

Multimodal Representation Learning Based on Personalized Graph-Based Fusion for Mortality Prediction Using Electronic Medical Records

Abdulrahman Al-Dailami, Hulin Kuang*, and Jianxin Wang

Abstract: Predicting mortality risk in the Intensive Care Unit (ICU) using Electronic Medical Records (EMR) is crucial for identifying patients in need of immediate attention. However, the incompleteness and the variability of EMR features for each patient make mortality prediction challenging. This study proposes a multimodal representation learning framework based on a novel personalized graph-based fusion approach to address these challenges. The proposed approach involves constructing patient-specific modality aggregation graphs to provide information about the features associated with each patient from incomplete multimodal data, enabling the effective and explainable fusion of the incomplete features. Modality-specific encoders are employed to encode each modality feature separately. To tackle the variability and incompleteness of input features among patients, a novel personalized graph-based fusion method is proposed to fuse patient-specific multimodal feature representations based on the constructed modality aggregation graphs. Furthermore, a MultiModal Gated Contrastive Representation Learning (MMGCRL) method is proposed to facilitate capturing adequate complementary information from multimodal representations and improve model performance. We evaluate the proposed framework using the large-scale ICU dataset, MIMIC-III. Experimental results demonstrate its effectiveness in mortality prediction, outperforming several state-of-the-art methods.

Key words: multimodal representation learning; Graph Neural Network (GNN); mortality prediction; Electronic Medical Records (EMR); Intensive Care Unit (ICU)

1 Introduction

In the Intensive Care Unit (ICU), mortality prediction

- Abdulrahman Al-Dailami is with Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China, and also with Faculty of Computer and Information Technology, Sana'a University, Sana'a 999101, Yemen. E-mail: al-dailami@csu.edu.cn.
- Hulin Kuang and Jianxin Wang are with Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China. E-mail: hulinkuang@csu.edu.cn; jxwang@mail.csu.edu.cn.

* To whom correspondence should be addressed.

Manuscript received: 2024-09-11; revised: 2024-12-12;
accepted: 2024-12-15

models are utilized to stratify patients into different risk groups and facilitate the assessment of their clinical status. With the advancement of medical technology, ICU patients are monitored by several vital sign measurements from bedside sensors, lab test reports, and clinical notes, which in turn leads to the need to process a large amount of heterogeneous data. Electronic Medical Records (EMRs) are multi-attribute and multimodal data that record patient clinical status from different sources. EMR data contain (1) structured data, including static demographic information (age, gender, etc.) and temporal data (lab tests, blood pressure, heart rate, etc.); and (2) unstructured data, such as free-text clinical notes and reports (e.g., Electrocardiogram (ECG),

Echocardiogram (Echo), Radiology, etc.), as shown in Fig. 1.

In medical informatics, predicting ICU patient mortality risk has evolved. Researchers have proposed several statistical scoring systems to evaluate the patient's health status (the disease severity and prognosis) and the efficacy of clinical treatments. These systems are usually used to evaluate patients' vital signs and symptoms. Knaus et al.^[1] proposed the first scoring system called Acute Physiology And Chronic Health Evaluation (APACHE). The APACHE-II^[2] is a widely used in severity scoring system that predicts ICU patient mortality risk and detects the abnormality in the acute physiology of ICU patients. However, such statistical scoring systems suffer from overestimating mortality risk because they estimate the risk based on predefined thresholds^[3].

In contrast, deep learning approaches show state-of-the-art performance in patient outcome prediction tasks, such as mortality prediction^[4], length of stay prediction^[5], and disease prognosis prediction^[6]. EMR data are multimodal and heterogeneous, and come from different sources in different formats, including structured static and temporal features and unstructured clinical notes. Structured EMR data, such as vital sign observations of ICU patients and clinical notes, are usually used as features to train models. Although these

methods demonstrate state-of-the-art performance, there are still several challenges and issues.

(1) How to effectively encode the different clinical notes?

Clinical notes offer valuable insights into a patient's medical history, symptoms, and treatment progress, ultimately aiding doctors in making informed decisions about their care^[7]. These clinical notes are written by various specialized physicians describing the patient's clinical status from different clinical perspectives (see Fig. 1). To effectively encode different types of clinical notes (such as ECG, Echo, radiology, and nursing notes), it is crucial to recognize that these notes describe the patient's clinical status from different perspectives and sources. In previous works, all notes were either merged in the input stage^[8] or treated as a single feature type by the model^[4, 9, 10]. However, given the distinct roles that each type of clinical note may play in mortality prediction, they should be encoded separately. This separation allows the model to capture the unique information and nuances provided by each type of clinical note, improving the overall prediction accuracy and performance.

(2) How to tackle variability and incompleteness of multimodal EMR data?

In healthcare, patients' varying clinical events and personalized treatment plans often result in diverse and

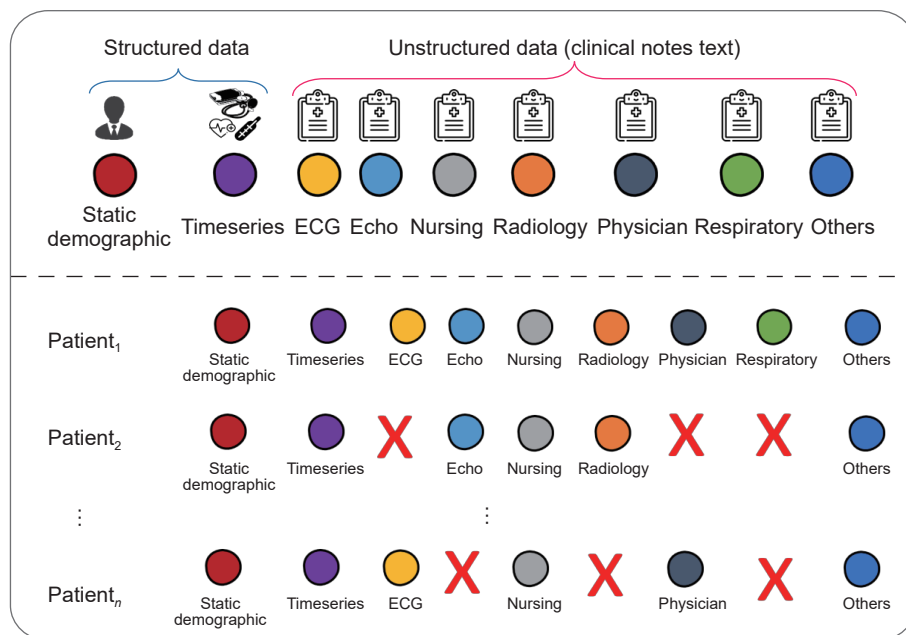


Fig. 1 Multimodal EMR data, comprising both structured data and unstructured clinical notes. Notably, many patients' records often lack certain types of clinical notes.

incomplete input features. Multimodal representation learning approaches typically assume uniformity of input features among patients^[4, 9, 10]. However, patients present with varying clinical events and properties, resulting in incomplete modalities and features, as shown in Fig. 1. For example, one patient might have vital sign time-series, medical imaging, and clinical notes, while another might lack certain modalities or have different types of notes. It is crucial for models to adapt to this variability and effectively fuse patient-specific multimodal features. This ensures the model reflects the real-world diversity of patient data and medical practice. Accommodating such variability enables the model to learn from the unique combination of modalities available for each patient, leading to more accurate and personalized predictions.

(3) How to capture effective complementary multimodal representation?

Utilizing multimodal features can enhance the accuracy of predicting patient mortality^[4, 9, 10]. However, in EMR data, repetitive patient information may originate from different data sources. For example, laboratory test results can be documented in clinical notes and presented as numerical values in time series variables. It is essential to guide the model in extracting distinct and complementary information from each data modality.

To address the abovementioned challenges, we propose a multimodal deep learning framework based on gated contrastive representation learning with a novel personalized graph-based fusion approach for predicting ICU patient mortality risk using multimodal EMR data. Our framework is designed to handle the inherent variability and incompleteness in ICU data by dynamically tailoring the model to each patient's specific context. Unlike conventional methods that treat all patients uniformly, our approach accounts for the unique clinical events and treatment plans associated with each patient.

Specifically, We partition clinical notes into seven distinct categories and encode each category individually to capture the rich information contained within these notes. This enables our model to better understand the diversity of clinical narratives and their relevance to the patient's current status. Building upon recent advancements in Graph Neural Networks (GNNs), which have proven effective in integrating multimodal heterogeneous data into graph

structures^[11, 12], we construct patient-specific modality aggregation graphs. These graphs represent the various modalities of patient data, and our proposed personalized graph-based fusion approach combines these multimodal representations. This fusion enables the generation of personalized clinical status representations that reflect the individual's unique data profile, thereby enhancing the predictive performance of our model.

To further improve the quality of the patient representations, we introduce a MultiModal Gated Contrastive Representation Learning (MMGCRL) module unlike conventional contrastive learning, which contrasts positive and negative sample pairs to learn feature representations, MMGCRL leverages contrastive loss to guide the model in extracting diverse and complementary information from multiple modalities. The gating mechanism in MMGCRL selectively emphasizes relevant features from each modality, ensuring that the model learns robust and informative patient representations that are crucial for accurate mortality risk prediction. The main contributions of this research are as follows:

- We propose a personalized graph-based deep learning framework for ICU patient mortality prediction, focusing on patient-specific multimodal representation learning from incomplete multimodal EMR data to reflect individualized clinical contexts and treatment plans.
- We introduce a multi-type clinical note encoding method for the diverse clinical notes to enable effective clinical note representation extraction.
- We propose a novel personalized graph-based fusion method based on patient-specific modality aggregation graphs to integrate the patient-specific representations from the different incomplete multimodal features and generate an effective clinical status representation of ICU patients.
- We develop a multimodal gated contrastive representation learning method to enhance multimodal clinical status representation learning.

2 Related Work

2.1 Methods based on structured EMR data

EMR data contains structured temporal features that record the patient's vital signs and clinical events during the ICU stay. Most existing research focuses on modeling structured temporal data and learning useful

representations for patient outcome prediction tasks, such as mortality prediction.

Harutyunyan et al.^[13] provided a preprocessing pipeline and benchmarking for the MIMIC-III dataset. They selected a set of clinical features listed in an acute physiology score table based on the APACHE II scoring system and used a Long Short-Term Memory (LSTM) encoder to extract useful representations for predicting patient outcomes, such as mortality, length of stay, decompensation prediction, and phenotyping. Several machine learning models have been proposed for mortality prediction, including Convolutional Neural Networks (CNN)^[5, 14], Recurrent Neural Networks (RNN)^[13, 15], stacked autoencoders^[16], and attention-based networks^[17–19].

2.2 Methods based on the unstructured EMR data

Clinical notes are mostly free-text notes written by clinicians describing several clinical events and observations, constituting 80% of EMR data^[20]. This crucial data are underutilized resources within healthcare informatics^[21]. Recently, there has been considerable interest among researchers in clinical notes^[22, 23]. Gehrmann et al.^[24] designed a CNN to extract useful phrases for phenotyping classification. Darabi et al.^[25] utilized a Transformer-based skip-gram model to extract the code representation and adopted a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to extract the medical notes representation. Wu et al.^[23] proposed an end-to-end Joint Attention Network (JAN) utilizing document-based and label-based attention for automatic International Classification of Diseases (ICD) coding. They utilized document-based and label-based attention to handle dense and sparse data in long-tailed label distribution. Wu et al.^[22] recently introduced FineEHR, a system that employs two representation learning techniques: metric learning and fine-tuning. This system refines clinical note embeddings while capitalizing on the intrinsic correlations among various health statuses and note categories. These studies show that using clinical notes can provide an adequate representation of the patient's prognosis. However, they did not consider other EMR data modalities, such as time-series and static demographic information.

2.3 Methods based on multimodal data

It has been demonstrated that multimodal learning is useful for EMR data mining^[26]. Khadanga et al.^[8]

incorporated clinical note embedding with clinical time series features to improve the model's performance. Similarly, Hashir and Sawhney^[27] combined clinical note embedding with structured time-series features to learn multimodal representation. They encoded each note of the first 48 hours of the patient's stay in the ICU individually. Then, they used a Gated Recurrent Unit (GRU) temporal encoder to aggregate the clinical notes' representations. Yang and Wu^[28] used a fusion method based on attention gating mechanism to fuse the time-series and clinical note representations. Deznabi et al.^[4] combined time-series features with clinical notes to predict mortality risk. They fine-tuned the Bio+Clinical BERT model^[29] to encode the clinical notes individually and used LSTM to encode time-series features. Lyu et al.^[9] introduced a multimodal transformer architecture for joint representation learning of clinical notes and time-series data to predict mortality risk. Zhang et al.^[30] adopted a discretized multi-time attention (namely mTAND)^[31] to model irregularities in multimodal EMR data for medical predictions. Huo et al.^[32] proposed a multimodal interactive fusion model for medication recommendation.

Some studies have proposed approaches to address the challenge of missing features in multimodal EMR data. Zhang et al.^[33] proposed an end-to-end model to compensate for missing patient information by leveraging patient similarity learned through a task-guided modality-adaptive similarity metric. Similarly, Liu et al.^[10] introduced a Transformer-based fusion model with modality-specific tokens to generate the missing values in each modality. However, these methods assume the uniformity of input features across patients and focus on filling the missing values. In contrast, our study takes a different approach by acknowledging the reality of the variability of multimodal features across patients and the variability of multi-type clinical notes in EMR data for effective representation extraction. Rather than assuming the existence of all input data modalities and imputing the missing ones, it is essential to focus solely on utilizing existing features instead of only imputing missing values that do not necessarily exist due to variations in patient treatment plans and clinical events, thereby reflecting the genuine clinical scenario observed in EMR data. Additionally, instead of considering clinical notes as one type, it is important to recognize multi-

type clinical notes for effective representation extraction.

3 Method

3.1 Input data notations and problem formulation

In this study, we utilize the multimodal EMR data, specifically the static demographics, the multivariate time-series observations, and the unstructured clinical notes. Formally, for each patient, we denote the input features as $P = \{S_p, X_p, N_p^c, A_p, y_p\}$, where p denotes the patient, and

$S_p \in \mathbf{R}^{D \times 1}$ denotes the static demographic features, where D is the number of static demographic features that include age, gender, weight, height, and others.

$X_p \in \mathbf{R}^{T \times F}$ denotes the multivariate time-series observations, where F is the number of time-series variables, encompassing clinical features like laboratory test measurements (e.g., serum creatinine, white cell count, etc.), and T signifies the prediction window size, set at 24-hour or 48-hour.

$N_p^c \in \mathbf{R}^C$ denotes the free-text clinical notes, where C is the number of clinical note categories.

$A_p \in \mathbf{R}^{M \times M}$ is the adjacency matrix of the modality aggregation graph. Here, M signifies the number of input data modalities to the model, including static demographics, time-series data, ECG notes, echo notes, nursing notes, physician notes, radiologist notes, respiratory notes, and other notes.

y_p represents the patient outcome label, specifically the patient's mortality (alive or died).

Given multimodal input features S_p , X_p , A_p , and N_p^c derived from the initial 24 or 48 hours of the patient's stay in the ICU, we aim to predict the patient's mortality risk.

3.2 Overview

This study proposes a multimodal representation learning approach based on a novel personalized graph-based fusion approach for mortality risk prediction. Figure 2 shows our proposed model architecture.

To learn clinical status representation from multimodal EMR features, we employ modality-specific encoders (Fig. 2a). Specifically, we use a FlatEncoder (Fig. 2b) to encode static demographic features, a GRUEncoder to encode time-series features, and a NoteEncoder (Fig. 2c) to encode clinical note texts.

The hidden representations produced by these

modality-specific encoders are subsequently fused via our personalized graph-based fusion method by passing these representations with the associated modality-aggregation graphs through a GNN encoder, resulting in a robust and interpretable representation of incomplete multimodal features (Fig. 2d).

Next, we process the obtained multimodal representation and the hidden representations of static demographic and time-series features using the MMGCRL module to produce the final personalized clinical status representation (Fig. 2e).

Finally, a Fully Connected (FC) layer with a Sigmoid activation function generates predictions from the final personalized clinical status representation. The technical details of our proposed method are elaborated upon below.

3.3 Modality-specific encoders

We employ distinct encoders for each input data modality to extract representations. The modality-specific encoders are outlined as follows.

3.3.1 Static demographic data encoder

Static demographic features typically consist of categorical and numerical data that do not change over time. Like previous works^[14, 17, 18], we employ an FC layer followed by a Rectified Linear Unit (ReLU) activation function, which is sufficient for encoding such features. This configuration enables effective learning of complex non-linear relationships between input and output variables. Formally, we encode static demographic features as follows:

$$SD = \text{ReLU}(\text{FC}(S)) \quad (1)$$

where S denotes the static demographic features, $\text{ReLU}(\cdot)$ is the rectified linear unit activation function defined as $\text{ReLU}(x) = \max(0, x)$, $SD \in \mathbf{R}^a$, and a denotes the dimension of the hidden representation of static demographic features.

3.3.2 Clinical time-series data encoder

We employ a bidirectional RNN based encoder to capture the temporal dependencies between time-series features, which resamples a chain of RNN cell modules to model sequential features efficiently^[34]. Our method uses the GRU variant^[35], BiGRU, due to its better performance and simple structure. We encode the time-series features as follows:

$$h = \text{BiGRU}(X, h_0) \quad (2)$$

where $\text{BiGRU}(\cdot)$ denotes the bidirectional GRU

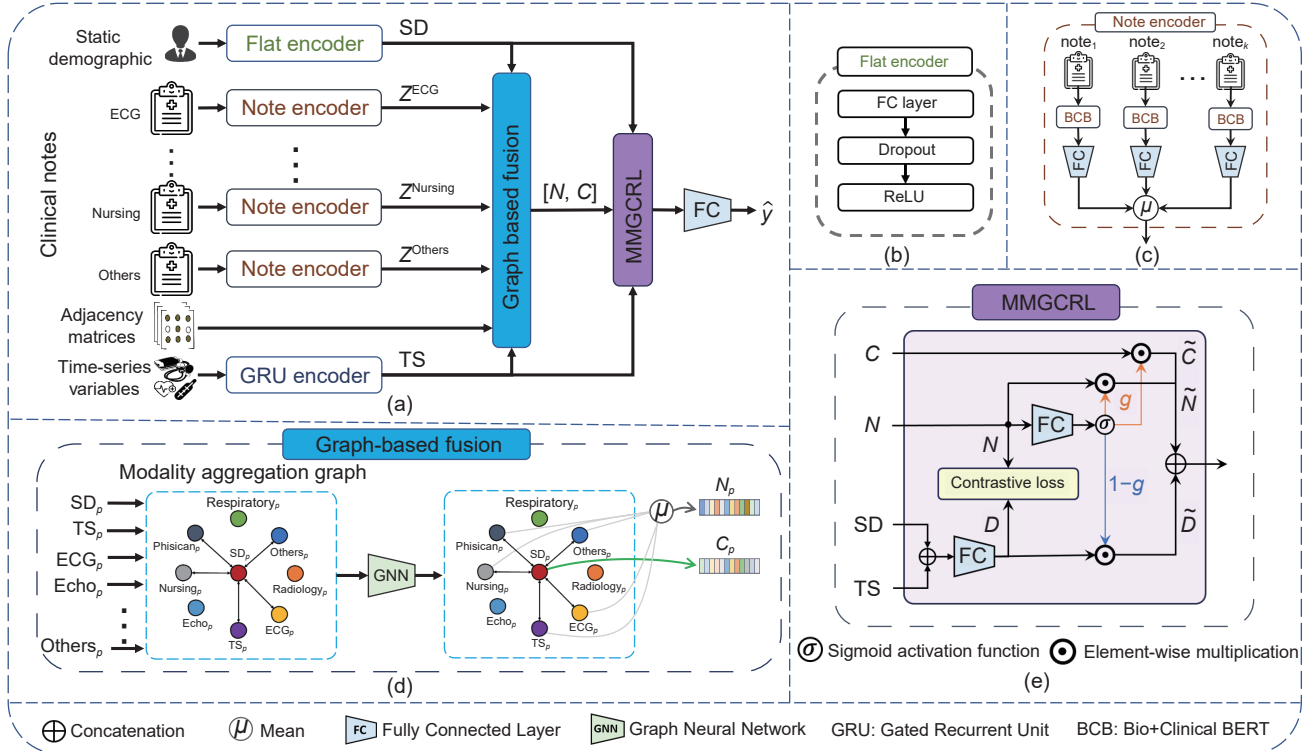


Fig. 2 Architecture of the proposed method. (a) Each input data modality is first encoded using a modality-specific encoder to obtain the modality-specific representation; (b) static features are encoded using flat encoder, (c) clinical notes are encoded using note encoder, and time-series features are encoded using GRU encoder; (d) our proposed graph based fusion module then processes these representations, which employs modality aggregation graphs to aggregate patient-specific multimodal representations, allowing for personalized multimodal representation fusion; and (e) the output of the graph-based fusion module, the static demographic, and the time-series features are processed by the MMGCRL module to obtain the final enhanced multimodal representation. Finally, the final representation is used to generate prediction using an FC classification layer.

encoder, X is the multivariate time-series input features, h_0 is the initial hidden state, and h is the resulting bidirectional hidden states ($h_1, h_2, \dots, h_T$).

The final time-series feature representation TS is formed by taking the last hidden state from the output hidden states h ,

$$TS = h_T \quad (3)$$

where $TS \in \mathbf{R}^b$, and b denotes the dimension of the output representation.

3.3.3 Clinical note preprocessing and encoding

In this study, we enhance patient mortality risk prediction by extracting multi-type clinical notes from the NOTEEVENT table of the MIMIC-III dataset. We extract seven note types per patient: ECG, Echo, Nursing, Physician, Radiology, Respiratory, and Others. The clinical notes text preprocessing follows Ref. [8].

Modeling clinical notes requires capturing intricate word dependencies^[36]. BERT-based architectures excel in this task, outperforming traditional Natural

Language Processing (NLP) models^[29, 36, 37]. Bio+Clinical BERT has demonstrated superior performance in a wide range of clinical NLP tasks, such as medical relation extraction and clinical concept extraction, making it well-suited for extracting meaningful representations from clinical notes^[29]. Besides, it is pre-trained on the same dataset i.e., MIMIC-III, that we use in this study. This domain-specific pre-training allows Bio+Clinical BERT to better capture the nuances of medical terminology and context, making it highly suitable for processing clinical text data in our prediction task. Therefore, we chose Bio+Clinical BERT^[29] as an encoder for clinical notes.

Our approach involves two phases: first, fine-tuning Bio+Clinical BERT for mortality prediction, then integrating it into our framework to encode each clinical note individually. Formally, the clinical notes are encoded as follows:

$$Z_c = \frac{1}{K} \sum_1^K \text{FC}_c(\text{BCB}(\text{nt}_c^k)) \quad (4)$$

where $\text{FC}_c(\cdot)$ denotes the fully connected layer for the note type c , $\text{BCB}(\cdot)$ denotes the fine-tuned Bio+Clinical BERT encoder, which is shared among all note types, $c \in \{\text{ECG, Echo, Nursing, Physician, Radiology, Respiratory, Others}\}$ denotes the clinical note category, K denotes the number of clinical notes within each type, nt_c^k denotes the k -th clinical note of category c , and $Z_c \in \mathbf{R}^b$ denotes the output notes representation.

3.4 Personalized graph-based multimodal feature fusion

Given the diversity in disease types and patient treatment plans, varying sets of input features are created for different patient groups, as shown in Fig. 1. To accommodate this variability, we propose a novel personalized graph-based fusion method that fuses patient-specific representations from multimodal EMR features, including multi-type clinical notes, static demographics, and time-series features, which can vary from one patient to another. We construct an individualized modality aggregation graph for each patient. These graphs provide the model with supplementary information regarding the actual multimodal inputs associated with each patient. They guide the model in aggregating the existing patient-specific multimodal features. Additionally, they provide insights into each patient's overall treatment plan. Then, we utilize a GNN encoder to fuse the multimodal representations, as shown in Fig. 2d. This approach enables the model to focus on patient-specific relevant features for effective multimodal patient-specific clinical status representation learning.

3.4.1 Modality aggregation graph construction

The modality aggregation graph aims to help the model learn from patient-specific features. Static demographic data, like age and gender, offer key insights into patients' basic condition and prognosis^[38]. Centering our framework around these features ensures personalized characteristics are considered in the modeling. Figure 3 illustrates the graph structure used for our method. Each node in the graph, except for the central node representing the patient's static demographic information, corresponds to a type of clinical note or other data modality feature, such as time-series features (which can be extended to

accommodate other data modalities, such as X-ray images). The graph is represented using an adjacency matrix $A \in \mathbf{R}^{M \times M}$. The values in the adjacency matrix are either 0 or 1. If the data modality feature exists, the value is 1, indicating a link to the central node. Otherwise, the value is 0, indicating no link to the central node. This process can be formally represented as follows:

$$A_{0,m} = \begin{cases} 1, & \text{if modality } m \text{ exists;} \\ 0, & \text{Otherwise} \end{cases} \quad (5)$$

where $A \in \mathbf{R}^{M \times M}$ is the adjacency matrix, and $m \in \{\text{Time-series, ECG, Echo, Nursing, Physician, Radiology, Respiratory, Others}\}$.

The construction process of the modality aggregation graph is outlined in Algorithm 1. Using these graphs, the model is guided to integrate the actual patient-specific multimodal features for each patient from the set of incomplete multimodal input features.

3.4.2 GNN encoder

Firstly, we utilize an FC layer to project the static demographic representation to align with the other representations' dimensions as follows:

$$\hat{S}\mathbf{D} = \text{FC}(\mathbf{S}\mathbf{D}) \quad (6)$$

where $\hat{S}\mathbf{D} \in \mathbf{R}^b$ is the projected hidden static demographic representation.

Then, because the adjacency matrix does not explicitly capture relationships among leaf nodes (other features), our graph-based fusion module utilizes two GNN layers. In the first layer, each node interacts with

Algorithm 1 Modality aggregation graph construction

function ModalityAggGraph (MultimodalData)

Input: Patient multimodal data

Output: Patient-specific modality aggregation graph

```

1: graph  $\leftarrow$  new ModalityAggGraph ( ); // Create a new graph object
2: central_node  $\leftarrow$  new Node (PatientDemographics); // Create a new node object for static demographic data
3: graph.add_node (central_node); // Adding the node to the graph
4: for each modality in patient_data do
5:   if patient_has_data (modality) then
6:     modality_node  $\leftarrow$  new Node (modality); // Create a new node object for the current data modality
7:     graph.add_edge (modality_node, central_node); // Connect the current node to the central node
8:   end if
9: end for
10: return graph

```

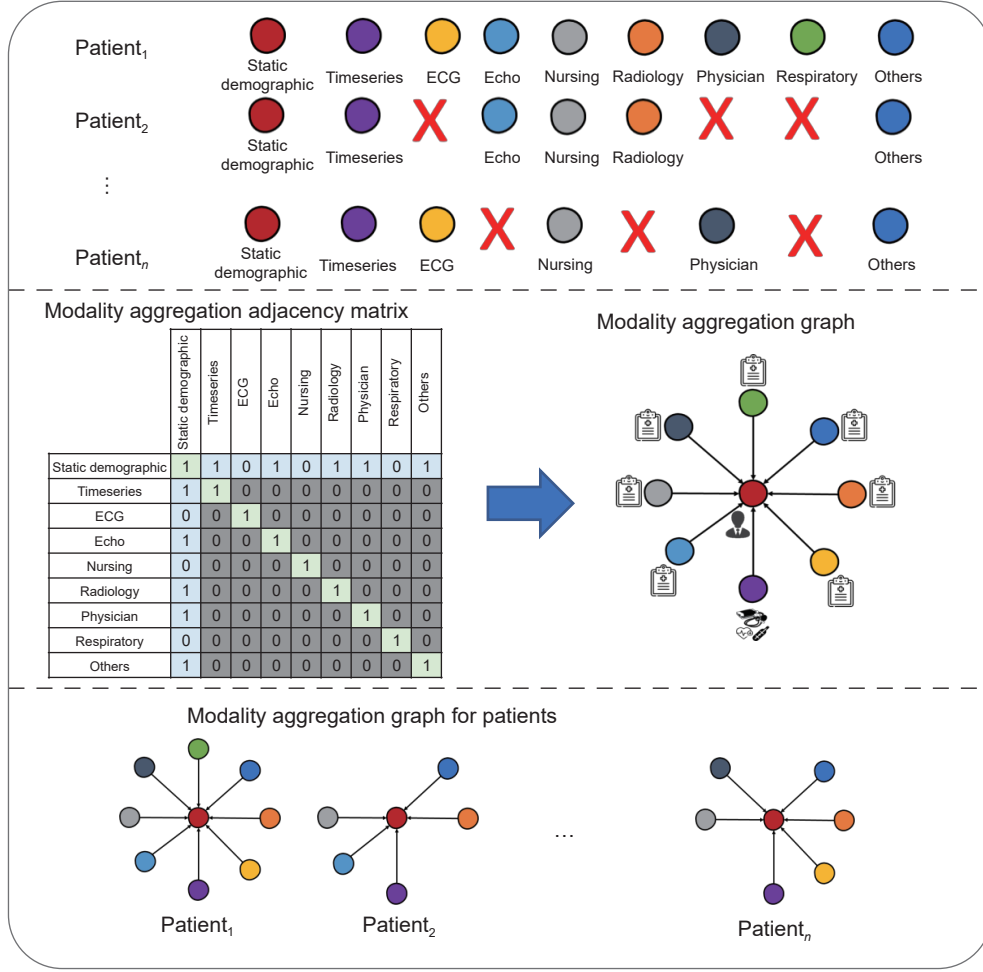


Fig. 3 Modality aggregation graph constructed based on the input data modalities for each patient. The central node represents the patient's demographic features, while the leaf nodes correspond to the available multimodal features associated with the patient. The edges of the graph are defined by an adjacency matrix, where links between the central node and leaf nodes indicate the presence of corresponding data modalities.

the directly connected node, which, in our case, is the central node, making the central node capture information from all connected leaf nodes. In the second GNN layer, the central node interacts with all connected leaf nodes, which broadcasts the captured interaction information from the previous GNN layer to all connected nodes, resulting in an indirect interaction among all nodes via the central node. This method is used to combine the multimodal features and obtain effective representations based on the modality aggregation graph for each patient,

$$\begin{aligned}
 G_p &= \text{GNN}_1(A_p, E_p), \\
 G_p &= \text{GNN}_2(A_p, G_p), \\
 C_p &= G_p[0], \\
 N_p &= \mu(E_p, A_{\text{leaf}, p})
 \end{aligned} \quad (7)$$

where $E_p = [\text{TS}_p, \hat{\text{SD}}_p, Z_p^{\text{ECG}}, Z_p^{\text{Echo}}, \dots, Z_p^{\text{Others}}]$ denotes

the hidden representations of the encoded multimodal input features for patient p , including the projected hidden representation of static demographic $\hat{\text{SD}}_p \in \mathbf{R}^b$, time-series $\text{TS}_p \in \mathbf{R}^b$, and the different types of clinical notes $[Z_p^{\text{ECG}}, Z_p^{\text{Echo}}, \dots, Z_p^{\text{Others}}]$. $N_p \in \mathbf{R}^b$ is the mean of the hidden representation of the leaf graph nodes for the patient p , $C_p \in \mathbf{R}^b$ is the representation of the center node for the patient p , μ denotes the mean function, $A_p \in \mathbf{R}^{M \times M}$ denotes the adjacency matrix for the patient p , $A_{\text{leaf}, p}$ is the index of leaf nodes of the patient p graph, and $G_p \in \mathbf{R}^{M \times b}$ is the hidden representation of the GNN-encoded multimodal input features for the patient p .

We employ the Graph Attention Network (GAT)^[39] as a GNN encoder to derive efficient multimodal representations from the incomplete multimodal input features. GAT enables nodes to focus on their

neighbors with varying weights, allowing them to gather diverse information from neighbors based on their significance.

3.5 Multimodal gated contrastive representation learning

We introduce a multimodal gated contrastive representation learning approach, the MMGCRL module, to improve clinical status representation learning from multimodal data, including static demographics, time-series, and clinical notes. This method aims to extract nuanced insights from each data modality, facilitating diverse information learning from structured and unstructured data. The MMGCRL uses contrastive loss optimization with a gated fusion mechanism to extract complementary and distinctive features from the multimodal structured and unstructured data representations. Unlike traditional contrastive learning, which focuses solely on distinguishing positive and negative pairs, MMGCRL is specifically designed to align and integrate multimodal representations effectively, as illustrated in Fig. 2e. Formally, MMGCRL is expressed as follows.

Gated fusion: First, the static demographic and time-series representations are concatenated and then processed by an FC layer to obtain the structured features representation as follows:

$$D = \text{FC}([SD, TS]) \quad (8)$$

where $D \in \mathbf{R}^b$ is the projected time-series and static demographic representations.

Then, the representation obtained from the personalized graph-based fusion module is processed by an FC layer followed by a Sigmoid activation function to obtain the gate vector g as follows:

$$g = \sigma(\text{FC}(N)) \quad (9)$$

where σ denotes the Sigmoid activation function.

Then, the personalized graph-based fusion module representations N and C are multiplied by the gate vector g to obtain the gated representations as follows:

$$\begin{aligned} \tilde{N} &= N \times g, \\ \tilde{C} &= C \times g \end{aligned} \quad (10)$$

Then, the structured features representation is multiplied by the complement of the g vector to obtain the gated structured features representation as follows:

$$\tilde{D} = D \times (1 - g) \quad (11)$$

Finally, the gated representations are concatenated to

form the enhanced multimodal representation R^{final} , which is then used in the following prediction layer to predict mortality risk,

$$R^{\text{final}} = [\tilde{D}, \tilde{N}, \tilde{C}] \quad (12)$$

Contrastive loss: We employ a contrastive loss approach by calculating the negative pairwise distance between the structured and unstructured data representations. Our goal is to maximize the distance similarity between the structured data representation and the unstructured data representation obtained from the personalized graph-based fusion module as follows:

$$\text{CLoss} = \mu(\text{dist}(D, N)) \quad (13)$$

where $\text{dist}(\cdot)$ denotes the negative pairwise distance calculation function.

3.6 Prediction layer

Finally, to implement the mortality risk prediction, the final multimodal representation R^{final} is processed by an FC layer followed by a Sigmoid activation function $\sigma(\cdot)$ to generate predictions $\hat{y}$ as follows:

$$\hat{y} = \sigma(\text{FC}(R^{\text{final}})) \quad (14)$$

3.7 Loss function

The mortality prediction task aims to predict whether patients die in the ICU using data from the first 24 or 48 hours after admission^[13, 15]. Mortality labels represent their status at the end of their ICU stay. Patients whose ICU stay lasts less than 24 or 48 hours depending on the prediction window size, are excluded from the analysis, which is the same as Refs. [13, 15]. To train the model, we use the binary cross-entropy loss function.

The total objective of our model training is the weighted sum of the binary cross entropy loss ℓ_{bce} and the contrastive loss (CLoss) obtained from Eq. (13),

$$\text{loss} = \ell_{\text{bce}} \times (1 - \beta) + \text{CLoss} \times \beta \quad (15)$$

where β is a regularization variable.

4 Data and Experiment

4.1 Data description

Our experiments evaluate the proposed approach using the MIMIC-III dataset^[40], a publicly available dataset containing EMR data for nearly 60 000 ICU patient admissions to six different ICUs at the Boston Hospital between 2001 and 2015. The dataset includes

information about patients' health conditions during their ICU stay, such as medications, vital signs, lab test results, and clinical notes.

We employ a semi-automated cohort selection method similar to the one used by Refs. [5, 14, 18] to select our cohort. We extract the data from the first 24 and 48 hours of ICU stay for adult patients (those older than 18 years) who have stayed in the ICU for at least 24 and 48 hours. We only include patients with at least one clinical note and one observation for all temporal variables, resulting in 39 097 and 20 070 ICU stays for 24-hour and 48-hour prediction window cohorts, respectively, as detailed in Tables 1 and 2. We only include time-series variables that are present in at least 25% of the ICU patients. The values of these variables are resampled hourly. To account for the irregularity of the time-series observations, we use a feedforward strategy to fill in the gaps between time-series observations. This approach is more factual than interpolation, as clinicians typically only consider the most recent values.

The clinical notes are provided by various clinical experts, including doctors, nurses, and radiologists, and offer information about patients from different perspectives. It is crucial to consider all available information and not overlook any details. To ensure comprehensive analysis, we include seven clinical note types for each patient: ECG, Echo, Nursing, Physician, Radiology, Respiratory, and Others, as detailed in

Table 3. Discharge summaries are excluded as they may reveal information about the patient's outcome. Any missing notes are filled with a predefined keyword, "NO-NOTE". We follow Ref. [8] for text preprocessing of the clinical notes. The note length is standardized to 512 for Nursing and Radiology note types and 128 for other note types, with any deviations addressed through padding or truncation.

In addition, we generate an adjacency matrix for the modality aggregation graph, unique to each patient, based on their available notes, static demographics, and time-series features. The adjacency matrix depicts a graph of nine nodes, with seven nodes corresponding to the note types, one representing the patient's static demographics, and the final denoting the time-series features. Details regarding the construction method for the graph are elucidated in the preceding sections.

Like Ref. [13], we allocate 70% for training, 15% for validation, and 15% for testing. Tables 1–3 provide details of the resulting cohorts and feature statistics.

4.2 Implementation details

Our model is developed using Pytorch and Python 3.8.0 and trained on a machine with an NVIDIA RTX 2080Ti GPU. We optimize the model's parameters using the Adam optimizer with a learning rate of 0.001 for the Flat and the BiGRU encoders, 0.0006 for the Bio+Clinical BERT encoder, and 0.0007 for other model layers. The batch size is 128. Dropout is

Table 1 Statistics of selected cohorts using 24-hour and 48-hour prediction window sizes.

Prediction window	Category	Total ICU stays	Train set	Validation set	Test set
24 hours	All	39 097	27 413	5829	5855
	Alive	34 709	24 365	5154	5190
	Died	4388	3048	675	665
48 hours	All	24 070	16 996	3546	3528
	Alive	20 713	14 638	3022	3053
	Died	3357	2358	524	475

Table 2 Statistics of selected features and the percentage of patients in each modality. The value in parentheses represents the percentage of patients with this data modality.

Prediction window	Note types	Time-series features	Static demographics features	Total features
24 hours	7	34 (100%)	12 (100%)	53
48 hours	7	42 (100%)	12 (100%)	61

Table 3 Number of notes and percentage of patients in each clinical note type. The value in parentheses represents the percentage of patients with this note type.

Prediction window	ECG	Echo	Nursing	Physician	Radiology	Respiratory	Others
24 hours	32 865 (32%)	2520 (8%)	147 798 (94%)	37 172 (71%)	95 522 (62%)	4864 (95%)	4147 (4%)
48 hours	27 389 (39%)	2611 (6%)	160 635 (76%)	37 431 (71%)	88 471 (44%)	6962 (86%)	5474 (6%)

implemented throughout the model to prevent overfitting, with a dropout rate of 0.45. The output hidden state dimensions are as follows: $a = 64$ and $b = 256$. The graph encoder is implemented using the Pytorch Geometric library. We train the model for 50 epochs, save the best model, and implement an early stopping with the patience of 5 epochs based on the model performance on the validation set to ensure convergence and avoid overfitting. Finally, we report the performance metrics of the best model on the test set. We run the experiments five times with different random seed initializations and report the metrics' mean and standard deviation (mean $\pm$ std). The source code of our proposed method is available on GitHub [©].

4.3 Baselines

To measure the effectiveness of our proposed method, we compare it to the following eight baseline methods:

(1) **MultiModal-1DCNN (MM-1DCNN)**^[8]: The model consists of a 1D CNN-based network to encode the clinical notes and an LSTM encoder to encode the time-series features. The encoded clinical notes and time-series features are concatenated and processed by a fully connected layer to obtain the final representation, which is then used to predict patient outcomes.

(2) **MMHCR-GRU**^[27]: This baseline method utilizes multimodal clinical data, including time-series and clinical notes. The notes are encoded using a CNN encoder, and the encoded clinical note representations are aggregated using a GRU encoder. The final representation is obtained by concatenating the encoded clinical notes and time-series representations, which are then used to predict the mortality risk.

(3) **MultiModal-BCB-LSTM (MM-BCB-LSTM)**^[4]: This baseline model uses a fine-tuned BioBERT model to generate embeddings for each patient's clinical notes. These embeddings are then aggregated using the mean function to obtain an overall embedding for text data. The text embeddings are concatenated with the output of an LSTM encoder that encodes the patients' temporal features to produce the final representation, which is used for mortality prediction.

(4) **MultiModal Transformer (MM-Transformer)**^[9]: In this baseline, a pre-trained clinical BERT model encodes clinical notes, and a Transformer

encoder is applied to the time-series combined with the notes' representation. The learned representations are utilized to predict mortality risk.

(5) **LstmBert**^[28]: This baseline utilizes a Multimodal Adaptation Gate (MAG)^[41] to combine LSTM and ClinicalBERT representations for learning multimodal representation.

(6) **ARMOUR**^[10]: This model employs a transformer with integrated modality-specific tokens to encapsulate information from individual modalities and impute missing values; it also utilizes a cross-attention mechanism to capture interactions between different modalities.

4.4 Evaluation metrics

Since our task is binary classification and the dataset used is unbalanced, we use the Area Under the Receiver Operating Characteristic curve (AUROC), the Area Under the Precision-Recall Curve (AUPRC), the minimum value between sensitivity and positive predictive value min (Se, P+)^[42], and the F1 score as performance metrics since they are effective evaluation metrics for the classification of the imbalanced datasets capturing the trade-off between precision and recall and balancing sensitivity and precision. The reported values in our study represent the mean $\pm$ standard deviation from five independent experiment runs with different random seeds.

5 Result

5.1 Performance comparison with baselines

To analyze the performance of our model, we compare it to six baseline methods commonly used in mortality prediction. As shown in Tables 4 and 5, and visualized in Fig. 4, our model achieves an AUROC of 0.9161 and 0.8957, an AUPRC of 0.6681 and 0.6496, a min (Se, P+) of 0.6114 and 0.5928, and an F1 of 0.7827 and 0.7300 on the 24 and 48 hours cohorts, respectively, all of which are the highest among all the compared baseline methods, signifying an outperformance over baselines. Specifically, our model outperforms the best baseline method (MM-BCB-LSTM^[4]) by 0.0146, 0.0440, 0.0399, and 0.0481 in the AUROC, AUPRC, min (Se, P+), and the F1 score, respectively, using the 24-hour prediction window cohort, and by 0.0072, 0.0333, 0.0211, and 0.0197 in the AUROC, AUPRC, min (Se, P+), and the F1 score, respectively, using the 48-hour prediction window cohort, indicating an

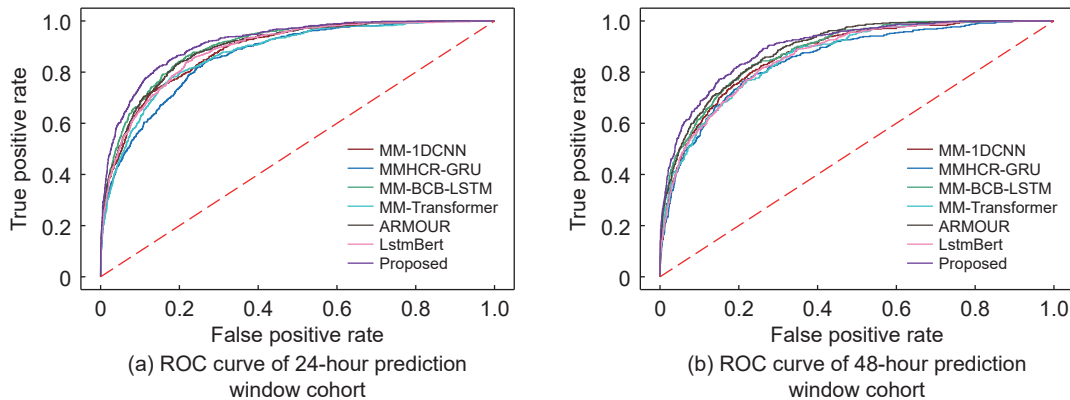
[©] https://github.com/Al-Dailami/MM_GFN

Table 4 Performance comparison with baseline models using 24-hour prediction window. The best results are highlighted in bold.

Method	AUROC	AUPRC	min (Se, P+)	F1
MM-1DCNN ^[8]	0.8878±0.0012	0.5814±0.0044	0.5435±0.0066	0.6754±0.0248
MMHCR-GRU ^[27]	0.8643±0.0015	0.5235±0.0078	0.4846±0.0082	0.6957±0.0113
MM-BCB-LSTM ^[4]	0.9015±0.0013	0.6241±0.0070	0.5715±0.0039	0.7346±0.0022
LstmBert ^[28]	0.8885±0.0012	0.5797±0.0090	0.5365±0.0101	0.7195±0.0088
MM-Transformer ^[9]	0.8734±0.0021	0.5442±0.0044	0.5081±0.0019	0.6813±0.0019
ARMOUR ^[10]	0.8934±0.0099	0.6030±0.0243	0.5482±0.0207	0.6988±0.0284
Our proposed	0.9161±0.0002	0.6681±0.0007	0.6114±0.0062	0.7827±0.0026

Table 5 Performance comparison with baseline models using 48-hour prediction window. The best results are highlighted in bold.

Method	AUROC	AUPRC	min (Se, P+)	F1
MM-1DCNN ^[8]	0.8718±0.0011	0.5740±0.0047	0.5292±0.0092	0.6750±0.0232
MMHCR-GRU ^[27]	0.8510±0.0036	0.5309±0.0064	0.5112±0.0048	0.6688±0.0149
MM-BCB-LSTM ^[4]	0.8885±0.0016	0.6163±0.0057	0.5717±0.0069	0.7103±0.0041
LstmBert ^[28]	0.8597±0.0023	0.5424±0.0044	0.5100±0.0124	0.6939±0.0098
MM-Transformer ^[9]	0.8609±0.0004	0.5474±0.0006	0.5103±0.0039	0.6644±0.0008
ARMOUR ^[10]	0.8737±0.0138	0.5909±0.0251	0.5456±0.0220	0.7222±0.0250
Our proposed	0.8957±0.0007	0.6496±0.0012	0.5928±0.0024	0.7300±0.0114

**Fig. 4** ROC curves of the proposed method and baselines.

improvement in mortality prediction accuracy and demonstrating its effectiveness in identifying patients at high mortality risk.

Among the baselines that incorporate both clinical notes and time-series features, MM-BCB-LSTM^[4] performs the best, achieving an AUROC of 0.9015, an AUPRC of 0.6241, a min (Se, P+) of 0.5715, and an F1 score of 0.7346 using the 24-hour prediction window cohort, and an AUROC of 0.8885, an AUPRC of 0.6163, a min (Se, P+) of 0.5717, and an F1 score of 0.7103 using the 48 hours prediction window cohort. This superior performance can be attributed to the model's individually encoding each clinical note using a fine-tuned BioBERT encoder and then averaging all

clinical note representations for each patient to derive the final text representation. This method effectively captures more detailed information from the clinical notes, resulting in improved mortality prediction performance.

MM-1DCNN^[8] exhibits lower performance than MM-BCB-LSTM, possibly due to its consolidation of all clinical notes into a single note, which may result in the loss of valuable information. Surprisingly, MMHCR-GRU^[27] performs worse than all baselines. MMHCR-GRU^[27] adopts a strategy of individually encoding all clinical notes for each patient, and subsequently aggregating them using a GRU encoder. This inferior performance could be attributed to the

note representation learning and aggregation method employed, indicating this method is less effective than the methods utilized in MM-1DCNN^[8] and MM-BCB-LSTM^[4]. These findings confirm that the finetuned BERT model encoder is superior for clinical notes representation, as demonstrated by the performance discrepancy among MM-BCB-LSTM^[25], MM-1DCNN^[8], and MMHCR-GRU^[27].

MM-Transformer^[9] and ARMOUR^[10] utilize Transformer encoders to learn multimodal representation but differ in their approach to processing multimodal representations. While the MM-Transformer combines structured time-series and clinical note representations before Transformer encoding, ARMOUR applies transformer encoding separately to each modality. Then, it employs cross-attention for inter-modality interaction. Results demonstrate ARMOUR's superior performance, suggesting the efficacy of individual modality processing. This superiority may stem from clinical notes' heightened sparsity compared to structured data, which can compromise joint representation learning. Therefore, jointly encoding multimodal representation using the transformer encoder yields inferior performance for the MM-Transformer model^[9] compared to the ARMOUR model^[10]. The LstmBert model encodes time-series features with LSTM and note texts with ClinicalBERT. It then employs an MAG fusion approach^[41] to fuse the multimodal representations. However, the LstmBert model performs worse than ARMOUR. This can be attributed to the fact that the ARMOUR model encodes a series of clinical notes instead of merging all notes into one note, as in LstmBert.

Overall, these results highlight the effectiveness of our proposed model, which leverages three input features: static demographics, time-series, and clinical notes. Unlike the baselines, we split the clinical notes into seven categories and encode each note type separately, capturing more fine-grained information and improving performance in mortality prediction. Additionally, our model considers the variability and incompleteness of multimodal input features using a

personalized graph-based fusion module, which effectively integrates patient-specific features from incomplete multimodal features. Furthermore, our proposed model utilizes the MMGCRL module to enhance multimodal representation learning.

5.2 Ablation study

5.2.1 Effectiveness of the personalized graph-based fusion

To assess the effectiveness of the personalized graph-based fusion module, we conduct an ablation experiment by removing this module from our model. The comparison reveals that the model without the personalized graph-based fusion (namely without graph in Table 5) exhibits lower performance, yielding an AUROC of 0.9138, an AUPRC of 0.6540, a min (Se, P+) of 0.6060, and an F1 score of 0.7548, as shown in Table 6. In contrast, our proposed model achieves an AUROC of 0.9161, an AUPRC of 0.6681, a min (Se, P+) of 0.6114, and an F1 score of 0.7827. These findings suggest that the personalized graph-based fusion module plays a crucial role in addressing the issue of incomplete multimodal input features, enabling the model to capture more latent information about each patient's clinical status through modality aggregation graphs.

5.2.2 Effectiveness of the MMGCRL module

We perform an ablation experiment to demonstrate the benefits of our MMGCRL module by removing it from our model. The comparison shows that our model without MMGCRL exhibits lower performance, yielding an AUROC of 0.9149, an AUPRC of 0.6613, and an F1 score of 0.7589, as shown in Table 6. In contrast, our proposed model achieves an AUROC of 0.9161, an AUPRC of 0.6681, a min (Se, P+) of 0.6051, and an F1 score of 0.7827. These results indicate that the MMGCRL module further enhances the model's prediction performance.

Additionally, We perform an ablation experiment to demonstrate the benefits of our MMGCRL module by comparing it with three alternative fusion methods: Concatenation, Tensor Fusion^[43], and Cross-attention Fusion, using three multimodal representations: static

Table 6 Ablation study of the proposed modules. The best results are highlighted in bold.

Method	AUROC	AUPRC	min (Se, P+)	F1
Without graph	0.9138±0.0002	0.6540±0.0012	0.6060±0.0020	0.7548±0.0045
Without MMGCRL	0.9149±0.0006	0.6613±0.0017	0.6051±0.0025	0.7589±0.0087
Our proposed	0.9161±0.0002	0.6681±0.0007	0.6114±0.0062	0.7827±0.0026

demographics, time-series, and clinical notes text.

For Concatenation, the three modalities are simply concatenated into a single feature vector. These concatenated representations are passed through the classification layer for prediction without explicit modeling of inter-modal relationships. In Tensor Fusion, each modality is treated as a separate tensor, and fusion is performed using tensor operations. In the Cross-attention Fusion method, the static demographic and time-series representations are first concatenated and then processed through a fully connected layer. After obtaining the fused representation of these modalities, we apply cross-attention between this resulting representation and the clinical text representation.

The results, presented in Table 7, show that MMGCRL outperforms all three methods across key metrics, such as AUROC, AUPRC, min (Se, P+), and F1 score. We can find that the concatenation fusion method performs better than tensor fusion and cross-attention fusion, but it fails to capture complex complementary information from the multimodal representations. Tensor fusion demonstrates lower performance across metrics due to its inability to effectively exploit complementary information among modalities. The cross-attention fusion method has the lowest results and is less consistent, as reflected in higher standard deviations, which we attribute to its dependence on modality alignment that is not explicitly optimized for diverse and heterogeneous ICU data.

In contrast, MMGCRL module consistently achieves the best performance across all metrics. Unlike the other methods, MMGCRL leverages its contrastive loss

optimization and gating mechanism to effectively extract diverse and complementary information from the multimodal representations, leading to a more robust and comprehensive feature integration.

5.2.3 Impact of clinical notes encoding and fusion method

To determine the optimal text encoding and fusion method for clinical notes, we conduct experiments using two encoders: CNN and the fine-tuned Bio+Clinical BERT, BCB for short. These experiments explore two fusion strategies for clinical notes: Early Fusion, where notes are aggregated into a single entity before encoding, and Late Fusion, where each note is encoded individually and then fused using the mean. Experimental results, shown in Table 8, demonstrate that employing Bio+Clinical BERT as a text encoder with the late fusion strategy yields the best performance, achieving an AUROC of 0.8727, an AUPRC of 0.5438, a min (Se, P+) of 0.5074, and an F1 score of 0.6880. Conversely, utilizing the early fusion strategy with BCB leads to degraded performance, with an AUROC of 0.8602, an AUPRC of 0.4947, a min (Se, P+) of 0.4799, and an F1 score of 0.6339. Meanwhile, employing the CNN encoder with the late fusion strategy outperforms the early fusion strategy, yielding an AUROC of 0.8679, an AUPRC of 0.5352, a min (Se, P+) of 0.5105, and an F1 score of 0.6869. These findings underscore the effectiveness of the late fusion strategy for encoding clinical notes. Consequently, we adopt the Bio+Clinical BERT encoder with the late fusion strategy for encoding clinical notes in our framework, in addition to segmenting the clinical notes into multiple categories.

Table 7 Comparison of MMGCRL module with other fusion methods. The best results are highlighted in bold.

Method	AUROC	AUPRC	min (Se, P+)	F1
Concatnation	0.9149±0.0006	0.6613±0.0017	0.6051±0.0025	0.7589±0.0087
Tensor fusion	0.9091±0.0011	0.6379±0.0102	0.5819±0.0043	0.7467±0.0046
Cross-attention fusion	0.9076±0.0231	0.6247±0.0636	0.5924±0.0343	0.7591±0.0451
Our proposed	0.9161±0.0002	0.6681±0.0007	0.6114±0.0062	0.7827±0.0026

Table 8 Results using clinical notes only with different encoding and fusion settings. The “✓” symbol denotes the selected settings, while the “✗” symbol represents the unselected settings.

Encoder		Fusion		AUROC	AUPRC	min (Se, P+)	F1
CNN	BCB	Early	Late				
✓	✗	✓	✗	0.8377±0.0019	0.4668±0.0103	0.4520±0.0143	0.6436±0.0152
✓	✗	✗	✓	0.8679±0.0025	0.5352±0.0075	0.5105±0.0145	0.6869±0.0128
✗	✓	✓	✗	0.8602±0.0005	0.4947±0.0011	0.4799±0.0026	0.6339±0.0073
✗	✓	✗	✓	0.8727±0.0006	0.5438±0.0026	0.5074±0.0026	0.6880±0.0037

5.2.4 Comparison of single modality and multiple modalities

We conduct experiments to compare the performance of our multimodal model with different combinations of input features. First, we establish a baseline model that uses all modalities, including static demographics, time-series, and clinical notes. Then, we gradually remove modules to create single-modality models and different combinations of modalities. For the single modality experiments, we use the encoder for that modality as in our model. We remove all other modules, including the encoders of the different modalities, the personalized graph-based fusion module, and the MMGCRL module. Experiment results are shown in Table 9.

Static demographics data only: In isolation, the static demographic features exhibit comparatively lower performance with an AUROC of 0.7369, an AUPRC of 0.2413, a min (Se, P+) of 0.2814, and an F1 score of 0.4746.

Time-series data only: The time-series modality emerges as an influential contributor, delivering better performance with an AUROC of 0.8640, an AUPRC of 0.5177, a min (Se, P+) of 0.4925, and an F1 score of 0.6647 when isolated.

Text data only: Our findings underscore the paramount importance of clinical notes, with remarkable AUROC performance of 0.8727, an AUPRC of 0.5438, a min (Se, P+) of 0.5074, and an F1 score of 0.6880 when deployed in isolation.

Multimodalities: The peak of performance is attained when all three modalities are seamlessly integrated, resulting in an AUROC of 0.9161, an AUPRC of 0.6681, a min (Se, P+) of 0.6114, and an F1 score of 0.7827. When limited to time-series and clinical notes, the model performance is an AUROC of

0.9031, an AUPRC of 0.6310, a min (Se, P+) of 0.5763, and an F1 score of 0.7476. In contrast, when limited to static demographic and clinical notes, the model performance is an AUROC of 0.8795, an AUPRC of 0.5613, a min (Se, P+) of 0.5250, and an F1 score of 0.7034.

Overall, experiments show that utilizing the three input data modalities can effectively improve prediction performance, indicating the potential value of leveraging diverse information sources. Moreover, our results consistently underscore the pivotal role of clinical notes in our predictive task, showcasing their dominance in individual performance.

6 Conclusion

Our study presents a novel representation learning framework based on a personalized graph-based fusion approach to address the challenges of predicting ICU mortality with incomplete and variable EMR data among patients. Our approach constructs patient-specific modality aggregation graphs to fuse incomplete multimodal features, and employs the MMGCRL method to capture complementary information. Evaluation on the MIMIC-III dataset demonstrates significant performance improvement in mortality prediction, outperforming several state-of-the-art methods. The strengths of our work include creating tailored patient representations, enhancing prediction accuracy and explainability, and effectively addressing the heterogeneity and incompleteness of EMR data. Through ablation experiments, we verify the benefits of each module in our framework and find that the combined use of all modules contributes to the model's overall performance. Our results also show that using all data modalities results in the best prediction performance.

Table 9 Ablation study of the input data modalities. “SD” represents static demographics, “TS” represents time-series, and “TXT” represents text. The “✓” symbol indicates selected modalities, while the “✗” symbol indicates unselected modalities. The best results are highlighted in bold.

Data modality			AUROC	AUPRC	min (Se, P+)	F1
SD	TS	TXT				
✓	✗	✗	0.7369±0.0017	0.2413±0.0033	0.2814±0.0080	0.4746±0.0013
✗	✓	✗	0.8640±0.0011	0.5177±0.0032	0.4925±0.0076	0.6647±0.0117
✗	✗	✓	0.8727±0.0006	0.5438±0.0026	0.5074±0.0026	0.6880±0.0037
✓	✓	✗	0.8735±0.0021	0.5286±0.0045	0.5030±0.0061	0.6665±0.0036
✓	✗	✓	0.8795±0.0004	0.5613±0.0016	0.5250±0.0033	0.7034±0.0041
✗	✓	✓	0.9043±0.0006	0.6453±0.0017	0.5763±0.0038	0.7575±0.0031
✓	✓	✓	0.9161±0.0002	0.6681±0.0007	0.6114±0.0062	0.7827±0.0026

We find that dividing clinical notes into multiple categories and using the personalized graph-based fusion approach to tackle the variability and incompleteness of the multimodal features among patients improve mortality prediction performance. It suggests that incorporating patient-specific and domain knowledge into the model can enhance its performance. Personalized features like static demographics can also improve the model's prediction performance. These findings highlight the importance of incorporating patient-specific characteristics in multimodal representation learning for mortality prediction. Furthermore, utilizing the MMGCRL module can further improve the model's prediction performance. It suggests that contrastive learning with gating guides the model in capturing distinct information from multimodal representations and learning effective clinical status representation. Our results indicate that our proposed framework can effectively leverage multiple data modalities and personalized features to improve mortality prediction.

One limitation of our study is that we only evaluate our proposed framework on the MIMIC-III dataset, primarily due to the lack of publicly available clinical notes data in other datasets. In the future, we will explore the generalization of our approach to other public datasets when clinical notes are available. Additionally, our study only focuses on mortality prediction and does not consider other clinical outcomes. Future research can investigate the effectiveness of our approach in predicting different patient outcomes. Furthermore, we intend to enhance our method by incorporating more advanced temporal encoders for time-series data modality.

Despite these limitations, our study represents a significant step towards effective patient-specific multimodal representation learning for mortality prediction. By acknowledging the variability of multimodal features among patients and guiding the model to learn effective patient-specific clinical status representation from these features, we contribute to advancing personalized healthcare.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. U24A20256) and the Science and Technology Major Project of Changsha (No.kh2402004). This work was carried out in part using

computing resources at the High Performance Computing Center of Central South University.

References

- [1] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence, APACHE-acute physiology and chronic health evaluation: A physiologically based classification system, *Crit. Care Med.*, vol. 9, no. 8, pp. 591–597, 1981.
- [2] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, APACHE II: A severity of disease classification system, *Crit. Care Med.*, vol. 13, no. 10, pp. 818–829, 1985.
- [3] E. Paul, M. Bailey, and D. Pilcher, Risk prediction of hospital mortality for adult patients admitted to Australian and New Zealand intensive care units: Development and validation of the Australian and New Zealand risk of death model, *J. Crit. Care*, vol. 28, no. 6, pp. 935–941, 2013.
- [4] I. Deznabi, M. Iyyer, and M. Fiterau, Predicting in-hospital mortality by combining clinical notes with time-series data, in *Proc. Findings of the Association for Computational Linguistics: ACL-IJCNLP*, Virtual Event, 2021, pp. 4026–4031.
- [5] A. Al-Dailami, H. Kuang, and J. Wang, Predicting length of stay in ICU and mortality with temporal dilated separable convolution and context-aware feature fusion, *Comput. Biol. Med.*, vol. 151, p. 106278, 2022.
- [6] M. Makino, R. Yoshimoto, M. Ono, T. Itoko, T. Katsuki, A. Koseki, M. Kudo, K. Haida, J. Kuroda, R. Yanagiya, et al., Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning, *Sci. Rep.*, vol. 9, no. 1, p. 11862, 2019.
- [7] A. Mathioudakis, I. Rousalova, A. A. Gagnat, N. Saad, and G. Hardavella, How to keep good clinical records, *Breathe*, vol. 12, no. 4, pp. 369–373, 2016.
- [8] S. Khadanga, K. Aggarwal, S. Joty, and J. Srivastava, Using clinical notes with time series data for ICU management, in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 6432–6437.
- [9] W. Lyu, X. Dong, R. Wong, S. Zheng, K. Abell-Hart, F. Wang, and C. Chen, A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction, in *Proc. AMIA Annu. Symp.*, Bethesda, MD, USA, 2022, pp. 719–728.
- [10] J. Liu, D. Capurro, A. Nguyen, and K. Verspoor, Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities, *J. Biomed. Inform.*, vol. 145, p. 104466, 2023.
- [11] M. Li, A. Micheli, Y. G. Wang, S. Pan, P. Lió, G. S. Gnecco, and M. Sanguineti, Guest editorial: Deep neural networks for graphs: Theory, models, algorithms, and applications, *IEEE Trans. Neural Network. Learn. Syst.*, vol. 35, no. 4, pp. 4367–4372, 2024.

- [12] M. Li, X. Zhuang, L. Bai, and W. Ding, Multimodal graph learning based on 3D Haar semi-tight framelet for student engagement prediction, *Information Fusion*, vol. 105, p. 102224, 2024.
- [13] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, Multitask learning and benchmarking with clinical time series data, *Sci. Data*, vol. 6, no. 1, p. 96, 2019.
- [14] E. Rocheteau, P. Liò, and S. Hyland, Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit, in *Proc. Conf. Health, Inference, and Learning*, Virtual Event, 2021, pp. 58–68.
- [15] S. Purushotham, C. Meng, Z. Che, and Y. Liu, Benchmarking deep learning models on large healthcare datasets, *J. Biomed. Inform.*, vol. 83, pp. 112–134, 2018.
- [16] M. Sushil, S. Šuster, K. Luyckx, and W. Daelemans, Patient representation learning and interpretable evaluation using clinical notes, *J. Biomed. Inform.*, vol. 84, pp. 103–113, 2018.
- [17] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, and J. Gao, ConCare: Personalized clinical feature embedding via capturing the healthcare context, in *Proc. 34th AAAI Conf. Artificial Intelligence*, New York, NY, USA, 2020, pp. 833–840.
- [18] A. Al-Dailami, H. Kuang, and J. Wang, Attention-based memory fusion network for clinical outcome prediction using electronic medical records, in *Proc. 2022 IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, Las Vegas, NV, USA, 2022, pp. 902–907.
- [19] W. Zhao, Z. Chen, P. Xie, J. Liu, S. Hou, L. Xu, Y. Qiu, D. Wu, J. Xiao, and K. He, Multi-task oriented diffusion model for mortality prediction in shock patients with incomplete data, *Information Fusion*, vol. 105, p. 102207, 2024.
- [20] T. B. Murdoch and A. S. Detsky, The inevitable application of big data to health care, *JAMA*, vol. 309, no. 13, pp. 1351–1352, 2013.
- [21] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis, *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1589–1604, 2018.
- [22] J. Wu, X. Ye, C. Mou, and W. Dai, FineEHR: Refine clinical note representations to improve mortality prediction, in *Proc. 2023 11th Int. Symp. Digital Forensics and Security (ISDFS)*, Chattanooga, TN, USA, 2023, pp. 1–6.
- [23] Y. Wu, Z. Chen, X. Yao, X. Chen, Z. Zhou, and J. Xue, Jan: Joint attention networks for automatic ICD coding, *IEEE J. Biomed. Health Inform.*, vol. 26, no. 10, pp. 5235–5246, 2022.
- [24] S. Gehrmann, F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Foote Jr, E. T. Moseley, D. W. Grant, P. D. Tyler, et al., Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives, *PLoS One*, vol. 13, no. 2, p. e0192360, 2018.
- [25] S. Darabi, M. Kachuee, S. Fazeli, and M. Sarrafzadeh, TAPER: Time-aware patient EHR representation, *IEEE J. Biomed. Health Inform.*, vol. 24, no. 11, pp. 3268–3275, 2020.
- [26] T. Shaik, X. Tao, L. Li, H. Xie, and J. D. Velásquez, A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom, *Information Fusion*, vol. 102, p. 102040, 2024.
- [27] M. Hashir and R. Sawhney, Towards unstructured mortality prediction with free-text clinical notes, *J. Biomed. Inform.*, vol. 108, pp. 103489, 2020.
- [28] B. Yang and L. Wu, How to leverage multimodal EHR data for better medical predictions? arXiv preprint arXiv:2110.15763, 2021.
- [29] E. Alsentzer, J. R. Murphy, W. Boag, W. H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, Publicly available clinical BERT embeddings, arXiv preprint arXiv:1904.03323, 2019.
- [30] X. Zhang, S. Li, Z. Chen, X. Yan, and L. R. Petzold, Improving medical predictions by irregular multimodal electronic health records modeling, in *Proc. 40th Int. Conf. Machine Learning*, Honolulu, HI, USA, 2023, pp. 41300–41313.
- [31] S. N. Shukla and B. M. Marlin, Multi-time attention networks for irregularly sampled time series, arXiv preprint arXiv:2101.10318, 2021.
- [32] J. Huo, Z. Hong, M. Chen, and Y. Duan, MIFNet: Multimodal interactive fusion network for medication recommendation, *J. Supercomput.*, vol. 80, no. 9, pp. 12313–12345, 2024.
- [33] C. Zhang, X. Chu, L. Ma, Y. Zhu, Y. Wang, J. Wang, and J. Zhao, M3Care: Learning with missing modalities in multimodal healthcare data, in *Proc. 28th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, Washington, DC, USA, 2022, pp. 2418–2428.
- [34] M. Schuster and K. K. Paliwal, Bidirectional recurrent neural networks, *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [35] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, in *Proc. 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, 2014, pp. 103–111.
- [36] K. Huang, J. Altsaier, and R. Ranganath, ClinicalBERT: Modeling clinical notes and predicting hospital readmission, arXiv preprint arXiv:1904.05342, 2020.
- [37] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, BioBERT: A pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [38] G. Levy, N. Schupf, M. X. Tang, L. J. Cote, E. D. Louis, H. Mejia, Y. Stern, and K. Marder, Combined effect of age and severity on the risk of dementia in Parkinson’s disease, *Ann. Neurol.*, vol. 51, no. 6, pp. 722–729, 2002.
- [39] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, Graph attention networks, arXiv preprint arXiv:1710.10903, 2018.
- [40] 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, p. 160035, 2016.
- [41] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao,

L. P. Morency, and E. Hoque, Integrating multimodal information in large pretrained transformers, in *Proc. 58th Annu. Meeting of the Association for Computational Linguistics*, Virtual Event, 2020, pp. 2359–2369.

- [42] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, Predicting in-hospital mortality of ICU patients: The PhysioNet/computing in cardiology challenge 2012,

in *Proc. 2012 Computing in Cardiology*, Krakow, Poland, 2012, pp. 245–248.

- [43] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, Tensor fusion network for multimodal sentiment analysis, in *Proc. 2017 Conf. Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 1103–1114.



Abdulrahman Al-Dailami received the BEng degree from Sana'a University, Yemen in 2011, the MEng and PhD degrees from Central South University, China in 2019 and 2024, respectively. He is currently a postdoctoral researcher at Hunan Provincial Key Lab on Bioinformatics, School of Computer

Science and Engineering, Central South University, China. His research interests include deep learning, multimodal representation learning, and medical informatics.



Hulin Kuang received the PhD degree from City University of Hong Kong, Hong Kong, China in 2016. He was a postdoctoral researcher at Department of Clinical Neurosciences, University of Calgary, Canada from 2017 to 2020. Now he is a distinguished associate professor at School of Computer Science and

Engineering, Central South University, China. His research interests include medical image analysis, and deep learning.



Jianxin Wang received the BEng, MEng, and PhD degrees in computer engineering from Central South University, China in 1992, 1996, and 2001, respectively. He is currently a professor at School of Computer Science and Engineering, Central South University, China. His research interests include algorithm

analysis and optimization, parameterized algorithm, bioinformatics, medical image analysis, and computer networks.