Medication Combination Prediction Using Temporal Attention Mechanism and Simple Graph Convolution

Haiqiang Wang, Yinying Wu, Chao Gao, Yue Deng, Fan Zhang, Jiajin Huang, and Jiming Liu, *Fellow, IEEE*

***Abstract*—Medication combination prediction can be applied to the clinical treatment for critical patients with multi-morbidity. The suitable medication combination can help cure patients and keep the treatment medication safe. However, the complexity and uncertainty of clinical circumstances limit the predictive accuracy of medication combination. Thus, this paper proposes a new medication combination prediction model based on the temporal attention mechanism (TAM) and the simple graph convolution (SGC), named as TAMSGC. More specifically, the TAM can capture the temporal sequence information in the medical records, and the SGC is implemented to acquire the medication knowledge from the complicated medication combination. Experiments in a real dataset show that TAMSGC surpasses the baseline models on the predictive accuracy of medication combination.**

***Index Terms*—Medication combination prediction, critical patients, attention mechanism, simple graph convolution, temporal sequence, medical records, medication knowledge.**

I. Introduction

PREDICTING the medication combination of critical patients can help experts improve the quality of guidelines and medical tools. Having made full use of electronic health record (EHR) dataset, deep learning techniques, which have shown powerful predictive capability in the health care domain [1]–[3], provide effective methods for us to predict the medication combination. However, patients with multi-morbidity could be hospitalized several times. Such behaviors result in the sequential medical records and complicated combination of medications in EHR dataset [4], as shown in Fig. 1. These temporal and complicated sequence information bring more difficulties for researchers to predict the medication combination accurately.

As to the temporal sequence information in the EHR dataset, it is caused by the historical diagnosis and treatment procedure records changing with the admission time of patients [5]. Due to such information, it is hard to extract the feature of patients [6], [7], which can improve the predictive accuracy of medication combination. Some models are proposed to capture the temporal sequence information, such as Medi-Care AI [8] and RETAIN [9]. These models obtain the feature hidden in the diagnoses and treatment procedures by exploring the relationship among disparate medical codes [8], [9]. However, patients of intensive care unit (ICU) are always with multi-morbidity. Such condition lowers the learning capacity of the aforementioned models due to their finite capabilities in extracting the significant feature of temporal sequence information from multi-morbidity patients.

In terms of the medication combination, it can help cure the multi-morbidity patients [10]. However, in view of patients with multiple complicated diseases, it is unavoidable that the ability of a doctor to identify effective medication combination is limited. The complex correlation of medications is the main reason that affects the decision of doctors. The safe medication correlations are helpful for the recovery of patients, but harmful correlations can cause polypharmacy side effects. Recently, some studies leverage graph networks to display the medication relationship of medication combination [6], [7], [11]. They attempt to use the graph data-structure to represent the medication correlation hidden in the medication combination. Based on the graph neural networks (GNNs), some models, such as GAMENet [6], CompNet [7] and G-BERT [11], perform well in learning the embedding vector of nodes in medication graph networks [12]. More specifically, each node is initialized with a vector in graphs, and the embedding vector of a medication node is the weighted sum of all its neighbor nodes [6]. Thus, each node aggregates the feature of neighbors. This method can describe the relationship

Manuscript received October 1, 2020; revised April 13, 2021; accepted May 17, 2021. Date of publication May 21, 2021; date of current version October 5, 2021. This work was supported by the National Natural Science Foundation of China under Grants 61976181, 11931015, and in part by the Key Technology Research and Development Program of Science and Technology-Scientific and Technological Innovation Team of Shaanxi Province under Grant 2020TD-013. *(Corresponding author: Chao Gao.)*

Haiqiang Wang, Yue Deng, and Fan Zhang are with the College of Computer and Information Science, Southwest University, Chongqing 400715, China (e-mail: W.HQ2019@outlook.com; 2511509874@qq.com; solo-zhang@foxmail.com).

Yinying Wu is with the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710072, China (e-mail: shadowless_111@163.com).

Chao Gao is with the College of Computer and Information Science, Southwest University, Chongqing 400715, China and also with the School of Artificial Intelligence, Optics, and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: cgao@swu.edu.cn).

Jiajin Huang is with the International WIC Institute, Beijing University of Technology, Beijing 100124, China (e-mail: jhuang@bjut.edu.cn).

Jiming Liu is with the Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong (e-mail: jiming@comp.hkbu.edu.hk).

Digital Object Identifier 10.1109/JBHI.2021.3082548

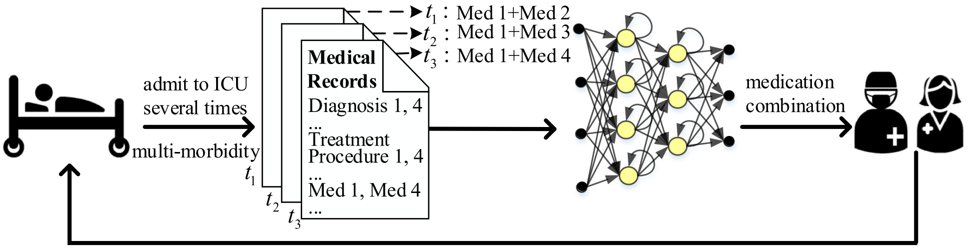


Fig. 1. The sequence information of a patient for deep learning to predict the medication combination. The dynamic medical records change with the different admission time (i.e., t_1, t_2, t_3) of the patient. In order to help experts improve the quality of guidelines and medical tools, medical records are input to the deep learning model for predicting the medication combination.

of medications, and it has been proved to be successful in the drug-drug interaction (DDI) graphs [13]. Nevertheless, the repetition of nonlinearity transformation between graph convolutional layers causes the high computational complexity [14], especially in dense graphs.

This paper focuses on the multi-morbidity patients who have been admitted to the ICU several times. For such patients, the temporal sequence information of medical records can affect the accuracy of predicting the medication combination. In addition, it is hard to extract the medication knowledge due to the medication correlation in the complicated medication combination. Thus, we propose a TAMSGC model based on the attention mechanism and the simple graph convolution (SGC) to improve the predictive accuracy of medication combination and learn the medication knowledge of curing patients, effectively and safely. In our proposed model, a temporal attention mechanism (TAM) is designed to capture the temporal sequence information of diagnoses and treatment procedures in medical records. More specifically, based on the recurrent neural networks (RNNs) [9], the TAM generates an attention parameter on each neural network layer to add weight to diagnoses and treatment procedures in different admission time. Then the TAM leverages overlapping weights to obtain the feature representation vector of patients. In terms of the complex relationship of medication nodes, the SGC can predigest the process of learning the representation of nodes and reduce the computational cost by removing the nonlinearity computation between layers [14]. Experiments verify the effectiveness of this model. The main contributions of this paper can be summarized as follows.

(1) Inspired by the reverse time attention mechanism [9], the TAM is designed to aggregate the temporal sequence information of diagnoses and treatment procedures.

(2) The application of SGC can predigest the process of learning the embedding vector of medication nodes in the complicated medication combination. The knowledge of medication correlation among edges can be combined more easily in the medication graph networks.

The rest of this paper is organized as follows. Section II reviews the related works. Section III proposes the TAMSGC model to predict the medication combination. Section IV evaluates the prediction results. Section V discusses the predictive accuracy and computational complexity based on the ablation study. Finally, Section VI concludes this paper.

II. Related Work

A. Rule-Based Medication Combination Prediction

As a kind of medication combination prediction methods, the rule-based method mainly relies on the clinical experience. It aims at recommending the optimal therapeutic medication according to the health state of patients and assisting doctors to prescribe for patients [15], [16].

Generally speaking, the rule-based method is defined by experts according to their experience or institution guidelines [15]–[17]. For instance, the adaptive treatment strategies (ATS), a sequence of decision rules, are adjusted by assessing the feasibility and acceptability of the treatment of adolescent depression [16]. The British Thoracic Society has made asthma guidelines that are developed through consensus but based on a combination of randomized trials and observational studies [18]. The U.K. National Institute for Health and Care Excellence has set up the guidelines for the prevention of venous thromboembolism after surgery [19]. There are nine drug-drug interaction screening tools to detect clinically relevant interactions for a large sample of oral oncolytics [20]. However, the establishment of rule-based method may consume a lot of energy of experts.

In summary, the rule-based method may not find some rules hidden in the large-scale medical data [21], which has been existed in the medical domain, such as the medical concept extraction [22], disease inference [23], [24] and medication combination prediction [6], [7], [11]. To make up for this shortcoming, this paper proposes a method to predict the medication combination to free experts from the analysis of large-scale medical data.

B. Neural Network Technologies for Medication Combination Prediction

Neural network technologies can be used to extract the knowledge from medical concepts automatically [25], which can compensate for the knowledge gap. Among such technologies, the recurrent neural networks (RNNs) and graph neural networks (GNNs) are widely used to predict the medication combination in recent years.

Concerning the RNN-based models, the recurrent structure of RNN determines that the next hidden state will be acquired based on its previous sequence information [26]. Thus, the RNN has been proved to be effective in learning sequence

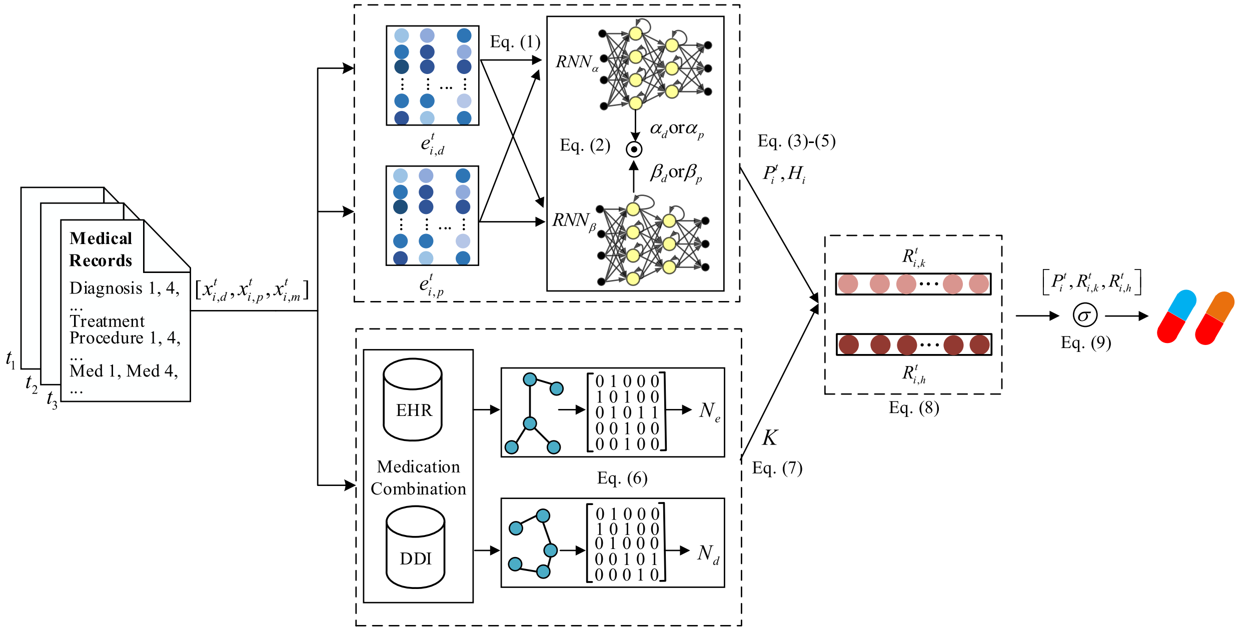


Fig. 2. The graphical illustration of TAMSGC. Firstly, for the i^{th} patient at the t^{th} admission time, multi-hot vectors of diagnoses and treatment procedures (i.e., $x^t_{i,d}$ and $x^t_{i,p}$) are translated into embedding vectors $e^t_{i,d}$ and $e^t_{i,p}$, respectively, based on Eq. (1). Then, RNNs (i.e., RNN_α and RNN_β) are used to generate attention parameters (i.e., α_d, β_d, α_p, β_p) based on Eq. (2). Such attention parameters can extract the feature representation vector of patients and represent their medical history vectors(i.e., P^t_i and H_i) based on Eqs. (3)- (5). Next, SGC can learn the medication knowledge K from the complicated medication combination based on Eqs. (6)- (7). The output vectors (i.e., $R^t_{i,k}$ and $R^t_{i,h}$) are computed for the prediction task based on Eq. (8). Finally, TAMSGC can predict the medication combination for patients based on Eq. (9).

information [27]. For instance, Choi *et al.* propose a temporal model based on the RNN, named Doctor AI, to leverage the time stamp knowledge in medical codes [28]. However, as one of the main missing patterns, the masking is not addressed in their models [29]. To obtain the knowledge hidden in the history of patients, Le *et al.* design a DMNC model based on the RNNs and the differentiable neural computers (DNC) [30]. Nevertheless, the medication correlation is not taken into consideration. The RETAIN model [9] and LEAP model [31] make full use of the advantage of RNN, but they do not learn the knowledge from the medication combination. The attention mechanism has shown great learning ability in many fields, such as image processing [32]–[34], natural language processing [35], [36], speech recognition [37] and health care [38]–[41]. This paper leverages RNNs to design a temporal attention mechanism (TAM), which is composed of a two-layer neural network. The TAM can be used to learn the temporal sequence information hidden in diagnoses and treatment procedures.

As for the GNN-based models, GNNs can learn the embedding vector of nodes in graphs and aggregate information from neighbor nodes in graph networks [12]. For example, Shang *et al.* find that drug-drug interactions (DDIs) might cause side effects [6]. They propose a GAMENet model and integrate the medical knowledge of EHR graphs and DDI graphs based on the graph convolutional networks (GCNs). However, other information of medical ontologies, such as diagnoses and treatment procedures [11], is not extracted sufficiently. Shang *et al.* utilize the pre-training techniques to build a G-BERT model in order to learn the knowledge of medical codes [11], but the problem of polypharmacy side effects is not addressed in their study. Although Zitnik *et al.* [42] and Asada [43] *et al.* achieve the DDI prediction, they do not consider the related correlations of DDIs in the knowledge graph [44]. Wang *et al.* consider the correlations between medicines in an order-free way and propose the CompNet model, but the learning process will be extremely hard and unstable in the large medicine space [7]. In some cases, graph networks tend to be large-scale with massive nodes and edges, such as drug-drug interaction graphs [6], protein-protein interaction graphs [45] and Reddit citation graphs [45]. As a kind of traditional GNNs, the high cost of GCN mainly comes from the massive nonlinearity computation between layers [14]. In this paper, to compensate the shortcoming of GCN, the simple graph convolution (SGC) is considered due to its superiority in predigesting the process of convolution computation.

III. Proposed TAMSGC Model

In this section, we propose a new medication combination prediction model, named TAMSGC based on the temporal attention mechanism (TAM) and the simple graph convolution (SGC). Section III-A describes the acquisition of feature representation vectors of patients using the TAM. Section III-B leverages the SGC to learn the knowledge from the complicated medication combination. Section III-C predicts the medication combination.

As shown in Fig. 2, the TAMSGC model is made up of three parts as follows. The patient representation module leverages diagnoses and treatment procedures to represent feature vectors of patients, which consist of medical history vectors of patients.

TABLE I
AN EXAMPLE OF HEALTH CARE EVENT (THE PATIENT ID IS 00 017)

Diagnosis Codes (d)	Diagnosis Type	Treatment Procedure Codes (p)	Treatment Procedure Type	Medication Codes (m)	Medication Type
4239	Pericardial Disease	3731	Pericardiectomy	51079087101	Sucralfate
5119	Pleural Effusion	8872	Dx ultrasound-heart	55390048101	Ketorolac
78551	Cardiogenic Shock	3893	Venous cath	74125901	Morphine Sulfate
4589	Hypotension	–	–	60977045101	Metoclopramide
311	Cutaneous Mycobacteria	–	–	406051262	Oxycodone Acetaminophen
7220	Cervical Disc Displacement	–	–	51079000220	Acetaminophen
71945	Joint Pain-Pelvis	–	–	904404073	Aspirin

The medication knowledge learning module aggregates the medication knowledge hidden in the complicated medication combination. Finally, the medication prediction module combines the representation of patients with the medication knowledge to predict the medication combination.

A. Patient Representation Module

For the i^{th} patient, the health care event S_i is made up of $x_i^1, x_i^2, \ldots, x_i^t$, and the t means the t^{th} admission time. An example is shown in Table I, we use d, p and m to denote treatment diagnosis codes, treatment procedure codes and medication codes, respectively. These medical codes are translated into the multi-hot vectors (i.e., $x_{i,d}^t$, $x_{i,p}^t$ and $x_{i,m}^t$), which can be applied to the proposed TAMSGC model in the same dimension. In order to capture the feature of each patient, TAMSGC translates diagnosis multi-hot vector $x_{i,d}^t$ and treatment procedure multi-hot vector $x_{i,p}^t$ into embedding vectors $e_{i,d}^t$ and $e_{i,p}^t$ based on Eq. (1), respectively.

$$\begin{cases} e_{i,d}^t = w_d x_{i,d}^t \\ e_{i,p}^t = w_p x_{i,d}^t \end{cases} \tag{1}$$

where w_d and w_p denote embedding matrices of diagnoses and treatment procedures, respectively. For the i^{th} patient, embedding vectors of diagnoses and treatment procedures at the t^{th} admission time are denoted as $e_{i,d}^t$ and $e_{i,p}^t$, respectively, which serve as inputs for recurrent neural networks (RNNs).

In order to learn the temporal sequence feature of diagnoses and treatment procedures, this paper designs a new temporal attention mechanism (TAM) which is based on the two-layer neural network (i.e., RNN_α, RNN_β), as illustrated in Fig. 3. More specifically, RNN_α and RNN_β are applied to learn the diagnosis embedding vector $e_{i,d}^t$ in order to generate two diagnosis attention parameters (i.e., α_d and β_d). Concerning the treatment procedure embedding vector $e_{i,p}^t$, its attention parameters α_p and β_p can be learned from RNN_α and RNN_β, respectively. Attention parameters are generated based on Eq. (2).

$$\begin{cases} o_\alpha = \mathrm{RNN}_\alpha\left(e_{i,*}^t\right), \alpha = tanh\left(w_\alpha o_\alpha + b_\alpha\right) \\ o_\beta = \mathrm{RNN}_\beta\left(e_{i,*}^t\right), \beta = tanhshrink\left(w_\beta o_\beta + b_\beta\right) \end{cases} \tag{2}$$

where $e_{i,*}^t$ stands for the embedding vector $e_{i,d}^t$ or $e_{i,p}^t$. The RNNs are used to learn $e_{i,*}^t$ and generate two output vectors (i.e., o_α and o_β). w_α, b_α, w_β and b_β are parameters requiring to be learned and updating during training process. Output vectors and updating parameters can calculate attention parameters (i.e., α_d, β_d, α_p and β_p) through the activation functions of neural network (i.e., $f_\alpha = \tanh()$ and $f_\beta = $ tanhshrink()).

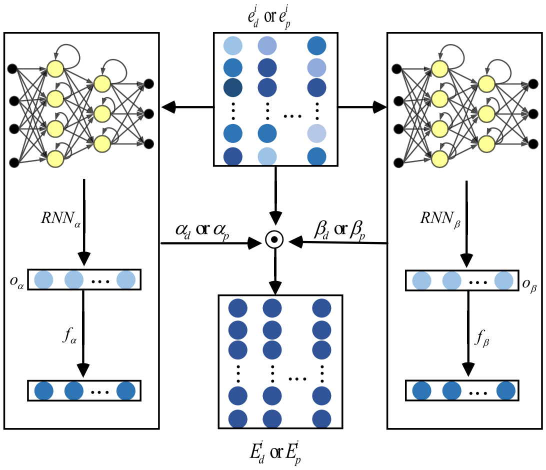


Fig. 3. The construction of temporal attention mechanism (TAM). RNNs (i.e., RNN_α and RNN_β) are used to learn the embedding vectors (i.e., $e_{i,d}^t$ and $e_{i,p}^t$). Based on Eq. (2), attention parameters (i.e., α_d, β_d, α_p, β_p) are calculated from output vectors of RNNs (i.e., o_α and o_β). Then, feature vectors of diagnoses and treatment procedures (i.e., $E_{i,d}^t$ and $E_{i,p}^t$) can be captured based on Eq. (3).

Based on these attention parameters, we can capture the significant information in the temporal sequence. More specifically, the diagnosis attention parameters α_d and β_d can obtain the feature vector of diagnoses $E_{i,d}^t$ based on Eq. (3). Analogously, the treatment procedure attention parameters α_p and β_p are used to learn the feature vector of treatment procedures $E_{i,p}^t$. The feature vectors $E_{i,d}^t$ and $E_{i,p}^t$ contain the temporal sequence information causing by the diverse admission time. Thus, the i^{th} patient at the t^{th} admission time can be represented as P_i^t, which is composed of $E_{i,d}^t$ and $E_{i,p}^t$ based on Eq. (4).

$$\begin{cases} E_{i,d}^t = \alpha_d \beta_d e_{i,d}^t \\ E_{i,p}^t = \alpha_p \beta_p e_{i,p}^t \end{cases} \tag{3}$$

$$P_i^t = \left[E_{i,d}^t, E_{i,p}^t\right] \tag{4}$$

P_i^t represents the feature of i^{th} patient at the t^{th} admission time in the medical records. Representation vectors of a patient in all admission time (i.e., $P_i^1, P_i^2, \ldots, P_i^t$) are combined and represented as $P_i^{t'}$ $(t' < t)$, together with the corresponding medication multi-hot vector $x_{i,m}^{t'}$, composing key-value pairs to represent the history of a patient H_i. Thus, the health care events of patients can be translated into the key-value vector pairs based on Eq. (5).

$$H_i = \left\{ P_i^{t'} : x_{i,m}^{t'} \right\}, for\ i = 1, \ldots, n \tag{5}$$

B. Medication Knowledge Learning Module

In the EHR dataset, prescriptions of ICU patients contain the complicated medication combination that can promote the recovery of patients [10]. And the DDI dataset describes the medication combination that may cause polypharmacy side effects [42]. These medication combinations can be represented by graph networks where nodes stand for drugs and edges represent relations between drugs [13]. The EHR graph and DDI graph are formulated as $\mathcal{G}_{ehr} = (\mathcal{V}, \delta_{ehr})$ and $\mathcal{G}_{ddi} = (\mathcal{V}, \delta_{ddi})$, respectively. The nodes $\mathcal{V}$ represent medications in graphs. The edges of EHR graph are denoted as δ_{ehr}, and δ_{ddi} denotes the known side effects between a pair of drugs in the DDI graph. The TAMSGC leverages SGC to learn the medication knowledge hidden in graph networks. For each node $\mathcal{V}$, the SGC can aggregate information from its neighbors. Based on Eq. (6), the normalized adjacency matrices N_e and N_d are computed from $\mathcal{G}_{ehr}$ and $\mathcal{G}_{ddi}$, respectively.

$$\begin{cases} N_{ehr} = D_{ehr}^{-\frac{1}{2}} \left(A_{ehr} + I_{ehr} \right) D_{ehr}^{-\frac{1}{2}} \\ N_{ddi} = D_{ddi}^{-\frac{1}{2}} \left(A_{ddi} + I_{ddi} \right) D_{ddi}^{-\frac{1}{2}} \end{cases} \tag{6}$$

where N_{ehr} denotes the normalized adjacency matrix of $\mathcal{G}_{ehr}$, and it describes the relations of beneficial medications. A_{ehr} means the initial adjacency matrix that can reflect the initial characteristic of the whole medication graph. D_{ehr} represents the diagonal matrix, and I_{ehr} means the identity matrix in the EHR graph. In the DDI graph, edges of two nodes are the medication combination that can affect the normal physical functioning of humans. The relation of polypharmacy side effects can be described by the normalized adjacency matrix N_{ddi} of $\mathcal{G}_{ddi}$, the initial adjacency matrix A_{ddi}, the diagonal matrix D_{ddi}, and the identity matrix I_{ddi}.

The SGC is used to learn the embedding vector of medications from the EHR graph (i.e., $\mathcal{G}_{ehr}$) and DDI graph (i.e., $\mathcal{G}_{ddi}$). Medication embedding vectors of $\mathcal{G}_{ehr}$ and $\mathcal{G}_{ddi}$ are fused by a parameter λ to obtain medication knowledge K based on Eq. (7).

$$K = N_{ehr} M_{ehr} W_{ehr} + \lambda (N_{ddi} M_{ddi} W_{ddi}) \tag{7}$$

where N_{ehr} denotes the normalized adjacency matrix, and M_{ehr} stands for the medication embedding vector in the EHR graph. W_{ehr} denotes the hidden weight parameter matrix, which is updated in the training process of model. Analogously, in the DDI graph, the normalized adjacency matrix, the medication embedding vector and the hidden weight parameter matrix are denoted as N_{ddi}, M_{ddi} and W_{ddi}, respectively. The medication knowledge of EHR graph and DDI graph are integrated by a parameter λ, which can affect the occurrence of drug-drug interactions.

We have got the feature vector of i^{th} patient at t^{th} admission time (i.e., P_i^t), and the history representation vector H_i that combines his health states and prescriptions (i.e., $P_i^{t'}$ and $x_{i,m}^{t'}$). Together with the medication knowledge K, the output vectors of TAMSGC (i.e., R_k^i and R_h^i) can be obtained for predicting medication combination based on Eq. (8). More specifically, R_k^i considers the feature of patients at each admission time, and R_h^i combines the comprehensive health status of patients from medical records.

$$\begin{cases} R_{i,k}^t = K \operatorname{soft\,max}(K P_i^t) \\ R_{i,h}^t = K x_{i,m}^{t'} \operatorname{soft\,max}(P_i^{t'} P_i^t) \end{cases} \tag{8}$$

where P_i^t is extracted from diagnoses and treatment procedures, and K is acquired from medication combinations of EHR dataset and DDI dataset. For a vector of patient history H_i as listed in Eq. (5), its key vector $P_i^{t'}$ and value vector $x_{i,m}^{t'}$ are used to represent a patient in all admission time and the proper medications for him, respectively.

C. Medication Prediction Module

We leverage the sigmoid function σ to decode the learning vectors. More specifically, the representation of patient P_i^t and output vectors (i.e., $R_{i,k}^t$ and $R_{i,h}^t$) are used to predict the multi-label drug $\hat{y}_i^t$ based on Eq. (9).

$$\hat{y}_i^t = \sigma \left(\left[P_i^t, R_{i,k}^t, R_{i,h}^t \right] \right) \tag{9}$$

The TAMSGC leverages the combined loss functions $\mathcal{L}$ to seek an optimal balance between the predictive accuracy and safety based on Eqs. (10)- (11). As Shang *et al.* have done, $\mathcal{L}_p$ is designed to make the predictive medication close to the true medication, and $\mathcal{L}_{ddi}$ is used to lower the risk of triggering the polypharmacy side effects with the soar of the DDI Rate (a metric based on Eq. (15)) [6].

$$\begin{cases} \mathcal{L}_p = -\pi[0] \sum_t^T \sum_i y_i^t \log \sigma \left(\hat{y}_i^t \right) + \left(1 - y_i^t \right) \log \left(1 - \sigma \left(\hat{y}_i^t \right) \right) \\ \qquad + \frac{\pi[1]}{L} \sum_t^T \sum_i^{|x_m|} \sum_j^{\left| \hat{Y}^t \right|} \max \left(0, 1 - \left(\hat{y}^t \left[\hat{Y}_j^t \right] - \hat{y}_i^t \right) \right) \\ \mathcal{L}_{ddi} = \sum_t^T \sum_{j,j'} \left(A_{ddi}[j, j'] \left(\hat{y}_j^t \hat{y}_{j'}^t \right) \right) \end{cases} \tag{10}$$

where y_i^t and $\hat{y}_i^t$ mean the real drug label and the predictive drug label of the i^{th} patient at the t^{th} admission time, respectively. σ is the sigmoid function. $\hat{y}^t[\hat{Y}_j^t]$ stands for the j^{th} predictive label indexed by a predictive label set $\hat{Y}^t$ at the t^{th} admission time. L is the size of y_i^t, and π represents the mixture weights ($\pi[0], \pi[1] \geq 0, \pi[0] + \pi[1] = 1$). A_{ddi} is an adjacency matrix of DDI graph, and $A_{ddi}[j, j'] = 1$ if medication i and j are a pair of medication combination in the DDI dataset.

By optimizing $\mathcal{L}_p$ and $\mathcal{L}_{ddi}$ interchangeably, we can alleviate the ineffective learning problem because the high-quality DDI representations can help guide $\mathcal{L}_p$ to learn effective parameters. In the combined loss functions, the choice of loss function depends on the DDI Rate of predictive medication combination s' and the expected DDI Rate s. If $s' \leqslant s$, we need to optimize $\mathcal{L}_p$ in order to ensure the effectiveness of the model. If $s' > s$, it means that the probability of triggering polypharmacy side effects is beyond our expectation. We use s', s and the decay index η to calculate p, and a higher p decides to optimize whether

Algorithm 1: TAMSGC.

```
Input: Diagnosis codes $d$, Treatment Procedures codes $p$, Medication codes $m$, the EHR graph $\mathcal{G}_{ehr}$, the DDI graph $\mathcal{G}_{ddi}$;
Output: Medications $\hat{y}_i^t$;
while training do
    translate medical codes into multi-hot vectors $[x_{i,d}^t, x_{i,p}^t, x_{i,m}^t]$;
    Compute the normalized adjacency matrices $N_{ehr} \in \mathcal{G}_{ehr}$ and $N_{ddi} \in \mathcal{G}_{ddi}$ based on Eq. (6);
    Fuse the medication knowledge $K$ based on Eq. (7);
    for $i \leqslant n$ do
        for $t \in admission\ time$ do
            Initialize attention parameters;
            Obtain the embedding vectors $e_{i,d}^t$ and $e_{i,p}^t$ based on Eq. (1);
            Calculate attention parameters $\alpha_d$, $\beta_d$, $\alpha_p$, $\beta_p$ based on Eq. (2);
            Generate the representation of a patient $P_i^t$ based on Eqs. (3)- (4);
            Obtain the history vector of a patient $H_i$ based on Eq. (5);
            Predict medications $\hat{y}_i^t$ based on Eqs. (8)- (9);
            Evaluate the current DDI Rate $s'$ based on Eq. (15);
            Updata the decay index $\eta$ in the combined loss functions based $\mathcal{L}$ on Eq. (11);
```

$\mathcal{L}_p$ or $\mathcal{L}_{ddi}$. So we have the final loss $\mathcal{L}$ based on Eq. (11).

$$\mathcal{L} = \begin{cases} \mathcal{L}_p & \text{if } s' \leq s \\ \mathcal{L}_{ddi}, \text{with } p = \exp\left(-\frac{s'-s}{\eta}\right) & \text{if } s' > s \\ \mathcal{L}_p, \text{with } p = 1 - \exp\left(-\frac{s'-s}{\eta}\right) & \text{if } s' > s \end{cases} \tag{11}$$

where s' shows the current DDI Rate of predictive medication combination, and s' can be calculated with DDI Rate equation based on Eq. (15). As the previous study [6], the expected DDI Rate s is 0.05, and the initial decay index η is set as 0.85. The details of TAMSGC model are shown in Algorithm 1.

IV. Experiments

To test the performance of TAMSGC model, this paper compares the TAMSGC with several baselines in the EHR dataset. Diverse metrics are used to evaluate the predictive accuracy and safety of medication combination for more credible results.

A. Datasets

This paper utilizes the medical information mart for intensive care (MIMIC-III) dataset, which is an openly available dataset developed by the computational physiology lab of Massachusetts Institute of Technology (MIT)[1]. It includes the information of patients who are admitted by the intensive care unit (ICU) [46]. In this dataset, medical concepts are encoded by standard coding systems, such as ICD codes for diagnoses and treatment procedures [47][2], NDC codes[3] for medications [48]. We follow the procedure similar to Shang *et al.* [6] to process the medical codes in this experiment. The NDC codes are transformed into ATC[4] codes for obtaining the hierarchical information of medications [49]. This paper chooses a set of medications prescribed by doctors during the first 24-hour, because such a time period tends to be the most critical time for patients to obtain the correct treatment [50], [51]. Moreover, some patients who visit a hospital more than one time are with temporal sequence feature. The medical records of these patients are selected from the MIMIC-III dataset. The clinical events of multi-morbidity account for approximately 99.72% among selected patients who visit a hospital repeatedly. The drug-drug interaction (DDI) knowledge is extracted from the TWOSIDES dataset [52], and the top-40 severe DDI types are used in our model. The reason is that they are representative with the usage of the most common DDI types based on previous studies [6], [7]. The statistical features of such data are shown in Table II.

TABLE II
The Statistical Information of Data

Statistical Items	Quantity
diagnoses	4557
treatment procedures	1428
medications	145
patients	6351
clinical events	15023
average of diagnoses	10.92
average of treatment procedures	3.84
average of medications	8.80
average of admission times	2.37
clinical events of multi-morbidity	14981

B. Experiment Setup

This paper randomly divides the dataset into training, validation and testing sets by the percentage of 2/3, 1/6, 1/6 as recommended in the previous study [6]. The TAMSGC selects the gated recurrent unit (GRU) [53], a variant of RNN, to learn the medical codes. The dropout is used to prevent the neural network from overfitting on the output of embedding [54]. According to the experience of previous study [6], the learning rate is set as 0.0002, and the convolutional layer has 64 hidden units. The TAMSGC has been implemented in PyTorch with Python 3.7 version, and trained for 40 epochs.

As for baseline models, this paper has tested the performance of RETAIN [9], RNN [28], LEAP [31] and DMNC [30] on the MIMIC-III dataset. Also, this paper directly leverages the experiment results of GAMENet [6], CompNet [7] and G-BERT [11], which use the same dataset to research the issue of medication combination prediction.

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

[2] https://en.wikipedia.org/wiki/List_of_ICD-9_codes

[3] https://www.drugfuture.com/fda-ndc

[4] http://www.whocc.no/atc/structure and principles

C. Metrics

In order to provide a fair comparison, this paper selects different metrics to estimate the predictive accuracy of medication combination. The Jaccard similarity score (Jaccard) is used to calculate the intersection size of predictive medications and true medications based on Eq. (12). Precision-recall area under curve (PRAUC) is the area enclosed by coordinate axes and the curve that is made up of Recall and Precision according to Eq. (13). Average F1 (F1) can comprehensively evaluate the predictive accuracy on the basis of Eq. (14).

The safety of medication combination is essential for the recovery of patients. Therefore, the drug-drug interaction rate (DDI Rate) is used to measure the safety of medication combination [6]. The DDI Rate shows the percentage of medication combination that contain DDIs based on Eq. (15).

$$\text{Jaccard} = \frac{1}{\sum\limits_{i}^{N}\sum\limits_{t}^{T_i} 1} \sum_{i}^{N}\sum_{t}^{T_i} \frac{|y_i^t \cap \hat{y}_i^t|}{|y_i^t \cup \hat{y}_i^t|} \tag{12}$$

$$\text{Recall} = \frac{|y_i^t \cap \hat{y}_i^t|}{|\hat{y}_i^t|}, \text{Precision} = \frac{|y_i^t \cap \hat{y}_i^t|}{|y_i^t|} \tag{13}$$

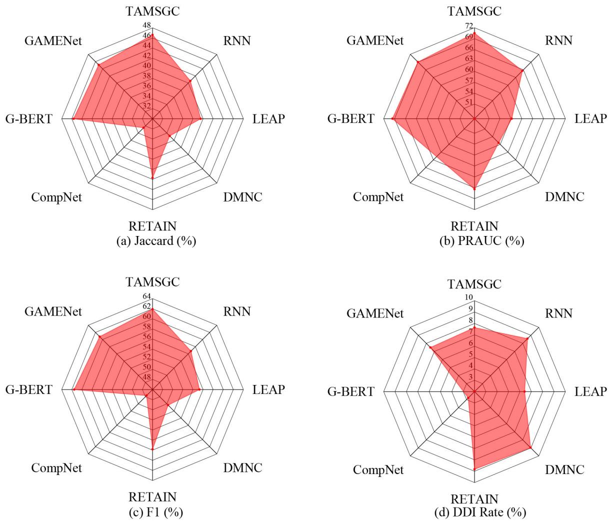
$$\text{F1} = \frac{1}{\sum_i^N \sum_t^{T_i} 1} \sum_{i}^{N}\sum_{t}^{T_i} \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

$$\text{DDI Rate} = \frac{\sum\limits_{i}^{N}\sum\limits_{t}^{T_i}\sum\limits_{a,b} |\{(c_a, c_b) \in \hat{y}_i^t| \, (c_a, c_b) \in \delta_{ddi}\}|}{\sum\limits_{i}^{N}\sum\limits_{t}^{T_i}\sum\limits_{a,b} 1} \tag{15}$$

where N is the number of patients, t means the t^{th} admission time, i means the i^{th} patient, and T_i denotes all admission times of the i^{th} patient. $\hat{y}_i^t$ is the predictive medication and y_i^t is the true medication of prescriptions in the EHR dataset. The predictive medication pairs (c_a, c_b) belong to the edge set δ_{ddi} in the DDI dataset.

D. Results

This paper compares the performance of various models for medication combination prediction. The predictive accuracy of medication combination is measured by Jaccard, PRAUC and F1 as shown in Figs. 4 (a), (b) and (c), respectively. The higher the scores, the greater the predictive accuracy of models. However, for the DDI Rate in Fig. 4 (d), a lower value indicates the safer medication combination.



(a) Jaccard (%)
(b) PRAUC (%)
(c) F1 (%)
(d) DDI Rate (%)

Fig. 4. The comparison of models on medication combination prediction task. Our proposed TAMSGC model is compared with baseline models (RNN [28], LEAP [31], DMNC [30], RETAIN [9], CompNet [7], G-BERT [11], GAMENet [6]) on (a) Jaccard, (b) PRAUC, (c) F1 and (d) DDI Rate. The high values on (a)-(c) and low values on (d) can reveal the predictive accuracy and safety of medication combination, respectively.

Concerning the predictive accuracy, our proposed TAMSGC model performs better than baselines on Jaccard, PRAUC and F1. Based on the recurrent neural networks (RNNs), the temporal attention mechanism (TAM) improves the predictive accuracy compared with the RNN based baselines (i.e., RETAIN [9], RNN [28], LEAP [31] and DMNC [30]). The RNN can make the best of recurrent structure to learn the representation of medical codes [28], [29], [55]. However, it cannot capture the significant information that affects the health state of patients in the highest degree. Attention parameters of TAM compensate for this shortcoming commendably. They add a high weight to the diagnosis or treatment procedure that repeat in medical records at the different admission time of patients. In this way, attention parameters capture the significant information from the temporal sequence. As for the graph neural networks (GNNs) based baselines (i.e., GAMENet [6], CompNet [7] and G-BERT [11]), the application of simple graph convolution makes the TAMSGC surpass them. In fact, the critical patients share some symptoms, which correspond to similar prescriptions. The smooth feature of SGC can integrate this similarity into the medication embedding vectors, and affects the predictive accuracy of medication. Moreover, our model is different from GAMENet [6] that integrates the RNN and GNN. As we know, the RNN does not distinguish the importance of patient features, and the smooth feature of GCN is not obtained to describe the similarity of the critical patients. These shortcomings limit the performance of GAMENet. In order to overcome these shortcomings, the TAM and SGC are applied in TAMSGC to learn the representations of patients and medications. On the predictive accuracy, the improvements of TAMSGC on Jaccard, F1 and PRAUC are 1.52%, 1.1% and 1.46%, respectively, compared with GAMENet. Such improvement shows the effectiveness of TAM and SGC in the TAMSGC.

When improving the accuracy of medication combination, we also expect a lower DDI Rate simultaneously, which indicates the possibility of triggering the polypharmacy side effects. The TAMSGC has a lower DDI Rate than RETAIN [9], RNN [28] and DMNC [30], and close safety performance towards GAMENet [6]. The DDI Rate of G-BERT is not compared, since the research of G-BERT involves patients with a single admission, which may result in deviation [11]. Even though the DDI Rate of our proposed TAMSGC is not the lowest

TABLE III
THE COMPARISON OF TAMSGC WITH THE TEMPORAL ATTENTION MECHANISM (TAM) OR SIMPLE GRAPH CONVOLUTION (SGC)

Method	Jaccard (%)	PRAUC (%)	F1 (%)	DDI Rate(%)
$TAMSGC_{TAM_-,SGC_-}$	44.51	68.89	60.26	8.26
$TAMSGC_{TAM_-,SGC_+}$	45.34	69.36	61.06	8.35
$TAMSGC_{TAM_+,SGC_-}$	45.73	69.83	61.39	7.71
$TAMSGC_{TAM_+,SGC_+}$	46.61	70.50	62.25	7.63

one, it can ensure the safety of medication combination to some extent.

V. DISCUSSION

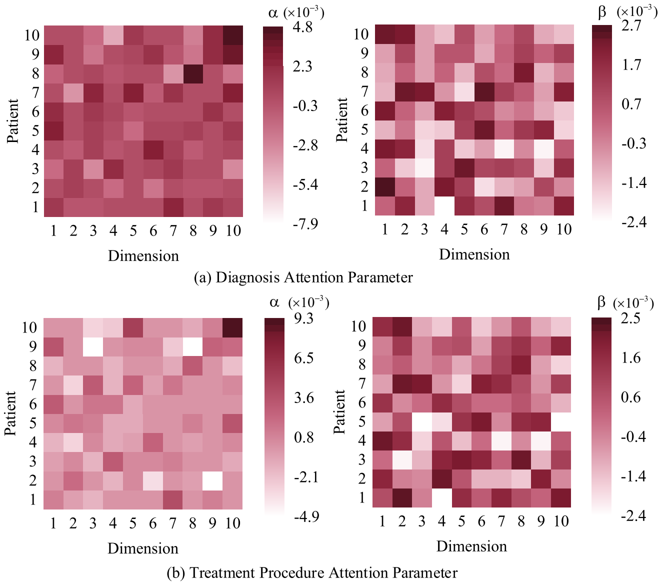
The proposed TAMSGC model mainly consists of the temporal attention mechanism (TAM) and the simple graph convolution (SGC). We do some ablation studies to analyze the influence of these two parts on the predictive accuracy and the computational complexity. If the TAMSGC takes TAM and SGC for medication combination prediction, TAM_+ and SGC_+ are added as a subscript (i.e., $TAMSGC_{TAM_+,SGC_+}$). Otherwise, the subscript will be TAM_- and SGC_- (i.e., $TAMSGC_{TAM_-,SGC_-}$).

A. The Predictive Accuracy Analysis

To analyze the corresponding influence of TAM and SGC on the predictive accuracy, the TAMSGC is tested by applying TAM or SGC one at a time. Table III has shown the comparison results.

As shown in Table III, on the predictive accuracy metrics (i.e., Jaccard, PRAUC and F1), the TAMSGC with TAM has a better performance than it without TAM. Attention parameters of TAM contribute to such results, since they extract the significant information from dynamic diagnoses and treatment procedures to obtain the representation vector of patients. In order to better understand the effect of attention parameters, we randomly select 10 patients as examples. Diagnoses and treatment procedures of them are represented as 10-dimension feature vectors for the evident comparison of attention parameters. As shown in the left subgraph of Fig. 5 (a), although the attention parameter α of diagnoses has a high value in some dimensions, the combination with the other attention parameter β in the right subgraph of Fig. 5 (a) can well differentiate the influential diagnosis features for these 10 patients. Different from diagnoses, the significant features of treatment procedures are captured by the attention parameter α and enhanced by β in Fig. 5 (b). Owing to the significant features extracted by attention parameters, TAM has the capability of improving the predictive accuracy of medication combination. Simultaneously, the predictive accuracy of TAMSGC is improved with the usage of SGC. It proves that the smooth feature of SGC makes nodes share similar features caused by the similar prescriptions of ICU patients.

The compared results show that the TAM and SGC are effective in predicting medication combination. In addition, the DDI Rate of $TAMSGC_{TAM_+,SGC_+}$ is the lowest with the application of TAM and SGC, which ensures the safety of the predictive medication combination.



(a) Diagnosis Attention Parameter

(b) Treatment Procedure Attention Parameter

Fig. 5. Attention parameters for 10 randomly selected patients. The diagnoses and treatment procedures are translated into 10-dimension feature vectors in order to compare the significant feature of different patients, which is captured by large attention parameters.

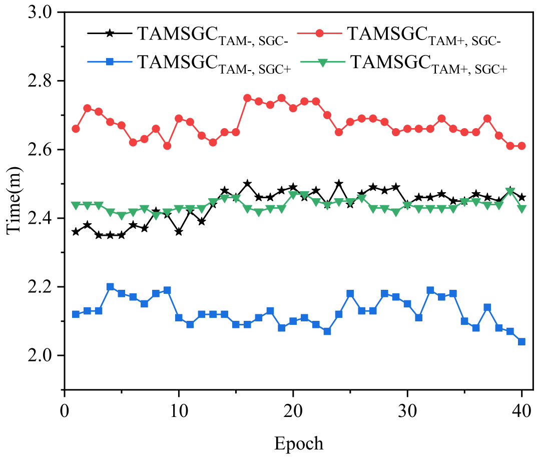


Fig. 6. The training time of TAMSGC with the temporal attention mechanism (TAM) and the simple graph convolution (SGC). The TAMSGC is trained 40 epochs to compare the training time of each epoch.

B. The Computational Complexity Analysis

To verify the computational consumption of the proposed TAMSGC model, we compare the training time of each epoch by ablation study. Then, we analyze the computational complexity in theory.

As shown in Fig. 6, the application of TAM (i.e., $TAMSGC_{TAM_+,SGC_+}$ and $TAMSGC_{TAM_+,SGC_-}$) can increase the training time. The TAM is made up of two-layer neural networks, and the time it takes to complete the training is more. However, the TAMSGC costs less time in each training epoch with the usage of SGC (i.e., $TAMSGC_{TAM_+,SGC_+}$

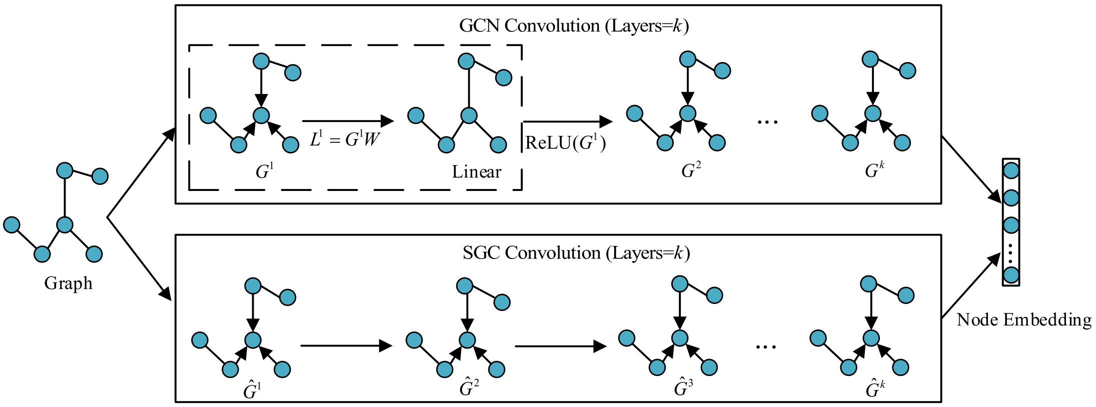


Fig. 7. The comparison of GCN and SGC. Compared with the GCN, the SGC gets rid of the linear transformation and nonlinear activation (i.e., ReLU function) to extract the feature hidden in the personal prescription.

and $TAMSGC_{TAM_-,SGC_+}$). The reason is that the SGC gets rid of the nonlinearity computation between layers compared with the traditional GCNs. Our proposed TAMSGC (i.e., $TAMSGC_{TAM_+,SGC_+}$) uses TAM and SGC to learn embedding vectors of medical codes. With a higher predictive accuracy, the training time of it between $TAMSGC_{TAM_+,SGC_-}$ and $TAMSGC_{TAM_-,SGC_+}$.

The fact is that the SGC attaches great importance to the averaging of local feature, but not the nonlinearity [14]. As shown in Fig. 7, the GCN firstly learns the local smooth feature of the graph (i.e., G) in each layer. Then, such feature is transformed into a linear one with a learned weight matrix W. In the last, the transformed feature L is activated by a nonlinear transition function ReLU before outputting the feature representation. Differently, the SGC predigests the entire procedure and learns the local smooth feature directly. The similarity of personal prescriptions causes the local smooth feature in the medication graph networks. Therefore, with the use of SGC, the local smooth feature can be extracted by our proposed TAMSGC (i.e., $TAMSGC_{TAM_+,SGC_+}$), and it achieves a satisfying performance on this EHR dataset.

In a word, the proposed TAMSGC has a reasonable computational consumption using the TAM and SGC. The TAM obtains attention parameters based on RNNs. We use $\hat{i}$ and $\hat{h}$ to denote the number of input neuron and hidden neuron, respectively. The computational complexity of TAM is $\mathcal{O}_1(\hat{i}\hat{h} + \hat{h}^2 + \hat{h})$ as the RNN. In graph neural networks, $\hat{k}$, $\hat{n}$, $\hat{m}$ and $\hat{d}$ represent the number of layers, nodes, edges and the dimension of nodes, respectively. The SGC has a lower computational complexity compared with the GCN (i.e., $\mathcal{O}_2(\hat{k}\hat{m}\hat{d} + \hat{k}\hat{n}\hat{d}^2)$). Therefore, the proposed TAMSGC can reduce the computational complexity to less than $\mathcal{O}_1(\hat{i}\hat{h} + \hat{h}^2 + \hat{h}) + \mathcal{O}_2(\hat{k}\hat{m}\hat{d} + \hat{k}\hat{n}\hat{d}^2)$.

VI. Conclusion

It is important for experts to design the clinic guidelines and expert system to predict the medication combination for curing the multi-morbidity patients in the intensive care unit (ICU), timely and accurately, which can provide the heuristic information for clinicians to make safe decision efficiently. This paper proposes a deep learning model, named TAMSGC, to help experts extract hidden knowledge efficiently and free from the large-scale clinical data analyses. More specifically, The temporal attention mechanism (TAM) aims at aggregating the temporal sequence information from medical records of patients. The simple graph convolution (SGC) predigests the process of learning the medication correlation compared with the graph convolutional networks (GCNs). The proposed TAMSGC model has been tested on a real dataset, and the results demonstrate that TAMSGC improves the predictive accuracy successfully compared with baseline models. Meanwhile, TAMSGC has a reasonable computational complexity to predict the medication combination. However, due to the uncertainty in the complex clinical environment, there is a gap between our current achievements and the ultimate goal, especially in the safety of medication combination. Therefore, in the future, we plan to integrate the expert knowledge and the relation of drugs and target proteins from the medical dataset. We expect to recommend the safe and effective medications in order to provide assistance for experts when facing patients with multi-morbidity, especially the uncommon morbidity.

References

[1] R. Yu, Y. Zheng, R. Zhang, Y. Jiang, and C. C. Poon, "Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 486–492, Feb. 2020.

[2] K. Yu and X. Xie, "Predicting hospital readmission: A joint ensemble-learning model," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 447–456, Feb. 2020.

[3] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019.

[4] 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.

[5] D. Jarrett, J. Yoon, and M. van der Schaar, "Dynamic prediction in clinical survival analysis using temporal convolutional networks," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 424–436, Feb. 2020.

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

[7] 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. Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 1623–1632.

[8] D. Liu, Y. L. Wu, X. Li, and L. Qi, "Medi-care AI: Predicting medications from billing codes via robust recurrent neural networks," *Neural Netw.*, vol. 124, pp. 109–116, Apr. 2020.

[9] 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. Adv. Neural Inf. Proces. Syst.*, 2016, pp. 3504–3512.

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

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

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

[13] T. Ma, C. Xiao, J. Zhou, and F. Wang, "Drug similarity integration through attentive multi-view graph auto-encoders," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 3477–3483.

[14] F. Wu, T. Zhang, A. H. d. SouzaJr, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6861–6871.

[15] D. Almirall, S. N. Compton, M. Gunlicks-Stoessel, N. Duan, and S. A. Murphy, "Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy," *Statist. Med.*, vol. 31, no. 17, pp. 1887–1902, Mar. 2012.

[16] M. Gunlicks-Stoessel, L. Mufson, A. Westervelt, D. Almirall, and S. Murphy, "A pilot SMART for developing an adaptive treatment strategy for adolescent depression," *J. Clin. Child. Adolesc. Psychol.*, vol. 45, no. 4, pp. 480–494, Mar. 2016.

[17] Z. Chen, K. Marple, E. Salazar, G. Gupta, and L. Tamil, "A physician advisory system for chronic heart failure management based on knowledge patterns," *Theory Pract. Log. Prog.*, vol. 16, no. 5-6, pp. 604–618, Sep. 2016.

[18] British Thoracic Society, "Guidelines for management of asthma in adults: I. - chronic persistent asthma," *BMJ.*, vol. 301, pp. 651–653, Sep. 1990.

[19] B. D. Lau and E. R. Haut, "Practices to prevent venous thromboembolism: A brief review," *BMJ*, vol. 23, pp. 187–195, Mar. 2014.

[20] L. A. Marcath *et al.*, "Comparison of nine tools for screening drug-drug interactions of oral oncolytics," *J. Oncol. Pract.*, vol. 16, no. 6, pp. e368–e374, Jun. 2018.

[21] K. H. Hoang and T. B. Ho, "Learning and recommending treatments using electronic medical records," *Knowl.-Based Syst.*, vol. 181, Oct. 2019, Art. no. 104788.

[22] M. Jiang *et al.*, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *J. Amer. Med. Inf. Assoc.* , vol. 18, no. 5, pp. 601–606, Sep. 2011.

[23] D. Zhao and C. Weng, "Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction," *J. Biomed. Inform.*, vol. 44, no. 5, pp. 859–868, Oct. 2011.

[24] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes," *J. Clin. Epidemiol.*, vol. 66, no. 4, pp. 398–407, Apr. 2013.

[25] P. Wallis and P. Danaee, "Learning semantic relationships from medical codes," in *Proc. Int. Florida Artif. Intell. Res. Soc. Conf.*, 2019, pp. 305–310.

[26] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[27] 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, Feb. 2019, Art. no. e0211057.

[28] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. Mach. Learn. Health. Conf.*, 2016, pp. 301–318.

[29] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, Apr. 2018, Art. no. 6085.

[30] H. Le, T. Tran, and S. Venkatesh, "Dual memory neural computer for asynchronous two-view sequential learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2018, pp. 1637–1645.

[31] 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. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2017, pp. 1315–1324.

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

[33] V. Mnih *et al.*, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2014, pp. 2204–2212.

[34] L. Yu, Y. Xiuyuan, and B. Qiliang, "Image inpainting algorithm based on neural network and attention mechanism," in *Proc. ACM Int. Conf. Proc. Ser.*, 2019, pp. 345–349.

[35] K. M. Hermann *et al.*, "Teaching machines to read and comprehend," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2015, pp. 1693–1701.

[36] Q. Li, X. Zhang, J. Xiong, W.-M. Hwu, and D. Chen, "Implementing neural machine translation with bi-directional GRU and attention mechanism on FPGAs using HLS," in *Proc. Asia South Pac. Des. Autom. Conf.*, 2019, pp. 693–698.

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

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

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

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

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

[42] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, Jul. 2018.

[43] M. Asada, M. Miwa, and Y. Sasaki, "Enhancing drug-drug interaction extraction from texts by molecular structure information," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2018, pp. 680–685.

[44] X. Lin, Z. Quan, and Z. J. Wang, "KGNN: Knowledge graph neural network for drug-drug interaction prediction," in *Proc. Conf. Artif. Intell.*, 2020, pp. 2739–2745.

[45] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2017, pp. 1024–1034.

[46] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, May 2016, Art. no. 160035.

[47] W. Wang *et al.*, "Graph-driven generative models for heterogeneous multi-task learning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 979–988.

[48] L. Simonaitis and C. J. McDonald, "Using national drug codes and drug knowledge bases to organize prescription records from multiple sources," *Amer. J. Health-Syst. Pharm.*, vol. 66, no. 19, pp. 1743–1753, Oct. 2009.

[49] L. Wang, W. Zhang, X. He, and H. Zha, "Personalized prescription for comorbidity," in *Proc. Lect. Notes Comput. Sci.*, 2018, pp. 3–19.

[50] D. J. Eveson, T. G. Robinson, and J. F. Potter, "Lisinopril for the treatment of hypertension within the first 24 hours of acute ischemic stroke and follow-up," *Amer. J. Hypertens.*, vol. 20, no. 3, pp. 270–277, Mar. 2007.

[51] G. C. Fonarow*et al.*, "Effect of statin use within the first 24 hours of admission for acute myocardial infarction on early morbidity and mortality," *Amer. J. Cardiol.*, vol. 96, no. 5, pp. 611–616, Sep. 2005.

[52] N. P. Tatonetti, P. Y. Patrick, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Sci. Transl. Med.*, vol. 4, no. 125, pp. 125ra 31–125ra31, Mar. 2012.

[53] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2018.

[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[55] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review," *J. Amer. Med. Inf. Assoc.*, vol. 25, no. 10, pp. 1419–1428, Oct. 2018.