

RPMF: In-hospital Mortality Risk Prediction based on Multimodal Fusion

Changtong Ding¹, Shichao Geng², Quanrun Song¹, Yalong Liu¹, Yu Zhao¹, Xiangwei Zhang³, and Lin Wang¹ *

¹ School of Information Science and Engineering, Shandong Normal University, Jinan 250014, Shandong, China 492508160@qq.com

² School of Journalism and Communication, Shandong Normal University, Jinan 250014, Shandong, China gengsc@hotmail.com

³ Department of Thoracic Surgery, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan 250014, Shandong, China

Abstract. In predicting mortality and disease risk, deep learning in clinical decision-support technology to analyze structured electronic health records (EHR) has been a highly scrutinized research area. However, despite the abundance of narrative clinical diagnostic records and ICU physiological indicators, we still believe there are two shortcomings in current disease risk prediction efforts. On the one hand, current models fail to utilize the available data fully and lack comprehensive modeling of patient characteristics. On the other hand, existing studies have not effectively captured potential correlations between multimodal data. In this paper, we introduced a pioneering design concept based on the "initial health state of the patient," which involves considering the patient's current health status (characterized by disease information) as a key element in sequence modeling. In addition, we have innovatively adopted the Informer model for processing time-series data of physiological indicators of ICU patients. More critically, we developed a multimodal feature interaction module that captures the interrelationships between different data modalities. Extensive experiments on real-world datasets show that our model has good results.

Keywords: Machine learning · Electronic health records · Initial health status · In-hospital mortality risk prediction · Multimodal fusion.

1 Introduction

The Intensive Care Unit (ICU) aims to provide comprehensive and reliable treatment for critically ill patients. In 2019, the worldwide mortality rate for ICU patients remained high, ranging between 10% and 20%. Therefore, predicting the risk of patient mortality is of utmost importance. The advancement of medical technology has raised higher demands for assessing patient mortality risk. A timely and accurate assessment

* Corresponding author.

of patient mortality risk and early identification of patients with poor prognoses can assist physicians in a comprehensive evaluation of the patient's condition. This is crucial for improving patient survival rates and mitigating adverse outcomes. Predicting ICU mortality rates can help doctors assess the patient's condition, enabling the use of more effective and cost-efficient diagnostic and treatment methods. Meanwhile, the rapid development of hospital informatics has propelled the digitization of medical records, leading to a wealth of electronic health records (EHR) available for medical research and applications. According to research by Julia Adler-Milstein and others (2015)[1], the widespread adoption of electronic health record (EHR) systems in the United States has generated a substantial volume of data, offering opportunities for improving clinical decision support through machine learning based predictive modeling. In particular, technologies based on deep learning, such as those mentioned by Theresa E. Fuller and others[4], have opened up new possibilities for disease prediction, as demonstrated by network structures proposed for predicting in-hospital mortality, such as the one by Junyu Luo and colleagues[10].

With the widespread adoption of electronic health record (EHR) systems, previous research efforts have mainly focused on single data modalities or simply mechanically connecting diagnostic records and physiological indicators. However, on the one hand, a patient's pre-ICU admission health status can significantly influence their health outcomes, which has not been considered in prior work. Additionally, the extraction of patient health status is a crucial concern. On the other hand, simply mechanically connecting two modalities also neglects the complexity of modality and temporal information. Although multimodal prediction models are becoming increasingly popular in various fields, the use of multimodal approaches in predictive modeling based on EHR data is still very limited. The lack of exploration of multimodal data modeling and multimodal relationship mining has become a bottleneck that has stalled progress in the field of risk prediction.

To solve the above problems, this paper proposes a new mortality risk prediction model, RPMF (Risk Prediction based on Multimodal Fusion). RPMF can be divided into three parts: feature extraction, feature interaction, and risk prediction. In the feature extraction phase, the Word2Vec model is applied to ICD codes for vectorization to capture the semantic relationships between diagnostic codes, and the Time Interval aware Self-Attention (TISA) mechanism is utilized to capture the temporal correlation within the ICD sequence. For clinical variables, the Informer model is used to extract features to respond to their time sensitivity and complexity. Meanwhile, the BERT model is used to extract information from clinical notes, generating context-rich embedded representations for each word in natural language with its powerful contextual understanding. In the feature interaction phase, the RPMF model maps the features of different modalities into a unified feature space and uses the Transformer structure to capture the feature relationships intra- and inter-modality. At the same time, with its self-attention mechanism, it handles sequences of different lengths and captures the dependencies between any elements within the sequences. Finally, in the risk prediction stage, the model comprehensively integrates all modal features for death risk prediction and learns the feature representation to predict the patient's death risk through a deep

learning method. The RPMF model not only considers the information from individual data sources, but also fully exploits the complementary relationship between multimodal data, which significantly enhances the accuracy and reliability of the prediction, and provides strong support for medical decision-making.

We harness the vast Medical Information Mart for Intensive Care (MIMIC-III) dataset to conduct thorough experiments, validating our proposed method. Our findings reveal that, within this dataset, our approach significantly surpasses the baseline, affirming the effectiveness of our model.

The key contributions of this paper are highlighted as follows:

1. We have innovatively designed a multimodal fusion risk prediction framework, PRMF, to significantly improve the accuracy of risk prediction by integrating features from different modalities. This method utilizes the complementary strengths of different data modalities to provide a more comprehensive view of patient health status.
2. An efficient ICD coding scheme combining Word2Vec and Time Interval aware Self-Attention is developed to model patients' health states more effectively. The method captures the semantic relationships between diagnoses and their temporal dynamics, providing nuanced insights into patient health trajectories.
3. A novel multimodal feature interaction approach was designed to facilitate the simultaneous mining of intra- and inter-modal feature relationships. This strategy significantly improves predictive performance by ensuring deep data integration from multiple sources, capturing a more comprehensive set of factors influencing patient risk.
4. We conducted extensive experimental analyses on real-world datasets to validate the effectiveness of our proposed model. These experiments demonstrate the superior performance of the model in health risk prediction, emphasizing the practical value and potential impact of our approach in healthcare analytics.

The organization of the remainder of this paper is as follows: In Section 2, we introduce deep learning methods applied to clinical medical prediction. Section 3 describes the approach to model construction. In Section 4, we provide detailed insights into the implementation of experiments. Finally, the conclusion is presented in Section 5.

2 Related Work

Electronic health records comprise structured physiological indicator data and unstructured clinical diagnostic records. Regarding structured physiological indicator data, vital signs are a key component of structured clinical variables, including heart rate, respiratory rate, temperature, and blood pressure. These variables can be easily transformed into time series data, providing researchers with a broad study area. With the development of deep learning[19], more and more researchers are beginning to apply

deep learning to the medical field. In previous works, Jocelyn Zhu et al.[18] and Xiaoran Li et al.[8] utilized deep learning artificial intelligence analysis of clinical variables to predict patient mortality. Research by Sanjay Purushotham et al.[12] indicated that deep learning models perform better when raw clinical time series data is used as input features for the model.

In general, unstructured clinical records contain rich and complementary background information, such as patient symptoms, medical history, and treatment details, often presented in the form of medical histories or orders. While conventional pre-trained natural language models, such as the Bidirectional Encoder Representations from Transformers (BERT) introduced by Devlin et al.[2], cannot directly process specific clinical records, various variants and approaches have been developed. For example, Yikuan Li et al.[9] incorporated age and location information to model clinical records, and its flexible architecture allows for the integration of multiple heterogeneous concepts (such as diagnoses, medications, measurements, etc.) to enhance predictive accuracy further.

The fusion of structured physiological indicator data and unstructured clinical diagnosis records is significant for improving medical decision-making and patient treatment effects. This fusion can provide more comprehensive and accurate medical information and provide doctors and clinical experts with better aids to diagnose disease, plan treatment better, and predict patient outcomes. Dhanesh Ramachandram et al.[14] pointed out that naively connecting features from different modalities may lead to poor results, and it is very challenging to fuse structured clinical variables and unstructured clinical record data. Sexual tasks because they represent two completely different fields.

In a comprehensive survey of deep representation learning for single and multiple domains from EHR data by Yuqi Si et al.[15], Extract features from structured data and then connect these two features. For example, Yuqing Wang et al.[16] proposed a multimodal model to predict early sepsis using laboratory measurements, vital signs, and patient demographic data from the patient's first 12, 18, 24, 30, and 36 hours in the ICU and clinical records to predict sepsis. Zhi Qiao et al.[13] proposed a hierarchical multi-label diagnostic prediction model (MHM) based on clinical multimodal data to accommodate structural information among diseases, time series data, and discrete medical codes. Yashodip R. Pawar et al.[11] combined a multimodal model with structured and unstructured data to predict 30-day all cause mortality in COVID-19 patients after admission.

However, these methods simply connect them together, but they do not consider the disease state information before hospitalization and largely ignore the complexity of modal and temporal information. Although multimodal transformers are gaining popularity in various fields, their application in EHR-based predictive modeling is limited.

3 METHODS

3.1 Overview of Architecture

Our proposed RPMF (Risk Prediction based on Multimodal Fusion) consists of three parts: Feature Extraction, Feature Interaction, and Risk Prediction. Among them, the feature extraction captures the features of ICD, Variables, and Notes; the feature interaction interacts with the features of different modalities; and the risk prediction module makes the prediction based on the above features. The overall architecture of the model is shown in Fig. 1.

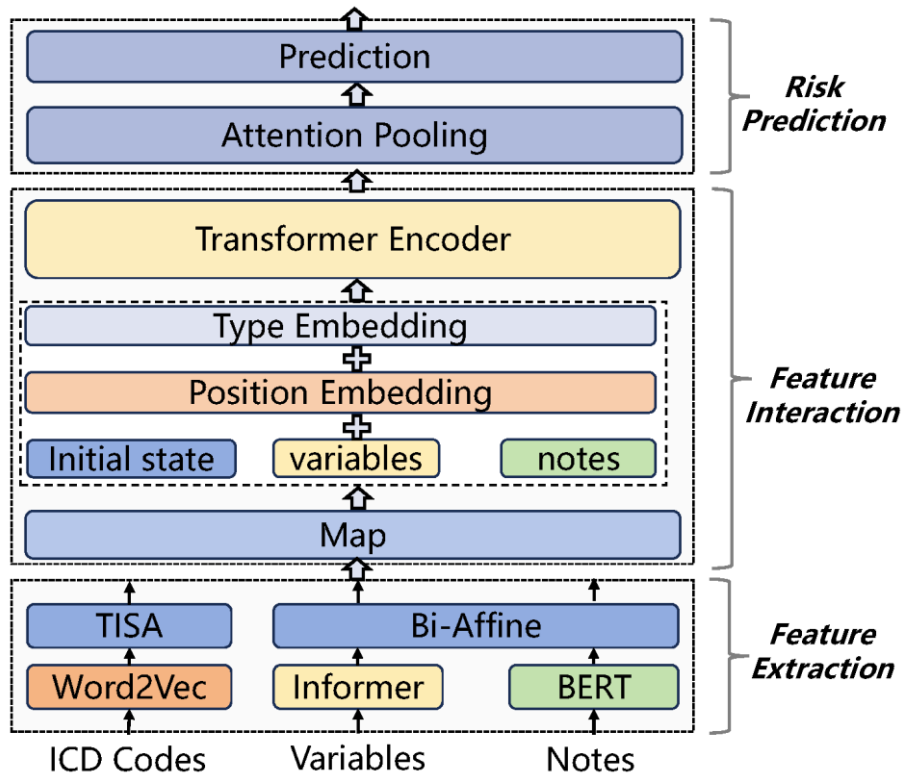


Fig. 1. Model Architecture of RPMF.

3.2 Feature extraction

In the feature extraction section, we intend to extract multimodal HER data to predict in-hospital mortality. Specifically, here we extract the ICD Codes features that represent the initial health state, as well as the clinical note features, and clinical variable features, respectively. Each feature extraction method is described below.

Initial health status For better mortality risk prediction, we consider the patient's current health status (represented using disease information), inspired by word2vec in the NLP field, since ICD codes usually consist of a series of codes representing specific diseases or medical concepts. We first use word2vec to encode the ICD code of the disease and regard the ICD code list of each patient's visit as a sentence in the text; each ICD code corresponds to a word, and we use word2vec pre-training to get the ICD code embedding. Using Word2Vec, these ICD codes can be mapped into low-dimensional vectors so that ICD codes with similar meanings or correlations are closer together in the vector space. In this way, we can measure the semantic association between ICD codes by calculating the cosine similarity between vectors and using other methods. The Word2Vec algorithm can learn word vectors by considering words within the context window, which means that when processing ICD encodings, it takes into account the contextual information between ICD encodings. This helps to capture better the semantic relationships and correlations between ICD codes rather than just the labels of the codes themselves. At the same time, Word2Vec can also link uncommon ICD codes with similar codes by sharing context information. This can solve the problem of some codes appearing less frequently in the data set, thereby overcoming the problem of data sparsity. We denote the above process as Eq.(1).

$$x_1, x_2, \dots, x_n = \text{Word2Vec}(\text{ICDs}) \quad (1)$$

where $x_i \in \mathbb{R}^d$ denotes the representation vector of the i th ICD code obtained via Word2Vec.

Second, the "time interval aware self-attention" structure is used to extract the patient's health status. Specifically, the vectors obtained by word2vec processing and ICD encoding are used as input x_1, x_2, \dots, x_n . After matrix mapping, these vectors are used as Q, K, and V of the self-attention model. The time interval of visits considered by the state is different. Generally speaking, the smaller the time interval, the higher the similarity of the disease conditions corresponding to the two time points. With this in mind, compared with the traditional self-attention model, our "time interval aware self-attention" has added the element of different time intervals. We use t_1, t_2, \dots, t_n to get the time interval $\Delta t_1, \Delta t_2, \dots, \Delta t_n$, and then pass the time interval Δt through the function to get the weight w_i of the corresponding time interval.

$$w_i = 1 - 0.3 \times \frac{\Delta t_i}{t_n - t_1} \quad (2)$$

$$\Delta t_i = t_n - t_i \quad (3)$$

We then multiply the result of q and k points. The corresponding weights are multiplied, and then they go through softmax and V for calculation. After "time interval aware self-attention", q_1, q_2, \dots, q_n are obtained.

$$q_i = K \cdot \text{soft max}\left(\frac{x_i @ K w_i}{\sqrt{d_1}}\right) \quad (4)$$

Clinical variable feature Clinical variable features contain many physiological indicators that are critical for risk prediction, such as heart rate, blood pressure, and blood analysis results. These features provide healthcare professionals with an immediate snapshot of a patient’s current health status and are commonly used for short- and long-term health risk assessment in clinical decision support systems. Due to the temporal dynamics and complex interrelationships of these variables, we chose the Informer (proposed by Haoyi Zhou et al.)[17] model to handle these critical clinical physiological metrics.

Informer introduced ProbSparse self-attention and self-attention distilling operations. ProbSparse self-attention is an improved self-attention mechanism dedicated to time series processing. Its input is a time series, and the output is a sequence of equal length, with each element encoded. The core idea is to sparse the attention weights in self-attention. Traditional self-attention computes the attention weight of each location with all other locations, resulting in an increase in computation as the sequence length increases. However, ProbSparse self-attention adopts probability sparsity, only considers some key positions, and localizes the calculation of attention weight, thereby reducing the computational complexity.

Finally, we extracted the clinical variable features using Informer.

$$v_1, v_2, \dots, v_m = \text{Informer}(\text{Variables}) \quad (5)$$

Where $v_i \in \mathbb{R}^d$ denotes the i th clinical variable feature obtained using Informer.

Clinical note feature Clinical notes, as a component of electronic health records, contain a wealth of patient health information such as symptom descriptions, treatment progress, and physicians’ professional assessments. These free-text data often contain clinical domain-specific terminology and complex sentence constructions, which pose specific challenges to natural language processing models. So, we chose Clinical BERT (Huang et al., 2019)[5] as the pretrained language model because it was trained on data containing all MIMIC-III clinical records, making it more suitable for our tasks. To better fit simulated data and fine-tune it for the in-hospital mortality prediction task, we refer to it as simulated BERT (MBERT). (proposed by Jacob Devlin et al.)[12]. With such fine-tuning, Clinical BERT can learn clinical-specific contextual embeddings in simulated data more accurately. We extracted hourly clinical note embeddings associated with temporal information for each patient to represent the clinical note data for a specific period.

Ultimately, we used BERT to obtain clinical note features.

$$b_1, b_2, \dots, b_k = \text{BERT}(\text{Notes}) \quad (6)$$

where $b_i \in \mathbb{R}^d$ denotes the i th clinical note feature obtained using BERT.

In order to mine the deep relationships between clinical note features and clinical variable features, we employ the Bi-Affine structure to perform feature interactions. The Bi-Affine structure was proposed by Timothy Dozat et al. [3] and allows us to

finely fuse features from different sources at the sequence element level. This structure was chosen based on its advantages in modeling complex relationships between two sets of features. These features can come from different embedding layers, different representation learning methods, or different tasks. This can capture more rich semantic information and help improve the performance of the model. On the other hand, the model can understand the input data more comprehensively through the feature interaction of vectors, thereby better learning the relationships and patterns between data and improving the model's representation learning ability.

We choose the Bi-Affine structure (proposed by Timothy Dozat et al.)[3] to perform feature interaction on Clinical Notes Embedding and Clinical Variables Embedding at the sequence element level. The common multimodal feature fusion operation first calculates the mean value of the single-modal sequence features to obtain a single-modal vectorized representation. Then, it aggregates the multimodal features using the sum or splicing concat operation. This paper uses a bi-affine attention mechanism (Bi-Affine) feature interaction at the sequence element level to perform multimodal information fusion better. Assuming that the diagnostic record text features are expressed as $V \in \mathbb{R}^{m \times d}$, after the above layers of processing, the physiological index features are expressed as $B \in \mathbb{R}^{k \times d}$, and the sequences V and B are fused and interacted through the dual-radiation attention mechanism. Among them, W_1 and W_2 are weight matrices.

$$A_1 = \text{soft max}(VW_1B^T) \quad (7)$$

$$A_2 = \text{soft max}(BW_2V^T) \quad (8)$$

$$V' = A_1V \quad (9)$$

$$B' = A_2B \quad (10)$$

With the above operations, we performed a preliminary interaction between the clinical variables and the clinical note features, and obtained the interacted features $V' \in \mathbb{R}^{m \times d}$, $B' \in \mathbb{R}^{k \times d}$.

3.3 Feature interaction

In the feature extraction stage, we extracted three types of features from different data sources: initial state features $Q' = [q'_1, q'_2, \dots, q'_m]$, clinical variable features $V' = [v'_1, v'_2, \dots, v'_m]$, and clinical note features $B' = [b'_1, b'_2, \dots, b'_m]$. Each type of feature represents a different perspective and dimension of the data, but they exist in their own vector space, which poses a challenge for subsequent unified processing and analysis. To address this problem and to fully utilize the intrinsic connections between these multimodal features, we designed a Transformer-based feature interaction module. This module realizes the effective fusion and interaction of these features through the combination of feature mapping, embedding generation, and Transformer network to capture the complex relationships within and across modalities.

The features of the three modalities are in different vector spaces and cannot be represented identically. First, we map the features in different vector spaces to a uniform vector space by means of a feature mapping function. This step is crucial because it not only solves the problem of inconsistent vector spaces but also lays the foundation for subsequent feature interaction and fusion.

$$\bar{Q}=\text{Map}_q(Q), \bar{V}=\text{Map}_v(V'), \bar{B}=\text{Map}_v(B') \quad (11)$$

$$\text{Map}_i(X) = W_m X + b_m \quad (12)$$

Where $W_m \in \mathbb{R}^{d \times d}$, $b_m \in \mathbb{R}^d$ are the mapping weight matrix and the bias vector, respectively.

In order to efficiently process these mapped features in Transformer and allow the model to recognize features of different modalities as well as features of different positions, we introduce position embeddings $P_i \in \mathbb{R}^d$ and type embeddings $T \in \mathbb{R}^{3 \times d}$. Position embeddings are used to encode the positional information among the features within a sequence to help the model understand the order of the features in the sequence, while type embeddings are used to differentiate between features of different modalities so that the model can recognize which modality each feature belongs to. We obtain the matrix used as input to the model by summing the feature matrix with these two embeddings.

$$H = \text{LayerNorm} \left(\begin{bmatrix} \bar{Q} + P_{1:n} + T_1 \\ \bar{V} + P_{1:m} + T_2 \\ \bar{B} + P_{1:k} + T_3 \end{bmatrix} \right) \quad (13)$$

Next, we utilize the Transformer network to process these features incorporating positional and type information. Through its self-attention mechanism, the Transformer can efficiently capture the dependency between any two positions in a sequence, no matter how far apart they are. This feature makes the Transformer well suited for processing sequence data with complex internal structures. In our scenario, the Transformer network captures feature relationships within and across modalities, enabling the modeling of deep connections between different data perspectives.

In order to facilitate the representation, we denote the above input sequence uniformly as H . We use Transformer Encoder to model intra- and inter-modal features:

$$\bar{H} = \text{TransformerEncoder}(H) \quad (14)$$

Through the above steps, the Transformer-based feature interaction module we constructed is not only able to handle multimodal data from different sources, but also able to deeply mine the intrinsic connection between these data, providing a richer and more detailed feature representation for subsequent risk prediction.

3.4 Risk prediction

After successful feature extraction and feature interaction, we have deeply explored and fully exploited multimodal features from different data sources. The next step is to construct a final representation of the patient by fusing these multimodal features and performing risk prediction based on this representation to support the medical decision-making process.

First, after obtaining the feature sequences that have been processed by the feature interaction module, we deploy an attention-based pooling operation to obtain the comprehensive representation features of an individual patient. Specifically, this step assigns a weight by calculating the importance of each feature vector in the patient representation, and then weights and sums all the feature vectors based on these weights to obtain the final patient representation vector r . This process can be expressed as Eq.(16):

$$r = \sum_i \alpha_i \cdot \bar{H}_i \quad (15)$$

where α_i is the attention score of each feature vector computed through an attention mechanism denoted as:

$$\alpha_i = q^T \sigma(W_1 \bar{H}_i + W_2 \bar{H}_{avg} + b) \quad (16)$$

Where $W_1, W_2 \in \mathbb{R}^{d \times d}$ denotes the weight matrix, $b \in \mathbb{R}^d$ denotes the bias vector, σ denotes the Sigmoid activation function, and $q \in \mathbb{R}^d$ denotes the attention parameter vector. An attention score α_i is computed by combining the overall features H_{avg} . The final patient representation vector r is obtained by weighting and summing all features by the attention score.

Next, we use a fully connected layer containing a Sigmoid activation function to predict the patient's risks. This layer receives as input the patient representation vector r obtained in the previous step. It outputs a prediction value \hat{y} , which represents the probability of a certain health risk for the patient. The output of the prediction model can be expressed as Eq.(18):

$$\hat{y} = \sigma(W_r r + b_r) \quad (17)$$

where $W_r \in \mathbb{R}^{d \times d}$ and $b_r \in \mathbb{R}^d$ denote the weight matrix and bias vector of the fully connected layer, respectively. With the Sigmoid function, we are able to compress the output value between 0 and 1 to obtain a probability value, which accomplishes the prediction of the patient's risk.

4 Results

4.1 Research Dataset

We used Johnson et al.[6]’s 2016 method to extract electronic health record (EHR) data from the MIMIC-III dataset, focusing on inpatients. We transformed clinical variables into time series data from ICU instruments, resulting in 17 variables such as capillary refill rate, blood pressure, and glucose levels. Additionally, following Khadanga’s approach[7], we extracted clinical record data, omitting records without time-related charts and patients without any clinical records. Unlike Khadanga et al., who only considered the first visit of each patient, we treated each visit as a separate sample. We summarize the statistical characteristics of the MIMIC-III dataset in Table 1.

Table 1. Statistics of the post-processed MIMIC III data for the in-hospital mortality prediction task

	Train	Validation	Test
Negative	12216	2682	2748
Positive	1852	404	359
Total	14068	3086	3107

4.2 Experimental details

Experimental software and hardware environment: We use the Python programming language (version 3.8). The model is implemented with the Pytorch (Paszke et al., 2019) deep learning framework. The server CPU used in the experiment is an Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60GHz, the GPU graphics card is an NVIDIA RTX 3090 (24GB), and the memory is 80GB. We adopt an early stopping strategy to train our model efficiently and avoid overfitting. Specifically, we monitor the model’s performance on the validation set; if the model performance does not improve further for three consecutive training epochs, we stop the training process. Meanwhile, we optimize the model using the AdamW optimization algorithm with batchsize 8.

To comprehensively evaluate the performance of our model, we compute three metrics: AUCROC, AUCPR, and F1. Other metrics, such as accuracy, recall, and precision, can be misleading due to the imbalanced nature of the dataset. For robustness, we conducted three sets of experiments, each of which was designed several times according to its purpose, each time using a different initialization. We reported the mean and standard deviation of the results.

4.3 Ablation experiment

To ensure the validity of the various parts of our proposed model and understand each module’s contribution to the final predicted performance, we performed a series of ablation experiments. Ablation experiments assess the importance of a part of the model by removing it to observe the change in performance. In our experiments, we set up the following four variants and compared them to our full model:

(-) **ICD**: In this variant, we removed initial health state features based on ICD coding to examine these features' impact on the model's predictive power.

(-) **Variables**: we excluded clinical variable features from the model to assess the role of these physiologic and examination data in overall performance.

(-) **Notes**: in this variant, clinical notes features were removed to help us understand the contribution of free text data to predictive results.

(-) **FI**: We removed the feature interaction module to assess the importance of this module in fusing multimodal data and improving prediction accuracy.

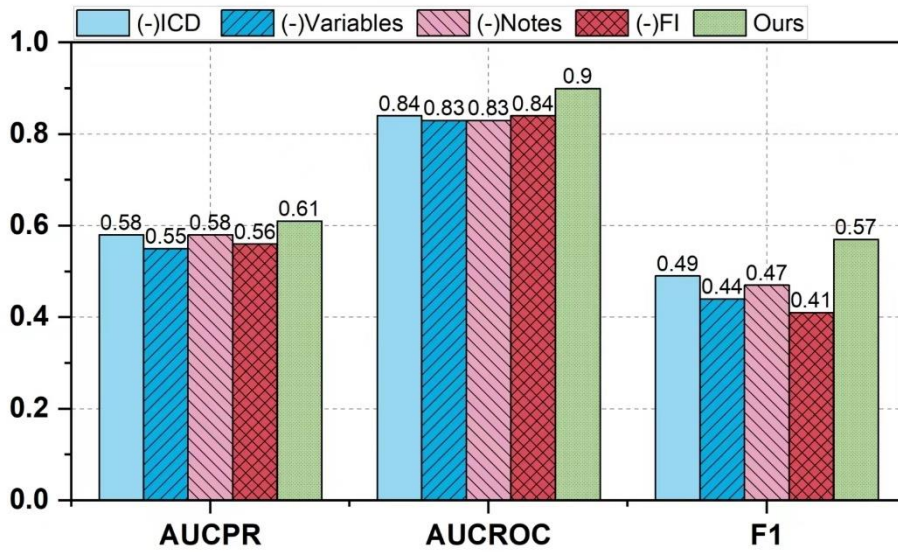


Fig. 2. Ablation experiment results.

The results of the ablation experiments are displayed in Figure 2. The results show that the model performance decreases whenever a feature or module is removed, which validates the importance of each part of our model. Specifically:

- (-) **ICD**: AUCPR, AUCROC, and F1 scores decreased when ICD-based features were removed, which suggests that ICD features play an important role in the predictive model, possibly because they contain important information about a patient's diagnostic history.

- (-) **Variables**: Similarly, the exclusion of clinical variables resulted in decreased performance for all three metrics, emphasizing the importance of these real-time physiological metrics in predicting patient risk.

- (-) **Notes**: Removing the clinical notes feature similarly showed a decrease in model performance, indicating that this text data provides valuable contextual information that is critical for a full understanding of the patient's state.

- (-) **FI**: The absence of the feature interaction module was similarly large for the model, as evidenced by the significant decrease in all metrics. This demonstrates the important role of this module in integrating different modal data and uncovering the

relationships between them, contributing significantly to the overall performance of the model.

Overall, the results of these ablation experiments not only demonstrate the critical contribution of the individual components of our model to the final performance but also reveal the combined value of multimodal data in healthcare risk prediction. These findings further emphasize the importance of considering multimodal feature fusion when designing predictive models.

4.4 Coding mode experiment

In order to study the effectiveness of our proposed ICD coding method in depth, we conducted meticulous experiments on coding styles. The purpose is to evaluate the impact of different coding strategies on the model performance. We designed the following three options for comparison experiments:

- Plan1 (Ours) refers to using the visit time interval of the query and key as the weight of the query@key result and then calculating softmax.
- Plan2 refers to using the visit time interval of query and key as a relative positional embedding, query@key + time_interval.
- Plan3 refers to mapping the visit time interval to a vector and using ICD code embedding + time embedding as self-attention input without changing its calculation method.

The experimental results of various plans are shown in Table 2.

Table 2. Coding mode experiment results.

PLAN	AUCPR	AUCROC	F1
Plan1(Ours)	0.608(±0.004)	0.897(±0.003)	0.570(±0.008)
Plan2	0.587(±0.002)	0.871(±0.004)	0.542(±0.011)
Plan3	0.591(±0.002)	0.884(±0.001)	0.551(±0.006)

From the experimental data, we can derive the following analysis:

Plan1(our method) shows the best performance among the three scenarios, with an AUCPR of 0.608, an AUCROC of 0.897, and an F1 score of 0.570. Plan1 achieves enhanced sensitivity to time relationships by adjusting the attention scores by directly utilizing the time intervals as weights. This direct time-weighting approach achieved the highest scores on all metrics, suggesting that it is effective in capturing important events and predicting risk more accurately when working with time-series data.

Although Plan2 attempts to embed time intervals as relative positions while keeping the attention calculation unchanged, its performance is slightly inferior to that of Plan1. This may be because the relative position information introduced in the model by Plan2 is not utilized as effectively.

The results of Plan3 are similar to those of Plan2. This option did not significantly outperform Plan2 in terms of performance and had a slight decrease in AUCROC, despite the fact that it was attempting to combine time embedding with ICD coding by means of vectorization. This may be due to the fact that merely using time information as another form of embedding input does not change the attentional mechanism itself and therefore does not fully utilize the time interval information.

In summary, Our proposed Plan1 shows the best performance in our experiments, indicating that directly utilizing the time interval as a weighting factor in the attention mechanism is an effective strategy for encoding temporal information, demonstrating the effectiveness of our design.

4.5 Model training experiment

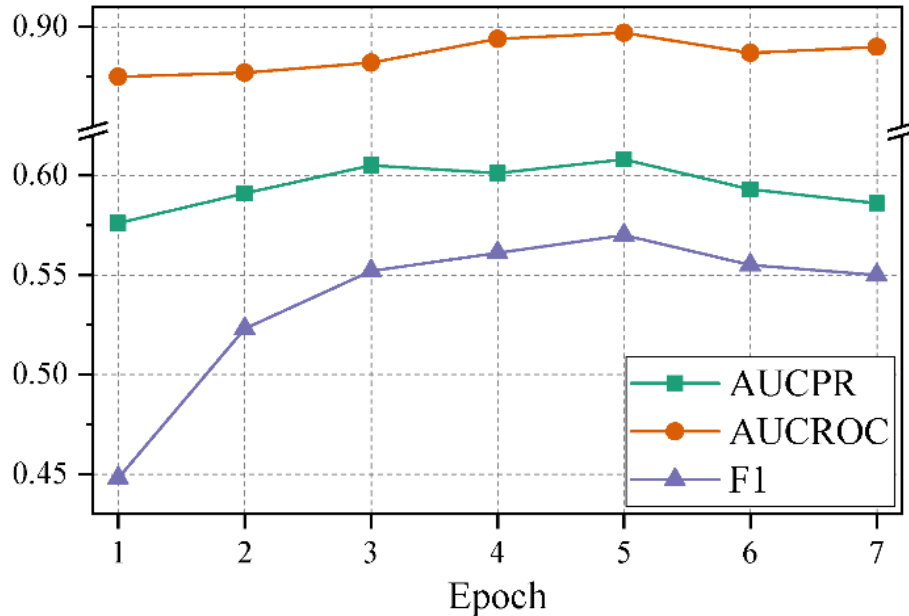


Fig. 3. Model training experiment

To delve into the stability and convergence capabilities of our model during the training process, we designed and conducted a series of model training experiments. The primary aim of these experiments was to observe and analyze the performance changes of the model across different training stages, thereby evaluating its learning efficiency and generalization ability. Specifically, our experimental setup involved seven training epochs, during which we monitored the model's performance on several key evaluation metrics. The changes in these metrics over the training epochs were visually presented through a series of charts (as shown in Fig. 3).

Through meticulous observation and analysis, it is evident that the model demonstrates commendable stability and convergence throughout the training process. As can be clearly seen from Figure 3, with the progression of training epochs, the model showed steady improvement across all the evaluation metrics. Notably, when the training reached the fifth epoch, all the key metrics achieved their optimum levels under the current experimental setup. This phenomenon indicates that the model had attained a high degree of mastery and generalization over the data features through ample learning.

However, it is important to note that, after undergoing sufficient training, the model began to exhibit signs of overfitting upon further training epochs. This was reflected in the slight decline across all performance metrics.

5 CONCLUSION

This paper presents a new multimodal model for risk prediction of in-hospital mortality in the ICU, RPMF. The model incorporates disease states that can utilize clinical notes and integrate clinical variables to better predict mortality risk. At the same time, the model utilizes an effective multimodal interaction technique to capture the characteristic relationships between modalities. To the best of our knowledge, we are the first multimodal model that can take into account a patient's initial health status, integrate both modalities of clinical records and clinical variables, and combine time-series information very effectively. We have conducted comprehensive experiments and the results show that our proposed multimodal model outperforms other state-of-the-art approaches.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 62102237.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adler-Milstein, J., DesRoches, C.M., Kralovec, P., Foster, G., Worzala, C., Charles, D., Searcy, T., Jha, A.K.: Electronic health record adoption in us hospitals: Progress continues, but challenges persist. *Health affairs* 34 12, 2174–80 (2015), <https://api.semanticscholar.org/CorpusID:207381986>
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019), <https://api.semanticscholar.org/CorpusID:52967399>
3. Dozat, T., Manning, C.D.: Deep biaffine attention for neural dependency parsing. *ArXiv abs/1611.01734* (2016), <https://api.semanticscholar.org/CorpusID:7942973>

4. Fuller, T.E., Pong, D.D., Piniella, N.R., Pardo, M., Bessa, N., Yoon, C.S., Boxer, R.B., Schnipper, J.L., Dalal, A.K.: Interactive digital health tools to engage patients and caregivers in discharge preparation: Implementation study. *Journal of Medical Internet Research* 22 (2020), <https://api.semanticscholar.org/CorpusID:213933640>
5. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv abs/1904.05342* (2019), <https://api.semanticscholar.org/CorpusID:119308351>
6. Johnson, A.E.W., Pollard, T.J., Shen, L., wei H. Lehman, L., Feng, M., Ghassemi, M.M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. *Scientific Data* 3 (2016), <https://api.semanticscholar.org/CorpusID:33285731>
7. Khadanga, S., Aggarwal, K., Joty, S.R., Srivastava, J.: Using clinical notes with time series data for icu management. *ArXiv abs/1909.09702* (2019), <https://api.semanticscholar.org/CorpusID:202719390>
8. Li, X., Ge, P., Zhu, J., Li, H., Graham, J.M., Singer, A.J., Richman, P.S., Duong, T.Q.: Deep learning prediction of likelihood of icu admission and mortality in covid-19 patients using clinical variables. *PeerJ* 8 (2020), <https://api.semanticscholar.org/CorpusID:226307358>
9. Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: Behrt: Transformer for electronic health records. *Scientific Reports* 10 (2019), <https://api.semanticscholar.org/CorpusID:198179603>
10. Luo, J., Ye, M., Xiao, C., Ma, F.: Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020), <https://api.semanticscholar.org/CorpusID:221191314>
11. Pawar, Y.R., Henriksson, A., Hedberg, P., Nauc ler, P.: Leveraging clinical bert in multi-modal mortality prediction models for covid-19. *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)* pp. 199–204 (2022), <https://api.semanticscholar.org/CorpusID:251957907>
12. Purushotham, S., Meng, C., Che, Z., Liu, Y.: Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics* 83, 112–134 (2018), <https://api.semanticscholar.org/CorpusID:46962892>
13. Qiao, Z., Zhang, Z., Wu, X., Ge, S., Fan, W.: Mhm: Multi-modal clinical data based hierarchical multi-label diagnosis prediction. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), <https://api.semanticscholar.org/CorpusID:220730133>
14. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine* 34, 96–108 (2017), <https://api.semanticscholar.org/CorpusID:38582742>
15. Si, Y., Du, J., Li, Z., Jiang, X., Miller, T.A., Wang, F., Zheng, W.J., Roberts, K.: Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of biomedical informatics* p. 103671 (2020), <https://api.semanticscholar.org/CorpusID:222140819>
16. Wang, Y., Zhao, Y., Callcut, R.A., Petzold, L.: Integrating physiological time series and clinical notes with transformer for early prediction of sepsis. *ArXiv abs/2203.14469* (2022), <https://api.semanticscholar.org/CorpusID:247762975>
17. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: *AAAI Conference on Artificial Intelligence* (2020), <https://api.semanticscholar.org/CorpusID:229156802>

18. Zhu, J., Ge, P., guo Jiang, C., Zhang, Y., Li, X., Zhao, Z., Zhang, L., Duong, T.Q.: Deep-learning artificial intelligence analysis of clinical variables predicts mortality in covid-19 patients. *Journal of the American College of Emergency Physicians Open* 1, 1364 – 1373 (2020), <https://api.semanticscholar.org/CorpusID:220971347>
19. Zhu, X., Zhang, Y., Wang, J., Wang, G.: Graph-enhanced and collaborative attention networks for session-based recommendation. *Knowledge-Based Systems* p. 111509 (2024)