PM^{2}F^{2}N: Patient Multi-view Multi-modal Feature Fusion Networks for Clinical Outcome Prediction

Ying Zhang[1,2], **Baohang Zhou**[1,2], **Kehui Song**[1,2], **Xuhui Sui**[1,2], **Guoqing Zhao**[3], **Ning Jiang**[3], **Xiaojie Yuan**[1,2,*]
[1] College of Computer Science, Nankai University, Tianjin, China
[2] Tianjin Key Laboratory of Network and Data Security Technology, Tianjin, China
[3] Mashang Consumer Finanace Co, Ltd
{zhangying,zhoubaohang,songkehui,suixuhui}@dbis.nankai.edu.cn
{guoqing.zhao02,ning.jiang02}@msxf.com, yuanxj@nankai.edu.cn

Abstract

Clinical outcome prediction is critical to the condition prediction of patients and management of hospital capacities. There are two kinds of medical data, including time series signals recorded by various devices and clinical notes in electronic health records (EHR), which are used for two common prediction targets: mortality and length of stay. Traditional methods focused on utilizing time series data but ignored clinical notes. With the development of deep learning, natural language processing (NLP) and multi-modal learning methods are exploited to jointly model the time series and clinical notes with different modals. However, the existing methods failed to fuse the multi-modal features of patients from different views. Therefore, we propose the patient multi-view multi-modal feature fusion networks for clinical outcome prediction. Firstly, from patient inner view, we propose to utilize the co-attention module to enhance the fine-grained feature interaction between time series and clinical notes from each patient. Secondly, the patient outer view is the correlation between patients, which can be reflected by the structural knowledge in clinical notes. We exploit the structural information extracted from clinical notes to construct the patient correlation graph, and fuse patients' multi-modal features by graph neural networks (GNN). The experimental results on MIMIC-III benchmark demonstrate the superiority of our method.

1 Introduction

With the development of information technology in medical area, an increasing number of devices are used for monitoring patients. And a large number of data is stored as electronic health records (EHR), which contain numerical results of physical examination in time series and clinical notes in text for patients' relevant information. The multi-type data can be utilized to predict the condition of patients, which can help in managing the resources in hospitals. Most previous works focused on modeling the problem using the time series data recorded by medical instruments (Ghassemi et al., 2015; Xu et al., 2018). However, the time series data gathered from medical devices only reflects physical status of patients in a one-sided way. Medical professionals need to utilize their expertise to analyze patients' data and make the diagnosis. The important analyses to patients' data are recorded in EHR as clinical notes.

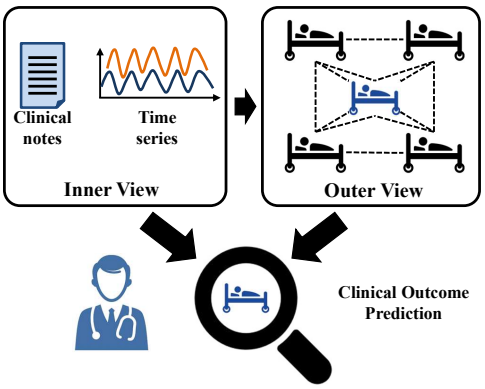

Figure 1: There are two views to analyze the observed patient. The inner view is to focus on the medical data of the observed patient. And the outer view is to exploit the medical correlation between patients for the observed one.

More recent work applied natural language processing (NLP) methods to take full advantage of medical information in clinical notes for prediction tasks (Boag et al., 2018; Lee et al., 2020; van Aken et al., 2021). They utilized pre-trained language models to extract text features of clinical notes and fed them into recurrent or convolution neural networks for clinical outcome prediction. Further more, considering to combine time series data with clinical notes for improved prediction on clinical outcome, some recent work proposed multi-modal learning methods to jointly model the two kinds of data (Khadanga et al., 2019; Bardak

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1985 - 1994

and Tan, 2021a; Deznabi et al., 2021). They used sequence models to extract features of time series and clinical notes respectively, and concatenated them for predicting clinical outcome. However, the existing methods do not consider that the features of time series data and clinical notes fuse different parts of each other with various weights. Besides, the multi-modal features of a single patient is not sufficient for clinical outcome prediction, and the medical correlation between patients has not been exploited for this task.

To overcome the above disadvantages of the existing methods, we propose the **p**atient **m**ulti-view **m**ulti-modal **f**eature **f**usion **n**etworks (PM^{2}F^{2}N) [1]for clinical outcome prediction. The model enhances the multi-modal feature fusion ability in two views. Firstly, from the patient inner view, we use the co-attention (Lu et al., 2016) module to enhance the fine-grained feature interaction between time series data and clinical notes. The co-attention module allows our model to attend to important parts of time series data as well as correlated medical information of clinical notes. Secondly, from the patient outer view, other patients' information is useful to predict the status of the observed one. We construct the patient correlation graph based on the structural medical information extracted from clinical notes, and fuse patients' multi-modal features by graph neural networks (GNN). With the multi-modal feature fusion from different views, our model can gain better generalization ability to predict clinical outcome. The contributions of this manuscript can be summarized as follows:

1. We analyze the disadvantages of the existing methods for clinical outcome prediction. To improve the ability to fuse the multi-modal features from different views, we propose the patient multi-view multi-modal feature fusion networks.

2. From the patient inner view, we extend the co-attention module to enhance the fine-grained feature fusion between time series data and clinical notes. Besides, from the outer view, we exploit the patient correlation graph to aggregate the multi-modal features between patients.

3. We evaluate the effectiveness of the proposed model on MIMIC-III benchmark. The experimental results demonstrate that our model outperforms the baseline approaches. And the further analysis to multi-modal features also shows the superiority of our model.

[1]When ready, the code will be published at https://github.com/ZovanZhou/PM2F2N.

2 Related Work

2.1 Time Series for Clinical Outcome Prediction

The earlier works on mortality prediction designed hand-crafted features and used traditional machine learning methods like logistic regression with various severity scores (Vincent et al., 1996). With the progress of the deep learning, the sequence models, such as: long-short term memory networks (LSTM) (Hochreiter and Schmidhuber, 1997) and gated recurrent units (GRU) (Cho et al., 2014), are utilized to tackle with time series data for clinical outcome prediction. Besides, some researchers exploited irregular sampling of the data over time in their prediction models (Zhang et al., 2021b). Furthermore, the self-attention mechanism is also used to capture the dependencies within the time series data for clinical outcome prediction (Song et al., 2018; Ma et al., 2020).

2.2 Clinical Notes for Clinical Outcome Prediction

Considering the time series data is limited in explicit medical information, some works focused on using clinical notes for outcome prediction. They utilized the pre-trained word embeddings (Zhang et al., 2019) as text features of clinical notes, and fed them into recurrent neural networks (RNN) or convolution neural networks (CNN) to extract hidden features for predicting results (Ghorbani et al., 2020). Besides, the external medical knowledge is useful to predict the physical status of patients. The clinical outcome pre-training method was proposed to integrate knowledge from multiple patient outcomes (van Aken et al., 2021).

2.3 Multi-modal Learning for Clinical Outcome Prediction

With the development of multi-modal learning, the above methods are limited in fusing various sources of available data when predicting medical outcomes. And the data of every modal can be enhanced with each other in multi-modal learning. The multi-modal learning for time series data and clinical notes showed the effectiveness on clinical outcome prediction (Khadanga et al., 2019;

Deznabi et al., 2021). They utilized RNN to extract hidden representations of time series data and CNN to acquire ones of clinical notes. The two hidden features were then concatenated and fed into feed-forward neural networks (FFNN) for predicting results. Besides, to model the robust representations of patients' multi-type data in EHR, the supervised deep patient representation learning framework was proposed for clinical outcome prediction (Zhang et al., 2021a). To make use of sparse medical information in clinical notes, the named entity recognition (NER) model was utilized to extract entities in texts and the representations of them were introduced into multi-modal learning model for making predictions (Bardak and Tan, 2021a). The existing methods do not consider to evaluate the status of patients from different aspects. Therefore, we propose PM^{2}F^{2}N model to fuse multi-modal features from various views for clinical outcome prediction with better generalization ability.

3 Model

We introduce the notations about clinical outcome prediction before getting into the details of the proposed model. The training set with N_s samples is denoted as $\{(\mathbf{X}^{(i)}, \mathbf{C}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N_s}$, where $\mathbf{X}^{(i)}$ and $\mathbf{C}^{(i)}$ are i-th patient's time series data and clinical note respectively, and $\mathbf{y}^{(i)}$ is the task-defined label. Given a time series with N_t time steps and N_v observed variables, the patient's vital signals can be formulated as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_t}\} \in \mathbb{R}^{N_t \times N_v}$. We denote the clinical note with N_w words as $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{N_w}\}$. After pre-processing the multi-modal data, we feed them into the proposed model.

The PM^{2}F^{2}N model is shown in Figure 2. For time series data, we utilize the bidirectional GRU to extract hidden representations. Considering to acquire the multi-grained features of clinical notes, we apply the NER model to extract medical entities as local features and use the term frequency-inverse document frequency (TF-IDF) method to extract global features of clinical notes. To combine the entity representations of clinical notes with hidden representations of time series data in a fine-grained way, we exploit the co-attention module to acquire the multi-modal fusion features with various attention weights. Based on the medical information of different patients, we build the patient correlation graph and exploit it to aggregate multi-modal features of various neighbors via GNN. The concatenation of global features of clinical notes, last hidden features of time series data and aggregation multi-modal features is fed into FFNN for outcome prediction.

3.1 Multi-modal Feature Extraction

Given the multi-modal data as input, we need to pre-processe them and map them into the dense representations for deep neural networks as shown in Figure 2. We denote time series data which has N_t time steps and N_v observed variables as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_t}\} \in \mathbb{R}^{N_t \times N_v}$. With the impressive performances of RNN, GRU and LSTM are utilized to extract the hidden representations of sequence data. Considering to capture the context information in forward and backward directions, we utilize the bidirectional GRU (BiGRU) to acquire the hidden features of time series data $\mathbf{X}$ (Bardak and Tan, 2021a). The extraction process is simplified as $\text{BiGRU}(\mathbf{X}; \theta_1) = \mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{N_t}\} \in \mathbb{R}^{N_t \times N_h}$ where N_h is the dimension number of hidden feature vector and θ_1 is the trainable parameters of BiGRU.

The clinical notes contain detailed information about patients and medical knowledge implicated in inference of doctors. Considering that clinical notes may contain redundant information, we need to extract the representative features to highlight critical patient information. Therefore, we propose to extract the multi-grained features of clinical notes. To make full use of unstructured clinical note $\mathbf{C}$, we utilize the TF-IDF to extract the global feature vector. With the advantage of TF-IDF, the important tokens in clinical notes can be represented by the global feature vector. However, the dimension of global feature vector is too high to represent the patient with a tight way and fit all into the memory. We then apply principal component analysis (PCA) to reduce the dimension of global feature vector and the dimension-reduced global feature of clinical note is defined as $\mathbf{C}_g \in \mathbb{R}^{N_g}$ where N_g is the dimension number of global feature vector.

Besides, there are various medical information defined as entities including: diseases, drugs, dosage and so on, in clinical notes (Kormilitzin et al., 2021). The structural medical knowledge, known as entities, is the most important information to represent the status of patients. The raw clinical notes contain lots of redundant free-text

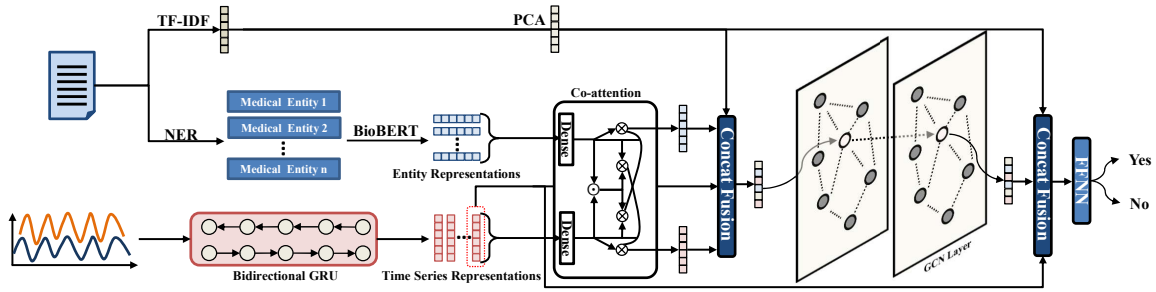

Figure 2: The patient multi-view multi-modal feature fusion networks (PM^2F^2N) for clinical outcome prediction.

and sparse medical knowledge. To exploit the medical entities as local features, we utilize the Med7 model (Kormilitzin et al., 2021) to extract the medical entities from the clinical note $\mathbf{C}$. The Med7 is the model trained with fully annotated EHR dataset and can recognize seven categories of medical entities. We feed the clinical note $\mathbf{C}$ into the Med7 model and acquire the medical entity set $\mathbf{C}_e = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{N_e}\}$ where N_e is the number of entities in $\mathbf{C}$. The medical entity $\mathbf{e}$ is the text span with discrete words in clinical note. To map the discrete words of medical entities into the dense representations, we use the pre-trained language model BioBERT (Lee et al., 2020) as feature extractor. Each medical entity span is fed into BioBERT and the distributed representations of medical entities in clinical note $\mathbf{C}$ are denoted as $\mathbf{C}_b = \{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_{N_e}\} \in \mathbb{R}^{N_e \times N_b}$ where N_b is the dimension number of medical entity features.

3.2 Multi-modal Feature Fusion with Multi-view

There are different views to evaluate the physical status of patients. From the inner view of the observed patient, the doctor analyzes the multi-modal data to make diagnosis. Based on accumulated clinical experience, the doctor can also dig into the correlation between patients to provide the diagnostic result to the observed patient. From the outer view, therefore, the multi-modal data of other patients, which are correlated with the observed one in medical knowledge, is also beneficial to the diagnostic results. With the target to enhance the representation ability of our model, we propose to improve the multi-modal feature fusion strategy in two different views.

3.2.1 Feature Fusion with Inner View

The time series data $\mathbf{X}$ contains various physiological signals changing over time. And the medical entity set $\mathbf{C}_e$ includes different medical knowledge representing the condition of patient. Some of the medical information in clinical note is relevant to the physical signals at certain times. For example, the observed patient was treated with certain drugs and the physical signals would change in some aspect. To capture the fine-grained correlation between multi-modal data, we propose to exploit co-attention module for fusing the multi-modal features. Although the co-attention achieved significant success in visual question answering area (Lu et al., 2016), this is the first time to expand it to the medical multi-modal data mining area. Given the extracted features $\mathbf{H}$ of time series data $\mathbf{X}$ and that $\mathbf{C}_b$ of clinical note $\mathbf{C}$, we unify the dimension number of both as: $\mathbf{H}' = \mathbf{H}\mathbf{W}_a$ and $\mathbf{C}'_b = \mathbf{C}_b\mathbf{W}_b$ where $\mathbf{W}_a \in \mathbb{R}^{N_h \times N_d}$ and $\mathbf{W}_b \in \mathbb{R}^{N_b \times N_d}$ are trainable weights. To calculate the correlation degree between time series and medical entities, the shared feature space $\mathbf{S} \in \mathbb{R}^{N_t \times N_e}$ is defined as $\mathbf{S} = \tanh\left(\mathbf{H}'\mathbf{W}_s\mathbf{C}'^T_b\right)$ where $\mathbf{W}_s \in \mathbb{R}^{N_d \times N_d}$ is the trainable weight in the module. The shared feature $\mathbf{S}$ is used to calculate the correlation between time series and medical entity features. Firstly, the multi-modal fusion features of time series data and clinical note are calculated as: $\mathbf{H}^f = \tanh\left(\mathbf{W}_o\mathbf{H}'^T + \mathbf{W}_e\mathbf{C}'^T_b\mathbf{S}^T\right)$ and $\mathbf{C}^f_b = \tanh\left(\mathbf{W}_e\mathbf{C}'^T_b + \mathbf{W}_o\mathbf{H}'^T\mathbf{S}\right)$ where $\mathbf{W}_e \in \mathbb{R}^{N_k \times N_d}$ and $\mathbf{W}_o \in \mathbb{R}^{N_k \times N_d}$ are the weight parameters. Secondly, the attention probabilities of each medical entity $\mathbf{b}_i$ and physiological hidden feature $\mathbf{h}_i$ of every time step are calculated as: $\mathbf{a}^o = \text{softmax}\left(\mathbf{W}^T_{ho}\mathbf{H}^f\right)$ and $\mathbf{a}^e = \text{softmax}\left(\mathbf{W}^T_{he}\mathbf{C}^f_b\right)$ where $\mathbf{W}_{ho} \in \mathbb{R}^{N_k}$ and $\mathbf{W}_{he} \in \mathbb{R}^{N_k}$ are trainable weights for attention prediction. After acquiring the attention scores of multi-modal features, the fine-grained fusion vectors are calculated as the weighted sum of the time series and medical entity features, i.e., $\hat{\mathbf{H}} = \sum_{i=1}^{N_t} \mathrm{a}^o_i\mathbf{h}_i$ and $\hat{\mathbf{C}}_b = \sum_{i=1}^{N_e} \mathrm{a}^e_i\mathbf{e}_i$. The multi-modal fusion features of time series data $\mathbf{X}$ and clinical note $\mathbf{C}$ from the inner view are de-

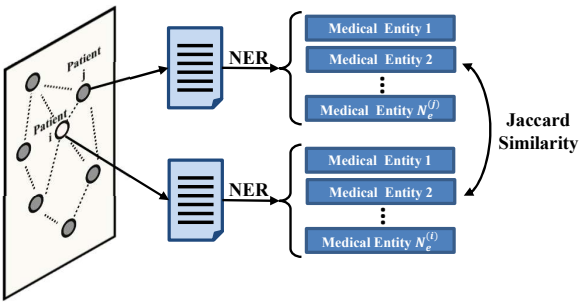

Figure 3: The detailed construction of the patient correlation graph. The degree of correlation between patients is defined as the jaccard similarity between medical entity sets in each patient's clinical note.

Item	Train	Development	Test
# samples	16760	2394	4790
T.S. length	24	24	24
# avg words	8533	8680	8436
# avg entities	38	40	39
Class Distribution (No%:Yes%)			
In-hospital Mortality	In-ICU Mortality	LOS >3	LOS >7
89.5%:10.5%	93%:7%	56.8%:43.2%	92.1%:7.9%

Table 1: The statistical information of the MIMIC-III dataset extracted by MIMIC-Extract. "T.S." is short for "time-series".

noted as $\hat{\mathbf{H}}$ and $\hat{\mathbf{C}}_b$.

3.2.2 Feature Fusion with Outer View

To analyze the physical status of the observed patient, the relevant patients' information is worthing referring to. The patients with the approximate physiological conditions are represented with similar multi-modal features. Therefore, we make an effort to construct the correlation graph between patients and aggregate the multi-modal features of them by their neighbors with medical knowledge relevance. Given the clinical notes $\{\mathbf{C}^{(i)}\}_{i=1}^{N_s}$ in training set, we have acquired the medical entity set $\{\mathbf{C}_e^{(i)}\}_{i=1}^{N_s}$ of them by Med7 model. The patient correlation graph (PCG) $\mathbf{A} \in \mathbb{R}^{N_s \times N_s}$ is initialized as an identity matrix. And the elements $\{\mathbf{A}_{ij} | i, j \in \{1, 2, \ldots, N_s\}\}$ in the PCG are the correlation degree between i-th patient and j-th one. Considering that the medical entity is the most important information to represent the patient, we exploit it to evaluate the correlation degree between each patient as shown in Figure 3. The jaccard similarity is the metric to evaluate the correlation of two sets, and the correlation degrees are calculated as follows:

$$\mathbf{A}_{ij} = \mathbf{A}_{ji} = \frac{\left|\mathbf{C}_e^{(i)} \cap \mathbf{C}_e^{(j)}\right|}{\left|\mathbf{C}_e^{(i)} \cup \mathbf{C}_e^{(j)}\right|}. \tag{1}$$

We concatenate the original extracted multi-modal features and the fusion ones as the patients' features $\mathbf{P} = \{\mathbf{p}^{(i)}\}_{i=1}^{N_s}$ where the i-th patient's multi-modal feature is calculated as $\mathbf{p}^{(i)} = \tanh\left(\mathbf{W}_p \left[\mathbf{h}_{N_t}^{(i)}; \hat{\mathbf{H}}^{(i)}; \mathbf{C}_g^{(i)}; \mathbf{C}_b^{(i)}\right] + \mathbf{b}_p\right)$, and $\mathbf{W}_p$ and $\mathbf{b}_p$ are trainable weights in the model.

To update the observed multi-modal features via the relevant patients, we utilize the graph convolution networks (GCN) (Kipf and Welling, 2017) to aggregate ones of neighbors. The calculation of the aggregation multi-modal features is simplified as $\hat{\mathbf{P}} = \sigma\left(\mathbf{APW}_g + \mathbf{b}_g\right)$ where $\mathbf{W}_g$ and $\mathbf{b}_g$ are trainable weights in GCN module, and σ is the nonlinear function. The patient multi-modal fusion feature with outer view is denoted as $\hat{\mathbf{P}} = \{\hat{\mathbf{p}}^{(i)}\}_{i=1}^{N_s}$ that contains various correlated ones.

3.3 Training Procedure

After acquiring the multi-modal fusion feature with multi-view, we utilize it to predict the target probabilities. The concatenation of multi-modal fusion features with multi-view and original extracted features is feed into the FFNN. The prediction probabilities are calculated as $\hat{\mathbf{y}}^{(i)} = \text{FFNN}\left(\left[\mathbf{h}_{N_t}^{(i)}; \mathbf{C}_g; \hat{\mathbf{P}}^{(i)}\right]; \theta_2\right)$ where θ_2 is the trainable weights in the FFNN module. To solve the classification task, we utilize the cross-entropy loss as follows:

$$\mathcal{L} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \hat{\mathbf{y}}^{(i)} \log \mathbf{y}^{(i)} \tag{2}$$

We feed the multi-modal data into the model and acquire the loss according to Equation 2. To train the parameter weights of the model, we use the stochastic gradient descent (SGD) method to update them according to the calculated loss.

4 Experiments

4.1 Dataset and Experiment Settings

We compare the proposed model with the existing methods on the medical benchmark dataset MIMIC-III (Johnson et al., 2016). The dataset contains the multi-type data collected from the real scenario including vital signals, clinical notes, ICD-9 code and so on. We follow the previous work (Bardak and Tan, 2021a) to extract the time series data

Method	In-hospital mortality		In-ICU mortality		LOS >3 days		LOS >7 days	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
(Khadanga et al., 2019)	86.50	52.50	88.66	50.59	64.67	57.00	70.70	16.59
(Park and Ho, 2020)	83.07	45.53	87.89	44.01	68.92	61.66	71.20	17.73
(Deznabi et al., 2021)	87.44	54.28	88.54	48.64	67.67	61.13	72.34	19.94
(Zhang et al., 2021a)	87.88	56.09	88.38	50.29	66.60	59.97	69.03	17.27
(Bardak and Tan, 2021a)	86.42	54.22	87.17	48.47	68.90	61.88	71.63	17.22
(Bardak and Tan, 2021b)	88.43	53.10	89.00	49.68	70.25	64.96	71.53	19.70
(Yang et al., 2021)	88.51	56.76	89.68	52.52	70.37	63.99	70.21	18.06
Ours	**90.22**	**62.76**	**90.71**	**56.71**	**71.72**	**65.24**	**75.06**	**21.26**

Table 2: The experimental results on four clinical outcome prediction tasks in macro-averaged % AUROC and % AUPRC. We run the experiments 5 times with different random seeds and report the average results. Our model outperforms the baseline methods on four tasks.

and clinical notes from the raw dataset with the publicly available tool MIMIC-Extract (Wang et al., 2020). The detailed statistical information of the dataset is shown in Table 1. The dataset is always used for two common targets: mortality and length of stay (LOS). And there are four binary classification tasks analyzed by the above works as follows:

1. **In-hospital Mortality**: This task targets to predict whether a patient dies before being discharged.

2. **In-ICU Mortality**: This task is defined to detect patients who are physically declining and predict the mortality of them within 24 hours.

3. **LOS >3**: This task targets to predict whether a patient stays in the ICU longer than 3 days.

4. **LOS >7**: This task is defined to detect patients who stay in the ICU longer than 7 days.

After extracting the dataset, we utilize the Python package `fancyimpute` [2] to impute the missing values in the time series data. We feed the clinical notes into the Med7 model to extract the medical entities and utilize the `BioBERT-Large` (Lee et al., 2020) version of the language model BERT to extract text features of the entities.

The dimension numbers N_d and N_k of hidden features in co-attention module are set to 128, and the others are set to 256 in our model. We set the dropout rate and learning rate to be 0.5 and 0.001 respectively. During the training process, we firstly train the model on the training set 300 epochs at most and test it on the development set. According to the early stopping strategy, we stop training the model when the loss on the development set does not decrease within 20 epochs. We use two different metrics including AUROC and AUPRC to evaluate the models on the imbalance tasks. All experiments are accelerated by a single NVIDIA GTX A6000 device.

[2]https://github.com/iskandr/fancyimpute

4.2 Compared Methods

We compare the proposed model with the existing machine learning methods. The models proposed by (Khadanga et al., 2019; Deznabi et al., 2021) were designed to combine the time series data with clinical notes with simple feature fusion strategy for outcome prediction. Besides, a new calibrated random forest (CaliForest) utilizing out-of-bag samples was proposed for the risk prediction (Park and Ho, 2020). Taking the structural medical information into account, the models proposed by (Bardak and Tan, 2021b,a) were implemented to combine the time series data with important medical mentions for clinical outcome prediction. The robust representations of patients' multi-model data in EHR are critical to the downstream tasks and the supervised deep patient representation learning framework was proposed for outcome prediction (Zhang et al., 2021a). The label aware attention mechanism was introduced into the multi-modal learning method (Yang et al., 2021) for the prediction task.

4.3 Experimental Results

We compare PM^{2}F^{2}N with the baseline methods on four classification tasks. The detailed experimental results on MIMIC-III are shown in Table 2. Our model can always achieve the best results on four tasks when compared with baseline methods. And the AUROC and AUPRC scores of the

Method	In-hospital mortality		In-ICU mortality		LOS >3 days		LOS >7 days	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
TS	88.79	57.23	88.86	50.83	69.62	63.34	71.60	19.00
TS + CN	89.99	61.47	89.31	52.37	71.30	64.68	74.26	21.11
TS + CN + PCG	90.10	62.49	90.25	55.02	71.50	64.86	74.90	21.18
TS + CN + PCG + CA	**90.22**	**62.76**	**90.71**	**56.71**	**71.72**	**65.24**	**75.06**	**21.26**

Table 3: The results for ablation study. "TS" and "CN" are short for time series data and clinical notes respectively. "PCG" represents the patient correlation graph which is introduced into our model. "CA" is a co-attention module to fuse features of time series and clinical notes.

proposed model increase by $1.1\% \sim 3.7\%$ and $0.4\% \sim 10.5\%$ over baselines on four tasks respectively. Compared with traditional method CaliForest (Park and Ho, 2020), the deep learning models can always gain better results than it on most classification tasks. Our model can outperform the multi-modal learning methods with simple feature fusion strategy (Khadanga et al., 2019; Deznabi et al., 2021) because the proposed one takes full advantage of patient multi-view multi-modal feature fusion. Although the models by (Bardak and Tan, 2021b,a) utilized the medical entities in clinical notes, they did not model the fine-grained features between multi-modal data. And the model by (Yang et al., 2021) incorporated the label information to enhance the text features of clinical notes and ignored fine-grained feature fusion. Our model gains better results over them with the use of co-attention module for effectively modeling multi-modal fusion features. Besides, the representation learning method (Zhang et al., 2021a) is beneficial to downstream risk prediction task. However, the method did not take the patient correlation in medical knowledge into account and model the relevant multi-modal features. Our model exploits the structural medical information for constructing patient correlation graph and fuses the multi-modal features by GCN based on the graph. Therefore, the proposed model gains better generalization ability for clinical outcome prediction.

4.4 Further Discussion

To dig into the model, we conduct the detailed analysis for presenting it in different aspects. The ablation study is performed to demonstrate the effectiveness of the different feature fusion strategies proposed in our model. Besides, to verify the effectiveness of the patient correlation graph, we compare the performances of the tasks that are conducted on the adjacency matrixes filled with different values. Eventually, we visualize the multi-modal features extracted from the proposed model for presenting the usefulness of the patient correlation information in the feature fusion aspect.

4.4.1 Ablation Study

As shown in Table 3, we conduct the ablation study to present the effectiveness of the proposed multi-modal feature fusion strategies. We utilize the single modal data (TS), multi-modal data (TS + CN) to train clinical outcome prediction models respectively as the base comparison methods. It proves the effectiveness of multi-modal learning that the model trained with multi-modal data achieves better results than that with single-modal data. When the patient correlation graph (PCG) is introduced into the base multi-modal learning method, the results on the four tasks are improved to vary degrees. The multi-modal feature fusion with outer view can aggregate that of various patients and improve the generalization ability of the proposed model for clinical outcome prediction. The model incorporated with co-attention (CA) is the proposed model PM^{2}F^{2}N and gets vary improvements on the four tasks. With the advantage of CA, our model can fuse the multi-modal features in a fine-grained way.

4.4.2 Effect of Patient Correlation Graph

As shown in Figure 4, we conduct the comparison experiments to demonstrate the effectiveness of the proposed patient correlation graph (PCG). The proposed model is fed with two distinct adjacency matrices filled with all 0s and 1s to replace the PCG. The "Adj=0" model utilizes the adjacency matrix filled with all 0s to disentangle GCN from our model as a baseline. And the "Adj=1" model exploits the adjacency matrix filled with all 1s to verify the effect of the patient correlation degrees. Compared with baseline "Adj=0" model, the "Adj=1" model gets various drops on AUROC and AUPRC while our model gains $0.5\% \sim 1.3\%$

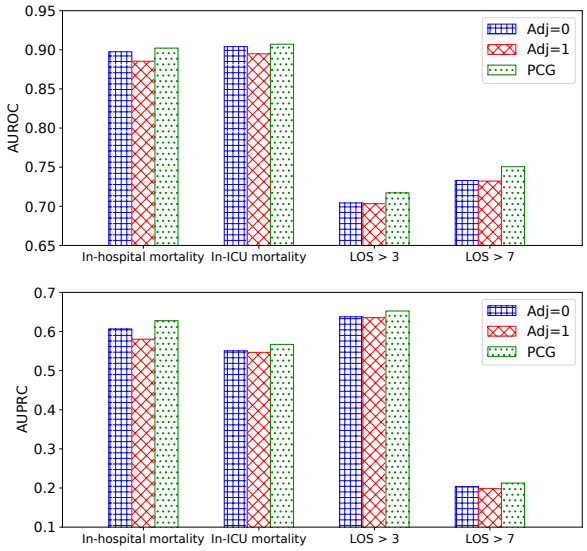

Figure 4: The effect of patient correlation graph (PCG). "Adj=0" indicates that the model which utilizes the adjacency matrix filled with all 0s for prediction. "Adj=1" indicates that the model exploits the adjacency matrix filled with all 1s for the tasks. "PCG" represents our proposed patient correlation graph which are introduced into the model as described in Section 3.2.2.

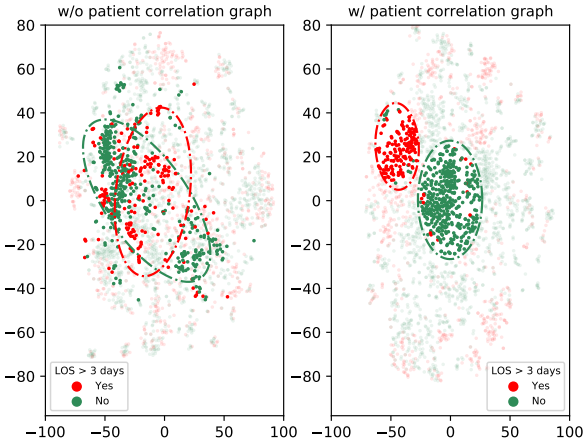

Figure 5: The t-SNE visualization results of the multi-modal features extracted from the models with and without patient correlation graph respectively. We evaluate the models on **LOS >3** task. And the highlighted points represent the same group of patients.

and $1.0\% \sim 2.1\%$ improvements over it on the four tasks. Because the adjacency matrix filled with all 1s is a fully-connected graph and introduces correlation noise into the model for worse results. In comparison, our proposed PCG takes patient medical correlation as constraints and exploits medical information in clinical notes to build it. The experimental results confirm that incorporating patient correlation information benefits the clinical outcome prediction.

4.4.3 Visualization Analysis

To verify the effectiveness of patient correlation graph (PCG) to the multi-modal fusion feature intuitively, we visualize the learned features extracted from the models with and without patient correlation graph as shown in Figure 5. We focus on the **LOS >3** task because of its balanced class distribution. After training the models, we utilize them to acquire the multi-modal features of samples in the test set. We visualize the patient multi-modal fusion features in Figure 5 where the dimension is reduce to two by t-SNE. Further more, we also select the same group of patients to highlight their feature points and circle them. As the whole patients observed, the multi-modal features of them with same label that are learned with PCG are more clustered. The selected patients' features learnt with PCG are clustered into two groups with a clear boundary, but that without PCG are scattered and intertwined. The comparison between the two results demonstrate the effectiveness of the patient correlation graph which connects the multi-modal features of relevant patients.

5 Conclusion

In this paper, we analyze the disadvantages of existing multi-modal learning methods for clinical outcome prediction. To enhance the multi-modal feature in different views, we propose the patient multi-view multi-modal feature fusion networks (PM^2F^2N) for the task. From the inner view, we extend the co-attention module to fuse the features of the time series data and structural medical knowledge in a fine-grained way. From the outer view, we exploit the correlation between patients to aggregate the multi-modal features between the similar patients. With the multi-view multi-modal feature fusion strategy, the proposed model can learn the general patient representations for clinical outcome prediction. Compared with the existing methods, our model can gain the best results on benchmark dataset MIMIC-III. And the further discussion including ablation study, effect of PCG and visualization analysis, verifies the effectiveness of the proposed strategy. Considering the heterogeneity of patients, we will try to adapt the heterogeneous graphs for modeling the correlation between patients in the future.

6 Limitations

The proposed model is limited to feeding the whole patient multimodal data into it and utilizing large memory to calculate aggregation multimodal features by GCN layer. Besides, the scalability of patient correlation graph (PCG) is poor because the PCG should be reconstructed when the new patients are added into the original patient set.

Acknowledgements

We thank the anonymous reviewers for the valuable comments on our manuscript. This research is supported by the Chinese Scientific and Technical Innovation Project 2030 (2018AAA0102100), the National Natural Science Foundation of China (No. 62272250, U1936206, 62002178).

References

Batuhan Bardak and Mehmet Tan. 2021a. Improving clinical outcome predictions using convolution over medical entities with multimodal learning. *Artif. Intell. Medicine*, 117:102112.

Batuhan Bardak and Mehmet Tan. 2021b. Prediction of mortality and length of stay with deep learning. In *SPCA*, pages 1–4.

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. 2018. What's in a note? unpacking predictive value in clinical note representations. *AMIA on Translational Science Proceedings*, 2018:26 – 34.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Iman Deznabi, Mohit Iyyer, and Madalina Fiterau. 2021. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of ACL-IJCNLP*, pages 4026–4031.

Marzyeh Ghassemi, Marco A. F. Pimentel, Tristan Naumann, Thomas Brennan, David A. Clifton, Peter Szolovits, and Mengling Feng. 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *AAAI*, pages 446–453.

Ramin Ghorbani, Rouzbeh Ghousi, Ahmad Makui, and Alireza Atashi. 2020. A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset. *IEEE Access*, 8:141066–141079.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Swaraj Khadanga, Karan Aggarwal, Shafiq R. Joty, and Jaideep Srivastava. 2019. Using clinical notes with time series data for ICU management. In *EMNLP-IJCNLP*, pages 6431–6436.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo J. Nevado-Holgado. 2021. Med7: A transferable clinical natural language processing model for electronic health records. *Artif. Intell. Medicine*, 118:102086.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, pages 289–297.

Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. 2020. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *AAAI*, pages 833–840.

Yubin Park and Joyce C. Ho. 2020. Califorest: calibrated random forest for health data. In *CHIL*, pages 40–50.

Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *AAAI*, pages 4091–4098.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *EACL*, pages 881–893.

J. L. Vincent, R. Moreno, J. Takala, S. Willatts, and L. G. Thijs. 1996. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710.

Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. 2020. Mimic-extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In *CHIL*, pages 222–235.

Yanbo Xu, Siddharth Biswal, Shriprasad R. Deshpande, Kevin O. Maher, and Jimeng Sun. 2018. RAIM: recurrent attentive and intensive model of multimodal patient monitoring data. In *KDD*, pages 2565–2573.

Haiyang Yang, Li Kuang, and FengQiang Xia. 2021. Multimodal temporal-clinical note network for mortality prediction. *J. Biomed. Semant.*, 12(1):3.

Xianli Zhang, Buyue Qian, Yang Li, Yang Liu, Xi Chen, Chong Guan, and Chen Li. 2021a. Learning robust patient representations from multi-modal electronic health records: A supervised deep learning approach. In *SIAM*, pages 585–593.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*, 6.

Ying Zhang, Baohang Zhou, Xiangrui Cai, Wenya Guo, Xiaoke Ding, and Xiaojie Yuan. 2021b. Missing value imputation in multivariate time series with end-to-end generative adversarial networks. *Inf. Sci.*, 551:67–82.