Patient Similarity via Medical Attributed Heterogeneous Graph Convolutional Network

Yi Li, Dan Yang, Xi Gong

Abstract—**Electronic medical records (EMR) record the whole process of patients' diagnosis and treatment in the hospital, which contains a lot of valuable information. Through this information, medical services can be provided to patients in more timely and convenient manner. Accurately identifying patients with similar diseases based on EMR is the key to personalized healthcare. Most patient similarity studies mainly utilize discrete medical entities (drugs, procedures) embedded as patient feature representation. But these structured data could be either incomplete or erroneous which has a significant impact on the final patient representation. And the previous studies rarely considered the structural and semantic information existing between medical entities. Therefore, we propose a patient similarity framework based on a medical attributed heterogeneous graph convolution network, named AHGCN-PS. Firstly, the framework leverages the patients' medical entity and incorporates the patients' medical text as the attributes of patients to obtain more integral patient information. Then, we construct a medical attributed heterogeneous information network from EMR, capturing the structural information in the network and the hidden semantic information between different nodes by selecting different meta-paths. Then, we adopt a graph convolutional neural network and a semantic attention mechanism to aggregate node neighbor information and meta-path semantic information. Finally, this paper uses the obtained patient node feature representation for patient similarity calculation. We use the real-world ICU patient dataset MIMIC-III to evaluate the experimental performance of AHGCN-PS, the experimental results demonstrate the effectiveness and feasibility of the patient similarity framework.**

Index Terms—**Patient Similarity, Heterogeneous Graph Neural Network, Meta-path, Electronic Medical Records**

I. Introduction

With the continuous development of medical technology, the research based on EMR has made great progress, which promotes the continuous growth of EMR data. Its widespread application has provided opportunities for patients to make individualized decisions. The study of patient similarity [1] provides effective help for medical health and further improves the doctors' diagnosis and treatment effect of patients. The patient similarity study has been applied to target patient detection [2], clinical pathway analysis [3], and other tasks. Patient similarity quantitatively analyzes the distance between concepts in the semantic space of complex concepts by selecting clinical concepts (such as procedures, drugs, diagnosis, family history, etc.) as the patients' features, to quantitatively describe the distance between patients and make similar patients cluster.

Most of the existing research on patient similarity learned the potential embedding of patient representation by extracting medical entities related to patients, but they often ignore the attribute information of medical entities and the rich semantic information between patients and their related medical entities. Therefore, effectively learning the structural information and semantic information of patients can improve the accuracy of patient similarity results. However, the complexity and heterogeneity of medical data bring difficulties to research in this field. For example, medical data contains multiple node types, and each node contains attribute information. Take the patient and drug nodes as an example, the attributes of patients are gender, age, height, weight, etc. The attributes contained in drugs are drug type, drug text description, and so on. This kind of graph comes with multiple different types of nodes, also widely known as heterogeneous information networks (HINs) [4]. It is challenging to effectively extract the rich and diverse structure information and attribute information of nodes and encode them into a low-dimensional vector space.

Most existing heterogeneous graph embedding methods are based on the idea of meta-paths. Meta-path [5] is a widely used structure to capture semantics, and it is a composite relation that connects two objects. For example, patient-drug-patient (PDP) and patient-drug-procedure-patient (PDTDP) are meta-paths that describe two different relationships between patients. Among them, the PDP meta-path describes two patients who used the same drug to treat the disease, while the PDTDP describes two patients who used different drugs but used the same procedure. It can be seen that according to different meta-paths, the relationships between nodes in heterogeneous graphs also have different semantics. In consequence, it is challenging to select meaningful meta-paths and integrate semantic information.

In view of the above challenges, this paper proposes a patient similarity framework based on an attributed heterogeneous graph convolution network, named AHGCN-PS. As shown in Fig. 1. Firstly, the framework extracts medical entity and patient text information from EMR to construct a medical attributed heterogeneous information network. Then, aggregate the neighbor information of the patient nodes through the medical heterogeneous graph convolutional network, and the important meta-paths are aggregated by combining the

Manuscript received March 2, 2022; revised September 7, 2022. This work was supported by the General Scientific Research Project of Liaoning Provincial Department of Education (2022).

Yi Li is a postgraduate student at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: asly_ustl@ 163.com).

Dan Yang is a professor at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: asyangdan@ 163. com).

Xi Gong, the corresponding author, is a lecturer at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (email: askdjy05gx@ 163. com).

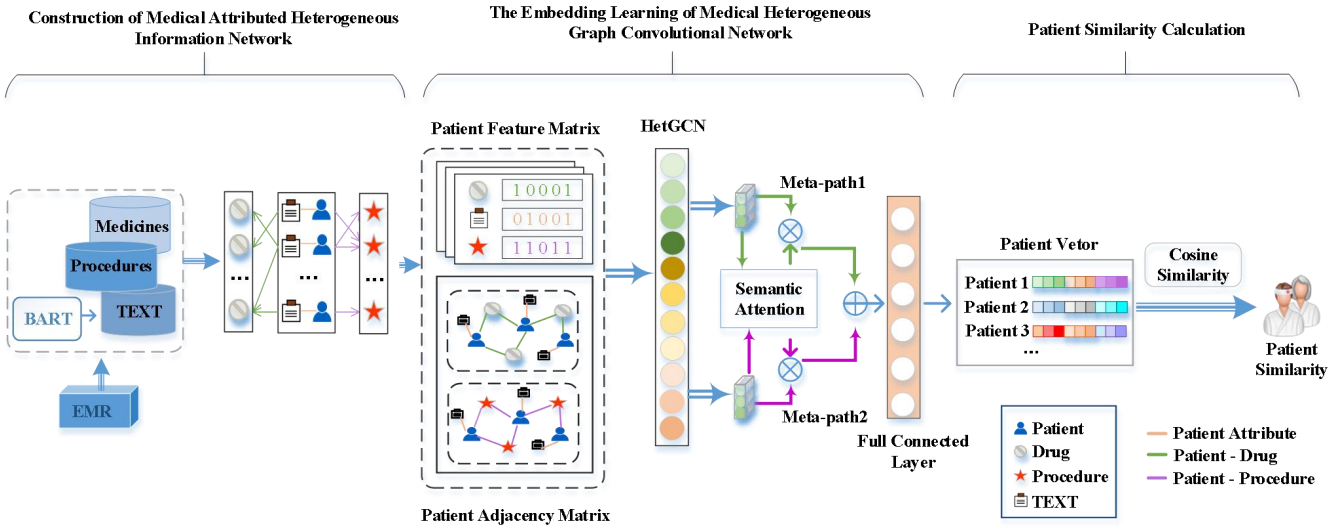

Fig. 1. Overview of AHGCN-PS

semantic attention mechanism to fuse the semantic information. Finally, we obtain the patient feature representation for patient similarity calculation, and the results show that our framework is effective.

The main contributions are summarized as follows:

- Medical entities are selected from EMR to construct a medical attributed heterogeneous information network, which retains the rich heterogeneous relationship between data. Considering that a single medical entity cannot effectively and comprehensively represent the patients' features, therefore, the medical text of patients is introduced as the attribute of patients to learn more accurate patient vector representation.
- We propose a patient similarity framework AHGCN-PS, to better aggregate patient related information. By selecting patient-drug-patient and patient-procedure-patient two different meta-paths, comprehensive structural information and rich semantic information among nodes in heterogeneous networks can be effectively learned to improve the patient similarity clustering effect.
- We evaluate the effectiveness of the AHGCN-PS on a real MIMIC-III dataset. Experimental results show the superiority of the proposed patient similarity framework by comparing it with the three state-of-the-art methods.

II. Related Work

A. Patient Similarity

In recent years, patient similarity has become a hot topic. More and more researchers regard patient similarity as one of the key steps to achieving precision medicine, which plays a huge role in medical research. For example, related work [6] constructed a temporal medical entity association graph to learn medical entity vector representation and combined it with the time decay function, to improve the patient similarity effect. Related work [7] proposed a locally supervised metric learning to effectively combine expert knowledge for patient similarity measurement. Related work [8] calculated the feature similarity respectively and compared all possible combinations of three disease codes, three laboratory test sets, and three weight allocation schemes. Related work [9] selected 30 patients' similarity scores and feature similarity to form a labeled sample set for semi-supervised learning (SSL) algorithm to learn patient similarity. Related work [10] used cosine similarity measure to identify similar patients predicted by 30-day mortality, all prediction variables were represented by a numerical vector to generate cosine similarity measure. Related work [11] proposed a novel framework PSE, with PSE integrated temporal information into the embedding of medical concepts for patient representational learning. Related work [12] proposed mtTSML. A multi-task triple constrained sparse metric learning method to monitor the similarity progress of patient pairs. Related work [13] proposed a generic framework for healthcare models, which found patients with similar conditions and structures in the dataset, extracted their valuable information and enhanced patient representation learning. The framework improved the performance of the health care model by combining the auxiliary information of similar patients. Related work [14] proposed a new deep learning model TDBNN, which used the triple structure, dynamic Bayesian network (DBN) and recurrent neural network (RNN) to study the fine-grained similarity between patients. However, as mentioned above, either the patients' attribute information is not taken into account, or the structural information and semantic information between nodes are ignored.

B. Graph Neural Network

The graph is a data structure consisting of nodes and edges. The data in many application scenarios have a natural graph structure, and the traditional deep learning method is difficult to apply to graph data. Therefore, in recent years, people have extended the deep learning algorithm to graph neural network, and many representatives have emerged. Such as related work [15-18] as pioneering work. Related work [16] proposed a scalable semi-supervised learning method for graph structured data, which is called graph convolution network (GCN). Related work [15] proposed a general inductive framework, which used node feature information to effectively embed nodes for data generation that have not been seen before. It extracts and transforms the local neighborhood of the target node through the aggregator function to train and generalize it to invisible nodes or graphs. Related work [18] proposed a propagation

model that propagates information to all nodes through gated recurrent units. With the wide application of attention mechanisms such as self-attention [19] and soft attention [20] in deep learning. The attention mechanism based on graph neural network is widely popular in various fields, such as recommendation [21-22]. Inspired by the attention mechanism, related work [17] proposed a graph attention network (GAT), which learns different nodes in the neighborhood through the self-attention layer and assigns different weights to nodes. However, the GNN mentioned above was constructed for homogeneous graphs and considered only a single node, the real world is mostly composed of multiple types of nodes. Therefore, related work [23] proposed a heterogeneous graph-based Transformer mechanism. It leverages the meta-relations of heterogeneous graphs to parameterize the weight matrices of heterogeneous mutual attention, message passing, and propagation steps. Related work [24] proposed a novel heterogeneous graph neural network based on hierarchical attention to learn heterogeneous node information. With the in-depth study of graph neural networks, GNNs are widely used in various fields. For example, text classification [25]. A corpus text graph was constructed based on word co-occurrence and document word relationship, and then proposed a text graph convolution network for the text classification task. User analysis [26], a semi-supervised method based on heterogeneous graph learning was used for user analysis and modeling. Anomaly detection [27] used graph convolution network to detect spam, which meets the efficiency requirements and reduces the impact of confrontational behavior. In this paper, graph neural network is used to study patient similarity which has achieved better results.

C. Heterogeneous Graph Embedding

The traditional network representation learning mainly focuses on the structural information in the network, so that the learned embedding is applied to the downstream work. The method is based on matrix decomposition. GraRep [28] is the model for learning vertex representations of weighted graphs which learns low dimensional vectors to represent vertices appearing in the graph. The method based on random walk. For example, node2vec [29] adopted the method of deep neural network and used random walk to obtain the nearest neighbor sequence of vertices. Related work [30] proposed a structured deep network embedding, which maintains both local network structure and global network structure. But these algorithms are based on homogeneous graphs and cannot fully learn nodes' rich structure information and semantic information.

Heterogeneous graph embedding is to map the nodes in the heterogeneous graph into the low-dimensional vector space to learn the rich potential information in the graph. Metapath2vec [31] conducted random walks through a single meta-path and used skip-gram [32] to learn node representations. HERec [33] converted the neighbors based on the meta-path into homogeneous graph by artificially defining a meta-path and made DeepWalk [34] learn the node embedding of the target type. Related work [35] proposed a projection metric embedding model called PME, which calculates the similarity between nodes by Euclidean distance and projects different types of nodes into the same relational space for heterogeneous link prediction. However, the above description ignores node attribute information or only considers a single meta-path.

III. PRELIMINARIES

In this section, we formally describe the key concepts in the patient similarity framework as follows.

Definition 1. Medical entity records. EMR contains medical entities such as patients, procedures, drugs, and medical text information for patients. Extract these data from it to generate medical entity records for patients $r_p = <P, e, O>$. Among them, P represents the patients' entity, and e represents the medical entity used by the patient including procedures and drugs. $O = \{O^{(p_1)}, O^{(p_2)}, \ldots, O^{(p_n)}\}$ is the set of attributes of patient P, where n represents the number of patients.

For example, Fig.2 shows the medical entity record P_1:$<e_1, O_{p1}>$ represents that patient P_1 uses Procedure Code (CPT) 3722, Drug Insulin, and other medical entities, at the same time, describe the basic information of patient P_1 through a medical text.

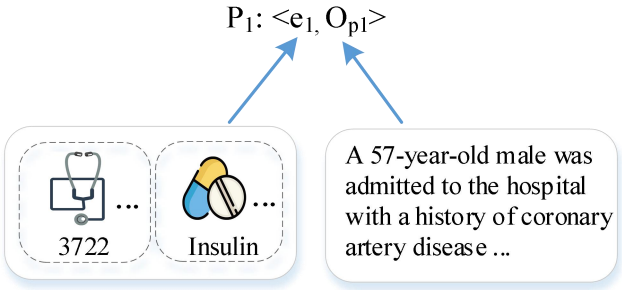

Fig. 2. Patient Related Medical Entity Record

Definition 2. Medical attributed heterogeneous graph. A medical attributed heterogeneous graph is defined as $G = (V, E)$. It includes patient sets $P = \{p_1, p_2, \ldots, p_n\}$, drug sets $D = \{d_1, d_2, \ldots, d_k\}$, and procedure sets $T = \{t_1, t_2, \ldots, t_m\}$ as nodes, that is $V = P \cup D \cup T$. Where k is the number of drugs and m is the number of procedures. The edge sets E represent the relationship between nodes. Nodes and edges are associated with a type mapping function respectively ϕ: $V \rightarrow \mathcal{A}$ and Ψ: $E \rightarrow \mathcal{R}$, $\mathcal{A}$ and $\mathcal{R}$ denote the predefined sets of node types and types, respectively, with $|\mathcal{A}|+|\mathcal{R}|>2$.

Definition 3. Meta-path. In medical heterogeneous graph, two nodes can be connected by different semantic paths called meta-path. A meta-path Φ is defined as a path in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$. The path describes the composite relation between $R = R_1 \circ R_2 \circ \cdots \circ R_l$ objects A_l to A_{l+1}, where $\circ$ denotes the composition operator on relations.

Definition 4. Medical attributed heterogeneous graph embedding. Medical attributed heterogeneous graph embedding is to map the nodes in the medical attributed heterogeneous graph to low dimensional vector space. Give a medical attributed heterogeneous graph $G = (V, E)$. A node attribute matrix $X_{A_i} \in R^{|V_{A_i}| \times d_{A_i}}$, where $A_i \in \mathcal{A}$, then embed the nodes of the graph into the d-dimensional vector space $h_v \in R^{|V| \times d}$ ($d \ll |V|$). Node embedding can be used for

various graph mining tasks, such as link prediction, multi-label classification and node clustering.

IV. Patient Similarity Framework

This section will describe the proposed patient similarity framework AHGCN-PS in detail. The framework is mainly divided into three modules. The first module uses EMR to build a medical attributed heterogeneous information network. The second module uses the medical heterogeneous graph convolution network to aggregate the neighbor information of the patients and then aggregates the semantic information of the meta-path through semantic attention to obtain the final feature representation of the patient. The third module uses the feature representation of the patient node to calculate the patient similarity.

A. Construction of Attributed Medical Heterogeneous Graph

1) Patient Medical Text Generation

Extracting the patients' medical text information from the EMR as the patients' attributes can obtain more accurate patient embedding representation. However, patient medical texts in EMR are lengthy and contain unnecessary information. In recent years, the pre-trained language model BART [36] has achieved the best performance in the summarization task. BART can better handle text content, it consists of two parts, bidirectional encoder and unidirectional autoregressive decoder. This paper adopts the BART model to generate the most important information related to patients. As shown in Fig. 3. We extract EMR medical text $X = \{x_1, x_2, ..., x_z\}$ as the input of BART decoder. Where $x_1, x_2, ..., x_z$ is a single word in medical text, z represents the number of medical texts. Randomly disrupt the order of the original words, and then use the autoregressive method to calculate through the decoder. Finally, obtain the medical text related to the patient $Y = \{y_1, y_2, ..., y_n\}$ (the final text length is less than the initial text length).

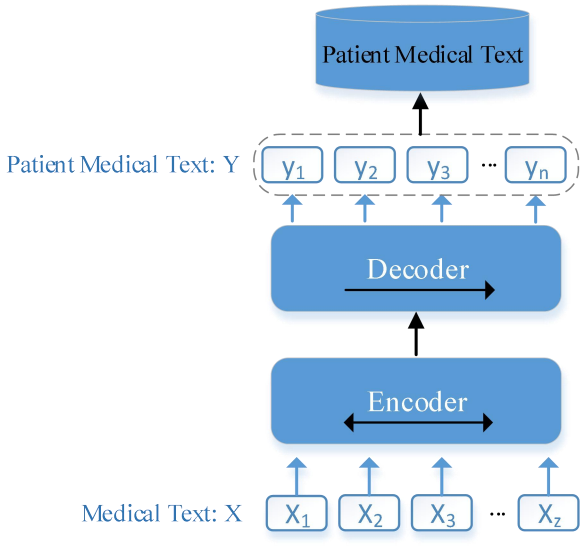

Fig. 3. Patient Medical Text Generation

2) Construct Medical Attributed Heterogeneous Graph

We extract patient-related medical entities (drugs, procedures) from EMR to construct a medical attributed heterogeneous information network. The network includes three types of nodes for patient, procedure, and drug, two types of edges for patient-procedure and patient-drug. AHGCN-PS uses medical text as the attribute information of each patient and uses a set of three types of nodes and different edge relationships between each node to connect nodes, to obtain a medical attributed heterogeneous information network.

B. The Embedding Learning of Medical Heterogeneous Graph Convolutional Network

1) Generation of patient adjacency matrix and feature matrix

Through the medical attributed heterogeneous information network, we construct the adjacency matrix and feature matrix for patients (P), procedure(T) and drug (D) in the network. For the patient adjacency matrix, we define two meta-paths, P-D-P and P-T-P to construct two adjacency matrices with different semantics, $A^{(1)}$, $A^{(2)}$ respectively. For $A^{(1)}$, if p_i uses d_j, there is an edge between p_i and d_j, where (i = 1, 2, …, n), (j = n+1, n+2, …, n+k), n is the number of patients, and k is the number of drugs. If p_i and p_j use the same drug, then there is an edge between p_i and p_j. For $A^{(2)}$, the same can be obtained. $A^{(1)}$ and $A^{(2)}$ are calculated as follows:

$$A_{ij}^{(1)} = \begin{cases} 1, p_i \text{ use } d_j \\ 1, p_i \text{ and } p_j \text{use } d_k \\ 0, otherwise \end{cases}$$
$$A_{ij}^{(2)} = \begin{cases} 1, p_i \text{ use } t_j \\ 1, p_i \text{ and } p_j \text{use } t_m \\ 0, otherwise \end{cases} \quad (1)$$

For the patient feature matrix, each patient p contains attributes $O=\{O^{(p_1)}, O^{(p_2)}, …, O^{(p_n)}\}$. In order to encode the patients' relevant medical entities (drugs, procedures) and the rich semantic information in the patients' medical text, this paper encodes each attribute information of patient p by one-hot. Firstly, number each word. Then we use one-hot to extract feature vectors for each paragraph to get the final vector representation, that is, the feature matrix X. $X \in \mathbb{R}^{|V| \times d}$ is a feature matrix containing node features, $x_v \in \mathbb{R}^d$ represents each row x_v as a feature vector of a node v, have d dimensional.

AHGCN-PS uses adjacency matrices $A^{(1)}$, $A^{(2)}$ and feature matrix X as input to medical heterogeneous graph convolutional network to learn potential entity representation.

2) Medical Heterogeneous Graph Convolutional Network

GCN (Graph Convolutional Network) [16] can learn graph-structured data and continuously update parameters through convolution, which improves the accuracy of the model and reduces computation time. Because traditional GCNs are used for homogeneous networks, considering the heterogeneity of our medical data, they cannot be directly applied to traditional GCNs. Therefore, this paper adopts the medical heterogeneous graph convolutional network, considering the diversity of different types of nodes. The specific structure is shown in Fig 4.

This paper uses HetGCN to consider different network structures, we transform the original large graph into patient-drug-patient and patient-procedure-patient subgraphs based on meta-path and project them with their respective

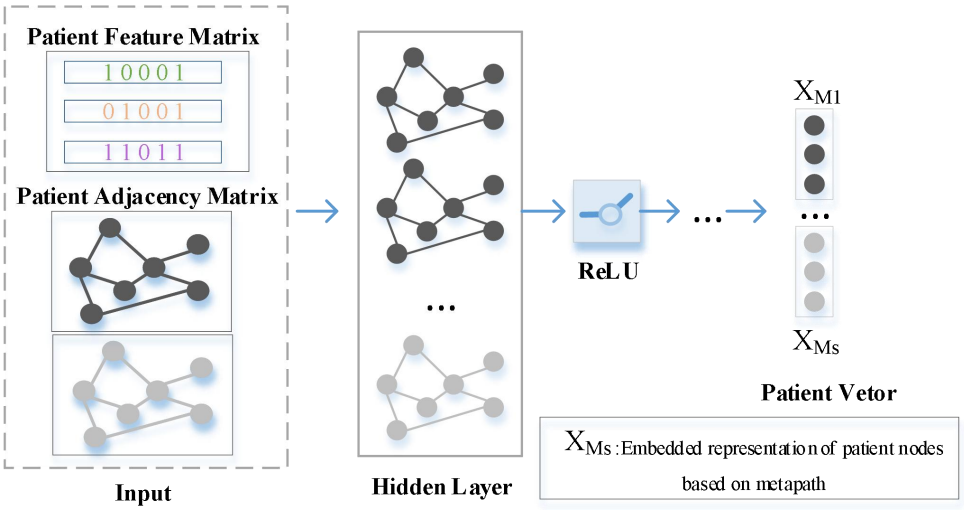

Fig. 4. Structure Hierarchy of Heterogeneous GCN Model

transformation matrices into a common feature space. Given a set of meta-paths $M = \{M_1, M_2, \ldots, M_S\}$, the corresponding adjacency matrix $A = \{A^{(1)}, A^{(2)}, \ldots, A^{(s)}\}$, where s represents the number of meta-paths and the number of adjacency matrices, and each meta-path corresponds to an adjacency matrix. Considering that the node itself also contains certain information, the adjacency matrix is processed into the following form.

$$\widetilde{\boldsymbol{A}} = A + I \tag{2}$$

where I is the identity matrix. Hierarchical propagation of medical heterogeneous graph convolutional networks as shown in the formula:

$$\boldsymbol{H}^{(l+1)} = \sigma\left(\sum_r \widetilde{\boldsymbol{A}}_r \cdot \boldsymbol{H}_r^{(l)} \cdot \boldsymbol{W}_r^{(l)}\right), (r = 1, 2, ..., s) \tag{3}$$

Where $\sigma(\cdot)$ represents the softmax activation function, $\widetilde{\boldsymbol{A}}_r \in \mathbb{R}^{|V|\times|V_r|}$ is the submatrix of $\widetilde{\boldsymbol{A}}$, V represents all nodes under type r, V_r represents the neighbor node of type r, $\boldsymbol{W}_r^{(l)} \in \mathbb{R}^{d^{(l)}\times d^{(l+1)}}$ represents the trainable parameter matrix. Consider that the patients' neighbors, that is, medical entities (procedures, drugs) have a certain effect on the patient. Therefore, we aggregate all patient neighbors and update patient nodes to obtain $\boldsymbol{H}^{(l+1)}$. We simultaneously learn the patients' attribute information and structural information in the GCN, so that the two complement each other and work together to influence the final patient node representation.

3) Semantic Layer Aggregation

Each node in the medical attributed heterogeneous information network contains different types of semantic information to learn more comprehensively structural and semantic information. This paper selects two meta-paths: patient-drug-patient and patient-procedure-patient. We obtain rich semantic information through these two meta-paths to obtain more precise patient similarity. The node embedding learned from the medical heterogeneous graph convolutional network is used as input, and the learned weight of each meta-path is set as β_{M_s}. The formula is as follows:

$$(\beta_{M_1}, \ldots, \beta_{M_s}) = \sum_{s=1}^{s} att_{sem}(\boldsymbol{X}_{M_s}) \tag{4}$$

Where att_{sem} represents a deep neural network for semantic-level attention, which is used to learn edge importance, and $\boldsymbol{X}_{M_s}$ is the final node embedding of heterogeneous graph convolution. Next, we obtain the importance e_{M_s} of each meta-path by averaging the potential representation vectors of all nodes $v \in V$ under a specific meta-path. The formula is as follows:

$$e_{M_s} = \frac{1}{|V|}\sum_{v\in V} \boldsymbol{q}^T \cdot tanh\left(\boldsymbol{W} \cdot \boldsymbol{X}_i^{M_s} + \boldsymbol{b}\right) \tag{5}$$

Where $\boldsymbol{W}$ is the weight matrix, $\boldsymbol{b}$ is the bias vector, and through the semantic level attention vector $\boldsymbol{q}$ to learn the importance of different semantic embedding. After getting the importance of each meta-path, we use the softmax function to normalize e_{M_s}. The final attention coefficient is obtained as:

$$\beta_{M_s} = \frac{exp(e_{M_s})}{\sum_{p=1}^{p} exp(e_{M_s})} \tag{6}$$

β_{M_s} represent the importance of meta-path s to nodes based on meta-path s. Finally, all nodes are weighted with the corresponding meta-path to obtain the final embedding representation:

$$p = \sum_{s=1}^{s} \beta_{M_s} \cdot \boldsymbol{X}_{M_s} \tag{7}$$

C. Patient Similarity Calculation

In this section, we will introduce the patient similarity calculation. Given a vector representation of a target patient, we use cosine similarity to calculate the similarity score between the patient and other patients. The similarity scores between patients are defined as follows:

$$score(p_i, p_j) = \frac{\hat{p}_i \cdot \hat{p}_j}{\|\hat{p}_i\|_2 \cdot \|\hat{p}_j\|_2} \tag{8}$$

Where p_i is the target patient, $\hat{p}_i$ is the embedding vector of the target patient, $\hat{p}_j$ is the vector representation of the query patient p_j. When the value between patient and patient is closer to 1, it means that the similarity between the two is higher, and vice versa. Sort according to the size of the similarity score to get the most similar patient to the target patient.

V. Experiments And Evaluation

In this section, we will evaluate the performance of the

TABLE I
DISEASE EXAMPLE OF PATIENT MEDICAL RECORDS

Patient ID	Procedures	Drugs	Patient medical text	Disease category
890	9604, 966, 9672, 110, 8751, 8754, 3891 ...	Insulin, Warfarin, Potassium Chloride, Magnesium Sulfate...	67-year-old male was admitted to hospital with abdomina lpain, hypotension, fever, ...	Septicemia
32779	3778, 8964, 966, 9672, 9960, 3891, 3893	Senna, Cisatracurium Besylate, D5W, PrediSONE...	A 69-year-old female with a history of diabetes, coronary artery disease, COPD who ...	Respiratory Failure
8911	3722, 8853, 8855, 3961, 3612, 3615	Atenolol, Simvastatin, Potassium Chloride, Simethicone, ...	The patient had been experiencing substernal chest pain starting early in ...	Coronary Disease

patient similarity framework through the experimental dataset. Firstly, introduce the experimental dataset, evaluation index and comparison method. Then perform patient disease classification, patient clustering, visualization tasks, Top-k similar patients and parameter sensitivity experiments to evaluate the effective performance of the framework. Finally, make a summary of the work. The experiment is based on python3.6 PyTorch in Intel (R) Xeon (R) E5-2620 v4 @ 2.10GHz hardware environment.

A. Dataset

MIMIC-III [37] (Medical Information Mart for Intensive Care III) is a large dataset of intensive care medical information that is freely available to the public, its purpose is to promote medical research and improve ICU decision support. This dataset records the medical information (such as vital signs, test results, medication characteristics, etc.) and demographic information (admission and discharge time, race, gender, medical orders, etc.) of ICU patients from Beth Israel Dikang Medical Center (BIDMC) from 2001 to 2012. All data resources in the MIMIC-III dataset are strictly de-identity information processing.

We selected five kinds of diseases from the MIMIC-Ⅲ dataset: respiratory failure, coronary disease, heart failure, sepsis and gastritis, and extracted drugs, procedures and patient medical text data from patients with these diseases. Deal with missing values and use preprocessed data for subsequent experiments. Table I is an example of patient medical records, and Table II is the statistical information of the datasets:

TABLE II
STATISTICS OF DATASETS

Number of nodes	Edges	Meta-paths	Node types
Patient (P)：7413 Drug (D)：1396 Procedure (T)：570	P-D：283760 P-T：66916	P D P P T P	3

B. Evaluation Metrics

After generating the feature representation of each patient, we use *Macro-F1* and *Accuracy* to evaluate the effect of disease classification, in addition, to evaluation of patient clustering using Rand Index (*RI*), *Purity*, and Normalized Mutual Information (*NMI*). Next, the detailed definition of five evaluation indicators are described as follows:

1) Macro-F1

In the multi-classification task, *TP* (true case), *FP* (false positive case), *FN* (false negative case) and *TN* (true negative case) are used to calculate the F1 value, where *n* is the number of disease categories, the formula is as follows:

$$Macro-F1=\frac{1}{n}\sum_{i=1}^{n}\frac{2TP(i)}{2TP(i)+FP(i)+FN(i)} \tag{9}$$

2) Accuracy

The accuracy rate is the proportion of positive samples after model training in the classification task to the total samples, where *S* is the total number of samples, defined as follows:

$$Acc=\frac{TP+TN}{S} \tag{10}$$

3) Rand Index (RI)

$$RI=(TP+TN)/\binom{n}{2} \tag{11}$$

Where *n* is the total number of patients, *TP* indicates that patients with the same type of disease are divided into the same cluster, *TN* indicates that different types of diseases are divided into different clusters. The higher the *RI* value, the better the patient clustering effect.

4) Purity

$$Purity=\sum_{i=1}^{k}\frac{x_i}{x}P_i \tag{12}$$

Where *k* is the total number of clusters and *x* is the number of members involved in the whole cluster partition.

5) Normalized Mutual Information (NMI)

NMI is usually used for data clustering to measure the similarity of the results of the two classes, the closer the *NMI* value is to 1, the better the patient clustering effect. The formula is defined as follows:

$$NMI(X,Y)=\frac{2\cdot I(X.Y)}{(H(X)+H(Y))} \tag{13}$$

Let the joint distribution of two random variables (*x*, *y*) be $p(x,y)$, and the edge distribution be $p(x)$ and $p(y)$ respectively Mutual information *I*(*X, Y*) is the relative entropy of joint distribution *P*(*x, y*) and *P*(*x*)(*y*), that is, the formula is:

$$I(X,Y)=\sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{14}$$

H(X) is the information entropy, and the formula as follows:

$$H(X)=-\sum_i p(x_i)\log p(x_i) \tag{15}$$

C. Baselines

In order to verify the effectiveness of the patient similarity framework, the experiments select homogeneous and heterogeneous graph neural networks for comparison:

- GCN. A homogeneous graph neural network, this paper tests the homogeneous graph based on meta-path, analyzes the structural relationship and learns the vector representation of nodes. Here we test all the meta-paths for GCN and report the best performance.
- GAT. A homogeneous graph neural network performs attention mechanism on homogeneous graphs and assigns different weights to different neighbors in different neighborhoods. Here we test all the meta-paths for GAT and report the best performance.
- HAN. A heterogeneous graph neural network based on GAT, by layering (node level attention and semantic level attention), aggregates neighbor nodes to generate node feature representation.
- AHGCN-PS$_{attr}$. This method is a kind of deformation of AHGCN-PS, and its difference from AHGCN-PS is that it does not consider the patients' attribute information for patient similarity learning.

D. Experimental Results and Analysis

For parameter settings, we set the number of iterations to 200, the dimension to 128, the learning rate to 0.001, the Adam optimizer, and the Dropout to 0.5. During the training process, Dropout randomly discards some nodes to prevent overfitting and improve the model effect. The parmeter settings of the comparative model are the same as those of AHGCN-PS.

1) Classification of Patient Diseases

We investigate the effectiveness of AHGCN-PS framework in patient disease classification tasks. When implementing GCN, GAT, HAN, AHGCN-PS$_{attr}$, and AHGCN-PS, we randomly selected 80 % of the data for learning and 20 % for testing. The results of the disease classification comparison are shown in Table III.

The experimental results show that AHGCN-PS has the best Accuracy and Macro-F1 values compared with other baseline models, the results can reach 0.873 and 0.845, respectively. The index values obtained by GCN and GAT are lower than those obtained by HAN and AHGCN-PS. This shows that heterogeneous graph neural networks may have better performance in graph neural networks. Leveraging heterogeneous node features helps improve embedding performance. Furthermore, in the absence of attribute (AHGCN-PS$_{attr}$), the experimentally obtained results are lower than AHGCN-PS. This shows that incorporating attribute information will further improve the performance of the framework. To sum up, AHGCN-PS achieves better performance in disease classification tasks compared with other baseline models.

TABLE III
PATIENT DISEASE CLASSIFICATION RESULTS

Model	Accuracy	Macro-F1
GCN	0.823	0.773
GAT	0.853	0.812
HAN	0.855	0.821
AHGCN-PSattr	0.869	0.840
AHGCN-PS	**0.873**	**0.845**

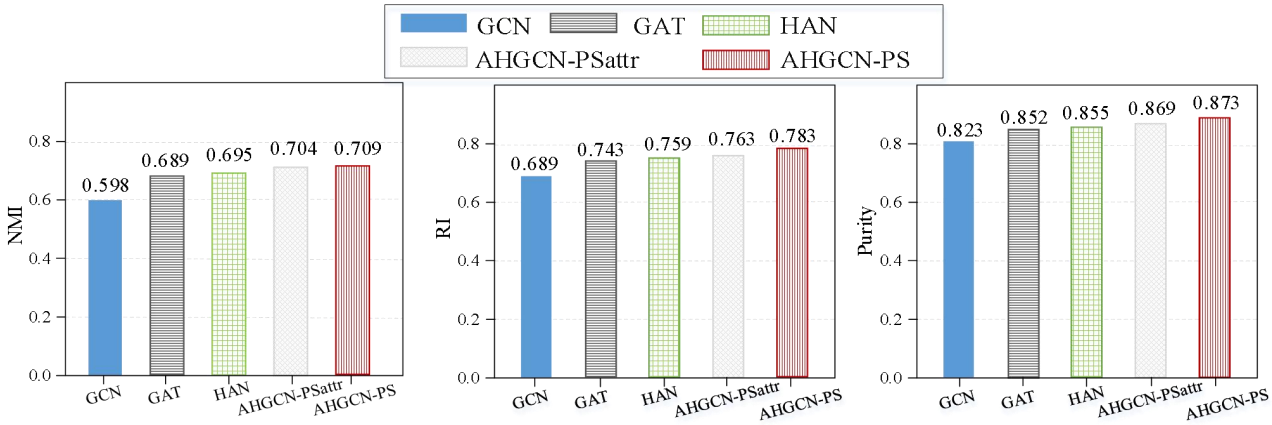

Fig. 5. Patient Clustering Results

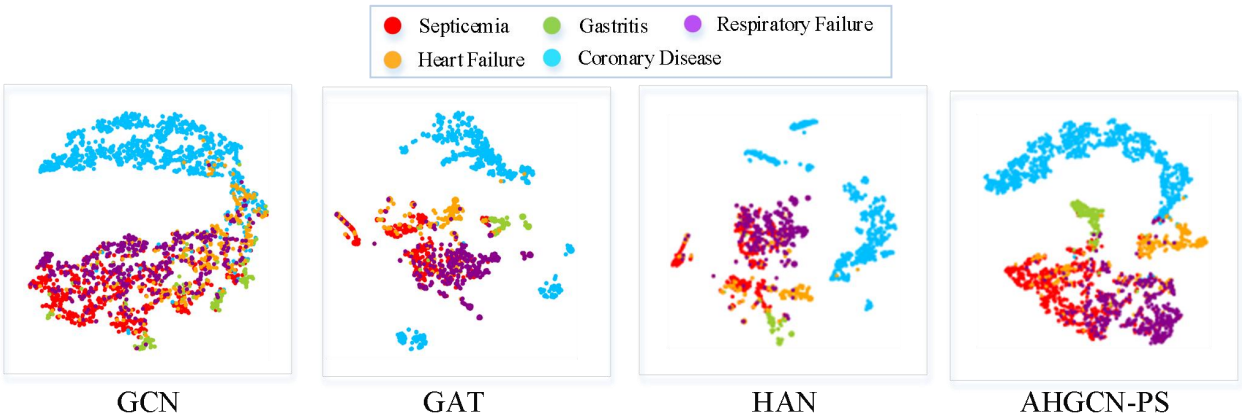

Fig. 6. Visualization Results of Patient

2) Patient Clustering Results

Patient clustering analysis can further observe the disease distribution of patients. This experiment evaluates the performance of patient clustering through *NMI*, *RI* and *Purity*, the clustering effect is shown in Fig.5. AHGCN-PS can achieve the best performance on *NMI*, *RI* and *Purity*, which are 0.709, 0.783 and 0.873 respectively. Followed by HAN, with indicators of 0.695, 0.759 and 0.855 respectively. The clustering gap between HAN and AHGCN-PS is not large, because both of them take into account the semantic information based on meta-path. Therefore, it is necessary to consider the rich semantic information of nodes when aggregating nodes, which can further improve the performance of patient clustering.

3) Visualization

Fig.6 shows the experimental results of patient visualization. In this experiment, the vector representation of the patient is dimensional reduced by t-SNE and displayed in two-dimensional space. We select five disease categories: Respiratory Failure, Coronary Disease, Heart Failure, Septicemia, and Gastritis, map each patient into a two-dimensional vector and finally Visualize each patient vector as a point in two-dimensional space. Points of different colors in the graph represent different types of diseases that patients suffer.

It can be seen from the figure that the visualization effect of patient vector representation obtained by GCN model is unsatisfactory. The dots corresponding to each disease are confounded with each other. For GAT and HAN, the clusters corresponding to each disease are formed, but the points of several diseases still overlap with each other. In terms of distribution, AHGCN-PS has better visualization effect than others.

4) Top-K Most Similar Patients

TABLE IV
TOP-3 SIMILARITY PATIENTS (GCN)

Patient ID	1st Patient ID	2nd Patient ID	3rd Patient ID	Disease category
620	828	2234	215	Respiratory Failure
929	1195	2197	635	Coronary Disease
335	446	2370	647	Heart Failure
1977	2289	932	2216	Septicemia
2161	1250	31	1113	Gastritis

TABLE V
TOP-3 SIMILARITY PATIENTS (GAT)

Patient ID	1st Patient ID	2nd Patient ID	3rd Patient ID	Disease category
620	2075	1675	250	Respiratory Failure
929	834	1408	71	Coronary Disease
335	595	116	1492	Heart Failure
1977	1839	960	2178	Septicemia
2161	1917	1413	1538	Gastritis

TABLE VI
TOP-3 SIMILARITY PATIENTS (HAN)

Patient ID	1st Patient ID	2nd Patient ID	3rd Patient ID	Disease category
620	1675	250	1112	Respiratory Failure
929	1930	2203	292	Coronary Disease
335	1583	1870	2036	Heart Failure
1977	1479	1347	392	Septicemia
2161	1538	18	2245	Gastritis

TABLE VII
TOP-3 SIMILARITY PATIENTS (AHGCN-PS)

Patient ID	1st Patient ID	2nd Patient ID	3rd Patient ID	Disease category
620	1675	2238	1112	Respiratory Failure
929	292	856	1861	Coronary Disease
335	1994	826	196	Heart Failure
1977	2293	2383	558	Septicemia
2161	1538	734	2245	Gastritis

In this experiment, the embedded representation of patients is obtained through GCN, GAT, HAN and AHGCN-PS, we randomly selected one patient (a total of 5) from each type of disease for testing, select Top-k (k = 3) patients for each patient, and the experimental results are shown in Table IV, V, VI, and VII. The results of AHGCN-PS are quite different from those of GCN and GAT, the main reason for this phenomenon is that they are homogeneous graph neural networks, not fully considering the structural and semantic information of nodes. For example, patient ID: 620 the Top-k results obtained through GCN and GAT learning are completely different from AHGCN-PS, but similar to HAN, but close to the similar patient ID number of HAN (such as patient ID: 929, 2161), HAN is a heterogeneous graph neural network like AHGCN-PS, which fully considers the heterogeneity of nodes and aggregates different semantic information.

5) Parameter Sensitivity Test

In this section, we mainly study the influence of parameters on the experimental results of AHGCN-PS and analyze them from the aspects of node embedding dimension *d* and meta-path.

a) Dimension

Fig 7 observes the three indicators of clustering with the change of node embedding dimension *d*, *d* set 16, 32, 64, 128, 200, respectively. The results show that the three indicators are constantly changing with the growth of dimension. When $d = 128$, the indicators maximize, and when $d > 128$, all indicators show a downward trend.

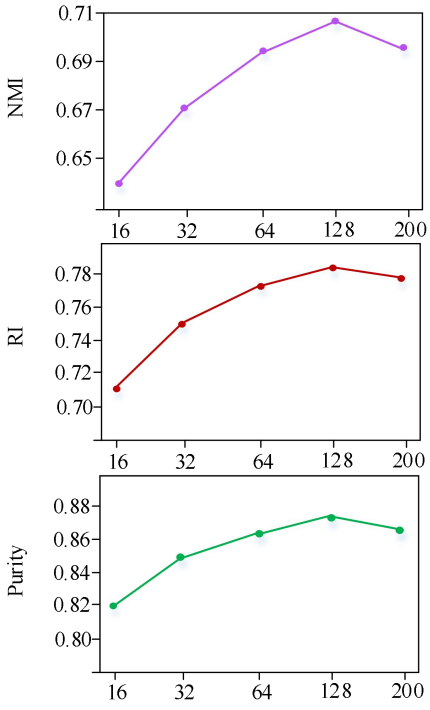

Fig. 7. The Sensitivity Experiments of Parameter

b) Meta-path

Different types of meta-paths perform differently on the patient similarity framework. As shown in Table VIII. The patient-procedure-patient meta-path indexes are higher than the patient-drug-patient indexes. However, the combination of the two has significantly improved its performance, the *NMI*, *RI* and *Purity* of AHGCN-PS are 0.709, 0.783 and 0.873 respectively. So meta-paths with rich semantics can improve the performance of framework.

TABLE VIII
THE PERFORMANCE OF DIFFERENT META-PATHS

Meta-path	NMI	RI	Purity
Patient - Drug - Patient	0.535	0.651	0.790
Patient - Procedure - Patient	0.688	0.765	0.858

VI. CONCLUSIONS AND FUTURE WORK

Previous patient similarity studies only conducted patient similarity analysis by learning related entities of patients, ignoring the attribute information contained in entities, and failing to make full use of patient feature information. Moreover, patients and their related medical entities contain a large amount of structural information and semantic information. How to make full use of this information is a major challenge in patient similarity research. In response to the above problems, this paper proposes a patient similarity framework AHGCN-PS. The framework uses BART to extract the patients' medical text as the patients' attribute information and selects two meta-paths (patient-drug-patient, patient-procedure-patient) with different semantics to improve the accuracy of patient similarity results. After using the medical heterogeneous graph convolutional network to aggregate the neighbor information in the medical attributed heterogeneous graph, using the attention mechanism to further aggregate the semantic information of the meta-path to learn richer information. Efficiently mining and aggregating potential relationships between patients to obtain patient node representation for patient similarity calculation. We conduct a patient similarity study by extracting medical entities from the dataset. The experimental results show that AHGCN-PS achieves better performance than other baseline models.

The next stage of research focuses on how to extract meta-paths of different types and lengths, and introduce multimodal patient feature data, such as drug description information, procedure text and picture information, etc. To further improve the clustering effect of patient similarity.

REFERENCES

[1] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk Prediction with Electronic Health Records: A Deep Learning Approach," In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 2016, pp. 432–440. doi: 10.1137/1.9781611974348.49.

[2] J. Sun, F. Wang, J. Hu, and S. Edabollahi, "Supervised patient similarity measure of heterogeneous patient records," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 1, pp. 16–24, 2012.

[3] Z. Zhu, C. Yin, B. Qian, Y. Cheng, J. Wei and F. Wang, "Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding," *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, pp. 749-758, doi: 10.1109/ICDM.2016.0086.

[4] C. Shi, Y. Li, J. Zhang, Y. Sun and P. S. Yu, "A Survey of Heterogeneous Information Network Analysis," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17-37, 1 Jan. 2017, doi: 10.1109/TKDE.2016.2598561.

[5] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks," *Proceedings of the Vldb Endowment,* vol. 4, no. 11, pp. 992-1003, 2011.

[6] H. Jiang, and D. Yang, "Learning Graph-based Embedding from EHRs for Time-aware Patient Similarity," *Engineering Letters,* vol. 28, no. 4, pp1254-1262, 2020.

[7] J. Sun, D. Sow, J. Hu, and S. Ebadollahi, "Localized Supervised Metric Learning on Temporal Physiological Data," in *Proc. 20th Int. Conf. Pattern Recognit*., Aug. 2010, pp. 4149–4152, doi: 10.1109/ICPR. 2010.1009.

[8] Huang Y, Wang N, Liu H, et al. "Study on Patient Similarity Measurement Based on Electronic Medical Records," *Studies in Health Technology and Informatics*, 2019 Aug;264:1484-1485. doi: 10.3233/shti190496. PMID: 31438193.

[9] N. Wang, Y. Huang, H. Liu, Z. Zhang, and H. Chen, "Study on the semi-supervised learning-based patient similarity from heterogeneous electronic medical records," *BMC medical informatics and decision making*, vol. 2021, no. Suppl 2, p. 58.

[10] L. Joon, D. M. Maslove, J. A. Dubin, and E. S. Frank, "Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric," *PLoS ONE,* vol. 10, no. 5, p. e0127428, 2015.

[11] Z. Lin, and D. Yang, "Medical Concept Embedding with Variable Temporal Scopes for Patient Similarity," *Engineering Letters*, vol. 28, no. 3, pp651-662, 2020.

[12] Q. Suo, W. Zhong, F. Ma, Y. Ye, M. Huai and A. Zhang, "Multi-task Sparse Metric Learning for Monitoring Patient Similarity Progression," *2018 IEEE International Conference on Data Mining (ICDM)*, 2018, pp. 477-486, doi: 10.1109/ICDM.2018.00063.

[13] Zhang, C., Gao, X., Ma, L., Wang, Y., Wang, J. and Tang, W. (2021). "GRASP: Generic Framework for Health Status Representation Learning Based on Incorporating Knowledge from Similar Patients," *Proceedings of the AAAI Conference on Artificial Intelligence*, 35, 1 (May 2021), 715-723.

[14] Y. Wang, W. Chen, B. Li, and R. Boots, "Learning Fine-Grained Patient Similarity with Dynamic Bayesian Network Embedded RNNs," DASFAA 2019. *Lecture Notes in Computer Science*, vol 11446. Springer, Cham. doi:10.1007/978-3-030-18576-3_35.

[15] W. L. Hamilton, R. Ying, and J. Leskovec. 2017. "Inductive Representation Learning on Large Graphs," In *NIPS*. 1024–1034.

[16] Thomas N Kipf and Max Welling. 2016. "Semi-Supervised Classification with Graph Convolutional Networks," In *ICLR*.

[17] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. "Graph Attention Networks, " In *ICLR*.

[18] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2016. "Gated Graph Sequence Neural Networks," In *ICLR*.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All you Need, " In *NIPS*. 5998–6008.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *Computer Science*, 2014.

[21] X. Han, C. Shi, S. Wang, P. S. Yu, and S. Li, "Aspect-Level Deep Collaborative Filtering via Heterogeneous Information Networks," In *Twenty-Seventh International Joint Conference on Artificial Intelligence {IJCAI-18}*, 2018.

[22] B. Hu, C. Shi, and W. Zhao, "Leveraging Meta-path based Context for Top- N Recommendation with A Neural Co-Attention Model," In *the 24th ACM SIGKDD International Conference*, 2018.

[23] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous Graph Transformer," In *Proceedings of The Web Conference 2020*: Association for Computing Machinery, 2020, pp. 2704–2710.

[24] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. "Heterogeneous Graph Attention Network,". In *WWW*. 2022–2032, 2019.

[25] L. Yao, C. Mao, and Y. Luo, "Graph Convolutional Networks for Text Classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7370-7377, 2019.

[26] Weijian Chen, Yulong Gu, Zhaochun Ren, Xiangnan He, Hongtao Xie, Tong Guo, Dawei Yin, and Yongdong Zhang. 2019. "Semi-supervised User Profiling with Heterogeneous Graph Attention Networks,". In *IJCAI*, Vol. 19. 2116–2122.

[27] A. Li, Z. Qin, R. Liu, Y. Yang, and D. Li, "Spam Review Detection with Graph Convolutional Networks," In *CIKM*. 2703–2711, 2019.

[28] S. Cao, L. Wei, and Q. Xu, "GraRep: Learning Graph Representations with Global Structural Information," *ACM*, 2015.

[29] A. Grover and J. Leskovec, "node2vec: Scalable Feature Learning for Networks," In *ACM,* 2016.

[30] D. Wang, P. Cui, and W. Zhu, "Structural Deep Network Embedding," In *the 22nd ACM SIGKDD International Conference*. 1225–1234, 2016.
[31] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable Representation Learning for Heterogeneous Networks," In *the 23rd ACM SIGKDD International Conference*, 2017.
[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Computer Science,* 2013.
[33] C. Shi, B. Hu, W. X. Zhao and P. S. Yu, "Heterogeneous Information Network Embedding for Recommendation," In *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 357-370, 1 Feb. 2019, doi: 10.1109/TKDE.2018.2833443.
[34] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online Learning of Social Representations," In *the 20th ACM SIGKDD International Conference*, 2014.
[35] H. Chen, H. Yin, W. Wang, H. Wang, Q. V. H. Nguyen, and X. Li, "PME: Projected Metric Embedding on Heterogeneous Networks for Link Prediction," In *the 24th ACM SIGKDD International Conference*, 2018.
[36] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *arXiv preprint arXiv:1910.13461,* 2019.
[37] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific data,* vol. 3, no. 1, pp. 1-9, 2016.