Medical-Knowledge-Based Graph Neural Network for Medication Combination Prediction

Chao Gao, Shu Yin, Haiqiang Wang, Zhen Wang, Zhanwei Du, and Xuelong Li, *Fellow, IEEE*

***Abstract*— Medication combination prediction (MCP) can provide assistance for experts in the more thorough comprehension of complex mechanisms behind health and disease. Many recent studies focus on the patient representation from the historical medical records, but neglect the value of the medical knowledge, such as the prior knowledge and the medication knowledge. This article develops a medical-knowledge-based graph neural network (MK-GNN) model which incorporates the representation of patients and the medical knowledge into the neural network. More specifically, the features of patients are extracted from their medical records in different feature subspaces. Then these features are concatenated to obtain the feature representation of patients. The prior knowledge, which is calculated according to the mapping relationship between medications and diagnoses, provides heuristic medication features according to the diagnosis results. Such medication features can help the MK-GNN model learn optimal parameters. Moreover, the medication relationship in prescriptions is formulated as a drug network to integrate the medication knowledge into medication representation vectors. The results reveal the superior performance of the MK-GNN model compared with the state-of-the-art baselines on different evaluation metrics. The case study manifests the application potential of the MK-GNN model.**

***Index Terms*— Heuristic medication features, medical knowledge, medication combination prediction (MCP), patient representation.**

I. Introduction

COMBINATION therapies have received great attention for the treatment of complex diseases [1], [2], such as diabetes and asthma. More and more research find that drug combinations have a better therapeutic efficacy compared with a single therapy [3], [4], [5], [6]. As shown in Fig. 1, repaglinide and metformin have synergistic effects in type 2 diabetes mellitus. Metformin resists glycosylation while repaglinide enhances insulin secretion, and the combination of the two drugs enhances the therapeutic effect due to their complementary mechanisms. However, the efficacy of other drug combinations is unknown. The existing medical knowledge about drug medications is not enough to find effective drug medications adequately [7], [8]. Electronic health records (EHRs) generate a mass of health data over time [9], [10]. The EHR data contain an amount of medical knowledge, which can be used in the medication combination prediction (MCP) task to find effective drug combinations. The challenge exists in the patient feature representation and the learning of medical knowledge.

Concerning the patient feature representation, some previous studies take into account the temporal evolution of personal medical records for the MCP. For instance, Wang et al. [11] adopt the gated recurrent units (GRUs) to encode the representation of patients from the diagnosis and treatment procedures. At each timestamp, they use two separate GRUs to learn the diagnoses and treatment procedures, respectively. However, their work does not distinguish the importance of different elements. Choi et al. [12] put forward a two-level attention mechanism on the basis of the recurrent neural network (RNN) to learn from multiple admissions. Such an attention mechanism can capture the significant features that affect the patient's condition, but it does not consider the correlations of different clinical events [13], [14]. These methods just focus on the patient feature representation to complete the MCP task without considering the medical knowledge. In fact, the medical knowledge from EHR which hides the implied drug–drug interaction relationship and accumulated experience of actual clinical treatment is also valuable for MCP. However, it is difficult to transform the complicated prior medical knowledge into a continuous numerical problem because some potential and hidden rules exist in the medical knowledge. It is challenging but necessary to incorporate the drug knowledge into inherent features of patients.

As for another challenge for the MCP, the learning of medical knowledge, it can be divided into the prior knowledge and the drug knowledge. With regard to the prior knowledge, it is from the clinical experience of doctors and is often used in the medical domain by the experienced clinicians. The prior knowledge has gradually been applied in the prediction task of medical domain in recent years [15], [16]. For example, Ma et al. [16] leverage the prior knowledge to predict the disease risk . But the calculation of prior knowledge is still a

Manuscript received 21 January 2022; revised 19 January 2023; accepted 3 April 2023. Date of publication 4 May 2023; date of current version 8 October 2024. This work was supported in part by the National Key Research and Development Program under Grant 2022YFE0112300; in part by the National Natural Science Foundation for Distinguished Young Scholars under Grant 62025602; in part by the National Natural Science Foundation of China under Grant 61976181, Grant 62261136549, and Grant U22B2036; in part by the Technological Innovation Team of Shaanxi Province under Grant 2020TD-013; and in part by the Tencent Foundation and XPLORER PRIZE. *(Corresponding author: Zhen Wang.)*

Chao Gao is with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China, and also with the College of Computer and Information Science, Southwest University, Chongqing 400715, China.

Shu Yin, Zhen Wang, and Xuelong Li are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: zhenwang0@gmail.com).

Haiqiang Wang is with the College of Computer and Information Science, Southwest University, Chongqing 400715, China.

Zhanwei Du is with the School of Public Health, Hong Kong University, Hong Kong.

Digital Object Identifier 10.1109/TNNLS.2023.3266490

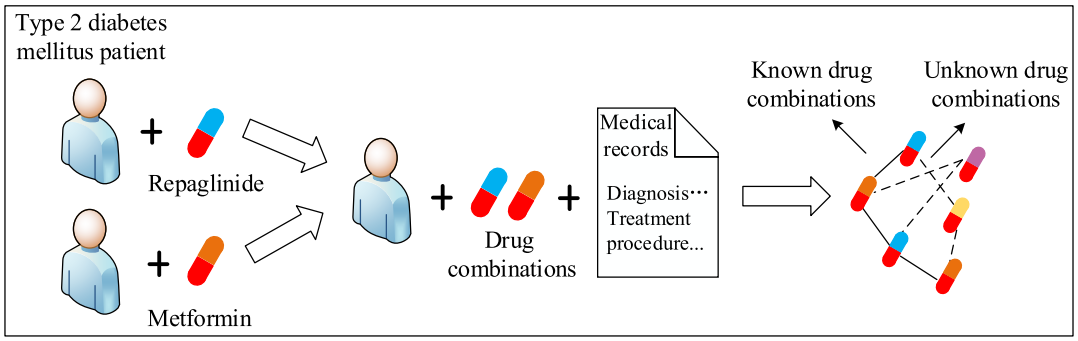

Fig. 1. Treatment example of type 2 diabetes mellitus. The combination of metformin and repaglinide can enhance the therapeutic effect, but the efficacy of other drug combinations is unknown.

challenge. Wang et al. [17] design a constraint rule to calculate heuristic medications via the prior knowledge. However, this article does not take into consideration the drug knowledge, which is hidden in the relationship between medications. Furthermore, it cannot trace the historical feature from the previous personal medical records of patients. In terms of the drug knowledge, it is hidden in drug combinations, which can be translated into the graph-structure data [18]. Specifically, two drugs will be linked with one edge if they are used together. Shang et al. [19] use the representation learning to learn the drug knowledge from a drug network. In clinical treatments, the medical history has a significant impact on the patient's current treatment. However, these methods do not take full advantage of the historical features of diagnoses and treatment procedures. Therefore, how to distinguish the importance of sequence-based EHR data on new diagnoses needs to be addressed.

In this article, we propose a novel model, denoted as medical-knowledge-based graph neural network (MK-GNN), for the MCP task. There are four main stages in MK-GNN. In the first stage, the prior medical knowledge is calculated from the EHR data to represent and map the relationship between the diagnosis and medication. Then the heuristic medication sequence can be obtained on the basis of the latest diagnosis consequences. The multihead attention is applied in the second stage to weigh the different historical treatment information under the sequence-based EHR data, which aims to represent the health condition of patients accurately. Among them, diagnoses and medical procedures are preserved in the personal health records of patients, and prior knowledge refers to the prescribed experience hidden in the EHR data. Attention weights extract the important feature that affects the patient's condition. In the third stage, the relationship between drug combinations is formulated as a drug network, which contains the drug–drug interaction relationship and accumulative drug knowledge in the clinic. The graph convolutional network (GCN) can learn each medication embedding vector that aggregates its neighboring medication information from the drug network. Each medication embedding vector has linked the knowledge between medications to guide the current treatment by further leveraging personal prescriptions. Finally, to trace the historical medication feature, we store the historical patient feature and the historical medication feature in key–value pairs. Therefore, in the process of predicting the medication combination, we can trace the historical feature from the previous personal medical records of patients. The contribution of this article can be generalized as follows.

1) This article designs the extraction rules of prior medical knowledge which match the features of historical patients with features of historical medications to provide the heuristic medication features for MCP.
2) The attention neural network is used to distinguish the importance of the intra-interaction of different historical features aiming to represent the health condition of patients accurately.
3) In addition to the prior knowledge, this article integrates the drug knowledge hidden in drug combinations into the proposed model based on GCN to better provide drug knowledge guidance for the new treatment.

This article is structured as follows. A review of related works is given in Section II. The details of MK-GNN are shown in Section III. In Section IV, we evaluate the MK-GNN model on different metrics and show the experimental result. Finally, the overall effort is concluded in Section V.

II. Related Work

This section briefly reviews the recent studies related to our work. The medical knowledge is summed up in Section II-A. Section II-B and II-C introduce some existing works about the attention mechanism and graph neural network, respectively.

A. Medical Knowledge

With the development of technology, the quality of EHR data improves constantly, which promotes the progress of medical research [20]. Some works have been done for the medical field, such as health risk prediction [21], disease diagnosis [22], and mortality prediction [23], [24]. An amount of medical knowledge is contained in drug combinations of EHR data [17], including the prior knowledge and the drug knowledge. They are able to provide some helpful medical knowledge for clinical decisions.

The prior knowledge is derived from the experience accumulated over time. It is often used in the medical domain by the experienced clinicians. For example, doctors usually prescribe according to their clinical experience which helps them decide which kind of medication is more appropriate. The prior knowledge has been successfully applied in deep

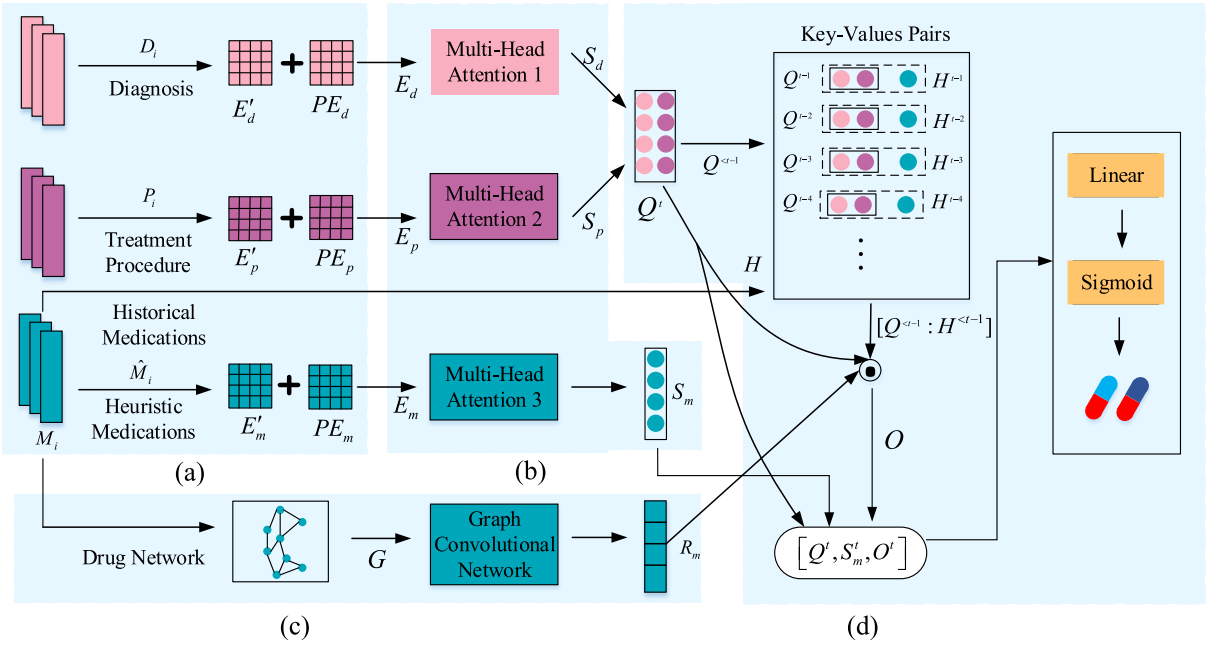

Fig. 2. Illustration of MK-GNN. (a) Medical sequences are converted into linear embedding vectors E'_d, E'_p, and E'_m and positional encoding vectors PE_d, PE_p, and PE_m. (b) Sum of linear embedding vectors and positional encoding vectors (i.e., E_d, E_p, and E_m) are imported into various multihead attention modules for the acquisition of vectors S_d, S_p, and S_m. (c) Drug combinations are formulated as a drug network. The GCN is applied to learn the representation vector of medications R_m, which is used to embed the drug combination in previous prescriptions. (d) Embedding vectors of diagnosis and procedure (i.e., S_d and S_p) are concatenated as Q, and H is the corresponding historical drug combination. Q and H comprise key–value pairs to query the historical features of patients. The patient feature vector Q^t, the queried output vector O^t, and the medication vector S_m^t are concatenated for the MCP.

learning which can assist clinicians, such as the medical image processing [25], [26], mortality prediction [15], and disease risk prediction [16]. As the heuristic feature, the prior knowledge is often used to help model fit parameters so that the model can perform better [27], [28].

The drug knowledge is from drug combinations of prescriptions. The medications used in the past can provide suggestions for the following treatment. With the consideration of medication relationship, it is used in the MCP, such as GAMENet [19] and GATE [13]. The drug combinations in prescriptions can be formulated as a drug network. The drug knowledge can be learned from such medication relationship, which is applied to the representation vectors of medications [19]. Such vectors can be used for predictive tasks in deep learning models.

B. Attention Mechanism

On account of the outstanding performance, the attention mechanism has a wide range of application in a lot of domains, such as the machine translation [29], image processing [30], [31], [32], speech recognition [33], and health care [34], [35], [36], [37], [38]. Due to the interpretability of attention mechanism, it is achievable to find the items that researchers are interested in based on the attention weights.

Recently, the multihead attention has shown a powerful learning ability in the medical domain [39]. For example, He et al. [40] use the multihead attention to complete the multiview learning task. Song et al. [38] use the multihead attention and positional encoding to capture the time-series features for multiple tasks. Wang et al. [41] calculate the correlation of various instances to achieve the single disease classification based on the multihead attention. Each self-attention layer of multihead attention defines the query, the key, and the value vector to learn the relationship between different elements [40]. It aims at learning the important features. Then, these features of diverse attention layers are concatenated to obtain the information in different subspaces.

C. Graph Neural Network

Recently, graph representation learning attracts many researchers [42]. Graph neural network (GNN) has shown great power in learning the representation of graph network in different domains [43], such as social networks [44], knowledge graphs [45], [46], and traffic networks [47]. It can aggregate the node information from neighbors in graph networks [48].

As a variant of GNN, GCNs are proposed to learn the representation of nodes. In graph networks, each node is related to its neighbors by edges [49]. The representation vector of each node should contain its own attributes and the features of its neighbors [18]. The GCN can be used to calculate the sum of weight from its neighbors. Nodes are updated together to achieve the forward propagation layer [13]. Each node disseminates the information of the whole graph network instead of the node attributes of its own only. Therefore, the GCN can be used to learn the drug knowledge from the drug network. For example, Shang et al. [19] use the GCN to model the interaction relationships between medications. Wang et al. [50] integrate the feature of medication relationship into the medication representation vector based on GCN. It is useful to learn the representation vectors of medications with GCN from the drug network [51]. Such vectors aggregate the knowledge

from the medication relationships, for which they can be used to embed the prescriptions of patients. Nevertheless, it does not consider the other knowledge that may also affect the result, such as the prior knowledge.

III. Framework of MK-GNN

In this section, how the MK-GNN model works in predicting medication combination will be introduced in detail. The framework of MK-GNN is shown in Fig. 2. In Section III-A, we illustrate the calculation part of embedding vectors (i.e., E_d, E_p, and E_m). The sequence data (i.e., diagnoses, treatment procedures, and heuristic medications) are embedded by means of linear embedding and positional encoding. Section III-B introduces the patient feature learning part (i.e., S_d, S_p, and S_m). The multihead attention module obtains the features of medical data from the embedding vectors. Section III-C describes how to represent medication vectors P from a drug network G, which is the medication embedding representation part. In the end, Section III-D elaborates the MCP part, including the storage and trace of historical patient features (i.e., $[Q : H]$).

A. Embedding of Medical Information

In this section, we will introduce the calculation part of embedding vectors in detail. We use D_i, P_i, and M_i to denote diagnosis sequence, treatment procedure sequence, and medication sequence, respectively. To obtain the prior medical knowledge, we extract it from the EHR data based on the relationship between the diagnosis and the medication. As shown in Fig. 3, we calculate heuristic medications according to the frequency of the used medication for a certain diagnosis. The high frequency of medications for a diagnosis may indicate the effectiveness of such medications. However, due to the difference in the probability of infection with different diseases, there is also a difference in used medications. Therefore, we balance such a difference based on the following:

$$\hat{M} = \begin{cases} m, & \text{if } f_m \geqslant \eta f_d \\ \text{null}, & \text{if } f_m < \eta f_d \end{cases} \tag{1}$$

where the frequency of diagnosis d and the corresponding medication m are denoted as f_d and f_m, respectively. If $f_m \geqslant \eta f_d$, this part of medications is defined as heuristic medication. It is indicated as $\hat{M} \rightarrow \hat{D}$ to represent the relationship between diagnosis and medication. Otherwise, if $f_m < \eta f_d$, such medications cannot be used as heuristic medications.

As shown in Table I, the diagnosis sequence $D_i \in \mathbb{R}^{L_d^i \times d_d}$ and the treatment procedure sequence $P_i \in \mathbb{R}^{L_p^i \times d_p}$ are inherent attributes, and the prescribed medication sequence M_i is the medication from the personal prescription. But the heuristic medication sequence $\hat{M}_i \in \mathbb{R}^{L_m^i \times d_m}$ is derived from the prior knowledge on the basis of the latest diagnosis results, where L_d^i, L_p^i, and L_m^i are the length of the diagnoses, procedures, and medications of the sample with index i, respectively. And d_d, d_p, and d_m are the standard medical codes with code space size in the MIMIC-III dataset. Note that $L_d^i \neq L_p^i \neq L_m^i$ and $d_d \neq d_p \neq d_m$ generally. As shown in Fig. 2(a), D_i, P_i, and $\hat{M}_i$ are converted into the linear embedding vectors E'^i_d, E'^i_p, and E'^i_m as shown in (2), respectively,

$$\begin{cases} E'^i_d = D_i W_d \\ E'^i_p = P_i W_p \\ E'^i_m = \hat{M}_i W_m \end{cases} \tag{2}$$

where the embedding weight matrices are denoted as $W_d \in \mathbb{R}^{d_d \times d}$, $W_p \in \mathbb{R}^{d_p \times d}$, and $W_m \in \mathbb{R}^{d_m \times d}$. d is the standard embedding which is used to unify the dimension of diagnoses, procedures, and medications. So $E'^i_d \in \mathbb{R}^{L_d^i \times d}$, $E'^i_p \in \mathbb{R}^{L_p^i \times d}$, and $E'^i_m \in \mathbb{R}^{L_m^i \times d}$.

Using the positional encoding, the recurrent features of sequences are acquired, which incorporate the sequence information [39]. For medical sequences, the calculative process of the positional encoding is shown as follows:

$$\begin{cases} \text{PE}(\text{pos}, 2j) = \sin\left(\dfrac{\text{pos}}{1000^{\frac{2j}{d}}}\right) \\ \text{PE}(\text{pos}, 2j+1) = \cos\left(\dfrac{\text{pos}}{1000^{\frac{2j+1}{d}}}\right) \end{cases} \tag{3}$$

where pos denotes the position index of every item in the sequence, and the even ones and the odd ones are expressed as $2j$ and $2j+1$, respectively.

The value of the embedding vector is equal to the addition of the linear embedding vector to the positional encoding vector. The calculative process is formulated as follows:

$$\begin{cases} E_d^i = E'^i_d + PE_d^i \in \mathbb{R}^{L_d^i \times d} \\ E_p^i = E'^i_p + PE_p^i \in \mathbb{R}^{L_p^i \times d} \\ E_m^i = E'^i_m + PE_m^i \in \mathbb{R}^{L_m^i \times d} \end{cases}. \tag{4}$$

B. Patient Feature Learning

In the part of patient feature learning, the multihead attention is added to the MK-GNN model so as to obtain the features of patients among the sequence data. The features of each self-attention layer in the multihead attention are concatenated from different subspaces to get the patient feature. A query vector (i.e., q) and a pair of key–value vectors (i.e., k and v) are components of the self-attention mechanism. The dot product of these two vectors is the weight matrix. To eliminate the effectiveness of dot products, all of them are divided by $\sqrt{d_k}$ and then applied to the softmax function to obtain attention weights. The attention vector can be formulated using attention weights and value vectors v as follows:

$$\text{Attention}(q, k, v) = \text{softmax}\left(\frac{qk^T}{\sqrt{d_k}}\right)v. \tag{5}$$

In every self-attention layer, the query q, the key k, and the value v are defined to obtain the inter-sequence feature [40]. The above q, k, and v vectors are linearly projected for n times, and then they are input to the n self-attention heads [39]. The calculation processing is as follows:

$$\begin{cases} q = E_* W_q \\ k = E_* W_k \\ v = E_* W_v \end{cases} \tag{6}$$

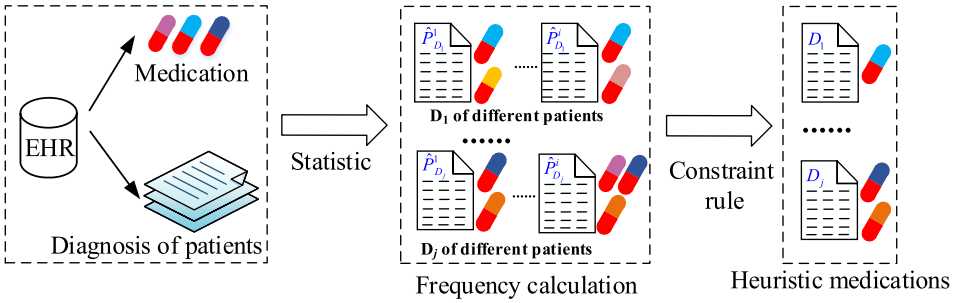

Fig. 3. Calculation of heuristic medications. For different patients $\hat{P}$ in the EHR data, the mapping relationship between medication and diagnosis is calculated to obtain the heuristic medications according to the frequency of the used medication for a certain diagnosis.

TABLE I
MEDICAL RECORDS OF ONE PATIENT WITH THE PRIOR MEDICATION INFORMATION

Visit Time	Diagnoses	Treatment Procedures	Prescribed Medications	Heuristic Medications
1	'4589', '2724', '40391' '2749', 'V4582', '2851' '5856', '99681', 'V4511' '25200', '515', '5533' '53550', '53560', '53085' '2767'	'3995', '4513', '4443'	'N02B', 'A01A', 'A02B' 'A06A', 'A12C', 'N01A' 'C07A', 'N02A', 'N06A' 'B01A', 'C10A', 'N05C' 'A07E', 'D07A', 'D06A' 'V03A', 'J01E', 'H05B'	'N02B', 'A01A', 'A02B' 'A06A', 'B05C', 'A12C' 'C07A', 'B01A', 'C10A'
2	'41071', '4280', '41401' '28521', '2749', '5856' '99681', '25200', '2767' '42833', '40311', '27800' 'V1005', '2811', 'E8780'	'3995'	'N02B', 'A01A', 'A06A' 'A12C', 'C07A', 'N06A' 'A02A', 'B01A', 'C10A' 'C01D', 'R03A', 'D07A' 'V03A', 'H05B', 'M04A'	'N02B', 'A01A', 'A02B' 'A06A', 'B05C', 'A12C' 'C07A', 'C03C', 'B01A' 'C10A', 'V03A'
3	'40391', '2749', 'V4582' '99673', '5829'	'3950', '3995', '9910' '3942', '8849'	'N02B', 'A12C', 'C07A' 'B01A', 'B03B', 'V03A'	'N02B', 'A01A', 'A02B' 'A06A', 'B05C', 'A12C' 'C07A', 'B01A', 'C10A'

where $W_q \in \mathbb{R}^{d\times d_q}$, $W_k \in \mathbb{R}^{d\times d_k}$ and $W_v \in \mathbb{R}^{d\times d_v}$ are the linear translated parameter matrices. d_q, d_k, and d_v are the respective embedding of query, key, and value, respectively. d is the output embedding of query, key, and value. The embedding vector of medical sequences is denoted as $E_*\in \mathbb{R}^{L_*\times d}$. Then $q \in \mathbb{R}^{L_*\times d_q}$, $k \in \mathbb{R}^{L_*\times d_k}$, and $v \in \mathbb{R}^{L_*\times d_v}$.

We use q_h, k_h, and v_h to denote the vectors computing from the hth self-attention head based on (5). The features of medical sequences are represented by the combination of vectors from n self-attention heads. As shown in Fig. 2(b), different multihead attention modules capture the patient features from diagnoses, treatment procedures, and heuristic medications, respectively. Attention vectors are computed as follows:

$$\begin{aligned} \mathrm{S}_* &= \mathrm{MultiHead}(q, k, v) \\ &= \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_n) \\ \text{where } \mathrm{head}_h &= \mathrm{Attention}(q_h, k_h, v_h) \end{aligned} \tag{7}$$

where S_* represents the encoding vectors of diagnosis S_d, the treatment procedure S_p, and the heuristic medication S_m.

C. Medication Embedding Representation

Drug combinations are efficient in the treatment of patients. However, the relationships between medications are complex and multiple. Thus, it is hard to exhibit the medication correlation from thousands of drug combinations. In this article, we formulate the relationships between drug combinations as the edges between nodes. As shown in Fig. 4, drug combinations in the EHR data are translated into a drug network $G = (\theta, \varepsilon)$, where θ and ε represent nodes and edges, respectively. Every node denotes a kind of medication, and the presence of the edge between two nodes means these two medications have been used for the treatment of patients in the clinical treatment. The drug knowledge is learned from a drug network using GCN to obtain the embedding vector of each medication. Following the research of Kipf et al. [52], the adjacency matrix A is normalized as follows:

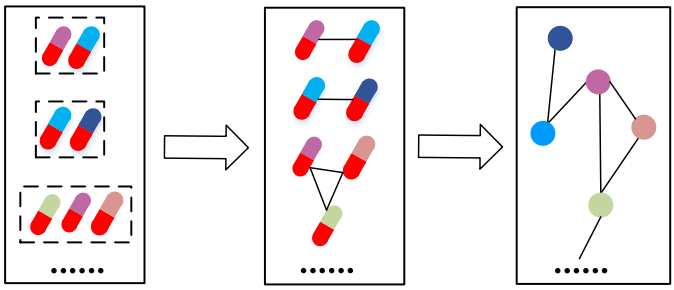

Fig. 4. Translation of drug combinations. The drug network is formulated based on the drug relationship between drug combinations.

$$\hat{A} = D_e{}^{-\frac{1}{2}}(A + I)D_e{}^{-\frac{1}{2}} \tag{8}$$

where the diagonal matrix D_e and adjacency matrix A represent the relationship between medications. The identity matrix I shows the one-hot encoder feature.

Based on the normalized matrix $\hat{A}$, two graph convolutional layers are used to learn the medication representation vectors from the drug network. On each layer, the convolutional computation is as follows:

$$R_m = \hat{A}\tanh(\hat{A}(W_{e1}M))W_{e2} \tag{9}$$

where $W_{e1} \in \mathbb{R}^{|M|\times d_r}$ and $W_{e2} \in \mathbb{R}^{d_r \times d_r}$ denote the medication embedding matrix and the hidden weight parameter matrix, respectively. M is the aforementioned medication sequence as shown in Section III-A. $\hat{A}$ is the normalized matrix from (8).

D. Medication Combination Prediction

Based on the above three stages, we use the extracted feature and knowledge (i.e., the storage and trace of historical patient features) to predict the appropriate drug combination. The features of patients are based on the inherent attributes to represent the health condition, such as the diagnoses and treatment procedures. The medical knowledge contains the prior knowledge and the drug knowledge. The prior knowledge is calculated from the EHR data based on the relationship between the diagnosis and the medication. More specifically, in terms of one diagnosis, the more frequently the medication appears, the more likely it is suitable for the treatment of this diagnosis [16]. As for the historical medications H in Fig. 2(d), it is used to embed the personal prescriptions into the MK-GNN. Embedding vectors of diagnoses and treatment procedures (i.e., S_d and S_p) are concatenated as Q to represent the inherent attribute feature of patients. As shown in Fig. 2(d), Q and H comprise key–value pairs $[Q : H]$ for the retrieval of historical records. The output vector of historical medication embedding is calculated in the following:

$$O^t = \text{softmax}\left(Q^t\left(Q^{<t-1}\right)^T\right)H^{<t-1}R_m \tag{10}$$

where the current health condition Q^t is matched with the historical health condition $Q^{<t-1}$ to retrieve the significant medication feature from the drug knowledge $H^{<t-1}$. Then the final output vector O^t can be obtained from the personal prescription feature vector R_m.

In Fig. 2(d), for each patient at the tth visit, the inherent attribute feature vector Q^t, the feature vector of heuristic medications from the prior knowledge S_m^t. and the output vector O^t are concatenated to obtain the features of patients and prescriptions. The concatenated vectors $[Q^t, S_m^t, O^t]$ are applied to predict the drug combinations based on the following:

$$\hat{y}^t = \text{sigmoid}\left(\text{Linear}\left[Q^t, S_m^t, O^t\right]\right). \tag{11}$$

To ensure the correct score in predicting the drug combinations, the binary cross-entropy loss $\mathcal{L}_{\text{binary}}$ and multilabel margin loss $\mathcal{L}_{\text{multi}}$ are constituted to denote the loss function $\mathcal{L}$, which is represented as follows:

$$\begin{cases} \mathcal{L}_{\text{binary}} = -\sum_{t}^{T}\sum_{i}^{|y|} y_i^t \log\sigma\left(\hat{y}_i^t\right) + \left(1-y_i^t\right)\log\left(1-\sigma\left(\hat{y}_i^t\right)\right) \\ \mathcal{L}_{\text{multi}} = \frac{1}{L}\sum_{t}^{T}\sum_{i}^{|y|}\sum_{j}^{\left|\hat{y}^t\right|}\max\left(0, 1-\left(pos(\hat{y}_i^t) - \text{pos}\left(\hat{y}_j^t\right)\right)\right) \\ \mathcal{L} = \lambda\mathcal{L}_{\text{binary}} + \gamma\mathcal{L}_{\text{multi}} \end{cases} \tag{12}$$

where the amount of real medications y and predictive medications $\hat{y}$ is denoted as $|y|$ and $|\hat{y}^t|$, respectively. T denotes the total treatment times of the patient, and L means the extent of medication labels. For the t^{th} visit time, $\text{pos}(\hat{y}_i^t)$ and $\text{pos}(\hat{y}_j^t)$ are the predictive medication positions of the real medication set and predictive medication set, respectively. λ and γ are parameters set to adjust the weight of the binary cross-entropy loss and multilabel margin loss, respectively. Algorithm 1 illustrates the MK-GNN model in detail.

Algorithm 1 MK-GNN

Input: Diagnosis D, Treatment Procedures P, drug network G, Training Epoch ep;
Output: Medications $\hat{y}$;

```
while ep do
    Calculate heuristic medications based on (1) ;
    Calculate embedding vectors E_d, E_p, and E_m based on (2)–(4);
    Normalize the adjacency matrix A based on (8);
    Learn medication embedding vectors based on (9);
    for ith patient do
        for tth admission time do
            Initialize the vectors of query q, key k, and value v based on (6);
            Calculate the self-attention vector based on (5);
            Concatenate attention vectors from different heads based on (7);
            Obtain the output vector of historical medication embedding O based on (10);
            Predict medications ŷ based on (11);
    Update the combined loss function L based on (12);
    Update ep = ep − 1
```

IV. Experiments

A. Dataset

Our experiment is based on a real medical dataset, MIMIC-III dataset[1], including the medical records of patients who stay in the intensive care units (ICUs) of the Beth Israel Deaconess Medical Center from 2001 to 2012 [53]. Clinical events in this dataset are used in our work, which contain

[1]https://mimic.physionet.org/

TABLE II
STATISTICAL INFORMATION OF DATA

Statistical Items	Amount
diagnoses	1958
treatment procedures	1426
medications	145
patients (multi-visit)	6350
clinical events	15016
average of diagnoses	10.51
average of treatment procedures	3.84
average of medications	8.80
average of admission times	2.36

the main information of patients, such as diagnoses, treatment procedures, and medications. The standard ICD-codes [54] are used for encoding diagnoses and treatment procedures, while NDC codes [55] for encoding medications. We follow the same strategy described in [19] to preprocess the data. More specifically, we only focus on the patients who have multiple visits to reduce the influence of occasionality. The NDC codes are mapped to the third-level ATC codes to label medications. The statistical information of the dataset is listed in Table II.

B. Experiment Setup and Evaluation Metrics

The MIMIC-III dataset is split into three sets, namely, the training set for 2/3, the validation set for 1/6, and the test set for 1/6. Two convolutional layers with 64 neural units on each are applied to learn the drug network. The MK-GNN model is trained for 15 epochs with the learning rate of 0.0002. The experiments are carried out using the PyTorch deep learning framework on Windows 10 with 8-GB memory and I5-CPU. The source codes of MK-GNN and the datasets are released in Github [2].

We use the frequently used metrics to evaluate the predictive accuracy, such as Jaccard similarity score (shorted as Jaccard) [19], [56], average F1 (denoted as F1) [19], and precision–recall AUC (named as PRAUC) [19]. The Jaccard is the ratio of the intersection size and the union size from the true medication set y_i^t and the predictive medication set $\hat{y}_i^t$. The PRAUC is the area under the precision–recall curve, while F1 is computed based on the precision and recall. The formulated definition is displayed in the following:

$$\begin{cases} \text{Jaccard} = \dfrac{1}{\sum_i^N \sum_t^{T_i} 1} \sum_i^N \sum_t^{T_i} \dfrac{\left|y_i^t \cap \hat{y}_i^t\right|}{\left|y_i^t \cup \hat{y}_i^t\right|} \\ \text{Precision} = \dfrac{\left|y_i^t \cap \hat{y}_i^t\right|}{\left|\hat{y}_i^t\right|}, \qquad \text{Recall} = \dfrac{\left|y_i^t \cap \hat{y}_i^t\right|}{\left|y_i^t\right|} \\ \text{Fl} = \dfrac{1}{\sum_i^N \sum_t^{T_i} 1} \sum_i^N \sum_t^{T_i} \dfrac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{cases} \tag{13}$$

where N represents the amount of patients, t stands for the current admission time, i stands for the sequence number of the patient, and T_i denotes the sum of admission times of the ith patient. $\hat{y}_i^t$ and y_i^t denote the predictive medication and the true medication in prescriptions, respectively. *Precision* is the rate of positive medication combination in predicted results. *Recall* is the rate of positive medication combination that is predicted successfully.

[2]https://github.com/cgao-comp/MK-GNN

C. Results

We conduct the comparative analysis of the results of the MK-GNN model and the baseline models evaluated by Jaccard, F1 and PRAUC. Fig. 5 shows the predictive precision of the MK-GNN model and the baselines assessed by three evaluation metrics.

As shown in Fig. 5, it is obvious that the MK-GNN model has the best precision in every evaluation metric. In Fig. 5(a), the precision of the MK-GNN model increases by a maximum of 14.4% and a minimum of 0.52%. Similarly, the improved range of F1 and PRAUC is 0.44%–13.93% and 0.28%–21.5% in Fig. 5(b) and (c), respectively. In contrast to the attention models based on the RNN (i.e., TAMSGC [50], RETAIN [12], LEAP [56], SARMR [27], and SafeDrug [28]), the models based on the multihead attention (i.e., AMANet [40] and our proposed MK-GNN) have a better performance. This may be due to that the multihead attention can capture relevant information on different subspaces by computing multiple times, thereby extracting more noteworthy historical diagnostic and treatment features. However, different from the AMANet model, our proposed MK-GNN model further takes into account the medical knowledge about prescriptions. The comparison results show that the utilization of medical knowledge can improve the predictive precision of drug combinations. As a kind of medical knowledge in our research, the prior knowledge is from the EHR data based on the relationship between the diagnosis and the medication. Therefore, it can guide the MK-GNN model to learn better optimal parameters and improve the predictive performance. Different from the medical knowledge used in the PKANet model [17], the MK-GNN model stores the historical patient feature and the historical medication feature in key–value pairs. It can trace the historical feature from the previous personal medical records of patients. Therefore, the MK-GNN model performs better than the PKANet model.

As for the models based on GCN (GAMENet [19], G-BERT [51], TAMSGC [50], and GATE model [13]), they leverage the GCN to learn the medication representation vector, but the TAMSGC model has a better performance compared with the GAMENet model and G-BERT model. This is because the GATE model takes the temporal information of different admissions into consideration. Different from TAMSGC, the GATE further improves the predictive accuracy via the graph-attention augmented mechanism. Ulteriorly, the predictive precision of the ARMR model [11] surpasses the GATE model due to the application of memory network besides the temporal information. However, these baselines use RNNs to represent the features of patients. Although the RNN can capture the temporal information with the recurrent structure, the whole admission features are easily lost. The next layer of RNN is calculated from its previous layer. The initial features of admission do not catch enough attention. Therefore, these baselines perform worse than the AMANet model [40] and our proposed MK-GNN model with the

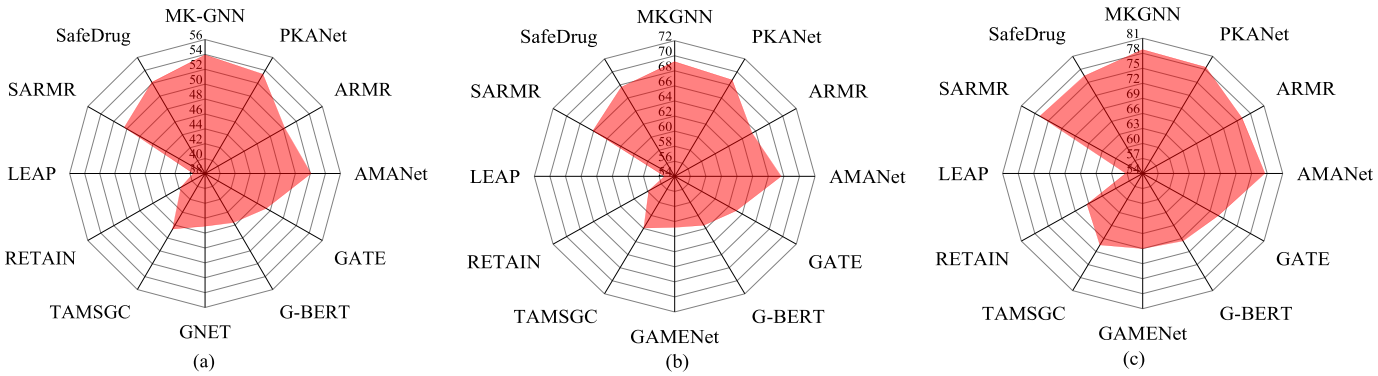

Fig. 5. Comparative analysis of predictive accuracy. The proposed MK-GNN has improved the predictive accuracy by 0.52%–14.4%, 0.44%–13.93%, and 0.28%–21.5% on (a) Jaccard, (b) F1, and (c) PRAUC, respectively, compared with baselines.

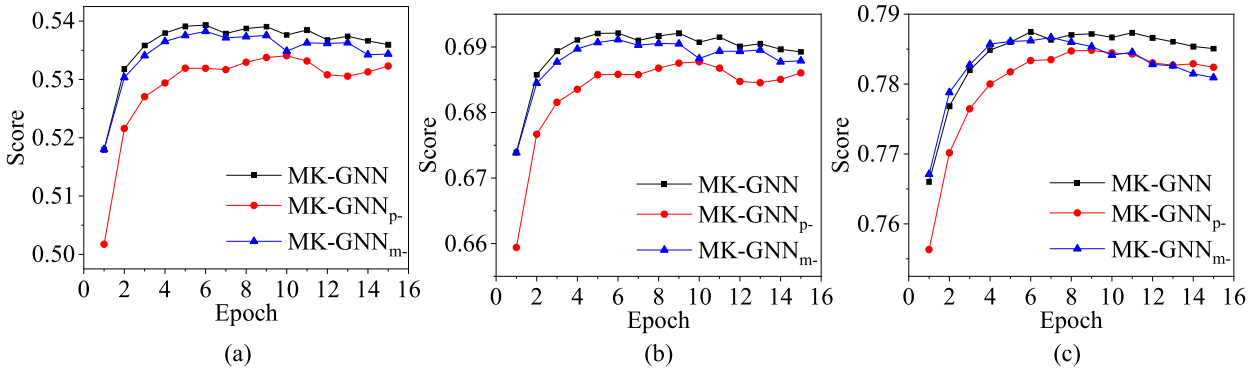

Fig. 6. Effectiveness of prior knowledge and drug knowledge. On the metrics of (a) Jaccard, (b) F1, and (c) PRAUC, the performance of MK-GNN is obviously better than that without the prior knowledge or drug knowledge.

TABLE III
COMPARISON OF STATISTICAL INFORMATION

Statistical Items	Original Quantity	Quantity after Processing	Percentage
patients	6350	343	5.4%
clinical events	15016	380	2.5%
diagnoses	1958	1261	64.4%
treatment procedures	1426	437	30.6%
medications	145	124	85.5%

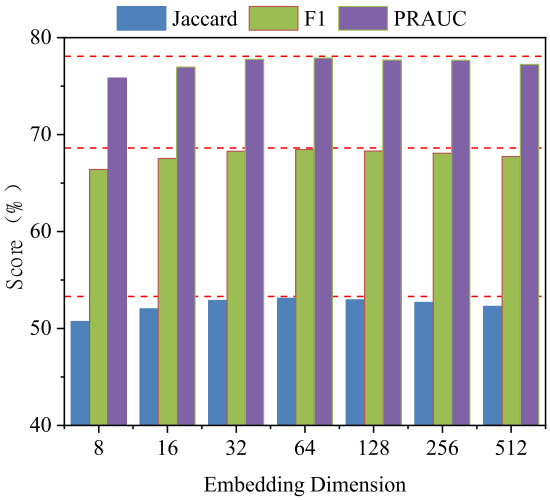

Fig. 7. Predictive precision of embedding dimension d on different evaluation metrics.

usage of the multihead attention. The multihead attention can distinguish the importance of different features and capture the global important feature from different subspaces. Aiming to acquire the sequence information, the position encoding is used to embed the medical data before inputting it into the multihead attention [39]. Therefore, the multihead attention can learn the whole admission features of patients to improve the predictive accuracy.

On the whole, compared with baselines, the improvement of the MK-GNN model on Jaccard, F1, and PRAUC is 0.52%–14.4%, 0.44%–13.93%, and 0.28%–21.5%, respectively. The MK-GNN model outperforms the state-of-the-art baselines on various evaluation metrics. The main reason is that MK-GNN designs a novel medical knowledge module to extract prior information from EHR data to correlate diagnosis and drug treatment and uses the multihead attention mechanism to extract more important features of historical patient data from complex multiview medical sequences.

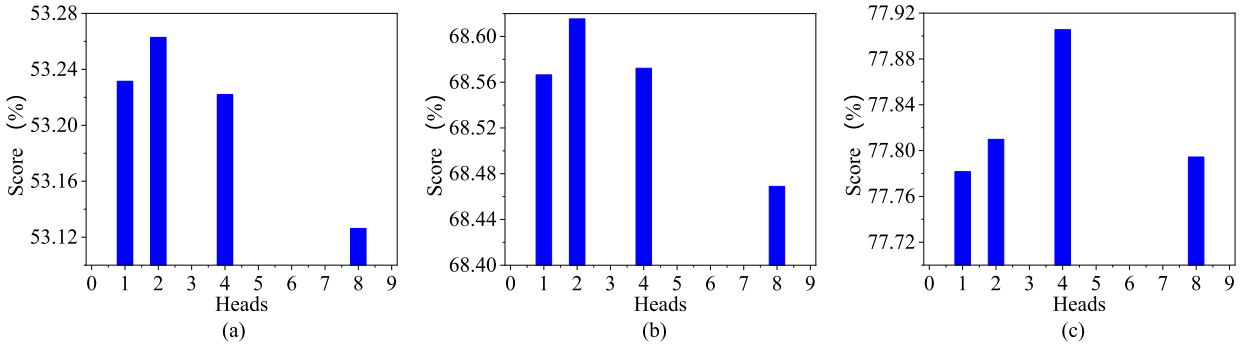

Fig. 8. Predictive precision of attention heads h on different evaluation metrics. (a) Jaccard. (b) F1. (c) PRAUC.

D. Ablation Study

In this article, the prior knowledge and the drug knowledge are applied to the MK-GNN model. To explore their importance, we show the results by removing the prior knowledge (i.e., MK-GNN$_{p-}$) and the drug knowledge (i.e., MK-GNN$_{m-}$).

As illustrated in Fig. 6, we test the change in predictive accuracy on evaluation metrics during the process of training the MK-GNN model. With the usage of prior knowledge, the predictive accuracy of MK-GNN improves 0.36%–1.62%, 0.3%–1.45%, and 0.22%–0.97% on Jaccard, F1, and PRAUC, respectively, compared with MK-GNN$_{p-}$. The reason is that the prior knowledge derives from the clinical experience over decades in the EHR data. Therefore, it can provide heuristic medication features based on the mapping relationship between diagnoses and medications to facilitate model fitting parameters.

Moreover, the performance of MK-GNN$_{m-}$ is also proved to be worse than MK-GNN. The relationship between drug combinations is translated into a drug network, in which GCN aggregates the drug knowledge. The historical personal prescriptions can be embedded into the MK-GNN model based on medication representation vectors learned by GCN. That may be the reason why the MK-GNN can display a relatively superior performance.

E. Parameter Analysis

In this section, we examine the embedding dimension d and the amount h of attention heads on the performance of the proposed MK-GNN model. We fix the others when studying one of the parameters.

In Fig. 7, we explore the influence of embedding dimension d by varying its value from 8 to 512. The predictive scores are increasing when the embedding dimension d is varying from 8 to 64. However, this trend turns to be opposite as the embedding dimension d continues increasing. That is to say, the predictive precision begins to decline when the embedding dimension is greater than 64. The fact indicates that the performance can be improved with a proper d which captures the features of patients, while an extremely large value of d results in overfitting.

The number of attention heads h can affect the predictive precision as shown in Fig. 8. The multihead attention applies the attention in d/h subspaces, where the embedding dimension d is 64 as analyzed in Fig. 7. In Fig. 8(a) and (b), two heads make the MK-GNN model achieve the best performance, while four heads make the MK-GNN model perform best in Fig. 8(c) PRAUC. The difference may be that the PRAUC relies on the precision–recall curve. Therefore, it is less sensitive than Jaccard and F1. In general, $h = 2$ can ensure that the MK-GNN perform better on different evaluation metrics to the maximum extent. In our case, the predictive precision will decline if heads are too many. It may due to the small d in our problem, which makes it not suitable to divide into smaller subspaces.

TABLE IV
ANALYSIS OF PREDICTIVE RESULTS

	Correct Medication	Unseen Medication
Percentage	70.10%	29.9%

F. Case Study

To verify the applicability of the MK-GNN model, we select the personal medical records that involve the diagnosis results appearing less than five times in the EHR data. In Table III, the statistical results show that these patients account for 5.4%, and the training data of clinical events cover only 2.5% of the original data.

The MK-GNN model is trained with the selected clinical events. We define the correct medication as one that is in both prescribed medications and predictive medications, the unseen medication as the one that is only in the prescribed medications. In Table IV, we find that the correct medication accounts for 70.10% on average. It proves that the MK-GNN can predict the most parts of drug combinations for these patients. In terms of the unseen medication, it takes up 29.9%. The reason may be that the amount of training data is not sufficient to avoid the misclassification of medications.

Three tested cases are listed in Table V, and the differences between the real drug combinations and the predictive drug combinations are emphasized with the bold font. As shown in Table V, the MK-GNN can predict most medications of three groups, and the correct medications account for 12/16, 21/29, and 21/26, respectively. However, a few real medications are incorrect or unseen in the predictive drug combinations. For

TABLE V
CASE STUDY ANALYSIS

Group	Diagnoses	Real Drug Combinations	Predictive Drug Combinations	Comparison Result
1	'2724', '4280', '41401' '25000', '4439', 'V4582' '5180', '3004', '4240' 'E8781', '79001', '28860' 'V103', '42823', '41072' 'E8889', 'V4364', '92232' '40291'	'N02B', 'A01A', 'A02B' 'A06A', 'B05C', 'C07A' 'C03C', 'A12B', 'N02A' 'N06A', 'B01A', 'C10A' **'C01B'**, **'N05C'**, **'C09A'** **'H04A'**	'N02B', 'A01A', 'A02B' 'A06A', 'B05C', **'A12C'** **'A07A'**, 'C07A', 'C03C' 'A12B', 'N02A', 'N06A' 'B01A', 'C10A', **'A04A'**	Correct: 12/16, Incorrect: 3/15, Unseen: 4/16
2	'4019', 'E8798', '07054' '99731', '04185', '30550' '8080', '8082', '80502' 'E8120', '8600', '86504' '86121', '80106'	'N02B', 'A01A', 'A02B' 'A06A', 'B05C', 'A12A' 'A12C', 'A07A', 'N01A' 'C07A', 'C03C', 'A12B' 'C02D', 'N02A', 'A02A' 'J01M', 'B01A', 'N05C' 'J01D', 'N05A', **'A04A'** 'M03A', 'R03A', **'N07B'** **'N05B'**, **'D06A'**, **'A03F'** 'S01E', **'H04A'**	'N02B', 'A01A', 'A02B' 'A06A', 'B05C', 'A12A' 'A12C', 'C01C', 'A07A' 'A10A', 'N01A', 'C07A' 'C03C', 'A12B', 'N02A' 'J01M', 'B01A', 'N05C' 'J01D', **'N03A'**, 'N05A' 'M03A', 'R03A', **'J01C'** 'S01E'	Correct: 21/29, Incorrect: 4/25, Unseen: 8/29
3	'45829', '9971', '2762' '570', '2767', '5570' '48241', '5845', '7850' '8052', '30500', '5718' '72888', '4264', '9584' '86803', '60886', '51851' '27669', '80704', '8602' '86389', 'E8150', 'V140' '8054', '86405', '86121' '80506', '86401', '86801'	'N02B', 'A01A', 'A02B' 'A06A', 'B05C', 'A12A' 'A12C', 'C01C', 'A07A' **'A10A'**, 'N01A', 'C03C' 'A12B', 'N02A', **'A02A'** 'J01M', 'B01A', 'N05C' 'J01D', 'N05A', 'R03A' **'D07A'**, 'D04A', 'S01E' **'H04A'**, **'B05X'**	'N02B', 'A01A', 'A02B' 'A06A', 'B05C', 'A12A' 'A12C', 'C01C', 'A07A' 'N01A', **'C07A'**, 'C03C' 'A12B', 'N02A', 'J01M' 'B01A', 'N05C', 'J01D' 'N05A', **'A04A'**, **'M03A'** 'R03A', 'N05B', **'R01A'** **'C01E'**, 'D04A', 'S01E'	Correct: 21/26, Incorrect: 6/27, Unseen: 5/26

example, in the first group, three medications are incorrect in the predictive drug combinations and four medications of real drug combinations are unseen in the predictive drug combinations. The reason may be that the selected clinical events account for only 2.5% and such small-scale data affect the predictive accuracy of drug combinations. Even so, the MK-GNN model can predict the most parts of drug combinations.

V. CONCLUSION

The clinic guidelines designed by medical experts can be used to cure patients in reality. To assist experts to relieve the workload of data analysis, we propose a novel deep learning model MK-GNN to predict the drug combinations for the treatment of patients. The MK-GNN model is able to learn the features of patients based on the multihead attention. The prior knowledge, which can help the MK-GNN model fit the optimal parameters, is derived from the EHR data based on the relationship between the diagnosis and the medication. With the purpose of obtaining the feature of personal prescriptions, the GCN is used to learn the medication representation vectors that aggregate the drug knowledge from the formulated drug network. Extensive experiments show the performance of the MK-GNN model surpasses the state-of-the-art baselines on various evaluation metrics. The medical data are an indispensable part for experts to comprehend the relationship between diseases and medications. Therefore, we expect our model can act as assistance for experts to analyze the massive amounts of medical data. However, in clinical medicine, due to the influence of external factors such as doctors' clinical experience and medical resources, there are differences in the MCP suggested by different doctors and different regions, which limits the generalization of drug combination prediction. Therefore, in future research, we will study the feature invariance of the drug combination, and then improve the generalization ability of the algorithm to approach the direction of actual clinical application.

REFERENCES

[1] A. Ianevski et al., "Prediction of drug combination effects with a minimal set of experiments," *Nature Mach. Intell.*, vol. 1, no. 12, pp. 568–577, Dec. 2019.

[2] J. Song et al., "Local–global memory neural network for medication prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1723–1736, Apr. 2021.

[3] C. Y. Lee and Y.-P.-P. Chen, "New insights into drug repurposing for COVID-19 using deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4770–4780, Nov. 2021.

[4] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston, "Cancer drug resistance: An evolving paradigm," *Nature Rev. Cancer*, vol. 13, no. 10, pp. 714–726, Oct. 2013.

[5] F. Cheng, I. A. Kovács, and A. L. Barabási, "Network-based prediction of drug combinations," *Nature Commun.*, vol. 10, no. 1, pp. 1–11, Mar. 2019.

[6] W. Jin et al., "Deep learning identifies synergistic drug combinations for treating COVID-19," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 39, pp. 1–7, Sep. 2021.

[7] H. Iwata, R. Sawada, S. Mizutani, M. Kotera, and Y. Yamanishi, "Large-scale prediction of beneficial drug combinations using drug efficacy and target profiles," *J. Chem. Inf. Model.*, vol. 55, no. 12, pp. 2705–2716, Dec. 2015.

[8] C. Yin, R. Zhao, B. Qian, X. Lv, and P. Zhang, "Domain knowledge guided deep learning with electronic health records," in *Proc. 19th IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 738–747.

[9] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Rev. Genet.*, vol. 13, no. 6, pp. 395–405, May 2012.

[10] A. Hoerbst and E. Ammenwerth, "Electronic health records," *Methods Inf. Med.*, vol. 49, no. 4, pp. 320–336, 2010.

[11] Y. Wang, W. Chen, D. Pi, and L. Yue, "Adversarially regularized medication recommendation model with multi-hop memory network," *Knowl. Inf. Syst.*, vol. 63, no. 1, pp. 125–142, Jan. 2021.

[12] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. 30th Conf. Neural Inf. Process. Syst.*, 2016, pp. 3504–3512.

[13] C. Su, S. Gao, and S. Li, "GATE: Graph-attention augmented temporal neural network for medication recommendation," *IEEE Access*, vol. 8, pp. 125447–125458, 2020.

[14] Y. Xu et al., "Time-aware context-gated graph attention network for clinical risk prediction," *IEEE Trans. Knowl. Data Eng.*, early access, Jun. 13, 2022, doi: 10.1109/TKDE.2022.3181780.

[15] L. Xiao, C. Zheng, X. Fan, Y. Xie, and R. Yu, "Predicting ICU mortality from heterogeneous clinical events with prior medical knowledge," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 55–59.

[16] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, "Risk prediction on electronic health records with prior medical knowledge," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1910–1919.

[17] H. Wang, X. Dong, Z. Luo, J. Zhu, P. Zhu, and C. Gao, "Medication combination prediction via attention neural networks with prior medical knowledge," in *Proc. 14th Int. Conf. Knowl. Sci., Eng. Manage.*, 2021, pp. 311–322.

[18] S. Wang, P. Ren, Z. Chen, Z. Ren, J. Ma, and M. De Rijke, "Order-free medicine combination prediction with graph convolutional reinforcement learning," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 1623–1632.

[19] J. Shang, C. Xiao, T. Ma, H. Li, and J. Sun, "GAMENet: Graph augmented memory networks for recommending medication combination," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 1126–1133.

[20] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[21] L. Ma et al., "AdaCare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 825–832.

[22] A. Shamsi et al., "An uncertainty-aware transfer learning-based framework for COVID-19 diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1408–1417, Apr. 2021.

[23] E. Jun, A. W. Mulyadi, J. Choi, and H. I. Suk, "Uncertainty-gated stochastic sequential model for EHR mortality prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 4052–4062, Sep. 2021.

[24] L. Ma et al., "ConCare: Personalized clinical feature embedding via capturing the healthcare context," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 833–840.

[25] V. Grau, A. U. J. Mewes, M. Alcaniz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 447–458, Apr. 2004.

[26] D. Kurrant, A. Baran, J. LoVetri, and E. Fear, "Integrating prior information into microwave tomography Part 1: Impact of detail on image quality," *Med. Phys.*, vol. 44, no. 12, pp. 6461–6481, 2017.

[27] Y. Wang, W. Chen, D. Pi, L. Yue, S. Wang, and M. Xu, "Self-supervised adversarial distribution regularization for medication recommendation," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 3134–3140.

[28] C. Yang, C. Xiao, F. Ma, G. Lucas, and J. Sun, "SafeDrug: Dual molecular graph encoders for recommending effective and safe drug combinations," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 3735–3741.

[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[30] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[31] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. 28th Conf. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[32] Y. Li, X. Yu, and Q. Bao, "Image inpainting algorithm based on neural network and attention mechanism," in *Proc. 2nd Int. Conf. Algorithms, Comput. Artif. Intell.*, 2019, pp. 345–349.

[33] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. 41st IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 4945–4949.

[34] W. Lee, S. Park, W. Joo, and I. C. Moon, "Diagnosis prediction via medical context attention networks using deep generative modeling," in *Proc. 18th IEEE Int. Conf. Data Mining*, 2018, pp. 1104–1109.

[35] H. Eom et al., "End-to-end deep learning architecture for continuous blood pressure estimation using attention mechanism," *Sensors*, vol. 20, no. 8, p. 2338, 2020.

[36] M. Yin, C. Mou, K. Xiong, and J. Ren, "Chinese clinical named entity recognition with radical-level feature and self-attention mechanism," *J. Biomed. Informat.*, vol. 98, May 2019, Art. no. 103289.

[37] D. A. Kaji et al., "An attention based deep learning model of clinical events in the intensive care unit," *PLoS ONE*, vol. 14, no. 2, 2019, Art. no. e0211057.

[38] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4091–4098.

[39] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[40] Y. He, C. Wang, N. Li, and Z. Zeng, "Attention and memory-augmented networks for dual-view sequential learning," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2020, pp. 125–134.

[41] Z. Wang, J. Poon, S. Sun, and S. Poon, "Attention-based multi-instance neural network for medical diagnosis from incomplete and low quality data," in *Proc. 32nd Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[42] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2020.

[43] R. A. Rossi, R. Zhou, and N. K. Ahmed, "Deep inductive graph representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 438–452, Mar. 2018.

[44] C. Gao, J. Zhu, F. Zhang, Z. Wang, and X. Li, "A novel representation learning for dynamic graphs based on graph convolutional networks," *IEEE Trans. Cybern.*, early access, Mar. 25, 2022, doi: 10.1109/TCYB.2022.3159661.

[45] X. Lin, Z. Quan, Z. Wang, T. Ma, and X. Zeng, "KGNN: Knowledge graph neural network for drug-drug interaction prediction," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 2739–2745.

[46] X. Zhao, L. Chen, and H. Chen, "A weighted heterogeneous graph-based dialog system," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 18, 2021, doi: 10.1109/TNNLS.2021.3124640.

[47] C. Chen et al., "Gated residual recurrent graph neural networks for traffic prediction," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 485–492.

[48] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Dec. 2008.

[49] S. Qi, X. Huang, P. Peng, X. Huang, J. Zhang, and X. Wang, "Cascaded attention: Adaptive and gated graph attention network for multiagent reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 10, 2022, doi: 10.1109/TNNLS.2022.3197918.

[50] H. Wang et al., "Medication combination prediction using temporal attention mechanism and simple graph convolution," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3995–4004, May 2021.

[51] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 5953–5959.

[52] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[53] A. E. Johnson et al., "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.

[54] P. Wu et al., "Mapping ICD-10 and ICD-10-CM codes to phecodes: Workflow development and initial evaluation," *JMIR Med. Inf.*, vol. 7, no. 4, 2019, Art. no. e14325.

[55] M. Pahor, E. A. Chrischilles, J. M. Guralnik, S. L. Brown, R. B. Wallace, and P. Carbonin, "Drug data coding and analysis in epidemiologic studies," *Eur. J. Epidemiol.*, vol. 10, no. 4, pp. 405–411, 1994.

[56] Y. Zhang, R. Chen, J. Tang, W. F. Stewart, and J. Sun, "LEAP: Learning to prescribe effective and safe treatment combinations for multimorbidity," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 1315–1324.

Chao Gao received the Ph.D. degree in computer science from the Beijing University of Technology, Beijing, China, in 2010.

He is currently a Professor with the School of Artificial Intelligence, OPtics, and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His main research interests include data-driven complex systems modeling, complex social networks analysis, and nature-inspired computing.

Shu Yin is currently pursuing the Ph.D. degree with the College of Computer Science, Northwestern Polytechnical University, Xi'an, China.

Her research focuses include data-driven modeling and statistical analysis.

Haiqiang Wang received the master's degree from the College of Computer and Information Science, Southwest University, Chongqing, China, in 2022.

His research focuses include data science and network intelligence.

Zhen Wang received the Ph.D. degree from Hong Kong Baptist University, Hong Kong, in 2014.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. He has authored/coauthored more than 100 research articles and four review articles with 16 000+ citations. His current research interests include network science, complex system, evolutionary game theory, and behavioral decision.

Zhanwei Du received the Ph.D. degree from the Department of Computer Science and Technology, Jilin University, Changchun, China, in 2015.

He is currently a Research Assistant Professor with the School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong. His current research interests include complex networks, smart city of traffic networks, and epidemic disease propagation.

Xuelong Li (Fellow, IEEE) is currently a Full Professor with the School of Artificial Intelligence, OPtics and ElectroNics, Northwestern Polytechnical University, Xi'an, China.