Knowledge-aware Attention Network for Medication Effectiveness Prediction

Yingying Zhang[1], Xian Wu[1] †, Yu Zhang[2], Yefeng Zheng[1]
[1] Jarvis Research Center, Tencent YouTu Lab
[2] Tencent
{ninzhang, kevinxwu, wrenzhang, yefengzheng}@tencent.com

Abstract

The first 24 hours' medication plan is critical to patients with serious or life-threatening illnesses and injuries. An appropriate medication can result in a lower mortality, a shorter length stay and a higher APACHE score. However, in clinical practice, the medication plan is often error-prone, especially when a decision must be made quickly for life-threatening situations in Intensive Care Unit (ICU). Therefore, predicting the effectiveness of the first 24 hours' medication plan is of great importance in assisting doctors to make proper decisions. Existing effectiveness prediction works usually focus on one specific medicine, one specific disease, or one specific lab test, making it hard to extend to general medicines and diseases in hospital/ICU scenarios. In this paper, we propose to predict medication effectiveness of the first 24 hours in hospital/ICU based on patients' information. Specifically, we use a knowledge enhanced module to incorporate external knowledge about medications and a medical feature learning module to determine the interaction between diagnosis and medications. To handle the data imbalance problem, we further optimize the proposed model with a contrastive loss. Extensive experimental results on a public dataset show that our model can significantly outperform state-of-the-art methods.

Keywords: contrastive learning, representation learning

1. Introduction

Patients with serious medical conditions need to stay in hospitals for immediate and overnight care, while emergent patients will be sent to an Intensive Care Unit (ICU) for special care. During a hospital/ICU admission, the medication plan of the first 24 hours is extremely important. For example, the management of burn cases in the first 24 hours is one of the greatest challenges and affect the mortality rate dramatically (Alharbi et al., 2012); For acute coronary syndrome patients, the first 24 hours' medication plan is deterministic to patients' lives (de Matos Soeiro et al., 2016).

However, in clinical practice, the medication plan is not always correct. The medication errors can cause disability and death up to 6.5% hospital admissions (Lisby et al., 2005), and occur more frequently in ICU than other departments. The Sentinel Events Evaluation (SEE) study reviewed errors recorded in more than 200 ICUs and found 10.5 medication errors per 100 patient days (Valentin et al., 2006). Therefore, an assessment of the effectiveness is quite important to guide doctors in establishing a more appropriate medication plan.

To tackle this problem, we propose to predict the effectiveness of medication plan for the first 24 hours in hospital/ICU. As shown in Figure 1, we formulate effectiveness prediction as a classification problem. Given a patient, we extract two type of information: (1) the medication plan of the first 24 hours; (2) the initial diagnosis given by doctors in admission. Based on these information, we predict the value of three core metrics: mortality, length of stay, and APACHE, i.e., Acute Physiology, Age, and Chronic Health Evaluation score (Zimmerman et al., 2006). Since a more appropriate medication plan can result in a lower mortality, a shorter length of stay and a higher APACHE score, we can use the prediction on these metrics to evaluate the effectiveness of the medication plan.

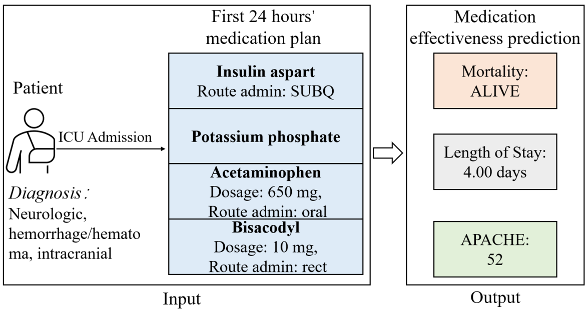


Figure 1: Tasks of first 24 hours' medication effectiveness prediction in hospitals/ICUs

Existing works on medication effectiveness prediction can be classified into two categories: 1) single medicine effectiveness prediction: these works either leverage genomics information (Alyass et al., 2015; Brown et al., 2017) or chemical-protein interactome (Luo et al., 2016) to infer the effectiveness. However, these works only focus on the information of a single medicine and ignore the specific condi-

† Corresponding author

LREC-COLING 2024, pages 9799–9809
20-25 May, 2024.

tions of a patient. In other words, these works focus on drug research rather than evaluate whether a medicine plan is suitable for a specific patient; 2) lab test effectiveness prediction: these works mainly use a linear fixed model to estimate the effects for a certain type of outcome (Ghalwash et al., 2017; Dey et al., 2019). For example, DELT (Ghalwash et al., 2017) aims to determine the effects of exposure of the medicine to laboratory measurements along with other confounding factors using a fixed effect model, and designs some penalties for regularization based on prior knowledge, such as drug sparsity and drug similarity; PerDREP (Dey et al., 2019) further provides personalized medication effectiveness prediction that incorporates a patient similarity graph as a network regularization. These methods require a lot of expert knowledge to identify the correlation between laboratory test results and the medicines. Therefore, they cannot be easily generalized to more diseases and drugs. In addition, patients in the hospital/ICU often suffer from multiple diseases at the same time. For example, hypertension and hyperlipidemia often occur together. Therefore, these drug effectiveness prediction methods could hardly be used for prediction in early hospital/ICU admission.

Since existing works cannot be directly adapted to predict the medication effectiveness of the first 24 hours, in this paper, we propose to predict the medication effectiveness using multiple sources of information: 1) current medication plan and prior knowledge of medicines; 2) the initial diagnosis upon hospital/ICU admission. Firstly, we use a knowledge-enhanced module to incorporate external knowledge about medicines in modeling. External knowledge contains information such as drug-drug interaction and what diseases the drug can treat. These external information are represented in the form of a knowledge graph; Secondly, we mine the interaction between diagnosis and medicines to conduct the prediction. To alleviate the data imbalance problem, we further propose a hybrid loss of a contrastive loss and a cross-entropy loss. The contrastive learning loss incorporates different within-class samples as positives for each sample; therefore, it can help learn better representation of medication. The major contributions of this paper can be summarized as follows:

- We propose a novel task: the effectiveness prediction of the medication plan for first 24 hours in hospital/ICU, which could help doctors to improve the treatment and in turn reduce the mortality rate.

- In this proposed model, we integrate external knowledge into data driven prediction which can provide credibility for prediction.

- We introduce a contrastive loss together with a cross-entropy loss to relieve the data imbalance problem. According to the experimental results on a public real dataset, our model outperforms state-of-the-art works.

2. Related Work

2.1. EHR-based Prediction

The Electronic Health Record (EHR) contains abundant patient information, which can be leveraged to predict the diagnosis (Peng et al., 2021; Ma et al., 2017), mortality (Hur et al., 2022b; Lyu et al., 2022), and length of stay (Lyu et al., 2022) of a specific patient. The main challenge in modeling medical data in the EHR is learning *unstructured* and *heterogeneous* information in the EHR. To deal with *unstructured* medical data, previous research uses deep learning architectures, such as convolutional neural networks (Nguyen et al., 2017; Yao et al., 2019) and recurrent neural networks (Choi et al., 2016a). They follow the ideas of processing sentences in documents to treat a patient's admission as a document and a medical record as a sentence. However, unlike normal sequential learning tasks, the intervals between two recordings can be different, which is an important factor in clinical studies. Attention-based models (Choi et al., 2016b; Song et al., 2018; Shang et al., 2019; Hur et al., 2022b) have been successfully used for healthcare tasks to model sequential EHR data. For example, SAnD (Song et al., 2018) uses an attention mechanism (Vaswani et al., 2017) to capture long-term dependencies for sequential medical events. BiteNet (Peng et al., 2020) uses a masked attention mechanism network to capture temporal information and an interval encoding module to encode the interval between recordings. To resolve *heterogeneous* challenge in EHR systems, DescEmb (Hur et al., 2022a) and UniHPF (Hur et al., 2022b) address heterogeneous information in medical codes with text embeddings. Furthermore, VGNN (Zhu and Razavian, 2021) and MedGTX (Park et al., 2022) construct a graph of diagnoses and medical events and utilize graph neural networks to acquire the embedding of the admission records.

2.2. Medication Effectiveness Prediction

Existing works on medication effectiveness prediction can be grouped in two categories: (1) *single medicine effectiveness prediction* and (2) *lab test effectiveness prediction*.

Single medicine effectiveness prediction works either leverage genomics information (Alyass et al., 2015; Brown et al., 2017) or chemical-protein interactome (Luo et al., 2016) to infer the effectiveness. However, these works focus only on the information of a single medicine and ignore the specific conditions of a patient. In other words, they are not designed for clinical use.

With increasing use of EHR, another research line is to *discover drugs associated with laboratory test results* (Ghalwash et al., 2017; Dey et al., 2019). DELT (Ghalwash et al., 2017) uses a fixed-effect model and considers prior knowledge as a penalty, such as drug sparsity, temporal smoothness, drug group structure, and drug similarity. PerDREP (Dey et al., 2019) also uses patient-specific time-invariant parameters to represent medication effects and learn personalized drug response predictions. Liu *et al.* (2021) performed an early prediction of mortality in patients with rhabdomyolysis in the ICU. However, these studies require expertise to determine the relationship between laboratory test results and medication effect and are difficult to generalize to other diseases. Our proposed medication effectiveness prediction model is generic and can be easily extended to different diseases.

3. Methodology

3.1. Problem Definition

In this subsection, we introduce the notation and formulate the targeted task. Considering the medication sequence, our goal is to predict the effectiveness of the medication. The medication that each patient takes within the first 24 hours after an admission can be represented as a sequence of medication records $[m_1, m_2, \ldots, m_N]$, where N is the total number of medications throughout the first 24 hours of admission, and the m_i's are sorted by the timing of the events. The i^{th} medication record contains multiple medication-related attributes, such as drug name, dosage, and drug infusion rate. Therefore, each attribute in m_i can be considered as a tuple of an attribute key n_i^k and its value v_i^k. The diagnoses corresponding to the inclusion are denoted as $[x_1, \ldots, x_L]$, where L is the number of the diagnoses, since a patient may have multiple diagnoses in an admission.

Different types of medications are recorded in different medical records and attributes have different schemas. For example, in the eICU dataset (Pollard et al., 2018), continuous infusions are recorded in the *infusionDrug* table and prescribed medications are recorded in the *medication* table. The rows in the *infusionDrug* table contain attributes such as the drug infusion rate and the total drug quantity, while the *medication* table contains drug dosage and frequency.

Additionally, to distinguish medications from different charts, each medication record has its corresponding medication type t_i. Thus, the single medication record m_i can be denoted as $\{t_i, \{(n_i^k, v_i^k), k \in \{1, \ldots, |m_i|\}\}\}$, where $|m_i|$ is the number of attributes of the medication m_i.

3.2. Overall Framework

In this section, as shown in Figure 2, we present a framework for predicting drug effectiveness that includes three parts:

- **Knowledge enhanced module**: Knowledge enhanced module is used to enhance external knowledge information for drugs.
- **Medication feature learning**: Medication feature learning module learns consistent drug embedding through text description and integrate diagnosis information in modeling.
- **Loss Function**: the model is optimized with a hybrid loss of contrastive loss and cross-entropy loss to alleviate the data imbalance problem.

3.3. Incorporating Knowledge Graph

As we assess the effect of medication during ICU admission, we mainly focus on the drug information in the medication. Among the two types of medications (*infusionDrug* and *medication*), there is a common attribute, namely "drug name" (n_i^0), the value of the attribute can be denoted as v_i^0.

To enrich the representation of the drug, we introduce external medical knowledge graph. We use BIOS (Yu et al., 2022) as our additional medical knowledge graph. It is an automatically generated comprehensive biomedical knowledge graph. It contains more than 54 million terms and 69 million triples that contain relation types such as "is a", "is part of", and "may treat".

Given the knowledge graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, we first use TorusE (Ebisu and Ichise, 2018) to obtain the entity's representation. In this manner, the entity representation contains the connection information in the knowledge graph. Then we retrieve related entities of the medicine name v_i^0 according to the textual similarity, their representations can be denoted as $\boldsymbol{e} \in \mathbb{R}^{M \times d}$, where M is the number of related entities, and d is the dimension of the embeddings. Then we generate the knowledge-enhanced drug representation $\boldsymbol{c}_{i,drug}$ corresponding to the related entities with the mean pooling operation:

$$\boldsymbol{c}_{i,drug} = \text{Pooling}(\boldsymbol{e}), \text{where } \boldsymbol{c}_{i,drug} \in \mathbb{R}^d. \quad (1)$$

3.4. Medication Representation Learning

Inspired by DescEmb (Hur et al., 2022a) and UniHPF (Hur et al., 2022b), we use text embedding to account for heterogeneity in the different charts. Given the single medication $m_i = \{t_i, \{(n_i^k, v_i^k), k \in$

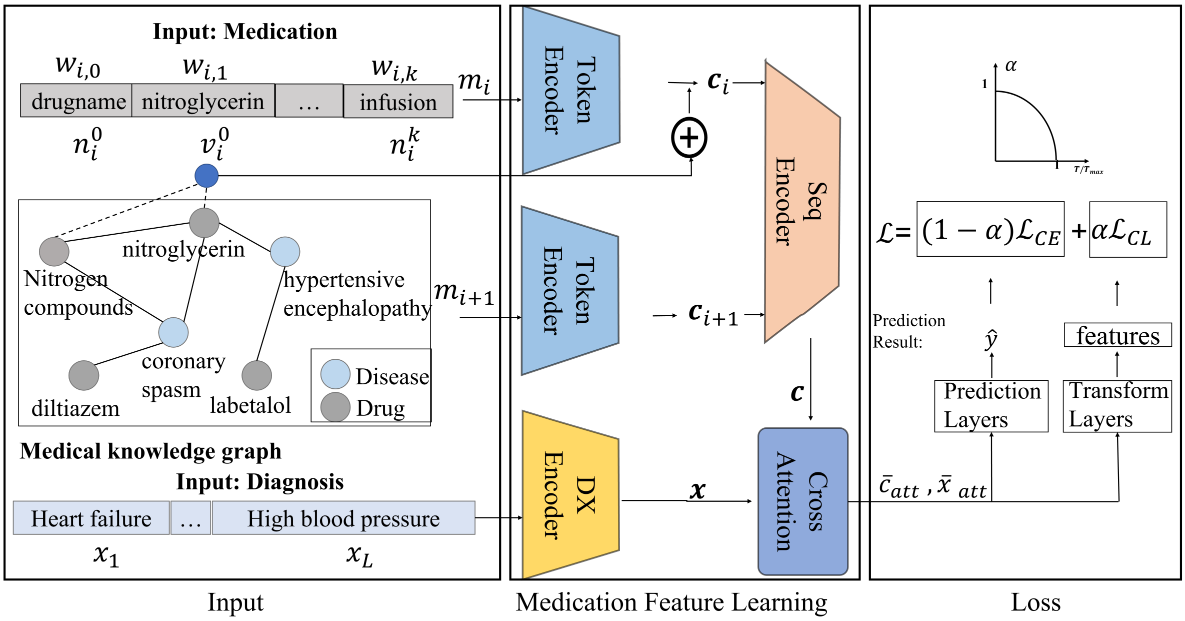


Figure 2: The overall framework. It includes three parts: (1) The input includes a medical knowledge graph, and a list of medications and initial diagnoses. A knowledge enhanced module is designed to enhance external knowledge information for drugs. (2) The medication feature learning module includes several encoders and a cross attention module. It learns consistent drug embedding through text description and integrates diagnosis information. (3) The model is trained with a hybrid loss of contrastive loss and cross-entropy loss to alleviate the data imbalance problem.

$\{1, \dots, |m_i|\}\}$, we can generate a description of the medication. The description is a sequence of words $(w_{i,1}, w_{i,2}, \dots, w_{i,s})$, consisting of tokens of attributes name n_i^k, value v_i^k and medication type t_i. As shown in Figure 2, the description of m_i can be represented as (*"drug name", "nitroglycerin", "drug rate",…,"infusion"*). We use a TokenEncoder to generate the representation of m_i corresponding to the word sequence:

$$\boldsymbol{c}_{i,token} = \text{TokenEncoder}(w_{i,1}, w_{i,2}, \dots, w_{i,s}). \quad (2)$$

The final representation of c_i is the combination of the token representation $\boldsymbol{c}_{i,token}$ and the knowledge-enhanced drug representation $\boldsymbol{c}_{i,drug}$:

$$\boldsymbol{c}_i = f([\boldsymbol{c}_{i,token} || \boldsymbol{c}_{i,drug}]), \quad (3)$$

where $\boldsymbol{c}_i \in \mathbb{R}^d$ and $f(\cdot)$ is a 3-layer linear transformation with the $\tanh$ activation, and $||$ denotes the concatenation operation. Then we use a SeqEncoder to encode the sequential information:

$$\boldsymbol{c} = \text{SeqEncoder}(\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_N), \text{where } \boldsymbol{c} \in \mathbb{R}^{L \times d}. \quad (4)$$

The TokenEncoder and SeqEncoder can be any sequential representation learning models, such as Bi-RNN and BERT (Devlin et al., 2019). Here, we use BERT-small (4-layers) as the backbone.

3.5. Incorporating Diagnosis Information

We first learn the representation of diagnoses $[x_1, \dots, x_L]$ using a DXEncoder:

$$\boldsymbol{x} = \text{DXEncoder}([x_1, \dots, x_L]), \text{where } \boldsymbol{x} \in \mathbb{R}^{L \times d}. \quad (5)$$

The initial embedding of x_j is derived from a randomly initialized embedding matrix. The DXEncoder can be any representation learning model. Here, we use a 3-layer linear transformation with the $\tanh$ activation.

To incorporate diagnosis information into prediction and find the relationship between diagnosis and medication, we use a cross-attention layer to introduce diagnostic information. The attention between the medication ($\boldsymbol{c}$) and the diagnosis ($\boldsymbol{x}$) helps the model figure out how the medication and the diagnosis are related.

The general attention mechanism (Vaswani et al., 2017) is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d}})V, \quad (6)$$

where Q, K, and V stand for query, key and value.

Then the cross-attention between the diagnosis and medication is as follows:

$$\boldsymbol{x}_{att} = \text{Attention}(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{c}), \text{where } \boldsymbol{x}_{att} \in \mathbb{R}^{N \times d}, \quad (7)$$

$$\boldsymbol{c}_{att} = \text{Attention}(\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{x}), \text{where } \boldsymbol{c}_{att} \in \mathbb{R}^{L \times d}. \quad (8)$$

Finally, we use mean pooling to make $\boldsymbol{x}_{att}$ and $\boldsymbol{c}_{att}$ the same shape

$$\bar{\boldsymbol{x}}_{att} = \text{Pooling}(\boldsymbol{x}_{att}), \text{where } \bar{\boldsymbol{x}}_{att} \in \mathbb{R}^{d}, \quad (9)$$

$$\bar{\boldsymbol{c}}_{att} = \text{Pooling}(\boldsymbol{c}_{att}), \text{where } \bar{\boldsymbol{c}}_{att} \in \mathbb{R}^{d}. \quad (10)$$

3.6. Optimization

After acquiring the attended diagnosis representation $\bar{\boldsymbol{x}}_{att}$ and the attended medication representation $\bar{\boldsymbol{c}}_{att}$, we perform a classification task to estimate the effectiveness of the drug:

$$a = \bar{\boldsymbol{x}}_{att} || \bar{\boldsymbol{c}}_{att}, \ \hat{y} = g(a), \quad (11)$$

where || is the concatenation operation and $g(\cdot)$ is a 3-layer linear transformation with the $\tanh$ activation. We adopt the cross entropy loss for the binary classifier:

$$\mathcal{L}_{CE} = \frac{-1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} [y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)], \quad (12)$$

where $\mathcal{B}$ denotes the samples in a batch.

Moreover, we use a contrastive loss to deal with the data imbalance problem. Contrastive learning can pull together samples from the same class in the normalized embedding space and push apart the samples from different classes (Wang et al., 2021). It incorporates different within-class samples as positives for each sample. Therefore, it can learn a better representation of the medication.

For the i^{th} sample in a batch $\mathcal{B}$, we first acquire its representation a_i according to Eq. (11). Then we generate k perturbed views of a_i using the 3-layer liner transformation with the $\tanh$ activation:

$$z_i^{(1)} = f_1(a_i), \ldots, z_i^{(k)} = f_k(a_i) \quad (13)$$

where $z_i^{(1)}$, ... ,$z_i^{(k)}$ denotes k different views of a_i and f_1, ... ,f_k denote k different functions without sharing any parameters. In this manner, the representations of sampels within a batch is changed from $\{a_1, a_2, \ldots, a_{|\mathcal{B}|}\}$ to $Z = \{z_1^{(1)}, \ldots, z_1^{(k)}, z_2^{(1)}, \ldots, z_2^{(k)}, \ldots, z_{|\mathcal{B}|}^{(1)}, \ldots, z_{|\mathcal{B}|}^{(k)}\}$ with $|\mathcal{B}| \times k$ elements. For each element z_i in this new batch set Z, we select the positive and negative pairs for constrastive learning. The set of positive sample of z_i is $z_i^+ = \{z_j | y_j = y_i, i \neq j\}$ (the same label) and the set of negative sample of z_i are $z_i^- = \{z_j | y_j \neq y_i, i \neq j\}$ (different labels). We set the number of views $k = 2$, to avoid $|z_i^+|$ being zero when there is only one sample of the long-tailed class in the batch. The contrastive loss can be formulated as follows:

$$\mathcal{L}_{CL} = \frac{-1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \frac{1}{|\boldsymbol{z}_i^+|} \sum_{\boldsymbol{z}_j \in \boldsymbol{z}_i^+} \log \frac{e^{\boldsymbol{z}_i \cdot \boldsymbol{z}_j / \tau}}{\sum_{\boldsymbol{z}_k \in \{z_i^+, z_i^-\}} e^{\boldsymbol{z}_i \cdot \boldsymbol{z}_k / \tau}}, \quad (14)$$

where τ is the temperature scalar. The final loss is the combination of the contrastive loss and the cross-entropy loss:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{CL} + (1 - \alpha) \cdot \mathcal{L}_{CE}, \quad (15)$$

where α is a weighting coefficient inversely proportional to the number of epochs.

4. Experiments

4.1. Dataset

We use a publicly available dataset: eICU (Pollard et al., 2018). It consists of ICU records from multiple US-based hospitals with up to 140,000 unique patients admitted between 2014 and 2015. We filter the admissions if their corresponding diagnoses or prediction targets (APACHE/ mortality/ LOS) are missing. Finally, there are 128,874 unique patients with 135,495 admissions. The dataset is divided into three sets with an 8: 1: 1 ratio, including 108,379 admissions for training, 13,544 for validation, and 13,572 for testing. The average number of tokens per medication is 59.59, the average number of medications per admission is 26.83, and the average number of diagnoses per admission is 3.55.

4.2. Prediction Tasks

To evaluate our framework on a variety of predictive tasks, we formulate three binary prediction tasks following (McDermott et al., 2021; Hur et al., 2022b) based on ICU admissions. The positive and negative ratio in three tasks is shown in Table 1. All evaluations are scored with micro-F1, macro precision, macro recall, and macro F1. There tasks are as follows: (1) *Mortality Prediction*: A sample is flagged as positive for mortality if the discharge state is "expired". The ratio between "alive" and "expired" is approximately 10: 1. (2) *Length-of-Stay Prediction (LOS7)*: Whether a given ICU admission has lasted longer than 7 days. The ratio between "LOS$\leq$ 7" and "LOS$>$7" is about 3: 2. (3) *APACHE* : APACHE is the abbreviation for Acute Physiology, Age, and Chronic Health Evaluation (Zimmerman et al., 2006), an instrument used to assess the risk of ICU patients for performance comparisons and quality improvement analysis in the ICU. The classification task is to assess whether the APACHE score is greater than 50. The ratio between two classes is approximately 1: 1.

	Training	Validation	Test
Mortality	98,326 : 10,053	12,235 : 1,209	12,266 : 1,306
LOS7	65,093 : 43,286	8,166 : 5,378	8,141 : 5,431
APACHE	53,557 : 54,882	6,883 : 6,661	6,690 : 6,882

Table 1: Statistics of training, validation and test sets. The numbers before and after ":" denote the numbers of positive and negative samples, respectively.

	Mortality		LOS7		APACHE	
	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
DescEmb	0.8369	0.6273	0.6672	0.6673	0.7184	0.7183
UniHPF	0.9116	0.6813	0.6775	0.6676	0.7169	0.7166
VGNN	0.8444	0.6693	0.6652	0.6586	0.6561	0.6759
MedGTX	0.9045	0.6199	0.6340	0.6310	0.6587	0.6589
GCT	0.8068	0.6594	0.6664	0.6538	0.6983	0.6983
BERT	0.7669	0.6337	0.6489	0.6151	0.6756	0.6756
UniHPF$_{cl}$	**0.9142**	0.6937	0.6850	0.6681	0.7231	0.7241
ours	0.9114	**0.6988**	**0.6974**	**0.6799**	**0.7288**	**0.7296**

Table 2: Prediction results for the hospital mortality prediction, the hospital length of stay within 7 days (LOS7), and APACHE score task.

4.3. Baselines

Our model is compared with several baselines to evaluate performance in medication effectiveness prediction tasks, including the text embedding methods BERT (Devlin et al., 2019), DescEmb (Hur et al., 2022a) and UniHPF (Hur et al., 2022b), the knowledge enhanced method MedGTX (Park et al., 2022) and the diagnosis-based methods GCT (Choi et al., 2020) and VGNN (Zhu and Razavian, 2021).

- **DescEmb** (Hur et al., 2022a): It utilizes text embedding for clinical descriptions linked to each medical diagnosis to overcome the heterogeneity of codes.
- **UniHPF** (Hur et al., 2022b): It handles heterogeneous EHR with a unified framework without feature selection.
- **VGNN** (Zhu and Razavian, 2021): It introduces variational regularization for node representation to ease the insufficiency of self-attention in graph-based models.
- **MedGTX** (Park et al., 2022): It uses a graph encoder to exploit the graphical nature of structured EHR data, a text encoder to handle unstructured text, and a cross-modal encoder to learn a joint representation space.
- **GCT** (Choi et al., 2020): It uses guided self-attention to learn the hidden graph structure of the EHR.
- **BERT** (Devlin et al., 2019): It concatenates all drug information and then feeds them to the BERT encoder.
- **UniHPF$_{cl}$**: It replace UniHPF's loss function by hybrid loss defined in Eq.(15) as a baseline to verify the effectiveness of contrastive loss.
- **Ours**: The full model contains diagnosis and knowledge graph information and trains with both $\mathcal{L}_{CE}$ and $\mathcal{L}_{CL}$.

4.4. Implementation Details

The embedding dim d is set to 768, and the dropout rate is set to 0.1. We use the Adam algorithm to optimize our model. The learning rate is set to 10^{-5} for all models. The batch size is 16. For the curriculum coefficient α, we use a parabolic decay with respect to the epoch number (Wang et al., 2021) and set $\alpha = 1 - (T/T_{max})^2$, where T denotes the current epoch number and T_{max} indicates the max epoch number, which is set to 40 for all methods. Methods except for ours (+cl) and ours (full) use an early stop strategy with 10-epoch patience. The temperature parameter τ in Eq. (14) is set to 0.07.

4.5. Results

4.5.1. Quantitative Analysis

Table 2 shows our experimental results for the three prediction tasks, respectively. Compared to the baselines, we can find that:

- Our method outperforms all baselines on all metrics for the APACHE and LOS7 prediction tasks. On the mortality prediction task,

Components			Mortality		LOS7		APACHE	
kg	cl	dx	mirco F1	marco F1	mirco F1	marco F1	mirco F1	marco F1
			0.9116	0.6813	0.6775	0.6676	0.7169	0.7166
✓			0.9131	0.6851	0.6896	0.6729	0.7217	0.7233
	✓		0.9142	0.6937	0.6850	0.6681	0.7231	0.7241
		✓	0.9147	0.6883	0.6934	0.6753	0.7222	0.7239
✓	✓		0.9101	0.6846	0.6851	0.6701	0.7233	0.7239
✓		✓	**0.9148**	0.6901	0.6912	**0.6836**	0.7262	0.7270
	✓	✓	0.9081	0.6881	0.6956	0.6753	0.7260	0.7273
✓	✓	✓	0.9114	**0.6988**	**0.6974**	0.6799	**0.7288**	**0.7296**

Table 3: Ablation study for three tasks.

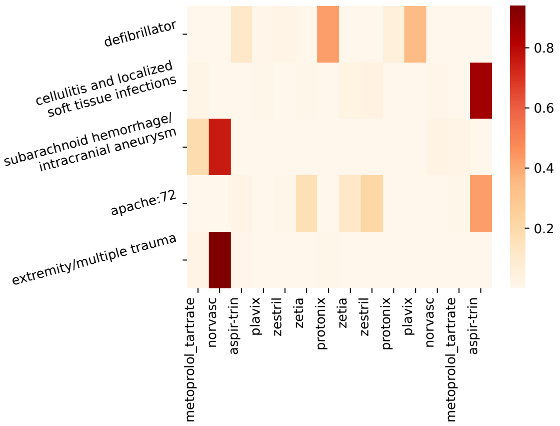


Figure 3: Attention weights between drugs and diagnoses.

our method and its variants outperform the baselines on both micro and macro F1 scores, achieving a better balance between precision and recall. There are several possible reasons. First, our model introduces external knowledge of drug by a knowledge-enhanced module, which enriches the representations of medicines with external knowledge like relations with diseases and other medicines; Second, our model mines the relationship between diagnosis and medications, thus capturing useful information about the diagnosis; Third, our model uses a hybrid of contrastive loss and cross-entropy loss, which can learn better features for long-tailed class, and thus learn a better classifier.

- Contrastive loss is more effective for dealing with imbalanced data. Compared to UniHPF, UniHPF$_{cl}$ gains 1.2% macro F1 on the mortality prediction task, while on APACHE and LOS7 prediction tasks, the performance gain are relatively smaller, 0.8% and 0.1%, respectively. Furthermore, our proposed model outperforms the contrastive learning-based baseline UniHPF$_{cl}$.

- BERT performs the worst among all methods, as it does not capture sequential information in the medications.

- Compared to knowledge-enhanced method MedGTX, our model gains about 4%–7% improvement on three tasks. This is probably due to our model utilizes the relation among entities in the external knowledge graph.

- Compared to methods that use graphs to embed diagnosis information like GCT and VGNN, our model obtains better performance. This shows that the co-attention layer can better capture the relationship between diagnosis and the related medication, thus learn better embeddings of an admission for drug effectiveness prediction.

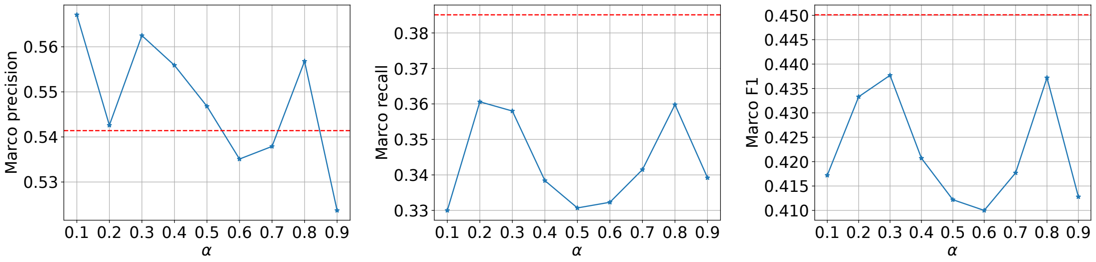
4.5.2. Ablation Study

To verify the effectiveness of each component in our model, we perform ablation study on all three tasks. Table 3 shows the results with the model equipped with different components, where “kg” stands for knowledge enhanced module, “cl” stands for contrastive loss and “dx” means the model are enhanced with diagnosis.

As shown in Table 3, the complete model outperforms all the variants. Models equipped with any component outperforms the baseline model. For most cases, models with two components are better than one components. These results demonstrate that the introduction of external knowledge, diagnosis information, and contrastive loss promotes the capacity to handle EHR data in medication effectiveness prediction.

4.5.3. Attention Visualization

Figure 3 shows the attention between the medication and diagnosis. The X-axis is the drug used in admission, and the Y-axis is the diagnosis. The deeper the color of the cell, the higher the attention score between diagnosis and medication. From the

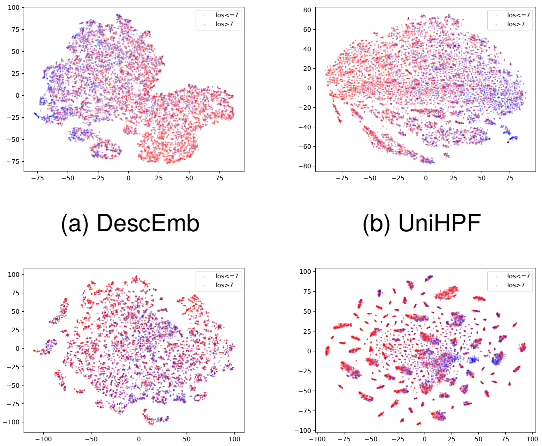


(a) Precision

(b) Recall

(c) F1

Figure 4: Performance with different weighting coefficient α on the mortality prediction task of long-tailed "expired" class. The red horizontal dashed line is the result of α with a strategy that decays with regard to the epoch number.



(a) DescEmb

(b) UniHPF

(c) MedGTX

(d) Ours

Figure 5: Embedding visualizations with t-SNE for different methods. Each dot represents one admission, and the dot color represents the target label for the LOS7 binary classification task.

figure, we can find out that the last cell "Aspirin" gets the highest score in the second row "cellulitis and localized soft tissue infections" since it is one of the antibiotics and is effective for curing infections. It shows our model can figure out which combination of medication is useful in early medication.

4.5.4. Parameter Analysis

Figure 4 shows the prediction results for the long-tailed "expired" class when the weighting coefficient α in Eq. (15) varies from 0.1 to 0.9 on the mortality prediction task. The red horizontal line is the result of α with a strategy that decays with regard to the epoch number. We can see that the best precision appears when α is 0.1, while the worst recall and F1 appears when α is 0.5 and 0.6. When α decays with regard to the epoch number, the recall and F1 is the best. This is probably because a well-trained representation or classifier layer can both achieve better results, since when $\alpha > 0.5$, it focuses on representation learning, while when $\alpha < 0.5$, it focuses on classifier learning. The decay of α helps the model learn better universal features first, then learn robust classifier. Therefore, the adopted α decay strategy can benefit the learning process.

4.5.5. Embedding Visualization

Figure 5 visualizes the learned embeddings of different methods. We take the last layer before the classification layer as input and use the t-SNE algorithm (van der Maaten and Hinton, 2008) to transform d-dim embedding to two dimensions. Each dot represents one admission, and the color of the dot represents the target label for the binary length-of-stay classification task. Compared with other baselines, the embeddings of different classes generated by our method are well clustered and separated, therefore easier to be classified. This shows that our model learns better embedding for each admission, thus can provide accurate prediction for medication effectiveness.

5. Conclusion

In this paper, we proposed to predict the effectiveness of first 24 hours' medication plan for patients. This is especially useful for patients with serious illness who needs immediate treatment. In modeling, we used a knowledge-enhanced module to incorporate external knowledge and a medication feature learning module to find the interaction between diagnosis and medication. To alleviate the data imbalance problem, the model is optimized with a contrastive loss to learn better features and a cross-entropy loss for classifier learning. Extensive experimental results on the eICU dataset demonstrated that our model could significantly outperform state-of-the-art methods. The proposed model can be applied to evaluate the medication plan for emergent patients and assists doctors in establishing a quick and probable treatment.

Ethics Statement

This study is primarily dedicated to research and is not intended to offer medical advice. The medical information utilized in this study is sourced from an open-access medical database. It is important to note that the accuracy of responses generated by the model cannot be guaranteed, and the medical knowledge utilized therein should not be construed as a substitute for professional medical advice.

Ziyad Alharbi, Andrzej Piątkowski, Rolf Dembinski, Sven Reckort, Gerrit Grieb, Jens Kauczok, and Norbert Pallua. 2012. Treatment of burns in the first 24 hours: simple and practical guide by answering 10 questions in a step-by-step form. *World Journal of Emergency Surgery*, 7(13).

Akram Alyass, Michelle Turcotte, and David Meyre. 2015. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics*, 8(33).

Adam S. Brown, Danielle Rasooly, and Chirag J. Patel. 2017. Leveraging population-based clinical quantitative phenotyping for drug repositioning. *CPT: Pharmacometrics & Systems Pharmacology*, 7:124 – 129.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016a. Doctor AI: predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning in Health Care*, volume 56 of *JMLR Workshop and Conference Proceedings*, pages 301–318. JMLR.org.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter F. Stewart. 2016b. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems 29*, pages 3504–3512.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew M. Dai. 2020. Learning the graphical structure of electronic health records with graph convolutional Transformer. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 606–613. AAAI Press.

Alexandre de Matos Soeiro, Pedro Gabriel Melo de Barros e Silva, Eduardo Alberto de Castro Roque, Aline Siqueira Bossa, Cindel Nogueira Zullino, Sheila Aparecida Simões, Mariana Yumi Okada, Tatiana de Carvalho Andreucci Torres Leal, Maria Carolina Feres de Almeida Soeiro, Carlos V Serrano, and Mucio Tavares de Oliveira. 2016. Mortality reduction with use of oral beta-blockers in patients with acute coronary syndrome. *Clinics*, 71:635 – 638.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*, pages 4171–4186. Association for Computational Linguistics.

Sanjoy Dey, Ping Zhang, Daby Sow, and Kenney Ng. 2019. PerDREP: Personalized drug effectiveness prediction from longitudinal observational data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1258–1268. ACM.

Takuma Ebisu and Ryutaro Ichise. 2018. Toruse: Knowledge graph embedding on a lie group. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1819–1826. AAAI Press.

Mohamed F. Ghalwash, Ying Li, Ping Zhang, and Jianying Hu. 2017. Exploiting electronic health records to mine drug effects on laboratory test results. In *Proceedings of the ACM on Conference on Information and Knowledge Management*, pages 1837–1846. ACM.

Kyunghoon Hur, Jiyoung Lee, Jungwoo Oh, Wesley Price, Younghak Kim, and Edward Choi. 2022a. Unifying heterogeneous electronic health records systems via text-based code embedding. In *Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 183–203. PMLR.

Kyunghoon Hur, Jungwoo Oh, Junu Kim, Min Jae Lee, Eunbyeol Cho, Jiyoun Kim, Seong-Eun Moon, Young-Hak Kim, and Edward Choi. 2022b. Unihpf: Universal healthcare predictive framework with zero domain knowledge. *CoRR*, abs/2207.09858.

Marianne Lisby, Lars Peter Nielsen, and Jan Mainz. 2005. Errors in the medication process: frequency, type, and potential clinical consequences. *International journal for quality in health care*, 17 1:15–22.

Chao Liu, Xiaoli Liu, Zhi Mao, Pan Hu, Xiaoming Li, Jie Hu, Quan Hong, Xiaodong Geng, Kun Chi, Feihu Zhou, Guangyan Cai, Xiangmei Chen, and Xuefeng Sun. 2021. Interpretable machine learning model for early prediction of mortality in ICU patients with rhabdomyolysis. *Medicine & Science in Sports & Exercise*, 53:1826 – 1834.

Heng Luo, Ping Zhang, Xianglian Cao, Dizheng Du, Hao Ye, Hui Huang, Can Li, Shengying Qin, Chunling Wan, Leming Shi, Lin He, and Lun Yang. 2016. DPDR-CPI, a server that predicts drug positioning and drug repositioning via chemical-protein interactome. *Scientific Reports*, 6.

Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. 2022. A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. *CoRR*, abs/2208.10240.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1903–1911. ACM.

Matthew B. A. McDermott, Bret Nestor, Evan Kim, Wancong Zhang, Anna Goldenberg, Peter Szolovits, and Marzyeh Ghassemi. 2021. A comprehensive EHR timeseries pre-training benchmark. In *ACM Conference on Health, Inference, and Learning*, pages 257–278. ACM.

Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2017. Deepr: A convolutional net for medical records. *IEEE Journal of Biomedical and Health Informatics*, 21(1):22–30.

Sungjin Park, Seongsu Bae, Jiho Kim, Tackeun Kim, and Edward Choi. 2022. Graph-text multi-modal pre-training for medical representation learning. In *Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 261–281. PMLR.

Xueping Peng, Guodong Long, Tao Shen, Sen Wang, and Jing Jiang. 2021. Sequential diagnosis prediction with Transformer and ontological representation. *2021 IEEE International Conference on Data Mining*, pages 489–498.

Xueping Peng, Guodong Long, Tao Shen, Sen Wang, Jing Jiang, and Chengqi Zhang. 2020. BiteNet: Bidirectional temporal encoder network to predict medical outcomes. In *20th IEEE International Conference on Data Mining*, pages 412–421. IEEE.

Tom J. Pollard, Alistair E. W. Johnson, Jesse Daniel Raffa, Leo Anthony Celi, Roger G. Mark, and Omar Badawi. 2018. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(180178).

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. GAMENet: Graph augmented memory networks for recommending medication combination. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4091–4098. AAAI Press.

Andreas Valentin, Maurizia Capuzzo, Bertrand Guidet, Rui Moreno, Lorenz Dolanski, Peter Bauer, and Philipp G. H. Metnitz. 2006. Patient safety in intensive care: results from the multinational sentinel events evaluation (SEE) study. *Intensive Care Medicine*, 32:1591–1598.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. 2021. Contrastive learning based hybrid networks for long-tailed image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 943–952. Computer Vision Foundation / IEEE.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*, 19-S(3):31–39.

Sheng Yu, Zheng Yuan, Jun Xia, Shengxuan Luo, Huaiyuan Ying, Sihang Zeng, Jingyi Ren, Hongyi Yuan, Zhengyun Zhao, Yucong Lin, Keming Lu, Jing Wang, Yutao Xie, and Heung-Yeung Shum. 2022. BIOS: an algorithmically generated biomedical knowledge graph. *CoRR*, abs/2203.09975.

Weicheng Zhu and Narges Razavian. 2021. Variationally regularized graph-based representation learning for electronic health records. In *ACM Conference on Health, Inference, and Learning*, pages 1–13. ACM.

Jack E. Zimmerman, Andrew A. Kramer, Douglas S. McNair, and Fern M. Malila. 2006. Acute physiology and chronic health evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Critical Care Medicine*, 34:1297–1310.