frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 08 March 2023
DOI 10.3389/fmolb.2023.1136071

OPEN ACCESS

EDITED BY
Xin Gao,
King Abdullah University of Science and Technology, Saudi Arabia

REVIEWED BY
Xi Luo,
First Affiliated Hospital of Xiamen University, China
Wenpeng Lu,
Qilu University of Technology, China

*CORRESPONDENCE
Ke Niu,
niuke@bistu.edu.cn
Naian Xiao,
wsxna@163.com

SPECIALTY SECTION
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 02 January 2023
ACCEPTED 24 February 2023
PUBLISHED 08 March 2023

CITATION
Niu K, Zhang K, Peng X, Pan Y and Xiao N (2023), Deep multi-modal intermediate fusion of clinical record and time series data in mortality prediction.
Front. Mol. Biosci. 10:1136071.
doi: 10.3389/fmolb.2023.1136071

Deep multi-modal intermediate fusion of clinical record and time series data in mortality prediction

Ke Niu[1]*, Ke Zhang[1], Xueping Peng[2], Yijie Pan[3,4] and Naian Xiao[5,6]*

[1]Computer School, Beijing Information Science and Technology University, Beijing, China, [2]Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW, Australia, [3]Department of Computer Science and Technology, Tsinghua University, Beijing, China, [4]EIT Institute for Advanced Study, Ningbo, China, [5]Department of Neurology, The Third Hospital of Xiamen, Xiamen, China, [6]Department of Neurology and Geriatrics, Fujian Institute of Geriatrics, Fujian Medical University Union Hospital, Fuzhou, China

In intensive care units (ICUs), mortality prediction is performed by combining information from these two sources of ICU patients by monitoring patient health. Respectively, time series data generated from each patient admission to the ICU and clinical records consisting of physician diagnostic summaries. However, existing mortality prediction studies mainly cascade the multimodal features of time series data and clinical records for prediction, ignoring thecross-modal correlation between the underlying features in different modal data. To address theseissues, we propose a multimodal fusion model for mortality prediction that jointly models patients' time-series data as well as clinical records. We apply a fine-tuned Bert model (Bio-Bert) to the patient's clinical record to generate a holistic embedding of the text part, which is then combined with the output of an LSTM model encoding the patient's time-series data to extract valid features. The global contextual information of each modal data is extracted using the improved fusion module to capture the correlation between different modal data. Furthermore, the improved fusion module can be easily added to the fusion features of any unimodal network and utilize existing pre-trained unimodal model weights. We use a real dataset containing 18904 ICU patients to train and evaluate our model, and the research results show that the representations obtained by themodel can achieve better prediction accuracy compared to the baseline.

KEYWORDS

mortality prediction, time series, clinical records, multimodal fusion, contextual information

1 Introduction

With the advancement of medical technology, patients in the Intensive Care Unit (ICU) are monitored by different instruments at the bedside that measure different vital signals Sun et al. (2021) about the patient's health. Such as heart rate, systolic blood pressure, temperature, *etc*. This type of data is called time series data. Time series data has two characteristics of irregularity under the aspects of intra-series and inter-series. Intra-series irregularity is the irregular time intervals between observations within a time series. Inter-series irregularity is the different sampling rates among time series. During their stay, doctors visit the patient intermittently for check-ups and make clinical record about the patient's health and physiological progress. These notes can be perceived as summarized expert knowledge about the patient's state. This type of data is called

clinical record data. Researchers based on these data can be used to make mortality predictions for patients, which facilitates hospitals to allocate medical resources appropriately and provide cost-effective resource services to patients who need them most. However, the above data have their own unique characteristics Niu et al. (2022b), such as temporality, high dimensionality, and heterogeneity, so it is a challenge to extract key features from the large amount of patient data for prediction tasks.

In recent years, deep learning techniques have been widely used in the medical field, enabling the task of extracting key features of patients to achieve predictive modeling. In particular, Recurrent Neural Networks (RNN) and its variants Long Short Term Memory (LSTM) Hochreiter and Schmidhuber (1997) and Gated recurrent Unit (GRU) Cho et al. (2014) have been used to process time series data. Lipton et al. (2015) used LSTM to model patient clinical data to classify 128 diagnoses. Che et al. (2018) used GRU-d Pei et al. (2021) (GRU-based recurrent neural network) to model patient time series data generated from ICU to analyze missing values in the time series. Convolutional neural networks (CNN) and BERT models have also been widely used in medical prediction tasks. Grnarova et al. (2016) used convolutional neural networks (CNN) to construct document representations for mortality prediction tasks. Lee et al. (2020) proposed the BioBert model for use in various biomedical text mining tasks. However, most previous work used a single data source for mortality prediction without aggregating information from multiple sources of data, making the representation learned by the model incomplete.

Most previous work has used time series data for mortality prediction tasks. For example, Zhu et al. (2018) combined a bidirectional LSTM model with supervised learning to capture temporal changes in time series data and used the extracted time series features for ICU mortality prediction. Some researchers have also attempted to use clinical records for mortality prediction tasks. For example, Darabi et al. (2020) used transformer networks and the recently proposed BERT language model to embed these clinical record data streams into a unified vector representation for downstream mortality prediction tasks. However, in a real-world clinical setting, time series data can be missing due to different collection frequencies Sun et al. (2020), and thus time series data may be incomplete. On the other hand, clinical records may suffer from spelling errors, non-standard abbreviations, and different writing styles Qiao et al. (2019), making it difficult to use clinical annotations for prediction.

To address the above issues, time series data and clinical records can be combined for mortality prediction, and these prediction tasks can benefit from complementary relationships in multimodal data Su et al. (2020). Huang et al. (2021) demonstrated mathematically that the quality of the latent representation space directly determines the effectiveness of multimodal learning models. And with sufficient training data, the richer the variety of modalities, the more accurate the estimation of the representation space and thus the better the multimodal learning model. Recently, several researchers have modeled both kinds of data for mortality prediction. For example, Yang et al. (2021) proposed a multimodal deep neural network that considers both time series data and more clinical records. In addition, chronic and non-chronic patients are classified when dealing with clinical records. Khadanga et al. (2019) proposed a multimodal neural network model that extracts features from both time series data and clinical records and derives multimodal features. For better processing of textual data information, Deznabi et al. (2021) proposed a fine-tuned Bert model to process textual data. Experiments showed that the model combining the two could improve the prediction accuracy compared to using only time series data or only clinical records.

However, existing multimodal models still have some limitations. Most models utilize different deep neural networks to model the modal data and then output feature representations of the modal data by connecting each modal feature representation Song et al. (2019) at the feature level, which is then used to predict the final outcome. Since different modal data usually have different properties, resulting in inconsistent spatial and temporal dimensions, this poses an obstacle to capturing potential interrelationships in the low-level feature space. The inclusion of the fusion module after extracting the data features also increases the complexity Yu et al. (2019) of the network and makes training more difficult.

To address the above issues and challenges, we propose a multimodal fusion neural network model that combines time-series data as well as textual information from clinical annotations for predicting mortality. This task is defined as predicting whether a patient will die before discharge from the hospital based on data from the first 2 days. Our model uses a fine-tuned Bert model Devlin et al. (2018) to process the text part and generate textual feature vectors, and feeds the time series data into a Long Short Term Memory (LSTM) network and generates time series feature vectors. We improve a lightweight fusion module Su et al. (2020) by allowing the feature vectors of two equal feature blocks to enter the fusion module before cascading the prediction, thus capturing the correlation and important feature information between the underlying modal data. Correlations and important feature information between features, and finally, the obtained multimodal features are modeled through the GRU layer to capture the dependencies of multimodal features.

To sum up, the main contributions of this paper are as follows:

1) We introduce a fusion module to fuse multimodal features to improve the performance of the model. It can capture individual modal contextual information as well as the underlying feature correlation between different modalities.
2) We propose an end-to-end model, to accurately predict patient mortality risk based on multimodal fusion features in time series data and clinical records.
3) We conducted experiments on the mortality prediction task using a real dataset and showed that the proposed model outperformed all comparative methods.

The rest of this paper is organized as follows: Section 2 reviews the related work. Section 3 then presents the details of our model. In Section 4, we present the results of experiments performed on real datasets. Finally, we conclude our work in Section 5.

2 Related works

There are three main areas of related work: time series data mining and clinical record mining and multimodal learning, and we briefly discuss the latest work in this area.

2.1 Clinical time series data mining

In the area of mortality prediction for critically ill patients, most of the previous work has focused on using time series data of patients for prediction. Due to the long-term dependency problem among time series data, the most advanced results have been achieved since the PhysioNet Challenge Silva et al. (2012) in 2012 for long short term memory (LSTM) networks to process medical data. Harutyunyan et al. (2019) provided a preprocessing standard for the MIMIC III Johnson et al. (2016) database using LSTM models to handle in-hospital mortality prediction, loss of compensation prediction, length of stay prediction and phenotypic typing for four tasks. In this paper, recurrent neural networks are used for prediction. Other models also use different forms of recurrent neural networks to predict mortality outcomes. Liu and Chen (2019) proposed a novel selective recurrent neural network with randomly connected gating units (SRCGUs), which not only reduces the number of parameters and saves time, but also dynamically adjusts their importance weights Niu et al. (2022a) to select a more appropriate neural network for prediction.

2.2 Natural language processing in medical text

For clinical records, researchers have attempted to apply natural language processing (NLP) to medical prediction tasks. Grnarova et al. (2016) proposed a convolutional document embedding method based on the unstructured textual content of clinical records and evaluated the mortality prediction task against the clinical records of MIMIC III. The open set of biomedical word vectors/embeddings proposed by Zhang et al. (2019) (BioWordVec) that integrates biomedical domain knowledge to better capture the semantics of specialized terms and concepts, and also improves the quality of word embeddings Zhang et al. (2020). However, one drawback of word embedding is that it cannot address multiple meanings because each word is represented by only one vector. Peters et al. (2018) proposed a context-based pre-training model (ELMO) that dynamically adjusts word embedding according to the context, which can address the problem of multiple meanings Liu et al. (2019) of a word. However, this model uses LSTM to extract features instead of Transformer, and many studies have demonstrated that Transformer is far more capable of extracting features than LSTM.

Recently, BERT-based models have been used in the medical field due to their phenomenal success in natural language processing (NLP). Many studies are based on BERT models for clinical applications. Darabi et al. (2020) trained BERT models on clinical records and produced results with timestamps. The obtained patient embeddings verified the validity of unstructured text data. Lee et al. (2020) investigated how the recently introduced pre-trained language model BERT can be applied to biomedical corpora and proposed a pre-trained biomedical language representation model (BioBERT) for biomedical text mining. The analysis results show that the pre-training of biomedical corpus helps BERT to understand complex biomedical texts.

2.3 Multi-modal learning

All these models use only one data source when making mortality predictions. However, Khadanga et al. (2019) showed that the combination of time series and clinical record is useful for ICU mortality prediction. They used a convolutional neural network (CNN) on top of the pre-trained word embedding Zhang et al. (2019) of to obtain a representation of clinical record and a long short term memory network (LSTM) to embed the time-series portion of the data. The two representations were then connected for prediction. Deznabi et al. (2021) also showed the usefulness of combining time series data with clinical record information. They used the LSTM model for the time series portion of the data and a fine-tuned BioBERT Lee et al. (2020) for the clinical record. The multimodal segmentation attention module proposed by Su et al. (2020) is able to fuse blocks of features in each channel direction and capture correlations Ilievski and Feng (2017) between feature vectors. The fusion module is designed to be compatible with features of various spatial dimensions and sequence lengths for both cnn and rnn. thus, the fusion module can be easily added to the fusion features of any unimodal network and utilize existing pre-trained unimodal model weights. Therefore, We will use a representation of the clinical record in conjunction with the time series portion of the patient data and incorporate the fusion module to improve the performance of in-hospital mortality prediction. In the next section, we will define the notation and method architecture.

3 Methods

In this section, we first define some notations to describe both data and mortality prediction tasks. Then we will describe in detail the composition of the model, which consists of three parts: unimodal feature modeling, multimodal feature fusion, and mortality prediction module. Figure 1 shows an overview of the model.

3.1 Notations

Each patient has a number of clinical records during his hospitalization, and the data are in the format of time t corresponding to one clinical record. For a patient p with k clinical records, we denote C(p) = Ct_1, Ct_2, . . ., Ct_k, We also have time-series data collected during the patient's stay as X(p) = X_1, X_2, X_3, . . ., X_t.

3.2 The overview of model

As shown in Figure 1, the model consists of a unimodal feature modeling module, a fusion module, and a mortality prediction module. We model the clinical records using a BERT-based fine-tuning model, and then feed each clinical record of the patient into the textual part of our model separately, and then combine the resulting embeddings to obtain a final representation of the textual part of the data.

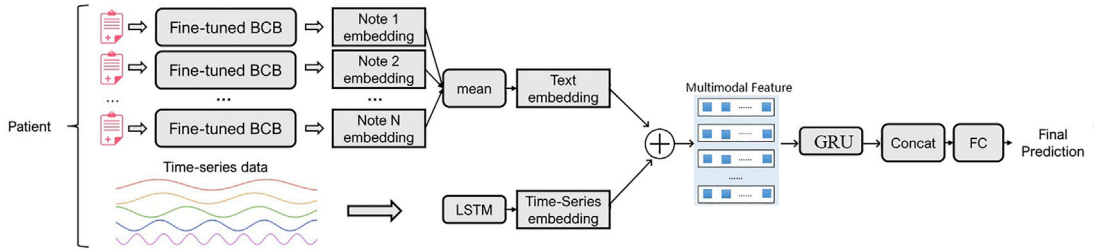

FIGURE 1
The overall architecture of the model.

We use a Long Short Term Memory (LSTM) network to model the pre-processed time series data and generate time series embeddings Shi et al. (2021). We then propose an improved fusion module for fusing features of single modalities, which partitions each modality into equal blocks of features on a channel and creates a joint representation for generating soft attention across feature blocks for each channel. The correlation information and important feature information between different modal data is captured through the attention mechanism. The unimodal features are passed through the fusion module to generate a multimodal feature matrix. Finally, the multimodal feature representation is modeled through the GRU layer to capture the dependencies between multimodal features. To more comprehensively represent the medical information of patients, we combine the modeled multimodal feature representation with the time series representation for prediction.

3.3 Single-modal feature modeling

The multimodal feature fusion module integrates different types of inputs. The feature extraction module of model includes time series data feature extraction and clinical record feature extraction.

3.3.1 Time-series feature modeling

For the time series part of the model, our preprocessing followed the work of Harutyunyan and Khadanga et al. We limited the collection of time series data to 48 h after admission, for which we used untimed sampling. If a variable had a missing value Tan et al. (2019) at the collection node, we used the preset value of the feature given by Harutyunyan et al. (2019) After preprocessing the time series data, we used the LSTM network. We input the whole time series data of the patient into the LSTM model. LSTM is an RNN for capturing long-term dependencies Yoon et al. (2018) in serial data. It takes a sequence of $\{x_t\}^T$ of length T as input and outputs a sequence of $\{h_t\}^T$ say hidden state vectors of length T using the following equation.

$$\mathbf{i}_t = \sigma(x_t\mathbf{W}^{(xi)} + \mathbf{h}_{t-1}\mathbf{W}^{(hi)}) \quad (1)$$

$$\mathbf{f}_t = \sigma(x_t\mathbf{W}^{(xf)} + \mathbf{h}_{t-1}\mathbf{W}^{(hf)}) \quad (2)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot c_{t-1} + \mathbf{i}_t \odot tanh(x_t\mathbf{W}^{(xc)} + \mathbf{h}_{t-1}\mathbf{W}^{(hc)} + \mathbf{b}^{(c)}) \quad (3)$$

$$\mathbf{o}_t = \sigma(x_t\mathbf{W}^{(xo)} + \mathbf{h}_{t-1}\mathbf{W}^{(ho)} + \mathbf{b}^{(o)}) \quad (4)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \sigma_h(\mathbf{c}_t) \quad (5)$$

In each step, the LSTM combines the current input x_t with the previous hidden state h_{t-1} to generate the current hidden state h_t; $h_t = \text{LSTM}(x_t, h_{t-1})$ for t = 1 to t = T.

3.3.2 Clinical record feature modeling

Modeling clinical records requires capturing the interactions between different words Huang et al. (2021), and the Word2vec Church (2017) approach was previously used to model text, which is no longer considered the most effective approach with the advent of BERT. Bert-based models started to be used to capture word interactions, and these models also outperformed traditional NLP models in the field of text modeling. So we use Bert-based models to fine-tune the clinical records to obtain embeddings Peng et al. (2019) of each clinical record of a patient. Specifically, we used the BioBert model proposed by Alsentzer et al. (2019) for the text part, which fine-tunes the clinical records in the MIMIC III dataset.

After initializing the text model to BioBert, we used the original BERT's sentence classifier, which uses a classification token (CLS) to classify relationships. We added a softmax layer to the output of the classification token (CLS) for each clinical record separately and further adjusted the parameters of the model to predict the in-hospital mortality rate of the patients. Finally, we took the average embedding value of all clinical records of the patient to obtain the overall textual embedding value of the patient. This is shown in Eq. 6.

$$\mathbf{H}_C = \frac{1}{N}\sum_{i=1}^{N} BCB(c_{t_i}) \quad (6)$$

3.4 Multi-modal feature fusion

In this section, we will detail the key part of the model, the multimodal feature fusion module. The flow chart of the algorithm is shown in Algorithm 1.

First, the time series and clinical record representations obtained in the modeling module are transformed into the same dimensions, and then their feature representations are

input to the intermediate fusion module. The feature matrix of a single modality is divided into equal channel feature blocks. The number of channels in each block is C. We denote the set of feature blocks of time series modality t as R_t, and we denote the set of feature blocks of clinical record modality c as R_c. That is, $R_c = (R_{c1}, R_{c2}, \ldots, R_{cn})$ and $R_t = (R_{t1}, R_{t2}, \ldots, R_{tn})$, and n is the number of characteristic blocks.

The intermediate fusion module performs a concatenation operation on the modal features of each channel block to learn the global contextual information inside each modality. Taking the modal feature R_c as an example, we perform a summation operation on the modal feature R_c level to generate a shared representation H_c, and then perform global average pooling in spatial dimension:

$$\mathbf{H}_c = \sum_{(R_{c1}, R_{c2} \ldots, R_{cn})} S(R_{c1}, \mathbf{R}_{c2}, \ldots, \mathbf{R}_{cn}) \tag{7}$$

H_c is the global representation of the clinical record, which summarizes the feature blocks of the clinical record. Then in the same way, we obtain the global representation of the time series data H_t. To obtain multimodal contextual information, we compute the sum of the elements of the global representation for each modality, which generates the multimodal global representation K_g:

$$\mathbf{K}_g = \text{elem}\left(\sum (H_c, H_t)\right) \tag{8}$$

The multimodal feature representation contains rich global contextual knowledge. Then, we apply a ReLU activation function on K_g to capture the dependencies between the multimodal features and map K_g to the joint representation Z, which helps the generalization of the model:

$$\mathbf{Z} = \mathbf{w}_z \mathbf{K}_g + \mathbf{b}_z \tag{9}$$

where w_z and b_z are the weights and biases. We generate the corresponding logits U_m^i by linearly transforming Z and obtain the block-level attention weights A_m^i by softmax activation. W_m^i and b_m^i are weights and bias of the corresponding fully connected layer.

$$\mathbf{U}_m^i = \mathbf{W}_m^i \mathbf{Z} + \mathbf{b}_m^i \tag{10}$$

$$\mathbf{A}_m^i = \frac{exp(U_m^i)}{\sum_k^M \sum_j^{|B_k|} exp(U_k^j)} \tag{11}$$

The soft attention in the original MSAF depends on the total number of feature blocks, and the extracted features are suppressed. In particular, the suppression effect is more pronounced in complex tasks. By analogy with the medical field, there is a close correlation between the change process of patients' health conditions and the clinical data generated during hospitalization, but soft attention suppresses this correlation and thus prevents accurate prediction. Therefore, we apply the soft attention mechanism at the modal data level to strengthen this correlation by weighting the average, while giving more weight to the more important parts of each block:

$$\mathbf{B}_m^i = [\lambda + (1 - \lambda) \times A_m^i] \odot B_m^i \tag{12}$$

$$\mathbf{F}_g = [\mathbf{B}_m^1, \mathbf{B}_m^2, \ldots, \mathbf{B}_m^{|B_m|}] \tag{13}$$

$\lambda \in (0,1)$ can be interpreted as a hyperparameter with control weights. We obtain an optimized feature block B_m^i using attention signals A_m^i and λ. Finally, the feature blocks belonging to modality m are merged by channel-wise concatenation to produce F_g.

```
Initialize clinical record, time series data, and feature blocks of clinical records and time series Rc, Rt
for i = 1; i < n; i + + do
Rc = (Rc1, Rc2, ..., Rcn) # clinical record embedding
Rt = (Rt1, Rt2, ..., Rtn) # time series embedding
The multimodal fusion module associated with Rc and Rt
The attention mechanisms learns to emphasize the important feature blocks
end for
Process the multimodal fusion matrix Fg
Fusion state information, Fg and Rt
Calculate the final risk score for death11: Perform prediction for test set and observe prediction performance
```

Algorithm 1: The multimodal fusion model.

3.5 Mortality prediction

To be able to further capture the dependencies between modalities, we input the multimodal fusion representation F_g to the GRU layer to obtain the feature representation F_{0g}. However, data offset may occur when data fusion is performed on different modal data. To compensate for this lost information and to represent patient information more comprehensively, the low-level feature representation extracted from the time series data is used for prediction along with the multimodal fusion features.

To fuse the multimodal features, we first connect the features p = concat (R_t, F_{0g}) to obtain a final multimodal fusion matrix for the patient. This matrix is then passed through a fully connected layer to predict the patient's mortality using sigmoid as the activation function. Finally the predicted results are output.

4 Experiment

In this section, we conduct experiments on the real-world medical dataset MIMIC III to evaluate the performance of model. Compared with the baseline prediction model, the model achieves better results under different evaluation strategies.

4.1 Data description

4.1.1 Dataset

We use a free and open, publicly available resource for intensive care unit research databases, MIMIC-III published on PhysioNet. The dataset includes 46,520 patients and contains identified comprehensive clinical monitoring device waveform data and rich clinical text records from the Intensive Care Unit (ICU) at Beth Israel Deaconess Medical

Center between 2001 and 2012. The content of MIMIC-III includes vital signs, laboratory and radiology reports, treatment information, and more.

4.1.2 Pre-processing

To prepare our in-hospital mortality dataset, we start with the root cohort and further exclude all ICU stays with unknown length of stay, length of stay less than 48 h, or no observed length of stay within 48 h. This yields a final training set and test set of 17,903 and 3,236 ICU stays, respectively. We determined in-hospital mortality by comparing the date of patient death with the time of admission and discharge. The resulting mortality rate was 13.23 (2,797 of 21,139 ICU stays). Because time-series data need to be combined with clinical records, clinical records collected within 48 h of the patient's admission are required. After these steps, our dataset consisted of 11,579 records in the training set, 2,570 records in the validation set, and 2,573 records in the test set.

For the time series part of the model, we first restricted the time series data to the first 48 h of the patient's hospitalization, and then we resampled the time series data to a 1-h interval. We performed forward imputation on the missing values if they were generated. If no previous values were recorded, we used the pre-set values of the features given in Harutyunyan et al. (2019). After preprocessing the time series data, we used the LSTM network. We input the entire time series data of the patient into the LSTM model. For the clinical record part of the model, we use a fine-tuned BERT-based model for modeling clinical record, then we feed each clinical record of a patient separately to the text part of our model, and then combine the resulting embeddings to get the final representation of text part of the data.

4.2 Experiment setup

In this section, we first describe the state-of-the-art methods used for mortality prediction as a baseline, and then outline the metrics used for evaluation. Finally, we detail the implementation details. Baseline models. We compare our method with the baseline method for mortality prediction. The baseline model is as follows:

1) LSTM: Time series data processing with LSTM.
2) Bert: Clinical record processing with Bert.
3) CNN + LSTM: Clinical record processing with CNN, time series data processing with LSTM.
4) CNN + GRU: Clinical record processing with CNN, time series data processing with GRU.
5) Bert + LSTM: Clinical record processing with Bert, time series data processing with LSTM.
6) BioBert + LSTM: Clinical record processing with BioBert, time series data processing with LSTM.

These unimodal baseline models were selected for LSTM, and Bert processed the time series and text separately. These multi-modal baseline models are all combined for mortality prediction by replacing the unimodal data processing methods, and all use the commonly concat method for fusion and prediction before entering the fully connected layer.

TABLE 1 Performance comparison on MIMIC-III data.

Model type	Model	AUCROC	AUCPR
Single-modal	LSTM	0.8333	0.4418
	GRU	0.8151	0.4931
	CNN	0.8374	0.4862
	Bert	0.8386	0.465
	BioBert	0.8553	0.5031
Multi-modal	CNN + LSTM	0.8486	0.5274
	CNN + GRU	0.8358	0.5136
	Bert + LSTM	0.8458	0.5215
	BioBert + LSTM	0.8724	0.5404
	BioBert + LSTM + Fusion	0.8835	0.5632

4.2.1 Evaluation strategies

Since in-hospital mortality prediction is a binary classification task with unbalanced categories (only about 10 of patients in this dataset suffered death), the area under subject working characteristics (AUC) was used to evaluate our model. We also report the area under the precision-recall curve (AUC-PR) metric because it can be more informative when dealing with highly skewed datasets.

4.2.2 Implementation details

We implemented all baselines and models using pytorch. For the training model, we used Adam with a batch size of 5, an initial learning rate of 2e-5 for training, and a weight decay of 0.01. Training was terminated after 10 epochs because the evaluation value no longer improved as the number of epochs increased. We randomly divided the dataset into a training set, a test set, and a validation set in the ratio of 7:1.5:1.5.

4.3 Performance evaluation

4.3.1 Overall performance

Table 1 demonstrates the performance of the proposed model and the baseline of predicted mortality in the MIMIC-III dataset. The results show that the model outperforms all baselines. We note that in the real public dataset, the AUCPR of LSTM is lower than other models and the AUCROC values are slightly lower, mainly because LSTM uses a model with padding for missing values, so it does not capture the dependencies and correlations between different time series. CNN, Bert, BioBert and our model can capture the dependencies between medical texts and take into account all patient visits.

TABLE 2 Performance of each model in the one clinical record.

Model	AUCROC	AUCPR
CNN + LSTM	0.8124	0.5026
CNN + GRU	0.8058	0.4847
Bert + LSTM	0.8186	0.5034
BioBert + LSTM	0.8327	0.5138
BioBert + LSTM + Fusion	0.8432	0.5256

TABLE 3 Performance of each model in the 4 clinical record.

Model	AUCROC	AUCPR
CNN + LSTM	0.8236	0.5095
CNN + GRU	0.8135	0.493
Bert + LSTM	0.8227	0.5089
BioBert + LSTM	0.8485	0.5214
BioBert + LSTM + Fusion	0.8637	0.5362

TABLE 4 Performance of each model in the 7 clinical record.

Model	AUCROC	AUCPR
CNN + LSTM	0.8367	0.5168
CNN + GRU	0.8287	0.5057
Bert + LSTM	0.8365	0.5136
BioBert + LSTM	0.8683	0.5367
BioBert + LSTM + Fusion	0.8752	0.5485

TABLE 5 Performance of each model in the 10 clinical record.

Model	AUCROC	AUCPR
CNN + LSTM	0.8486	0.5274
CNN + GRU	0.8358	0.5136
Bert + LSTM	0.8454	0.5215
BioBert + LSTM	0.8725	0.5404
BioBert + LSTM + Fusion	0.8835	0.5632

In the multi-modal fusion method, early fusion may contain a large number of redundant input vectors, while late fusion cannot capture the correlation between the underlying features of different modes. As can be seen from Table 1, compared with CNN + LSTM, CNN + LSTM and Bert + LSTM models that use concat to concatenate the data of the two modes before the decision, the middle fusion module we use makes the model perform better. Compared to the best-performing BioBert + LSTM, AUCROC improved by 1.1%, AUCPR improved by 2.28%. This is because the heterogeneity of multi-modal data makes the parameter space too complex to be determined by simple heterogeneous splicing. Our model can model different types of input and use different types of data to learn important characteristics, and the fusion module used can capture the correlation between different modes to help prediction. Compared with the simple splicing of multi-modal data for prediction (CNN + LSTM), our fusion method significantly improves the prediction performance of the model.

In order to prove the effectiveness of the performance improvement brought by multi-modal feature fusion, we also conducted experiments based on the original single-modal feature used by each model. It can be seen from Table 1 that the model performance is improved after the data of other modes are introduced. For example, in the BioBert + LSTM model, the ARUROC, AUCPR were improved by 2.8%, 3.7% respectively, after the clinical records were combined with the time-series data for prediction. This proves that information complementarity between multi-modal features can indeed enable models to better understand patient health and achieve higher predictive performance. This is also verified by the model based on GRU. The performance of the model proposed by us is significantly improved compared with that of GRU based on a single mode.

4.3.2 Robustness for number of clinical-record

Tables 2–5 show the prediction results of mortality for four different numbers of clinical records. Compared with the baseline model, our proposed model performed best in terms of AUCROC, AUCPR performance indicators. The multimodal model always works better than the unimodal model in different numbers of clinical records.

To assess the value of using multiple clinical records for each patient, we tried including a different number of clinical records for each patient in the different model. Figure 2 and Figure 3 shows the experimental results of the model when predicting different numbers of clinical records. It can be observed that the performance of the model in terms of both AUCROC and AUCPR rises as the number of clinical records increases, with the best performance being achieved when all clinical records are used.

4.3.3 Data ablation study

We conducted data ablation experiments based on the proposed model to examine the contribution of various data to the prediction task. First, as seen in Table 6, it is higher than a single model in all two evaluation metrics, suggesting that combining multiple models yields richer information about

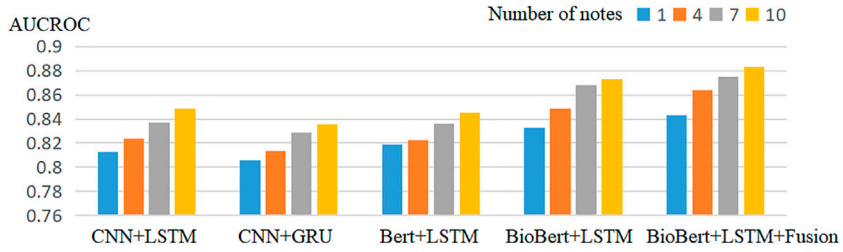

FIGURE 2
Models' robustness comparison of AUCROC.

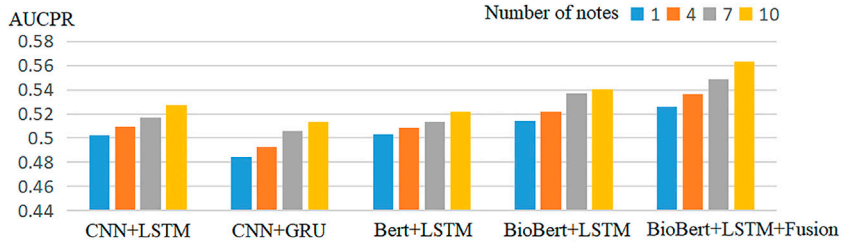

FIGURE 3
Models' robustness comparison of AUCPR.

TABLE 6 Data Ablation Results (T for time-series, C for clinical records).

Model	AUCROC	AUCPR
BioBert + LSTM + Fusion-T	0.8257	0.5048
BioBert + LSTM + Fusion-C	0.8465	0.5283
BioBert + LSTM + Fusion	0.8835	0.5632

the patient and allows better monitoring of the patient's condition. It is logical that the combination of different kinds of data can be complementary information. However, the predictive performance of the pure time series model is low because time series data have many missing values due to irregular sampling. This can have some impact on model learning. Clinical records are highly correlated with the patient's physical condition, and models based on clinical records perform better than time series.

5 Conclusion

Clinical records written by healthcare professionals include important information about the patient's history and current status in the hospital, which can be used to significantly improve mortality prediction. In this work, this paper utilizes time series data and clinical records for mortality prediction to improve the monitoring of patients' clinical health status. To overcome the drawbacks of unimodal data, We propose a multimodal fusion model that integrates multimodal data into the same architecture by improving the fusion module. Our proposed improved module partitions the features of each modal channel into feature blocks of equal size and then learns to emphasize the important feature blocks by generating attention values. Subsequently, the augmented feature blocks are reconnected for each modality to obtain an optimized feature space for understanding multimodal contexts. It captures the global information of a single modality and the relationship between multiple modalities, which helps to improve the predictive performance of the model. The performance of the model is validated on a real electronic medical record dataset, and the results show that our proposed model has good performance.

Semantic conflict repetition in current deep learning multimodal fusion models. The problems such as noise are still unresolved. Although attention mechanisms can solve part of the problem. But they work implicitly. Not actively controlled. Studying an active control can better combine logical reasoning and deep learning. How to solve the problem of unobstructed interoperability of information across modalities? How to solve the model generalization ability? These questions are the shortcomings of current multimodal models.

In the future, we want to solve the semantic conflict repetition, noise problem in deep learning multimodal fusion models.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

Conceptualization, KN; methodology KN, KZ, XP, YP, and NX; data curation and software KN; formal analysis, KN, KZ, XP, YP, and NX; writing—original draft preparation KN, KZ, XP, YP, and NX; writing—review and editing, KN, KZ, XP, YP, and NX; All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported in part by the Beijing Information Science and Technology University Qin Xin Talents Cultivation Program under grant no. QXTCPC202112, Research Level Improvement Project under grant no. 2020KYNH214, and 2025 Key Technological Innovation Program of Ningbo City under grant no. 2022Z080.

Conflict of interest

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

Publisher's note

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

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., et al. (2019). *Publicly available clinical bert embeddings. arXiv preprint arXiv: 1904.03323*.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* 8, 1–12. doi:10.1038/s41598-018-24271-9

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). *Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078*.

Church, K. W. (2017). *Word2vec.* Nat. Lang. Eng. 23, 155–162. doi:10.1017/s1351324916000334

Darabi, S., Kachuee, M., Fazeli, S., and Sarrafzadeh, M. (2020). Taper: Time-aware patient ehr representation. *IEEE J. Biomed. health Inf.* 24, 3268–3275. doi:10.1109/JBHI.2020.2984931

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805*.

Deznabi, I., Iyyer, M., and Fiterau, M. (2021). "Predicting in-hospital mortality by combining clinical notes with time-series data," in *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 4026–4031.

Grnarova, P., Schmidt, F., Hyland, S. L., and Eickhoff, C. (2016). *Neural document embeddings for intensive care patient mortality prediction. arXiv preprint arXiv: 1612.00467*.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Sci. data* 6, 96–18. doi:10.1038/s41597-019-0103-9

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

Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., and Huang, L. (2021). What makes multi-modal learning better than single (provably). *Adv. Neural Inf. Process. Syst.* 34, 10944–10956.

Ilievski, I., and Feng, J. (2017). Multimodal learning and reasoning for visual question answering. *Adv. neural Inf. Process. Syst.* 30.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., et al. (2016). Mimic-iii, a freely accessible critical care database. *Sci. data* 3, 160035–160039. doi:10.1038/sdata.2016.35

Khadanga, S., Aggarwal, K., Joty, S., and Srivastava, J. (2019). *Using clinical notes with time series data for icu management. arXiv preprint arXiv:1909.09702*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2020). Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240. doi:10.1093/bioinformatics/btz682

Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzel, R. (2015). *Learning to diagnose with lstm recurrent neural networks. arXiv preprint arXiv:1511.03677*.

Liu, J., and Chen, S. (2019). "Non-stationary multivariate time series prediction with selective recurrent neural networks," in *Pacific rim international conference on artificial intelligence* (Springer), 636–649.

Liu, N., Lu, P., Zhang, W., and Wang, J. (2019). "Knowledge-aware deep dual networks for text-based mortality prediction," in *2019 IEEE 35th international conference on data engineering (ICDE)* (IEEE), 1406–1417.

Niu, K., Guo, Z., Peng, X., and Pei, S. (2022a). P-resunet: Segmentation of brain tissue with purified residual unet. *Comput. Biol. Med.* 151, 106294. doi:10.1016/j.compbiomed.2022.106294

Niu, K., Lu, Y., Peng, X., and Zeng, J. (2022b). Fusion of sequential visits and medical ontology for mortality prediction. *J. Biomed. Inf.* 127, 104012. doi:10.1016/j.jbi.2022.104012

Pei, S., Niu, K., Peng, X., and Zeng, J. (2021). "Readmission prediction with knowledge graph attention and rnn-based ordinary differential equations," in *International conference on knowledge science, engineering and management* (Springer), 559–570.

Peng, X., Long, G., Shen, T., Wang, S., Jiang, J., and Blumenstein, M. (2019). Temporal self-attention network for medical concept embedding. In *2019 IEEE international conference on data mining (ICDM)* (IEEE), 498–507.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). "Deep contextualized word representations," in *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies* (New Orleans, Louisiana: Association for Computational Linguistics), 1, 2227–2237. (Long Papers). doi:10.18653/v1/N18-1202

Qiao, Z., Wu, X., Ge, S., and Fan, W. (2019). Mnn: Multimodal attentional neural networks for diagnosis prediction. *Extraction* 1, A1.

Shi, Z., Wang, S., Yue, L., Pang, L., Zuo, X., Zuo, W., et al. (2021). Deep dynamic imputation of clinical time series for mortality prediction. *Inf. Sci.* 579, 607–622. doi:10.1016/j.ins.2021.08.016

Silva, I., Moody, G., Scott, D. J., Celi, L. A., and Mark, R. G. (2012). "Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012," in *2012 computing in cardiology* (IEEE), 245–248.

Song, L., Cheong, C. W., Yin, K., Cheung, W. K., Fung, B. C., and Poon, J. (2019). Medical concept embedding with multiple ontological representations. *IJCAI* 19, 4613–4619.

Su, L., Hu, C., Li, G., and Cao, D. (2020). *Msaf: Multimodal split attention fusion. arXiv preprint arXiv:2012.07175*.

Sun, C., Hong, S., Song, M., and Li, H. (2020). *A review of deep learning methods for irregularly sampled medical time series data. arXiv preprint arXiv:2010.12493*.

Sun, C., Hong, S., Song, M., Zhou, Y., Sun, Y., Cai, D., et al. (2021). *Te-esn: Time encoding echo state network for prediction based on irregularly sampled time series data*. *arXiv preprint arXiv:2105.00412*.

Tan, Q., Ma, A. J., Ye, M., Yang, B., Deng, H., Wong, V. W.-S., et al. (2019). "Ua-crnn: Uncertainty-aware convolutional recurrent neural network for mortality risk prediction," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 109–118.

Yang, H., Kuang, L., and Xia, F. (2021). Multimodal temporal-clinical note network for mortality prediction. *J. Biomed. Semant.* 12, 3–14. doi:10.1186/s13326-021-00235-3

Yoon, J., Zame, W. R., and van der Schaar, M. (2018). Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Trans. Biomed. Eng.* 66, 1477–1490. doi:10.1109/TBME.2018.2874712

Yu, R., Zheng, Y., Zhang, R., Jiang, Y., and Poon, C. C. (2019). Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients. *IEEE J. Biomed. health Inf.* 24, 486–492. doi:10.1109/JBHI.2019.2916667

Zhang, X., Dou, D., and Wu, J. (2020). Learning conceptual-contextual embeddings for medical text. *Proc. AAAI Conf. Artif. Intell.* 34, 9579–9586. doi:10.1609/aaai.v34i05.6504

Zhang, Y., Chen, Q., Yang, Z., Lin, H., and Lu, Z. (2019). Biowordvec, improving biomedical word embeddings with subword information and mesh. *Sci. data* 6, 52–59. doi:10.1038/s41597-019-0055-0

Zhu, Y., Fan, X., Wu, J., Liu, X., Shi, J., and Wang, C. (2018). "Predicting icu mortality by supervised bidirectional lstm networks," in *AIH@ ijcai*.