.703)	0.702 (0.699 - 0.705)
	Deep-CFL (ours)	**0.928 (0.921 - 0.934)**	**0.745 (0.740 - 0.758)**	**0.748 (0.746 - 0.754)**
Next 16 h	Deep-LL	0.812 (0.808 - 0.816)	0.685 (0.680 - 0.688)	0.689 (0.685 - 0.704)
	Deep-CL	0.863 (0.861 - 0.867)	0.697 (0.685 - 0.701)	0.702 (0.696 - 0.703)
	Deep-FedAvg	0.844 (0.839 - 0.845)	0.682 (0.676 - 0.685)	0.685 (0.681 - 0.688)
	MEWS	0.814 (0.810 - 0.816)	0.689 (0.686 - 0.691)	0.687 (0.682 - 0.688)
	Deep-CFL (ours)	**0.894 (0.885 - 0.895)**	**0.741 (0.737 - 0.744)**	**0.743 (0.740 - 0.745)**
Next 24 h	Deep-LL	0.796 (0.790 - 0.799)	0.625 (0.621 - 0.626)	0.623 (0.620 - 0.626)
	Deep-CL	0.813 (0.804 - 0.814)	0.656 (0.652 - 0.659)	0.658 (0.655 - 0.662)
	Deep-FedAvg	0.809 (0.802 - 0.811)	0.652 (0.650 - 0.658)	0.653 (0.650 - 0.655)
	MEWS	0.787 (0.782 - 0.789)	0.651 (0.647 - 0.653)	0.654 (0.651 - 0.659)
	Deep-CFL (ours)	**0.882 (0.878 - 0.886)**	**0.705 (0.701 - 0.707)**	**0.706 (0.704 - 0.707)**

TABLE 10. Results of future-time prediction of in-ICU mortality by number of future prediction intervals on eICU (the observation window size d = 8h for clients).

Future prediction interval (k)	Method	Metrics (95% CI)		
		AUROC	AUPRC	AP
Next 8 h	Deep-LL	0.801 (0.798 - 0.804)	0.733 (0.725 - 0.734)	0.734 (0.724 - 0.736)
	Deep-CL	0.853 (0.851- 0.857)	0.748 (0.742 - 0.751)	0.746 (0.743 - 0.747)
	Deep-FedAvg	0.849 (0.846 - 0.853)	0.736 (0.734 - 0.739)	0.737 (0.734 - 0.739)
	MEWS	0.799 (0.795 - 0.803)	0.726 (0.723 - 0.729)	0.726 (0.721 - 0.728)
	Deep-CFL (ours)	**0.855 (0.848 - 0.856)**	**0.762 (0.759 - 0.765)**	**0.766 (0.761 - 0.767)**
Next 16 h	Deep-LL	0.726 (0.723 - 0.731)	0.685 (0.682 - 0.686)	0.689 (0.685 - 0.692)
	Deep-CL	0.798 (0.796 - 0.805)	0.732 (0.729 - 0.734)	0.736 (0.733 - 0.738)
	Deep-FedAvg	0.787 (0.784 - 0.792)	0.711 (0.705 - 0.716)	0.713 (0.709 - 0.715)
	MEWS	0.729 (0.725 - 0.732)	0.691 (0.688 - 0.695)	0.704 (0.699 - 0.705)
	Deep-CFL (ours)	**0.820 (0.817 - 0.824)**	**0.754 (0.750 - 0.758)**	**0.756 (0.752 - 0.759)**
Next 24 h	Deep-LL	0.715 (0.711 - 0.718)	0.648 (0.642 - 0.652)	0.647 (0.643 - 0.651)
	Deep-CL	0.736 (0.732 - 0.739)	0.689 (0.684 - 0.693)	0.692 (0.688 - 0.695)
	Deep-FedAvg	0.726 (0.719 - 0.727)	0.659 (0.653 - 0.760)	0.657 (0.654 - 0.659)
	MEWS	0.708 (0.702 - 0.710)	0.637 (0.632 - 0.638)	0.639 (0.638 - 0.645)
	Deep-CFL (ours)	**0.784 (0.781 - 0.786)**	**0.722 (0.718 - 0.725)**	**0.726 (0.721 - 0.727)**

For ICU mortality prediction (as shown in Table 9 and 10 for MIMIC III and eICU, respectively), performance tends to degrade when expanding the value of k is depicted for all experimental methods. Despite that trend, Deep-CFL

exhibits stable performance and outperforms all comparison methods in all prediction interval value settings. Typically, with $k = 24$ h in MIMIC III, Deep-CFL achieves 0.882 in AUROC, 0.705 in AUPRC, and 0.706 in AP, decreasing by 1.2%, 3.6%, and 3.7%, respectively, compared to its performance in the case for $k = 16$ h. In contrast, Deep-CL reduces AUROC, AUPRC, and AP by 5.0%, 4.1%, and 4.4%, respectively, indicating much greater reductions than the proposed technique in the same configuration. Similar results were observed in the eICU dataset, especially in situations in which clinical outcomes were predicted 24 h in advance. The performance metrics for AUPRC and AP fall below 0.700 for all comparison methods except Deep-CFL. The proposed method consistently maintains performance levels over 0.720 for both of these metrics. Based on these results, particularly in the context of prediction across various intervals, Deep-CFL demonstrates remarkable resilience to performance dips associated with extended prediction horizons and consistently outperforms comparative methods in accuracy and reliability, confirming its superior adaptability and effectiveness in future prediction tasks.

In future-time prediction tasks, Deep-CFL maintains high performance across different future prediction intervals, effectively managing the increased uncertainty and complexity associated with longer-term predictions through its contrastive learning component and efficient communication framework. In contrast, Local Learning (Deep-LL) and FedAvg show greater performance declines over longer prediction intervals, indicating their limitations in dealing with extended future predictions and fragmented data. MEWS continues to underperform compared to Deep-CFL, especially in predicting minority-class events, demonstrating the advantage of using advanced machine learning techniques.

These results validate the robustness of the Deep-CFL system in predicting future clinical states over varying intervals. Deep-CFL's superior performance across different future prediction intervals can be attributed to its advanced FL framework combined with contrastive learning and supervised learning techniques. The minimal performance degradation observed in Deep-CFL, even as the prediction interval increases, highlights its ability to effectively leverage historical data to make accurate future predictions. This resilience is due to Deep-CFL's ability to balance the contributions of local models with the global model through contrastive learning, ensuring each local model benefits from the broader data distribution without direct data sharing. The system's incorporation of imbalance learning techniques helps maintain high precision and recall, which is particularly crucial in handling the imbalanced nature of clinical data. The system's use of imbalanced learning approaches helps to retain high precision and recall, which is especially important when dealing with the imbalanced nature of clinical data. Deep-CFL's steady performance as the prediction interval increases implies that the system effectively controls the inherent uncertainty and complexity of longer-term predictions. In contrast to comparison models, which exhibit more significant performance reductions with longer prediction intervals, Deep-CFL's ability to retain accuracy illustrates its robustness and practical usefulness in real-world healthcare situations. Deep-CFL's consistent performance across multiple scenarios demonstrates its ability to give reliable and secure predictive insights, making it an important tool for assisting clinical decision-making in ICU settings. Overall, Deep-CFL's robust performance across varied prediction intervals demonstrates its capacity to handle the challenges of long-term prediction.

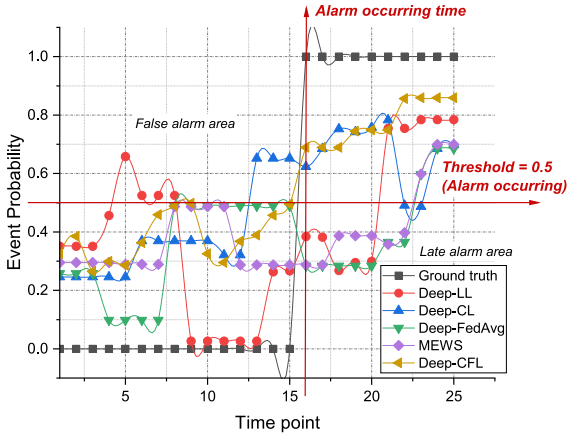


FIGURE 16. Probability sample results of comparison methods in the RRT dataset for the future 24-h clinical outcome prediction.

For a clearer understanding of the outcomes within the future-prediction framework, we visualized the probability predictions of events for a patient sample, as illustrated in Figures 16 and 17. Figure 16 illustrates the prediction probability of future 24-h clinical outcomes of a testing sample in the RRT dataset for five clients. This pattern tends to perform a normal state in the first time steps and an event occurs in the later time steps. Based on the ground truth of this sample, we placed a vertical line to mark the alarm occurrence time and a horizontal line to mark the threshold probability value (threshold = 0.50 for all cases in this study) such that any time point that is returned with a higher probability than the threshold is considered an event time. The two lines divide the resulting sample illustration into four regions. The top left area of the figure is the false alarm area, where false prediction points for the nonevent class are placed. The bottom right part is the late alarm area, containing false prediction points for the event class. The Deep-LL (red curve) and Deep-CL (blue curve) methods distribute the most prediction samples in the false alarm area among the compared methods. This result demonstrates that the excessive sensitivity due to the influence of Li causes these methods to raise unnecessary alarms during the prediction process. Both Deep-FedAvg (green curve) and the MEWS (violet curve), without implementing L_i, return the latest alarm samples in the late alarm region. This outcome is a particularly fatal error with prediction in the ICU, where

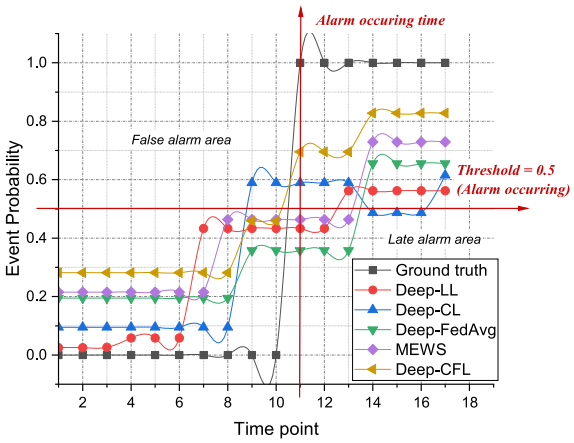


FIGURE 17. Probability sample results of comparison methods in the eICU dataset for the future 16-h clinical outcome prediction.

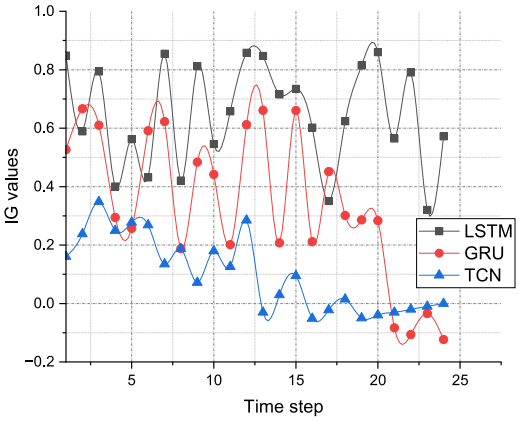


FIGURE 18. IG values of the extracted representations from three temporal-based model structures over 24 time steps in a sample of the RRT dataset.

the prediction model cannot recognize the time points that require intervention from the intensive care team. In contrast, Deep-CFL (brown curve) displays exceptional efficiency in applying L_i while ensuring a balance between sensitivity and specificity, with none of its prediction points falling within the two error regions. The proposed method returns predicted points with an event probability higher than the threshold of 0.50 when the ground truth is an event and vice versa.

Figure 17 presents the prediction probability of future 16-h clinical outcomes of a testing sample in the eICU dataset for five clients. This pattern can be considered easy, as most methods avoid false alarm regions in prediction errors. Therefore, we compared the performance of methods based on the late alarm area and true alarm area (top right region of the figure). In the late alarm area, the MEWS, Deep-FedAvg, and Deep-CL methods exhibit more instances of late prediction (three samples), whereas Deep-CFL guarantees performance with zero samples of late alarm. In the results of the true alarm area, the more efficient method returns a higher probability and a more stable curve. In addition, Deep-CFL excels in this criterion by returning the highest probability and a stable curve at the last time points. In contrast, although displaying a stable curve at the early time points from the alarm occurring time (time point 11), Deep-CL fluctuates strongly from the 14^{th} time point before returning to equilibrium at the last time point.

2) BASE DEEP LEARNING MODEL STRUCTURE VALIDATION

To evaluate the contribution of IG in the proposed method, we performed an additional experiment on the RRT dataset for splitting with five clients. We conducted experiments using Deep-CFL to predict clinical outcomes for 8 and 16 h into the future. However, in Deep-CFL, different structures of the base model were used as comparison methods. For the time-series context, typical and widely used RNN-based models were selected, including the LSTM, TCN, and GRU. We applied the discussed encoders to derive the representation z for each input sample x. For example, $z_{lstm} = E_{lstm}(x)$, where E_{lstm} refers to an LSTM-based encoder. Then, the IG values of each extracted z were calculated and compared to select the most optimal structure for the system.

Figure 18 presents the IG values of the extracted representation from three temporal-based model structures over 24 time steps in a sample from the RRT dataset. The IG values of the extracted features from LSTM always reach the highest values across time steps compared to the GRU and TCN-based encoders. The IG values for GRU have relatively high values but gradually decrease and have negative values when running until the last time steps, revealing that the influence of the extracted representation of this method gradually decreases over time. Based on this result, choosing the LSTM structure for the base model of the system is considered the choice of IG. In contrast, choosing the TCN or GRU structure for Deep-CFL can be considered a random selection.

Table 11 presents the experiment results of three model-selected cases for Deep-CFL in the RRT dataset for predicting the future 8- and 16-h clinical statuses for five clients. Corresponding to the results of comparing IG values, LSTM (choice based on IG) achieves superior results compared to the other two random-choice structures. These results demonstrate the importance of IG in improving model selection, interpretability, and performance within the prediction system. By directing the selection to LSTM as the most successful architecture for future event prediction, IG displays its utility as a reliable criterion for choosing models that capture meaningful information. This technique improves model interpretability, allowing for deeper insight into the decision-making process, and establishes IG as a useful benchmarking tool for comparing DL structures. The higher performance of LSTM, as demonstrated by IG, underlines its capacity to manage complicated temporal contexts, which is vital for ICU predictions. Overall, the application of IG lays a solid platform for future research into XAI, with the potential to advance the construction of reliable and effective prediction systems in hospitals.

TABLE 11. Results for deep learning structures for the base model.

Future prediction interval	Method	Metrics (95% CI)		
		AUROC	AUPRC	AP
Next 8 h	TCN (random choice)	0.874 (0.872 - 0.877)	0.829 (0.825 - 0.833)	0.831 (0.828 - 0.835)
	GRU (random choice	0.886 (0.882 - 0.892)	0.851 (0.847 - 0.853)	0.848 (0.844 - 0.849)
	LSTM (IG choice)	**0.911 (0.906 - 0.915)**	**0.855 (0.852 - 0.858)**	**0.853 (0.850 - 0.856)**
Next 16 h	TCN (random choice)	0.843 (0.841 - 0.848)	0.744 (0.738 - 0.745)	0.746 (0.738 - 0.748)
	GRU (random choice)	0.871 (0.867 - 0.874)	0.749 (0.746 - 0.751)	0.752 (0.749 - 0.755)
	LSTM (IG choice)	**0.892 (0.886 - 0.893)**	**0.759 (0.752 - 0.764)**	**0.757 (0.754 - 0.759)**

V. DISCUSSION

The results in Section IV demonstrate the superior performance of Deep-CFL over comparative methods across a variety of case studies and prediction contexts. We conducted ablation studies to evaluate the contribution of CFL, imbalanced learning, and IG techniques in solving the challenges of EHR data and alarm error problems of predictive tasks in the ICU.

Regarding interpretability, IG offers a robust and effective XAI solution for DL in healthcare application systems. This method evaluates the impact of extracted features from a certain DL structure on the classification model, changing its structure and increasing prediction performance. The ability to evaluate IG values using magnitude gradients boosts physicians' trust in the rationale for selecting a DL network architecture for a medical system. The ablation trial results for IG show its efficacy. This study illustrates that using IG yields superior outcomes while needing fewer training and validation operations than the typical hyperparameter approach.

Deep-CFL contributes to the solution of the problem of alarm errors in prediction systems by addressing the underlying problems. To address security concerns that may hinder adoption and data sharing between healthcare institutions, Deep-CFL combines contrastive learning with the FedAvg algorithm to ensure security while keeping strong prediction performance. The proposed method outperforms CL approaches (Deep-CL), which profit from breaking data security rules. Compared to the standard FedAvg approach (Deep-FL), Deep-CFL addresses the often overlooked element of the model performance difference between local and global models, resulting in more stable and superior performance. Contrastive loss improves local models before the FL algorithm combines them. Initially, contrastive loss enables local models to produce representations that closely mirror those of the global model. As a result, it enables local models to learn more refined representations than their predecessors, steadily improving until they reach a satisfactory performance level. Furthermore, Deep-CFL ensures that each local update adds significantly to the global knowledge base by encouraging a more sophisticated update process that highlights similarities with the global model as well as contrasts from previous local iterations. This methodological innovation improves model accuracy while reducing the danger of overfitting to local variables, a common problem in federated contexts.

Imbalance learning is used in the system as a primary strategy for addressing data quality issues (data imbalance), which have a direct impact on the alarm error rate. This experiment was carried done without modifying the datasets' class distribution, such as through under- or oversampling, to simulate practical situations in which ICU data retain their intrinsic qualities. This strategy assures that physicians do not change data settings, which limits the development of prediction biases that could jeopardize the validity of the results. Extensive analysis provides a clear knowledge of the impact of imbalance loss on comparison methods. This research highlights the importance of imbalance learning in correcting data imbalances, lowering late alarms, and providing early interventions to ICU patients.

A. BENEFITS FOR REAL-WORLD APPLICATION

From an application perspective, our experiments focus on two tasks corresponding to three datasets: predicting clinical deterioration (RRT dataset) and predicting ICU mortality (MIMIC III and eICU datasets). For clinical deterioration prediction using the RRT dataset, Deep-CFL can be applied to monitor patients in ICUs and predict clinical deterioration, allowing for timely interventions. For ICU mortality prediction using the MIMIC III and eICU datasets, Deep-CFL is effective in predicting ICU mortality, aiding in the early identification of high-risk patients. FL approach is ideal for decentralized healthcare systems where data privacy and security are paramount. The method enables collaboration across multiple healthcare centers without the need for data sharing, making it scalable and practical for large healthcare networks. By addressing these critical areas, Deep-CFL offers a robust, accurate, and interpretable solution for early prediction of clinical status in ICU settings, supporting effective patient management and improving clinical outcomes.

The proposed system has practical advantages for practical integration in the emergency department. Firstly, it dramatically improves data security by keeping EHR data localized within each healthcare facility, reducing the danger of data breaches, and protecting patient privacy without the requirement for data exchange across institutions. Second, combining contrastive learning and federated averaging enhances clinical alarm accuracy, significantly lowering false and late alarms. This results in more timely and consistent clinical treatments, which are crucial in ICU settings. Third, using explainable AI techniques improves the interpretability

of model predictions, giving healthcare practitioners clear insights into the AI system's decision-making process. Finally, the FL framework facilitates scalability, allowing the system to be deployed across different healthcare centers without the need for a centralized data source. This decentralized method protects data privacy while also allowing for more institutional adoption and collaboration.

Several practical considerations must be addressed to implement a Deep-CFL system effectively. First, participating healthcare centers must ensure adequate computing infrastructure to handle local model training and updates, including sufficient CPU/GPU resources and connectivity Reliable Internet. The two main teams in the system are healthcare practitioners and IT staff. The IT team takes on the role of integrating proposed techniques and developing the current system. Doctors monitor, verify, and make decisions about the intervention process. When Deep-CFL is integrated into RRS, it plays an alarming role through DL risk score. A risk score exceeding a certain threshold means that doctors need to intervene and provide treatment based on their experience. Pilot tests should be conducted in a small number of healthcare centers to identify and address potential challenges before scaling up implementation.

B. LIMITATION

As part of the RRS project, the ultimate goal of this research was to implement the proposed algorithm in a real-world hospital monitoring system. Despite the encouraging results, we acknowledge that this study still has limitations that must be addressed. First, data diversity can limit the usefulness of the system. Data collection and representation standards vary by hospital or healthcare center, necessitating a more flexible and generally applicable preprocessor with numerous pre-input features when incorporating them into the prediction model. Moreover, IG improves system interpretability; however, converting these scientific discoveries into practical clinical information might be challenging. This approach necessitates establishing an efficient correlation system between AI and medicine. Finally, for an ICU prediction system, a sensitivity of more than 80% ensures reliable decision-making ability. Despite achieving higher performance on all three datasets, Deep-CFL exhibits sensitivity concerns in specific samples. Improving and maintaining system performance is the primary focus. However, the novel concept of integrating contrastive learning and FL suggests a bright future for medical AI application systems.

The computational burden is also considered a shortcoming of our study. The computational load of Deep-CFL is mostly due to the combination of federated learning, contrastive learning, and supervised learning approaches. During our tests, we discovered that this burden is greatest in the local training phase of each communication round, where both contrastive and supervised learning losses are applied concurrently. This combination greatly increases the training length when compared to other approaches in similar experimental conditions. In actual implementations, this computational cost has little impact on the prediction system's real-time performance, but it does present issues during the development phase. Because of this increased complexity, the development process takes longer and requires more resources. Potential options include using advanced optimization techniques and studying incremental training methodologies, to balance the computational load across centers.

VI. CONCLUSION

In this investigation, we introduced Deep-CFL, an innovative approach to predicting clinical status in the ICU. This novel approach combines FL, contrastive learning, and an XAI technique to address the interpretability and alarm error problems of the ICU prediction system. The critical concept of Deep-CFL is to improve the learning ability of local models by teaching them essential capabilities from the global model, which consistently delivers more accurate predictions. Through comprehensive experiments on multiple datasets and analyses in various case studies and prediction contexts, Deep-CFL demonstrated outstanding performance in predicting clinical outcomes, effectively managing the imbalance between classes, and enhancing model interpretability without compromising patient data privacy. The application of Deep-CFL in the RRS project demonstrates its potential to transform patient monitoring systems by enabling timely and accurate clinical condition prediction. This study enables further research into more sophisticated FL models, improving model interpretation and integrating advanced AI technology into healthcare procedures. Finally, the success of Deep-CFL in resolving the complexity of ICU patient data analysis is a critical step forward in the application of FL and AI in healthcare.

REFERENCES

[1] J. C. Hurley, "Trends in ICU mortality and underlying risk over three decades among mechanically ventilated patients. A group level analysis of cohorts from infection prevention studies," *Ann. Intensive Care*, vol. 13, no. 1, p. 62, Jul. 2023.

[2] S. Richardson et al., "Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area," *J. Amer. Med. Assoc.*, vol. 323, no. 20, p. 2052, May 2020.

[3] P. Quah, A. Li, and J. Phua, "Mortality rates of patients with COVID-19 in the intensive care unit: A systematic review of the emerging literature," *Crit. Care*, vol. 24, no. 1, pp. 1–4, Dec. 2020.

[4] S. C. Auld, K. R. V. Harrington, M. W. Adelman, C. J. Robichaux, E. C. Overton, M. Caridi-Scheible, C. M. Coopersmith, and D. J. Murphy, "Trends in ICU mortality from coronavirus disease 2019: A tale of three surges," *Crit. Care Med.*, vol. 50, no. 2, pp. 245–255, 2022.

[5] L. L. Weed, "Medical records that guide and teach," *New England J. Med.*, vol. 278, no. 11, pp. 593–600, 1968.

[6] P. Carayon, R. Cartmill, M. A. Blosky, R. Brown, M. Hackenberg, P. Hoonakker, A. S. Hundt, E. Norfolk, T. B. Wetterneck, and J. M. Walker, "ICU nurses' acceptance of electronic health records," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 6, pp. 812–819, Nov. 2011.

[7] S. Asefzadeh, M. H. Yarmohammadian, A. Nikpey, and G. Atighechian, "Clinical risk assessment in intensive care unit," *Int. J. Preventive Med.*, vol. 4, no. 5, pp. 592–598, 2013.

[8] B. Gholami, W. M. Haddad, and J. M. Bailey, "AI in the ICU: In the intensive care unit, artificial intelligence can keep watch," *IEEE Spectr.*, vol. 55, no. 10, pp. 31–35, Oct. 2018.

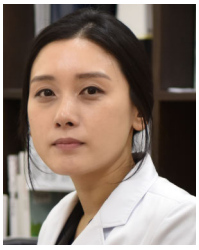
[9] I. Keshta and A. Odeh, "Security and privacy of electronic health records: Concerns and challenges," *Egyptian Informat. J.*, vol. 22, no. 2, pp. 177–183, Jul. 2021.
[10] S. Bhartiya and D. Mehrotra, "Threats and challenges to security of electronic health records," in *Proc. 9th Int. Conf. Quality, Rel., Secur. Robustness Heterogeneous Netw.*, Greader Noida, India. Cham, Switzerland: Springer, Jan. 2013, pp. 543–559.
[11] G. Tejay, G. Dhillon, and A. G. Chin, "Data quality dimensions for information systems security: A theoretical exposition," in *Proc. Secur. Manage., Integrity, Internal Control Inf. Syst.* Boston, MA, USA: Springer, 2005, pp. 21–39.
[12] P. Vimalachandran, H. Wang, Y. Zhang, B. Heyward, and F. Whittaker, "Ensuring data integrity in electronic health records: A quality health care implication," in *Proc. Int. Conf. Orange Technol. (ICOT)*, Dec. 2016, pp. 20–27.
[13] M. Talha, A. A. El Kalam, and N. Elmarzouqi, "Big data: Trade-off between data quality and data security," *Proc. Comput. Sci.*, vol. 151, pp. 916–922, Jan. 2019.
[14] A. K. Saiod, D. van Greunen, and A. Veldsman, "Electronic health records: Benefits and challenges for data quality," in *Handbook of Large-Scale Distributed Computing in Smart Healthcare*. Switzerland: Springer, 2017, pp. 123–156.
[15] S. Downey, M. Indulska, and S. Sadiq, "Perceptions and challenges of EHR clinical data quality," in *Proc. ACIS*, 2019, pp. 1–12.
[16] M. Botha, A. Botha, and M. Herselman, "Data quality challenges: A content analysis in the e-health domain," in *Proc. 4th World Congr. Inf. Commun. Technol. (WICT)*, Dec. 2014, pp. 107–112.
[17] D. E. Leisman, "Rare events in the ICU: An emerging challenge in classification and prediction," *Crit. Care Med.*, vol. 46, no. 3, pp. 418–424, 2018.
[18] M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, and K. Borgwardt, "Early prediction of sepsis in the ICU using machine learning: A systematic review," *Frontiers Med.*, vol. 8, May 2021, Art. no. 607952.
[19] J. Liu, X. X. Chen, L. Fang, J. X. Li, T. Yang, Q. Zhan, K. Tong, and Z. Fang, "Mortality prediction based on imbalanced high-dimensional ICU big data," *Comput. Ind.*, vol. 98, pp. 218–225, Jun. 2018.
[20] A. E. W. Johnson and R. G. Mark, "Real-time mortality prediction in the intensive care unit," in *Proc. AMIA Annu. Symp.*, 2017, p. 994.
[21] J. Norrie, "The challenge of implementing AI models in the ICU," *Lancet Respiratory Med.*, vol. 6, no. 12, pp. 886–888, Dec. 2018.
[22] M. J. Patton and V. X. Liu, "Predictive modeling using artificial intelligence and machine learning algorithms on electronic health record data," *Crit. Care Clinics*, vol. 39, no. 4, pp. 647–673, Oct. 2023.
[23] P. Mathur and M. L. Burns, "Artificial intelligence in critical care," *Int. Anesthesiol. Clinics*, vol. 57, no. 2, pp. 89–102, 2019.
[24] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, Dec. 2019.
[25] A. I. F. Poon and J. J. Y. Sung, "Opening the black box of AI-medicine," *J. Gastroenterol. Hepatol.*, vol. 36, no. 3, pp. 581–584, 2021.
[26] A. Adadi and M. Berrada, "Explainable AI for healthcare: From black box to interpretable models," in *Proc. Embedded Syst. Artif. Intell. (ESAI)*, Fez, Morocco. Cham, Switzerland: Springer, 2019, pp. 327–337.
[27] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, vol. 54, 2017, pp. 1273–1282.
[28] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.
[29] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.
[30] P. S. Chan, R. Jain, B. K. Nallmothu, R. A. Berg, and C. Sasson, "Rapid response teams: A systematic review and meta-analysis," *Arch. Internal Med.*, vol. 170, no. 1, pp. 18–26, 2010.
[31] H. Gao, A. McDonnell, D. A. Harrison, T. Moore, S. Adam, K. Daly, L. Esmonde, D. R. Goldhill, G. J. Parry, A. Rashidian, C. P. Subbe, and S. Harvey, "Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward," *Intensive Care Med.*, vol. 33, no. 4, pp. 667–679, Mar. 2007.
[32] I. D. Mapp, L. L. Davis, and H. Krowchuk, "Prevention of unplanned intensive care unit admissions and hospital mortality by early warning systems," *Dimensions Crit. Care Nursing*, vol. 32, no. 6, pp. 300–309, 2013.
[33] M. E. B. Smith, J. C. Chiovaro, M. O'Neil, D. Kansagara, A. R. Quiñones, M. Freeman, M. L. Motu'apuaka, and C. G. Slatore, "Early warning system scores for clinical deterioration in hospitalized patients: A systematic review," *Ann. Amer. Thoracic Soc.*, vol. 11, no. 9, pp. 1454–1465, 2014.
[34] D. A. Jones, M. A. DeVita, and R. Bellomo, "Rapid-response teams," *New England J. Med.*, vol. 365, no. 2, pp. 139–146, 2011.
[35] K. Strand and H. Flaatten, "Severity scoring in the ICU: A review," *Acta Anaesthesiologica Scandinavica*, vol. 52, no. 4, pp. 467–478, Apr. 2008.
[36] M. Larvin and M. Mcmahon, "APACHE-II score for assessment and monitoring of acute pancreatitis," *Lancet*, vol. 334, no. 8656, pp. 201–205, Jul. 1989.
[37] D. Wagner, E. Draper, and W. Knaus, "Development of APACHE III," *Crit. Care Med.*, vol. 17, no. 2, pp. S199–S203, 1989.
[38] T. Olsson, A. Terent, and L. Lind, "Rapid emergency medicine score: A new prognostic tool for in-hospital mortality in nonsurgical emergency department patients," *J. Internal Med.*, vol. 255, no. 5, pp. 579–587, May 2004.
[39] K. J. Rhee, C. J. Fisher, and N. H. Willitis, "The rapid acute physiology score," *Amer. J. Emergency Med.*, vol. 5, no. 4, pp. 278–282, Jul. 1987.
[40] C. P. Subbe, R. G. Davies, E. Williams, P. Rutherford, and L. Gemmell, "Effect of introducing the modified early warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions," *Anaesthesia*, vol. 58, no. 8, pp. 797–802, Aug. 2003.
[41] J. Gardner-Thorpe, N. Love, J. Wrightson, S. Walsh, and N. Keeling, "The value of modified early warning score (MEWS) in surgical in-patients: A prospective observational study," *Ann. Roy. College Surgeons England*, vol. 88, no. 6, pp. 571–575, Oct. 2006.
[42] S. Uppanisakorn, R. Bhurayanontachai, J. Boonyarat, and J. Kaewpradit, "National early warning score (NEWS) at ICU discharge can predict early clinical deterioration after ICU transfer," *J. Crit. Care*, vol. 43, pp. 225–229, Feb. 2018.
[43] J. Kwon, Y. Lee, Y. Lee, S. Lee, and J. Park, "An algorithm based on deep learning for predicting in-hospital cardiac arrest," *J. Amer. Heart Assoc.*, vol. 7, no. 13, Jul. 2018, Art. no. e008678.
[44] F. J. R. Catling and A. H. Wolff, "Temporal convolutional networks allow early prediction of events in critical care," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 3, pp. 355–365, Mar. 2020.
[45] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010.
[46] F. E. Shamout, T. Zhu, P. Sharma, P. J. Watkinson, and D. A. Clifton, "Deep interpretable early warning system for the detection of clinical deterioration," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 437–446, Feb. 2020.
[47] A. Tversky and D. J. Koehler, "Support theory: A nonextensional representation of subjective probability," *Psychol. Rev.*, vol. 101, no. 4, pp. 547–567, 1994.
[48] T.-C. Do, H.-J. Yang, G.-S. Lee, S.-H. Kim, and B.-G. Kho, "Rapid response system based on graph attention network for predicting in-hospital clinical deterioration," *IEEE Access*, vol. 11, pp. 29091–29100, 2023.
[49] T. K. Dang, K. C. Tan, M. Choo, N. Lim, J. Weng, and M. Feng, "Building ICU in-hospital mortality prediction model with federated learning," in *Federated Learning: Privacy and Incentive*. Switzerland: Springer, 2020, pp. 255–268.
[50] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. 3rd Mach. Learn. Syst. Conf.*, vol. 2, 2020, pp. 429–450.
[51] T. Deng, H. Hamdan, R. Yaakob, and K. A. Kasmiran, "Personalized federated learning for in-hospital mortality prediction of multi-center ICU," *IEEE Access*, vol. 11, pp. 11652–11663, 2023.
[52] C. T. Dinh, N. H. Tran, and T. D. Nguyen, "Personalized federated learning with Moreau envelopes," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 21394–21405.
[53] A. Georgoutsos, P. Kerasiotis, and V. Kantere, "Federated learning performance on early ICU mortality prediction with extreme data distributions," in *Proc. Int. Conf. Web Inf. Syst. Eng.* Heidelberg, Germany: Springer, 2023, pp. 483–495.
[54] L. Mondrejevski, I. Miliou, A. Montanino, D. Pitts, J. Hollmén, and P. Papapetrou, "FLICU: A federated learning workflow for intensive care unit mortality prediction," in *Proc. IEEE 35th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2022, pp. 32–37.

[55] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[56] R. Atangana, D. Tchiotsop, G. Kenne, and L. Chanel, "EEG signal classification using LDA and MLP classifier," *Health Informat. Int. J.*, vol. 9, no. 1, pp. 14–32, Feb. 2020.
[57] J. J. Bird, J. Kobylarz, D. R. Faria, A. Ekárt, and E. P. Ribeiro, "Cross-domain MLP and CNN transfer learning for biological signal processing: EEG and EMG," *IEEE Access*, vol. 8, pp. 54789–54801, 2020.
[58] M. Ramkumar, C. G. Babu, K. V. Kumar, D. Hepsiba, A. Manjunathan, and R. S. Kumar, "ECG cardiac arrhythmias classification using DWT, ICA and MLP neural networks," *J. Phys., Conf. Ser.*, vol. 1831, no. 1, Mar. 2021, Art. no. 012015.
[59] A. H. Khine, W. Wettayaprasit, and J. Duangsuwan, "Ensemble CNN and MLP with nurse notes for intensive care unit mortality," in *Proc. 16th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2019, pp. 236–241.
[60] M. F. Begum and S. Narayan, "Comprehensive performance assessment of multi-neural ensemble model for mortality prediction in ICU," *IEEE Access*, early access, Oct. 13, 2024, doi: 10.1109/ACCESS.2023.3324459.
[61] N. El-Rashidy, S. El-Sappagh, T. Abuhmed, S. Abdelrazek, and H. M. El-Bakry, "Intensive care unit mortality prediction: An improved patient-specific stacking ensemble model," *IEEE Access*, vol. 8, pp. 133541–133564, 2020.
[62] M. A. Rehmat, M. A. Hassan, M. H. Khalid, and M. Dilawar, "Next level of hospitalisation through smart ICU," *Intell. Syst. Appl.*, vol. 14, May 2022, Art. no. 200080.
[63] H. Maharlou, S. R. N. Kalhori, S. Shahbazi, and R. Ravangard, "Predicting length of stay in intensive care units after cardiac surgery: Comparison of artificial neural networks and adaptive neuro-fuzzy system," *Healthcare Informat. Res.*, vol. 24, no. 2, p. 109, 2018.
[64] J. Deasy, P. Liò, and A. Ercole, "Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or curation," *Sci. Rep.*, vol. 10, no. 1, p. 22129, Dec. 2020.
[65] B. Adhikari, A. Shrestha, S. Mishra, S. Singh, and A. K. Timalsina, "EEG based directional signal classification using RNN variants," in *Proc. IEEE 3rd Int. Conf. Comput., Commun. Secur. (ICCCS)*, Oct. 2018, pp. 218–223.
[66] H. Al-Askar, N. Radi, and Á. MacDermott, "Recurrent neural networks in medical data analysis and classifications," in *Applied Computing in Medicine and Health*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 147–165.
[67] T. Najafi, R. Jaafar, R. Remli, and W. A. W. Zaidi, "A classification model of EEG signals based on RNN-LSTM for diagnosing focal and generalized epilepsy," *Sensors*, vol. 22, no. 19, p. 7269, Sep. 2022.
[68] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[69] H. Zheng and D. Shi, "Using a LSTM-RNN based deep learning framework for ICU mortality prediction," in *Proc. 15th Int. Conf. Web Inf. Syst. Appl. (WISA)*, Taiyuan, China. Heidelberg, Germany: Springer, Sep. 2018, pp. 60–67.
[70] S. Maheshwari, A. Agarwal, A. Shukla, and R. Tiwari, "A comprehensive evaluation for the prediction of mortality in intensive care units with LSTM networks: Patients with cardiovascular disease," *Biomed. Eng./Biomedizinische Technik*, vol. 65, no. 4, pp. 435–446, Aug. 2020.
[71] K. Yu, M. Zhang, T. Cui, and M. Hauskrecht, "Monitoring ICU mortality risk with a long short-term memory recurrent neural network," in *Proc. Pacific Symp. Biocomputing*, Dec. 2019, pp. 103–114.
[72] S. Liu, X. Wang, Y. Xiang, H. Xu, H. Wang, and B. Tang, "Multi-channel fusion LSTM for medical event prediction using EHRs," *J. Biomed. Informat.*, vol. 127, Mar. 2022, Art. no. 104011.
[73] S. Liu, J. J. Schlesinger, A. B. McCoy, T. J. Reese, B. Steitz, E. Russo, B. Koh, and A. Wright, "New onset delirium prediction using machine learning and long short-term memory (LSTM) in electronic health record," *J. Amer. Med. Inform. Assoc.*, vol. 30, no. 1, pp. 120–131, Dec. 2022.
[74] A. C. Kiser, K. Eilbeck, and B. T. Bucher, "Developing an LSTM model to identify surgical site infections using electronic healthcare records," in *Proc. AMIA Summits Transl. Sci.*, 2023, p. 330.
[75] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1857–1865.
[76] C. Wang, C. Deng, and S. Wang, "Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," *Pattern Recognit. Lett.*, vol. 136, pp. 190–197, Aug. 2020.
[77] K. Pasupa, S. Vatathanavaro, and S. Tungjitnob, "Convolutional neural networks based focal loss for class imbalance problem: A case study of canine red blood cells morphology classification," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 11, pp. 15259–15275, Nov. 2023.
[78] D. Sarkar, A. Narang, and S. Rai, "Fed-focal loss for imbalanced data classification in federated learning," 2020, *arXiv:2011.06283*.
[79] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," 2019, *arXiv:1911.02855*.
[80] M. Beaussart, F. Grimberg, M.-A. Hartley, and M. Jaggi, "WAFFLE: Weighted averaging for personalized federated learning," 2021, *arXiv:2110.06978*.
[81] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, May 2016.
[82] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Sci. Data*, vol. 5, no. 1, pp. 1–13, Sep. 2018.
[83] Y. Shif, P. Doshi, and K. F. Almoosa, "What CPR means to surrogate decision makers of ICU patients," *Resuscitation*, vol. 90, pp. 73–78, May 2015.
[84] A. Higgs, B. A. McGrath, C. Goddard, J. Rangasami, G. Suntharalingam, R. Gale, and T. M. Cook, "Guidelines for the management of tracheal intubation in critically ill adults," *Brit. J. Anaesthesia*, vol. 120, no. 2, pp. 323–352, Feb. 2018.
[85] A. E. W. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard, "The MIMIC code repository: Enabling reproducibility in critical care research," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 1, pp. 32–39, Jan. 2018.
[86] MIT Lab. Comput. Physiol. (2016). *EICU Collaborative Research Database Code*. [Online]. Available: https://github.com/MIT-LCP/eicu-code
[87] D. Kozen and M. Timme, "Indefinite summation and the Kronecker delta," Cornell Univ., Ithaca, NY, USA, Comput. Inf. Sci. Tech. Rep., 2007. [Online]. Available: https://ecommons.cornell.edu/browse/title?scope=98835b1a-305f-462e-9406-d7141c714266&bbm.page=1&startsWith=Indefinite%20summation%20and%20the%20Kronecker%20delta
[88] O. Kramer and O. Kramer, "Scikit-learn," in *Proc. Mach. Learn. Evol. Strategies*, 2016, pp. 45–53.
[89] P. Esling and C. Agon, "Time-series data mining," *ACM Comput. Surv.*, vol. 45, no. 1, pp. 1–34, 2012.
[90] S. Ratnavale, C. Hepp, E. Doerry, and J. R. Mihaljevic, "A sliding window approach to optimize the time-varying parameters of a spatially-explicit and stochastic model of COVID-19," *PLOS Global Public Health*, vol. 2, no. 9, Sep. 2022, Art. no. e0001058.
[91] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced big data," *J. Big Data*, vol. 10, no. 1, p. 42, Apr. 2023.
[92] A. Singh, P. Vepakomma, O. Gupta, and R. Raskar, "Detailed comparison of communication efficiency of split learning and federated learning," 2019, *arXiv:1909.09145*.
[93] J. Mills, J. Hu, and G. Min, "Communication-efficient federated learning for wireless edge intelligence in IoT," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 5986–5994, Jul. 2020.
[94] J. Bopaiah and R. Kavuluru, "Precision/recall trade-off analysis in abnormal/normal heart sound classification," in *Proc. Int. Conf. Big Data Anal.* Cham, Switzerland: Springer, 2017, pp. 179–194.
[95] H. F. Nadrous, B. Afessa, E. A. Pfeifer, and S. G. Peters, "The role of autopsy in the intensive care unit," *Mayo Clinic Proc.*, vol. 78, no. 8, pp. 947–950, Aug. 2003.
[96] R. V. Craiu and L. Sun, "Choosing the lesser evil: Trade-off between false discovery rate and non-discovery rate," *Statistica Sinica*, vol. 18, no. 3, pp. 861–879, 2008.
[97] R. J. Walley and A. P. Grieve, "Optimising the trade-off between type I and II error rates in the Bayesian context," *Pharmaceutical Statist.*, vol. 20, no. 4, pp. 710–720, Jul. 2021.

TRONG-NGHIA NGUYEN received the M.S. degree from the Department of Computer Vision, Hanoi University of Science and Technology, Vietnam, in 2020. He is currently pursuing the Ph.D. degree with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include medical image processing and time-series analysis rapid response systems.

HYUNG-JEONG YANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Chonbuk National University, Gwangju, South Korea. She is currently working as a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University. Her research interests include multimedia data mining, medical data analysis, social network service data mining, and video data understanding.

BO-GUN KHO received the master's degree in internal medicine from Chonnam National University. He is a currently working as a Doctor with Chonnam National University Hospital, South Korea, where he is also a Clinical Assistant Professor. His current research interests include critical care and rapid response systems.



SAE-RYUNG KANG (Member, IEEE) received the M.S., Ph.D., and M.D. degrees from Chonnam National University Medical School, South Korea. She is currently working as a Clinical Assistant Professor with the Department of Nuclear Medicine, Chonnam National University Hwasun Hospital. Her research interests include molecular imaging, radiomics, and artificial intelligence.

SOO-HYUNG KIM (Member, IEEE) received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology, in 1988 and 1993, respectively. He is currently working as a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include pattern recognition, document image processing, medical image processing, and deep learning.