Artificial Intelligence In Medicine 150 (2024) 102805

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

GCNGAT: Drug–disease association prediction based on graph convolution neural network and graph attention network

Runtao Yang [a], Yao Fu [a], Qian Zhang [b], Lina Zhang [a,*]

[a] School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai, 264209, China
[b] Heze Institute of Science and Technology Information, Heze, 274000, China

ARTICLE INFO

Keywords:
Drug–disease association
Graph convolutional network
Graph attention network
Heterogeneous graph structure

ABSTRACT

Predicting drug–disease associations can contribute to discovering new therapeutic potentials of drugs, and providing important association information for new drug research and development. Many existing drug–disease association prediction methods have not distinguished relevant background information for the same drug targeted to different diseases. Therefore, this paper proposes a drug–disease association prediction model based on graph convolutional network and graph attention network (GCNGAT) to reposition marketed drugs under the distinguishment of background information. Firstly, in order to obtain initial drug–disease information, a drug–disease heterogeneous graph structure is constructed based on all known drug–disease associations. Secondly, based on the heterogeneous graph structure, the corresponding subgraphs of each group of drug–disease association pairs are extracted to distinguish different background information for the same drug from different diseases. Finally, a model combining Graph neural network with global Average pooling (GnnAp) is designed to predict potential drug–disease associations by learning drug–disease interaction feature representations. The experimental results show that adding subgraph extraction can effectively improve the prediction performance of the model, and the graph representation learning module can fully extract the deep features of drug–disease. Using the 5-fold cross-validation, the proposed model (GCNGAT) achieves AUC (Area Under the receiver operating characteristic Curve) values of 0.9182 and 0.9417 on the PREDICT dataset and CDataset dataset, respectively. Compared with other predictors on the same dataset (PREDICT dataset), GCNGAT outperforms the existing best-performing model (PSGCN), with a 1.58% increase in the AUC value. It is anticipated that this model can provide experimental reference for drug repositioning and further promote the drug research and development process.

1. Introduction

New drug research and development is an important solution for disease treatment, which determines the development level of a country's pharmaceutical industry and affects the living and health level of its people [1]. Although the national investment in the new drug research and development continues to increase, the number of new drugs approved by the country gradually decreases [2]. In order to solve the dilemma of high investment and low success rate of the new drug development, a new method (drug repositioning) has been widely used to redefine the therapeutic indications of drugs that have been marketed or failed to be marketed [3–6]. For example, based on the drug repositioning, a joint research team composed of Professor Li Hua from Huazhong University of Science and Technology and Professor Chen Lixia from Shenyang Pharmaceutical University conducted screening research on potential anti-COVID-19 drugs at the time of COVID-19 virus outbreak [7]. To discover drugs for the treatment of SARS-CoV-2, Serafin et al. conducted a repositioning study on seven drugs [8].

Drug repositioning can effectively save cost and time while ensuring safety by expanding the application scope of marketed drugs to find new application objects of unmarketed drugs [9]. It is estimated that the cost of repositioning a drug to the market is 1/10 of the cost of developing a new compound drug, and the time spent on drug repositioning is 1/3 of the time spent on new drug development [10,11]. At present, successfully repositioned drugs have been widely used in practical medical treatment. For example, sildenafil, originally used to treat cardiovascular diseases, is one of the common drugs used to treat erectile dysfunction [12]. Zidovudine was initially studied as a chemotherapeutic drug, but clinical trials have consistently failed. After drug repositioning, it became the first anti AIDS drug approved by

* Correspondence to: School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai, Wenhuaxi Road No. 180, Shandong Province, China.
E-mail addresses: yrt@sdu.edu.cn (R. Yang), fuyao1207@foxmail.com (Y. Fu), 151314964@qq.com (Q. Zhang), zln@sdu.edu.cn (L. Zhang).

https://doi.org/10.1016/j.artmed.2024.102805
Received 31 March 2023; Received in revised form 22 January 2024; Accepted 8 February 2024
Available online 17 February 2024

the US food and drug administration [13]. Thalidomide, used to treat leprosy, has also been found to be able to treat erythema nodosum and multiple myeloma through drug repositioning [14]. Metformin, used to treat mild diabetes, can significantly inhibit tumor growth when used in a fasting state [15].

As a type of drug repositioning, drug–disease association prediction can provide guidance and direction for biological experiments to detect new uses of drugs, screen treatable candidate diseases for each drug, and discover new therapeutic targets for related diseases, thereby improving the success rate of new drug research and development [16,17]. Currently, drug–disease association prediction algorithms are mainly based on matrix decomposition, traditional machine learning, or deep learning.

The drug–disease association prediction algorithms based on matrix decomposition project drugs and diseases into shared potential space, represent drug and disease characteristics through potential vectors, and calculate the association similarity through inner product multiplication. Based on the drug–disease heterogeneous network, Luo et al. reconstructed the drug–disease association matrix by combining low-rank matrix approximation and singular value thresholding algorithm, and proposed an association prediction model named DRRS by taking the inner product of the new matrix as the drug–disease similarity [18]. Yang et al. concatenated multiple drug–drug similarity matrices and disease–disease similarity matrices, decomposed drug–disease association matrices into drug feature matrices and disease feature matrices based on non-negative factorization, and proposed the association prediction model MSBMF [19]. Through an improved similarity network fusion algorithm to integrate drug and disease similarities, Sadeghi et al. adopted the idea of recommendation systems and non-negative matrix factorization to obtain the prediction scores of drug–disease adjacency matrix and potential drug–disease association pairs [20].

The drug–disease association prediction algorithms based on traditional machine learning firstly preprocess drug and disease information into corresponding features, and then construct a model to predict the resulting features. Through three methods: direct mapping to drugs, MetaMap based mapping to UMLS, and indication based mapping, Gottlieb et al. constructed drug–disease features, and then spliced them into a logistic regression classifier to obtain prediction results [21]. Building a drug–disease heterogeneous network through known associations, Yang et al. obtained features through network embedding, and introduced support vector machines to train and classify drug and disease features [22]. Taking protein as the mediator of drug–disease, Kitsiranuwat et al. extracted interaction features containing drug-protein-disease information, and applied random forest to predict drug–disease associations [23].

The drug–disease association prediction algorithms based on deep learning can automatically extract deep effective features and achieve end-to-end learning. CBPred [24] is a deep learning model built by Ye et al. to predict drug–disease associations. To represent the overall properties of drugs, diseases and the topological structure of drug–disease associations, CBPred employed long short-term memory networks and convolutional neural networks to capture path features and local features, respectively. The deepDR model developed by Zeng et al. integrated various heterogeneous network information through a multimodal deep autoencoder to obtain high-level drug–disease features, and infered potential associations between drugs and diseases through decoding by a variational autoencoder [25]. By fusing drug chemical structure and disease phenotypic features with drug–disease topological structure, HNet-DNN constructed drug–disease heterogeneous network and extracted drug–disease interaction features through deep neural network [26]. Based on five drug characteristics, Yu et al. constructed drug similarity, integrated them with disease similarity, extracted drug–disease related features using an input layer attention graph convolutional neural network, and ultimately applied a bilinear decoder to predict drug–disease associations [27]. The CTST model proposed by Gao et al. integrated multiple drug networks and disease networks, extracted deep features through a graph convolution autoencoder with shared parameters, and obtained the final drug–disease prediction results through an attention integration module [28]. Combining the biological knowledge of drugs and diseases with drug-protein-disease topology, Zhao et al. presented HINGRL based on heterogeneous information network and graph representation learning [29]. Sun et al. proposed PSGCN based on drug–disease heterogeneous graph structure, which combines a graph convolutional neural network and a layer attention mechanism to automatically capture context information of multi-level drug–disease association pairs [30].

The above-mentioned methods have their own merits and achieve satisfactory results. However, some inherent gaps still exist. Firstly, the methods based on matrix decomposition rely on existing data for shallow low-rank decomposition, which cannot deeply learn abstract features of drugs and diseases to capture complex structures. Secondly, traditional machine learning based methods need to manually screen and construct drug–disease features, and cannot guarantee the effectiveness of the obtained features. Thirdly, the action mechanisms of the same drug on different diseases are different. Most of the existing methods cannot automatically extract the corresponding background information based on different drug–disease association pairs, resulting in the failure to obtain a more accurate initial feature representation of drug–disease. To sum up, the pain point in this field is how to simply obtain high-quality drug–disease interaction vectors, which is also the motivation of this study.

Compared with the algorithms based on matrix decomposition and traditional machine learning, the algorithms based on deep learning have stronger learning ability, and their prediction performance can be improved with the increase of data. In addition, this kind of algorithms does not require manual feature selection, and can automatically extract deep interaction features of drugs and diseases, making the operation simpler and more efficient. Due to the similarity between drug–disease association prediction and link prediction in graph structures, a deep learning algorithm based on graph convolutional neural network and graph attention network in deep learning (GCNGAT) is proposed in this study to obtain the deep drug–disease interaction vectors.

As shown in Fig. 1, the workflow of GCNGAT is as follows. Firstly, based on known drug–disease association pairs, a drug–disease heterogeneous graph structure containing all association information is constructed as the initial information of this association prediction model. Secondly, taking each drug–disease association pair as the center, the corresponding neighborhood information is extracted to obtain the corresponding subgraph of each association pair. Finally, a GnnAp model that includes a graph representation learning module, a global average pooling layer, and a prediction module is designed to learn and predict drug–disease associations. The graph representation learning module is composed of two graph convolution layers and one graph attention layer, which can obtain the representation of drug–disease interaction features contained in the corresponding subgraph. The global average pooling layer is used to unify the dimensionality of the obtained drug–disease feature vector. The prediction module consists of two fully connected layers and a ReLU activation function to determine whether there is a association between drugs and diseases.

The main contributions of this study can be summed up in the following aspects. (i) Given the different action mechanisms of the same drug on different diseases, the corresponding subgraphs of each drug–disease association pair are extracted based on the heterogeneous graph structure to obtain different background information for each association pair. (ii) To capture the global and multi-scale features of the corresponding subgraphs efficiently, a GnnAp model containing a graph representation learning module, a global average pooling layer, and a prediction module is constructed.

The paper is organized as follows. In the Section 2, the dataset involved is first described, followed by an introduction to drug–disease subgraph extraction. Then, the principles of graph convolutional neural

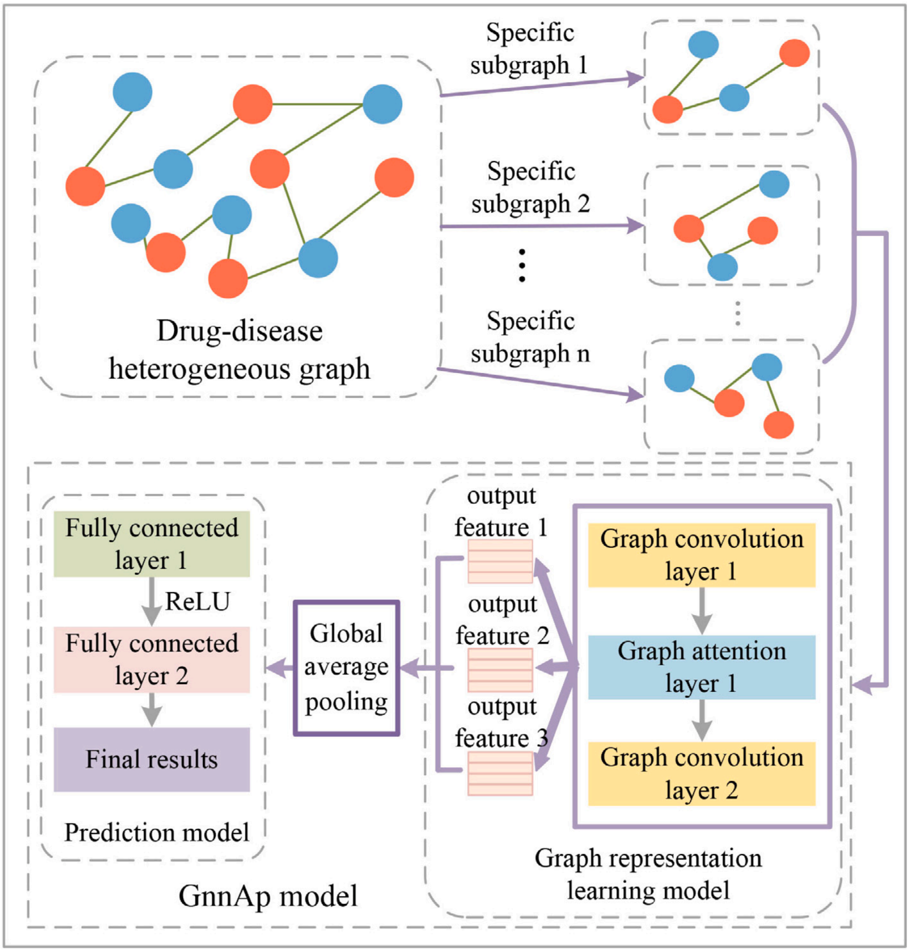


Fig. 1. The flowchart of the proposed method GCNGAT.

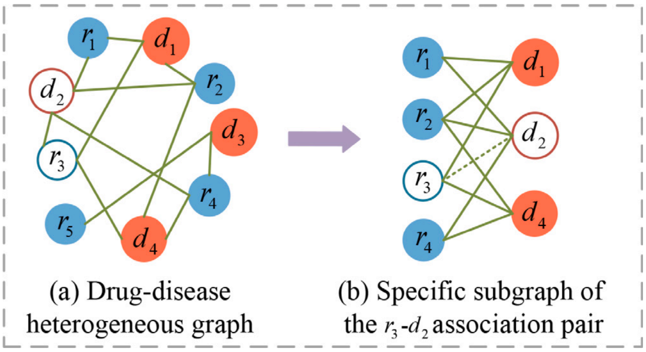
network and graph attention network are explained, and finally, the drug–disease association prediction model is introduced in detail. In the Section 3, we first describe the model evaluation measures and the experimental setup, then introduce the selection of hyperparameters, and finally provide experimental results and analyze them. The conclusions are presented in the Section 4.

2. Materials and methods

2.1. Extraction of drug–disease subgraphs

To obtain specific background information for each drug–disease association pair, this study extracts the drug–disease subgraphs. Firstly, a drug–disease graph structure is constructed. Secondly, adjacent nodes are extracted with each drug–disease association pair as the center. Finally, the corresponding nodes and adjacent nodes of each drug–disease association pair are integrated to form a specific subgraph of the drug–disease association pair. The specific processes to extract the drug–disease subgraphs are as follows.

As indicated in [30], the more indirect associations between nodes, the greater the likelihood of associations between nodes. Based on this prior knowledge, the process of create a graph is as follows. Based on known drug–disease associations, a drug–disease graph structure $G = (R, D, E)$ is constructed, where $R = \{r_1, r_2, \ldots, r_m\}$ represents all drug nodes, and m is the number of drug nodes; $D = \{d_1, d_2, \ldots, d_n\}$ represents all disease nodes, and n is the number of disease nodes; $E = \{(r_i, d_j) | r_i \in R, d_j \in D\}$ represents the set of all known drug–disease associations. The adjacency matrix of the drug–disease graph structure is represented by A with a dimension of $M \times N$. If there is an association edge between drug r_i and disease d_j, $A_{ij} = 1$, otherwise $A_{ij} = 0$. To prevent information leakage, test set data has been removed.



(a) Drug-disease heterogeneous graph

(b) Specific subgraph of the r_3-d_2 association pair

Fig. 2. The specific subgraph extraction of the $r_3 - d_2$ association pair.

Taking the drug-disease association pairs ($r_i - d_j$ association pair) as the center, the adjacent nodes of drug r_i and disease d_j are extracted. Then, taking the drug r_i, disease d_j and their surrounding adjacent nodes as nodes, and taking all known associations between the above nodes except the $r_i - d_j$ association as edges, a specific subgraph of the $r_i - d_j$ association pair is constructed. As shown in Fig. 2, taking the specific subgraph extraction of the $r_3 - d_2$ association pair as an example, the adjacent nodes of the drug r_3 include d_1 and d_4, and the adjacent nodes of the disease d_2 include r_1, r_2, and r_4. Therefore, the nodes in the subgraph are $\{r_1, r_2, r_3, r_4, d_1, d_2, d_4\}$, and the edges of the subgraph are

$\{(r_1, d_1), (r_1, d_2), (r_2, d_1), (r_2, d_2), (r_2, d_4), (r_3, d_1), (r_3, d_4), (r_4, d_2), (r_4, d_4)\}$. Through the above process, we can know that each drug-disease association pair corresponds to a subgraph. Therefore, the number of subgraphs is equal to the number of drug–disease association pairs.

Finally, based on the above-mentioned processes, specific subgraphs of all samples in the dataset are obtained as initial feature representations for each group of samples.

2.2. Construction of association prediction model

This section elaborates the association prediction models involved in this study. Firstly, the working principles of graph convolution neural network and graph attention network are introduced in detail. Secondly, to learn the drug–disease interaction information in the subgraph, the GnnAp model with a graph representation learning module and a prediction module is constructed.

2.2.1. Graph convolution network

Graph Convolutional Network (GCN) is a neural network that can directly perform convolution operations on the graph structure data without transitional invariance. It can extract potentially important features from the graph structure for downstream tasks such as node classification, graph classification and link prediction [31]. Previous study suggests that in addition to learning features of individual subjects, the graph convolutional network can simultaneously can introduce the pairwise similarity between subjects [32]. There are two main types of GCN, spatial domain based GCNs and spectral domain based GCNs [33]. The spatial domain based GCN defines a graph convolution operator based on the spatial relationship of a node, that is, it simulates the convolution operation of a Euclidean space in the local region of the graph structure. The spectral domain based GCN maps the graph structure to the frequency domain space by using the graph Fourier transform, and the convolution in the time domain is achieved by taking the product in the frequency domain space. Finally, the product result is mapped back to the time domain space through the inverse Fourier transform.

The GCN layer used in the GnnAp model belongs to the spectral domain based GCN. Taking the feature matrix X and the adjacency matrix A of the drug–disease heterogeneous graph structure as inputs, the propagation mode between the graph convolutions can be obtained from Eq. (1).

$$X^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X^{(l)} W^{(l)}), \tag{1}$$

where $\tilde{A} = A + I$. I represents the identity matrix of the node. $\tilde{D}$ is the degree matrix derived from $\tilde{A}$. $W^{(l)}$ is the weight matrix of the lth layer, and σ represents the nonlinear activation function.

2.2.2. Graph attention network

Graph Attention Network (GAT) introduces the attention mechanism into the spatial domain based graph neural network, updating node features only through the representation of first-order adjacent nodes [34]. GAT preserves the first-order approximate positioning characteristics of graph convolution while increasing adaptive edge weight coefficients. It uses attention mechanisms to learn different weights for different adjacent node features, and pays more attention to local feature extraction [35]. Previous study demonstrates the outstanding performance of the GAT in predicting molecular properties [36].

The graph attention layer transforms the input node features into new feature vectors containing adjacent node information through a shared weight matrix and a self attention mechanism [37]. Assuming that there is a graph structure with N nodes, whose node feature matrix is represented as H with a dimension of $N \times F$, and F represents the feature vector dimension of each node, then the input of the graph attention layer is H. Assuming that the dimension of the new node feature vector is F', then the output of the graph attention layer is H' with a dimension of $N \times F'$. As shown in Fig. 3, the specific transformation of the graph attention layer is as follows. First, for all input nodes, a weight matrix W with a dimension of $F' \times F$ representing the relationship between the F input features and the output F' features is trained. Secondly, the self-attention mechanism is implemented for each node, and the attention coefficient that represents the importance of node j's features to node i is defined as Eq. (2).

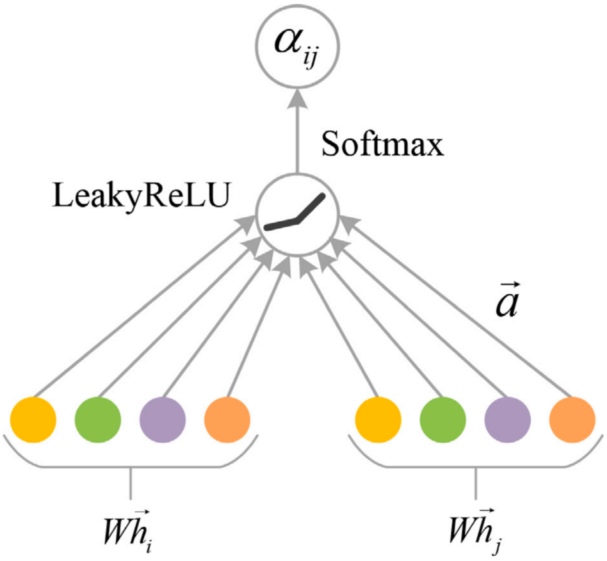


Fig. 3. The calculation process of the attention coefficient.

$$e_{ij} = a(W\vec{h}_i, W\vec{h}_j) = \text{LeakyReLU}(\vec{a}^T[W\vec{h}_i \parallel W\vec{h}_j]), \tag{2}$$

where $\text{LeakyReLU}(x) = \begin{cases} x, & x > 0 \\ \lambda x, & x \leq 0 \end{cases}$. $\vec{h}_i$ represents the feature vector of node i with a dimension of $1{\times}F$. a is a single-layer feedforward neural network parameterized by the weight vector $\vec{a}$. $\vec{a}^T$ represents the transpose of $\vec{a}$, and $\parallel$ represents the join operation. $\lambda = 0.2$ can regulate the range of the ReLU function.

In order to make the attention coefficient easy to compare between different nodes, the Softmax function is used to regularize the adjacent nodes of all nodes. The final normalized attention coefficient is calculated by Eq. (3).

$$\alpha_{ij} = \text{Softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k\in N_i} \exp(e_{ik})}, \tag{3}$$

where N_i represents the set of adjacent nodes of node i.

Finally, based on the attention coefficients between different nodes after regularization, the output features of each node are obtained as Eq. (4).

$$\vec{h}'_i = \sigma(\sum_{j\in N_i} \alpha_{ij} W \vec{h}'_j), \tag{4}$$

where W represents the shared weight matrix, and σ represents the nonlinear activation function.

To enable the self attention mechanism to stably represent nodes, the multi-headed attention mechanism [38] is introduced to improve the representation ability of the model. Specifically, K independent attention mechanisms execute Eq. (4), then average all attention results, and delay the application of the nonlinear activation function. The final output feature can be calculated by Eq. (5).

$$\vec{h}'_i = \sigma(\frac{1}{K} \sum_{k=1}^{K} \sum_{j\in N_i} \alpha_{ij}^k W^k \vec{h}'_j), \tag{5}$$

where α_{ij}^k represents the normalized attention coefficient calculated by the kth attention mechanism, and W^k represents the weight matrix in the kth attention mechanism.

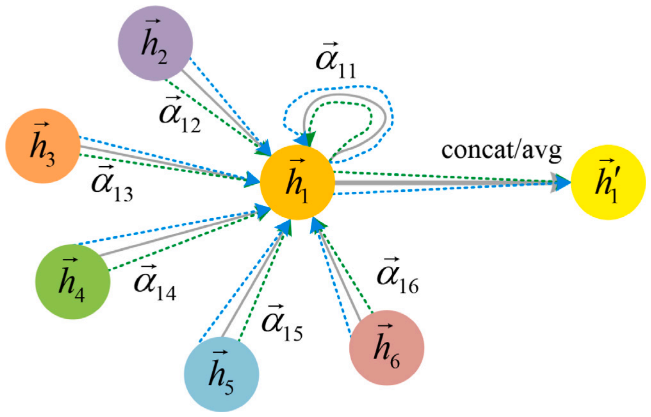


Fig. 4. The aggregation process of the three-headed graph attentional layer.

To more clearly explain the aggregation process of the multi-headed attention layer, the output feature $\vec{h}'_1$ of node 1 is taken as an example to show the aggregation process of the three-headed graph attention layer. As shown in Fig. 4, $\vec{h}_1$ is the feature of the center node. $\vec{h}_2$, $\vec{h}_3$, $\vec{h}_4$, $\vec{h}_5$, and $\vec{h}_6$ are the features of the surrounding adjacent nodes. The impact of the adjacent nodes on the center node is controlled by the weight $\vec{\alpha_{12}}$, $\vec{\alpha_{13}}$, $\vec{\alpha_{14}}$, $\vec{\alpha_{15}}$, and $\vec{\alpha_{16}}$. Three arrows with different colors and styles represent three groups of independent self-attention mechanisms, which can respectively be used to obtain three groups of feature vectors containing the information of surrounding neighbors. The final output feature $\vec{h}'_1$ is obtained through a concatenated operation or an averaging strategy.

2.2.3. GnnAp model

Given that GCN is good at extracting global features of graphs and GAT is good at extracting local features of graphs, the GnnAp model combines graph convolution layer and graph attention layer to ensure the comprehensiveness of the obtained information. As shown in Fig. 5, the graph representation learning module consists of two graph convolution layers and a graph attention layer to learn specific subgraphs of the drug–disease association pairs. In the module, the relationship between layers is progressive, that is, the output of graph convolution layer 1 is the input of graph attention layer 1, and the output of graph attention layer 1 is the input of graph convolution layer 2. The final output of the graph representation learning module is concatenated by the feature vectors obtained at each layer.

From the construction process of subgraphs, it can be seen that the number of each node contained in all subgraphs varies. If feature extraction is performed on each subgraph, we will obtain multiple sets of features for a given node, where the number of sets is equal to the number of the node contained in all subgraphs. The role of the global average pooling layer is to integrate these multiple sets of features, and ultimately reduce the dimensionality of all node features to the same dimension. The final feature representation of the drug–disease association pair obtained after global average pooling layer is expressed as Eq. (6).

$$H' = \text{average\_pool}(H^{(1)} \| H^{(2)} \| H^{(3)}), \tag{6}$$

where

$$\begin{cases} H^{(1)} = \text{GCN}_1(A, H) = \text{ReLU}(\tilde{D}^{-\frac{1}{2}}(A+I)\tilde{D}^{-\frac{1}{2}} H W_0) \\ H^{(2)} = \text{GAT}_1(A, H^{(1)}) = \text{ReLU}(\frac{1}{8}\sum_{k=1}^{8}\sum_{j\in N_i}\alpha_{ij}^{k} W_1^{k} H_j^{1}) \\ H^{(3)} = \text{GCN}_2(A, H^{(2)}) = \text{ReLU}(\tilde{D}^{-\frac{1}{2}}(A+I)\tilde{D}^{-\frac{1}{2}} H^{(2)} W_2) \end{cases},$$

A and H are the adjacency matrix and the feature matrix of the specific subgraph of the drug–disease pair, respectively. H' is the final feature representation of the drug–disease pair, and then input into the prediction module composed of two fully connected layers and a ReLU function to obtain the prediction results of drug–disease association pairs.

3. Results and discussions

3.1. Experimental setup

The proposed model is implemented on NVIDIA GTX 1080Ti. In terms of software settings, Ubuntu=18.04, Cuda=10.2, Python=3.7, pyTorch=1.13. For specific model parameter settings, please refer to Section 3.4.

3.2. Benchmark datasets

The databases involved in this paper are the PREDICT database [21] and the CDataset database [39]. The PREDICT database, compiled by Gottlieb et al. is a common baseline database containing 1933 known drug–disease associations with high confidence. It covers 593 marketed drugs in DrugBank23 [40] and 313 diseases in the OMIM database [41]. The CDataset database, another public benchmark database compiled by Luo et al. contains 2532 known drug–disease associations, covering 663 nationally approved drugs and 409 diseases. To construct a balanced dataset, the known drug–disease associations in the PREDICT database and the CDataset database are taken as positive samples of their corresponding datasets, and the randomly selected unknown associations are taken as negative samples of the corresponding datasets. Finally, the PREDICT dataset containing 1933 positive and 1933 negative samples and the CDataset containing 2532 positive and 2532 negative samples are formed. Given that all drug–disease associations in the PREDICT dataset have been validated in clinical trials, the PREDICT dataset is used as the baseline dataset to evaluate the model's comprehensive prediction performance. The benchmark datasets adopted in this study are available in Table S1.

3.3. Performance evaluation

In this paper, the 5-fold cross-validation (5-fold CV) [42] is adopted to evaluate the performance of drug–disease association predictors. During the process of the 5-fold CV, the traning dataset is randomly split into 5 disjoint subsets with roughly equal size. Each subset is in turn taken as a test set, and the remaining subsets are combined to train the predictor. The final result of the 5-fold CV is the average of the 5 test results, thereby avoiding deviations in the model prediction results caused by a single partition of the small dataset.

The Accuracy, Sensitivity, Specificity, and F1-score are used to quantitatively measure the prediction performance. They are respectively defined as Eq. (7), Eq. (8), Eq. (9), and Eq. (10).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{7}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \tag{8}$$

$$Specificity = \frac{TN}{TN+FP}, \tag{9}$$

$$\text{F1-score} = \frac{2TP}{2TP+FP+FN}. \tag{10}$$

where TP represents the number of positive samples that are correctly predicted; TN represents the number of negative samples that are correctly predicted; FN represents the number of positive samples that are incorrectly predicted; FP represents the number of negative samples that are incorrectly predicted.

In addition to the above evaluation metrics, the Receiver Operating Characteristic (ROC) curve and the Precision–Recall (PR) curve are adopted to further evaluate model performance. The horizontal axis of ROC curve is the False Positive Rate (FPR), focusing on all negative

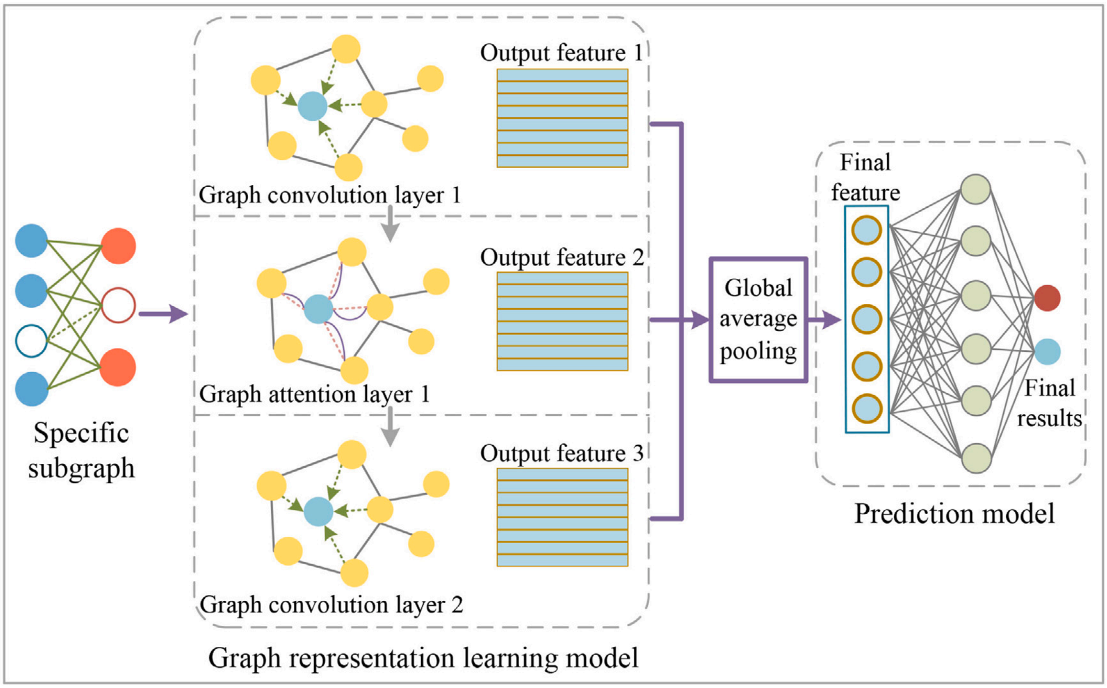


Fig. 5. The architecture of the GnnAp model.

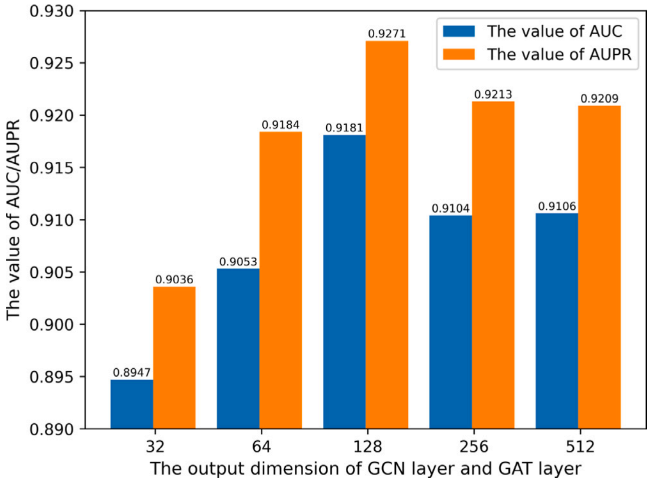


Fig. 6. The AUC/AUPR values against different output dimensions of GCN layer and GAT layer.

samples; The vertical axis is the True Positive Rate (TPR), focusing on all positive samples. FPR and TPR will not change with the change of category distribution, so ROC curves are commonly used to evaluate the comprehensive prediction performance of models. The area under the ROC curve (AUC) can more intuitively reflect the prediction performance of the model. The closer the AUC value is to 1, the better the generalization performance of the model is. The PR curve reveals the relationship between the recall rate (the samples predicted to be positive among the actual positive samples) and the precision rate (the actual positive samples among the samples predicted to be positive). The larger the area under the PR curve (AUPR), the better the prediction performance of the model.

3.4. Selection of hyperparameters

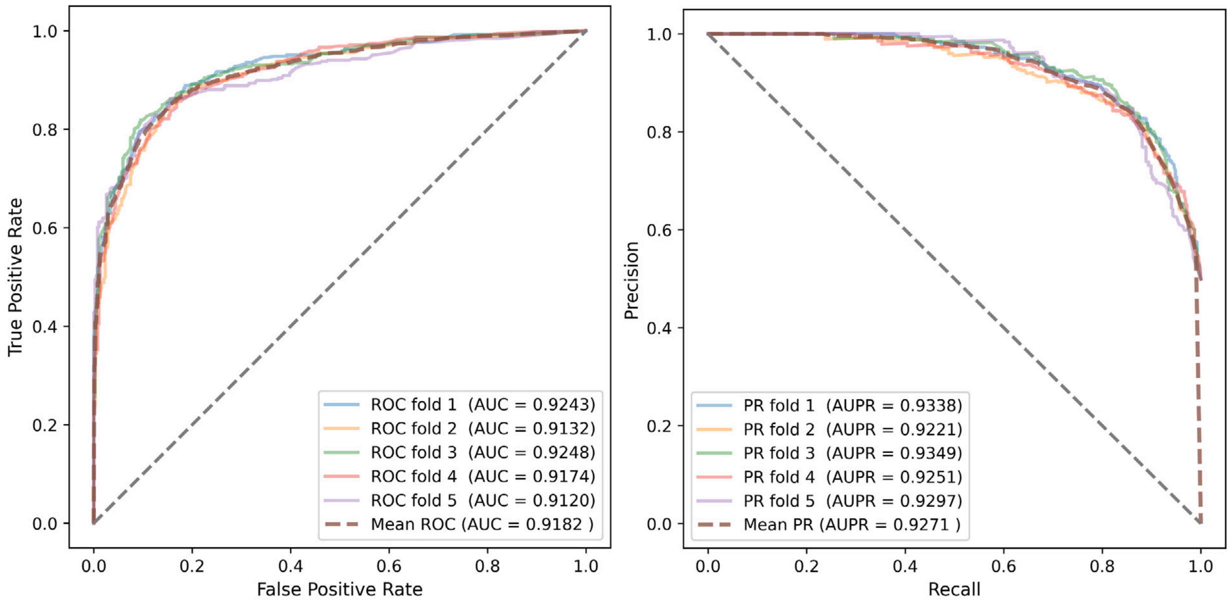
This section introduces the hyperparameters involved in the graph embedding model. The output feature dimensions of graph convolution layer and graph attention layer can affect the prediction performance of GCNGAT, so the prediction performance of the model is evaluated based on different output dimensions. As shown in Fig. 6, the optimal AUC value and AUPR value are obtained when the output dimension is 128. Therefore, the output dimension of the graph convolution layer and the graph attention layer is set to 128. To ensure that the three layer features in the graph representation learning module account for the same proportion, the input and output dimensions of the average pooling layer are set to 384∗128, respectively. The input and output dimensions of the fully connected layer 1 are 128 and 32, and the input and output dimensions of the fully connected layer 2 are 32 and 2, so as to obtain binary results. After multiple comparative experiments, the optimal learning rate of the model is 0.001, and the optimal optimizer is Adam. Since the training loss of the model tends to be stable when the epoch is 30, the epoch is set to 30.

Table 1
The prediction results achieved by GCNGAT on the PREDICT dataset using the 5-fold CV.

Test dataset	Accuracy	Precision	Sensitivity	F1-score
1	0.8516	0.8505	0.8527	0.8516
2	0.8426	0.8591	0.8191	0.8386
3	0.8439	0.8197	0.8811	0.8493
4	0.8422	0.875	0.7979	0.8347
5	0.8512	0.8672	0.829	0.8477
Mean	0.8463	0.8543	0.8360	0.8444

3.5. Prediction performance analysis

Firstly, the prediction performance of GCNGAT is evaluated using the 5-fold CV on the PREDICT dataset. As listed in Table 1, the mean values of accuracy, sensitivity, precision, and F1 score achieved by the 5 test datasets in the 5-fold CV are 0.8463, 0.8543, 0.8360, and 0.8444, respectively, all higher than 83%, indicating that the model has good comprehensive prediction performance. Fig. 7 shows the ROC curve and PR curve of GCNGAT on the PREDICT dataset. The areas under the corresponding curves, i.e. AUC value and AUPR value, reach



(a) (b)

Fig. 7. ROC curve and PR curve achieved by GCNGAT on the PREDICT dataset using the 5-fold CV. (a) ROC curve achieved by GCNGAT on the PREDICT dataset using the 5-fold CV. (b) PR curve achieved by GCNGAT on the PREDICT dataset using the 5-fold CV.

Table 2
The prediction results achieved by GCNGAT on the CDataset dataset using the 5-fold CV.

Test dataset	Accuracy	Precision	Sensitivity	F1-score
1	0.8660	0.8763	0.8521	0.8640
2	0.8749	0.906	0.8363	0.8697
3	0.8796	0.865	0.8992	0.8818
4	0.8815	0.8799	0.8834	0.8797
5	0.8539	0.8455	0.8656	0.8555
Mean	0.8712	0.8745	0.8673	0.8701

Table 3
Prediction performance of GCNGAT compared with existing methods on the PREDICT dataset.

Predictor	Accuracy	Precision	Sensitivity	F1-score
HNet-DNN [26]	0.7921	0.7889	0.8066	0.7961
HINGRL [29]	0.8138	0.8668	0.7545	0.8028
PSGCN [30]	0.8132	0.8494	0.7672	0.8036
GCNGAT	0.8463	0.8543	0.8360	0.8444

0.9182 and 0.9271, respectively, further demonstrating the excellent prediction ability of the model.

To demonstrate the stability of the prediction results for drug–disease associations, the prediction performance of GCNGAT is evaluated on another dataset, the CDataset, based on the 5-fold CV. As listed in Table 2, the mean values of accuracy, sensitivity, precision, and F1 score achieved by the 5 test datasets in the 5-fold CV are 0.8712, 0.8745, 0.8673, and 0.8701, respectively, showing good prediction performance. The same conclusion can be deduced from the ROC curve and PR curve obtained by GCNGAT on the CDataset as shown in Fig. 8, as well as the corresponding AUC value (0.9417) and AUPR value (0.9488).

Comparing the prediction results on the PREDICT dataset with those on the CDataset dataset, it can be found that GCNGAT exhibits better prediction performance on the CDataset datasets containing more data, that is, GCNGAT achieves better prediction results with more abundant data.

3.6. Effectiveness analysis of subgraph extraction

To verify the effectiveness of extracting the corresponding subgraphs of drug–disease association pairs, a comparative experiment with subgraph extraction as the variable is constructed. Specifically, a model called GCNGat-all that has not undergone subgraph extraction but is identical to GCNGAT in other parts, is constructed. The performance of GCNGAT-ALL is compared with that of GCNGAT on the PREDICT dataset. As shown in Fig. 9, all evaluation metrics of the GCNGAT-All model are lower than those of GCNGAT, especially the F1-score of GCNGAT is 9.96% higher than that of GCNGAT-All, indicating that the robustness and prediction ability of the GCNGAT model is much higher than that of GCNGAT-All.

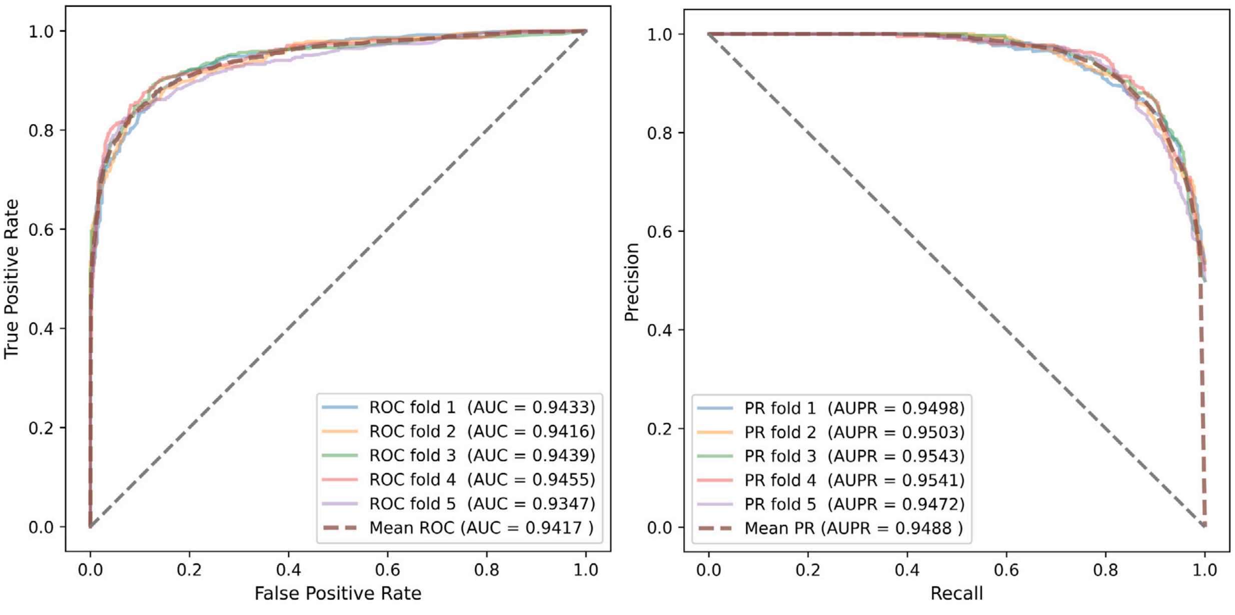
3.7. Effectiveness analysis of graph representation learning module

In order to demonstrate the good performance of the graph representation learning module proposed in this paper due to the combination of the graph convolution layer and the graph attention layer, it is compared with the model containing only the graph convolution layer (3-GCN) and the model containing only the graph attention layer (3-GAT) on the same dataset, respectively. The graph representation learning module in the 3-GCN model consists of three graph convolution layers, while the graph representation learning module in the 3-GAT model consists of three graph attention layers. To ensure the fairness, except for changing the graph convolution layer and the graph attention layer, the remaining parts, including dimensions, are consistent.

As shown in Fig. 10, except for Sensitivity, the evaluation metrics of 3-GCN and 3-GAT models are not significantly different. All evaluation metrics of the slightly better performing 3-GCN model are lower than those of the GCNGAT model, indicating that combining the graph convolution layer with the graph attention layer can extract complementary information to improve the prediction ability of the model.

3.8. Comparison with existing methods

Many prediction models have already existed in the field of drug-disease association prediction. To further evaluate the drug–disease



(a)

(b)

Fig. 8. ROC curve and PR curve achieved by GCNGAT on the CDataset dataset using the 5-fold CV. (a) ROC curve achieved by GCNGAT on the CDataset dataset using the 5-fold CV. (b) PR curve achieved by GCNGAT on the CDataset dataset using the 5-fold CV.

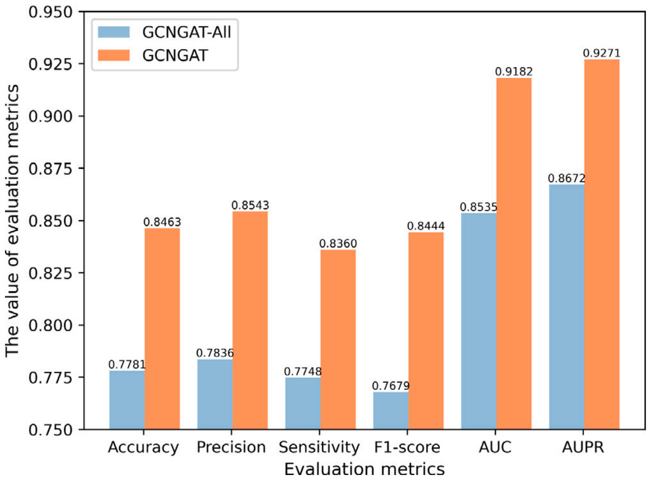


Fig. 9. Evaluation metrics of GCNGAT and GCNGAT-All on the PREDICT dataset.

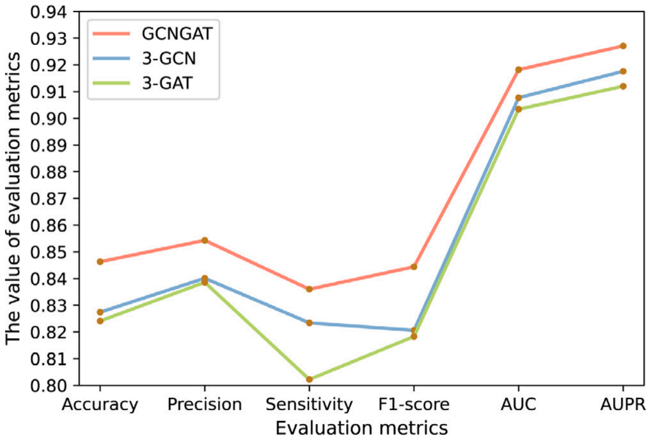


Fig. 10. Evaluation metrics of GCNGAT, 3-GCN and 3-GAT models on the PREDICT dataset.

association prediction performance of GCNGAT, GCNGAT is compared with novel deep learning models based on drug–disease heterogeneous information, including HNet-DNN [26], HINGRL [29], and PSGCN [30]. In order to ensure the fairness and objectivity, the above models are all evaluated on the PREDICT dataset. As shown in Table 3, although HNet-DNN achieves better Sensitivity than HINGRL and PSGCN, the remaining evaluation metrics are around 79%, ranking the lowest in overall performance. With the exception of Precision, the overall performance of HINGRL and PSGCN is not significantly different. Although the Precision of HINGRL is 0.0125 higher than that of GCNGAT, the Accuracy, Sensitivity and F1-score of GCNGAT are higher than those of HINGRL. In addition, as shown in Fig. 11, based on the same dataset, GCNGAT achieves the highest AUC value. Based on the criterion that the AUC value is closer to 1, the overall prediction performance of the model is better, it can be concluded that GCNGAT is an excellent predictor in the field of drug–disease association prediction.

3.9. Limitations

Compared with existing methods, GCNGAT achieves excellent performance. However, the model still has some limitations. Firstly, GCNGAT only performs prediction tasks based on the association between drugs and diseases, without considering the clinical phenotype of diseases and the positive and negative effects of drugs. The prior knowledge used in this study is not comprehensive enough. Inspired by the Ref. [43], more prior knowledge about the association between diseases and drugs should be extracted from public databases, thereby improving the quality of graph construction. Secondly, as a black box model, GCNGAT cannot provide enough interpretability for mechanistic insights on why a drug–disease pair associates or not. Finally, we have not conducted in-depth research on data enhancement, feature selection, and network architecture design.

4. Conclusions

In view of the different action mechanisms of the same drug on different diseases, a drug–disease association prediction model based on graph convolution neural network and graph attention network

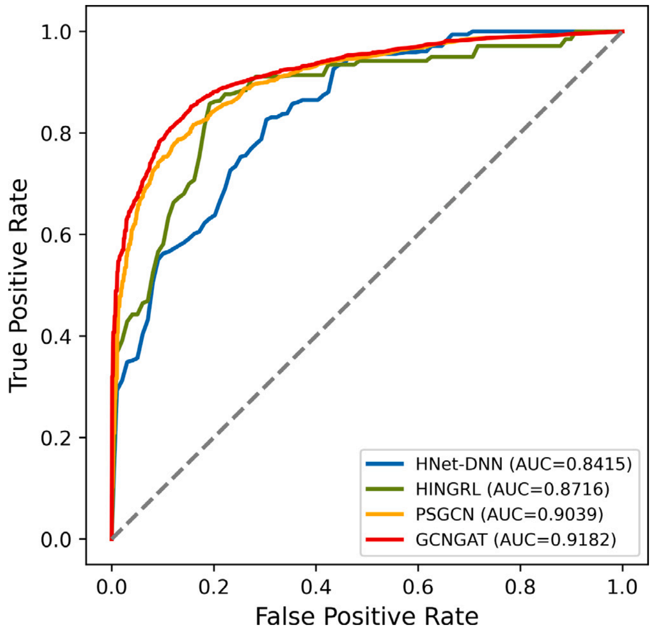


Fig. 11. ROC curves of GCNGAT and existing methods on the PREDICT dataset.

(GCNGAT) is proposed. Firstly, a drug–disease heterogeneous graph structure is constructed based on known drug–disease association pairs to obtain initial feature information of drugs and diseases. Secondly, the corresponding subgraphs of each drug–disease association pair are extracted from the heterogeneous graph structure to obtain different background information of each association pair. Finally, a GnnAp model is constructed by combining the graph convolution layer with the graph attention layer to learn the drug–disease interaction feature representations and predict the drug–disease associations. Based on the PREDICT dataset and the CDataset dataset, the AUC values achieved by GCNGAT are 0.9182 and 0.9417, respectively, indicating that the model has good prediction performance for drug–disease associations. The effectiveness analysis of subgraph extraction fully proves the rationality of extracting subgraphs based on drug–disease association pairs. The comparison experiments based on the graph representation learning module verify the effectiveness of the combination of graph convolution layer and graph attention layer. Compared with other existing drug–disease association prediction algorithms, the excellent performance of GCNGAT is fully demonstrated. In the future work, we will attempt to incorporate the clinical phenotypes of diseases and the positive and negative effects of drugs to construct a directed and multilateral heterogeneous graph structure.

CRediT authorship contribution statement

Runtao Yang: Project administration, Supervision, Writing – original draft, Writing – review & editing. **Yao Fu:** Data curation, Formal analysis, Methodology, Software. **Qian Zhang:** Methodology, Software, Validation, Visualization. **Lina Zhang:** Formal analysis, Investigation, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by the Natural Science Foundation of Shandong Province of China under the grant number ZR2020QF037 and ZR2020MF077.

Appendix A. Supplementary data

Table S1. The benchmark datasets. The PREDICT dataset contains 1933 positive and 1933 negative samples and the CDataset dataset contains 2532 positive and 2532 negative samples.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2024.102805.

References

[1] Mak KK, Pichika MR. Artificial intelligence in drug development: present status and future prospects. Drug Discov Today 2019;24(3):773–80. http://dx.doi.org/10.1016/j.drudis.2018.11.014.

[2] Réda C, Kaufmann E, Delahaye-Duriez A. Machine learning applications in drug development. Comput Struct Biotechnol J 2020;18:241–52. http://dx.doi.org/10.1016/j.csbj.2019.12.006.

[3] Jarada TN, Rokne JG, Alhajj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. J Cheminform 2020;12(1):1–23. http://dx.doi.org/10.1186/s13321-020-00450-7.

[4] Altay O, Mohammadi E, Lam S, et al. Current status of COVID-19 therapies and drug repositioning applications. iScience 2020;23(7):101303. http://dx.doi.org/10.1016/j.isci.2020.101303.

[5] Won JH, Lee H. The current status of drug repositioning and vaccine developments for the COVID-19 pandemic. Int J Mol Sci 2020;21(24):9775. http://dx.doi.org/10.3390/ijms21249775.

[6] Ge Y, Tian T, Huang S, et al. An integrative drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. Signal Transduct Target Therapy 2021;6(1):165. http://dx.doi.org/10.1038/s41392-021-00568-6.

[7] Wu C, Zheng M, Yang Y, et al. Furin: a potential therapeutic target for COVID-19. iScience 2020;23(10):101642. http://dx.doi.org/10.1016/j.isci.2020.101642.

[8] Serafin MB, Bottega A, Foletto VS, et al. Drug repositioning is an alternative for the treatment of coronavirus COVID-19. Int J Antimicrob Ag 2020;55(6):105969. http://dx.doi.org/10.1016/j.ijantimicag.2020.105969.

[9] Hua Y, Dai X, Xu Y, et al. Drug repositioning: Progress and challenges in drug discovery for various diseases. Eur J Med Chem 2022;234:114239. http://dx.doi.org/10.1016/j.ejmech.2022.114239.

[10] Luo H, Li M, Yang M, et al. Biomedical data and computational models for drug repositioning: a comprehensive review. Brief Bioinform 2021;22(2):1604–19. http://dx.doi.org/10.1093/bib/bbz176.

[11] Chen Z, Liu X, Hogan W, et al. Applications of artificial intelligence in drug development using real-world data. Drug Discov Today 2021;26(5):1256–64. http://dx.doi.org/10.1016/j.drudis.2020.12.013.

[12] Yella JK, Yaddanapudi S, Wang Y, et al. Changing trends in computational drug repositioning. Pharmaceuticals 2018;11(2):57. http://dx.doi.org/10.3390/ph11020057.

[13] Armando RG, Mengual Gómez DL, Gomez DE. New drugs are not enough-drug repositioning in oncology: An update. Int J Oncol 2020;56(3):651–84. http://dx.doi.org/10.3892/ijo.2020.4966.

[14] Badkas A, De Landtsheer S, Sauter T. Topological network measures for drug repositioning. Brief Bioinform 2021;22(4):bbaa357. http://dx.doi.org/10.1093/bib/bbaa357.

[15] Meshkani SE, Mahdian D, Abbaszadeh-Goudarzi K, et al. Metformin as a protective agent against natural or chemical toxicities: A comprehensive review on drug repositioning. J Endocrinol Investig 2020;43:1–19. http://dx.doi.org/10.1007/s40618-019-01060-3.

[16] Sakate R, Kimura T. Drug repositioning trends in rare and intractable diseases. Drug Discov Today 2022;22(7):1789–95. http://dx.doi.org/10.1016/j.drudis.2022.01.013.

[17] Datti A. Academic drug discovery in an age of research abundance, and the curious case of chemical screens toward drug repositioning. Drug Discov Today 2023;28(5):103522. http://dx.doi.org/10.1016/j.drudis.2023.103522.

[18] Luo H, Li M, Wang S, et al. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. Bioinformatics 2018;34(11):1904–12. http://dx.doi.org/10.1093/bioinformatics/bty013.

[19] Yang M, Wu G, Zhao Q, et al. Computational drug repositioning based on multi-similarities bilinear matrix factorization. Brief Bioinform 2020;22(4):bbaa267. http://dx.doi.org/10.1093/bib/bbaa267.

[20] Sadeghi S, Lu J, Ngom A. A network-based drug repurposing method via non-negative matrix factorization. Bioinformatics 2021;38(5):1369–77. http://dx.doi.org/10.1093/bioinformatics/btab826.

[21] Gottlieb A, Stein G, Ruppin E, et al. PREDICT: A method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 2011;7(1):496. http://dx.doi.org/10.1038/msb.2011.26.

[22] Yang K, Zhao X, Waxman D, et al. Predicting drug-disease associations with heterogeneous network embedding. Chaos 2019;29(12):123109. http://dx.doi.org/10.1063/1.5121900.

[23] Kitsiranuwat S, Suratanee A, Plaimas K. Integration of various protein similarities using random forest technique to infer augmented drug-protein matrix for enhancing drug-disease association prediction. Sci Progress 2022;105(3). http://dx.doi.org/10.1177/00368504221109215, 00368504221109215.

[24] Xuan P, Ye Y, Zhang T, et al. Convolutional neural network and bidirectional long short-term memory-based method for predicting drug-disease associations. Cells 2019;8(7):705. http://dx.doi.org/10.3390/cells8070705.

[25] Zeng X, Zhu S, Liu X, et al. deepDR: A network-based deep learning approach to in silico drug repositioning. Bioinformatics 2019;35(24):5191–8. http://dx.doi.org/10.1093/bioinformatics/btz418.

[26] Liu H, Zhang W, Song Y, et al. HNet-DNN: Inferring new drug-disease associations with deep neural network based on heterogeneous network features. J Chem Inf Model 2020;60(4):2367–76. http://dx.doi.org/10.1021/acs.jcim.9b01008.

[27] Yu Z, Huang F, Zhao X, et al. Predicting drug-disease associations through layer attention graph convolutional network. Brief Bioinform 2021;22(4):bbaa243. http://dx.doi.org/10.1093/bib/bbaa243.

[28] Gao L, Cui H, Zhang T, et al. Prediction of drug-disease associations by integrating common topologies of heterogeneous networks and specific topologies of subnets. Brief Bioinform 2021;23(1):bbab467. http://dx.doi.org/10.1093/bib/bbab467.

[29] Zhao BW, Hu L, You ZH, et al. HINGRL: Predicting drug-disease associations with graph representation learning on heterogeneous information networks. Brief Bioinform 2022;23(1):bbab515. http://dx.doi.org/10.1093/bib/bbab515.

[30] Sun X, Wang B, Zhang J, et al. Partner-specific drug repositioning approach based on graph convolutional network. IEEE J Biomed Health Inf 2022;26(11):5757–65. http://dx.doi.org/10.1109/JBHI.2022.3194891.

[31] Sun M, Zhao S, Gilvary C, et al. Graph convolutional networks for computational drug development and discovery. Brief Bioinform 2020;21(3):919–35. http://dx.doi.org/10.1093/bib/bbz042.

[32] Li H, Li Z, Du K, et al. A semi-supervised graph convolutional network for early prediction of motor abnormalities in very preterm infants. Diagnostics 2023;13(8):1508. http://dx.doi.org/10.3390/diagnostics13081508.

[33] Zhang S, Tong H, Xu J, et al. Graph convolutional networks: A comprehensive review. Comput Soc Netw 2019;6(1):1–23. http://dx.doi.org/10.1186/s40649-019-0069-y.

[34] Zheng X, Du H, Luo X, et al. BioByGANS: Biomedical named entity recognition by fusing contextual and syntactic features through graph attention network in node classification framework. BMC Bioinformatics 2022;23(1):501. http://dx.doi.org/10.1186/s12859-022-05051-9.

[35] Chen G, Liu ZP. Graph attention network for link prediction of gene regulations from single-cell RNA-sequencing data. Bioinformatics 2022;38(19):4522–9. http://dx.doi.org/10.1093/bioinformatics/btac559.

[36] Lv Q, Chen G, Yang Z, et al. Meta learning with graph attention networks for low-data drug discovery. IEEE Trans Neural Netw Learn Syst 2023. http://dx.doi.org/10.1109/TNNLS.2023.3250324.

[37] Guo MH, Xu TX, Liu JJ, et al. Attention mechanisms in computer vision: A survey. Comput Vis Media 2022;8(3):331–68. http://dx.doi.org/10.1007/s41095-022-0271-y.

[38] Ding X, Nie W, Liu X, et al. Compact convolutional neural network with multi-headed attention mechanism for seizure prediction. Int J Neural Syst 2023;33(3):2350014. http://dx.doi.org/10.1142/S0129065723500144.

[39] Luo H, Wang J, Li M, et al. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics 2016;32(17):2664–71. http://dx.doi.org/10.1093/bioinformatics/btw228.

[40] Zong N, Wen A, Moon S, et al. Computational drug repurposing based on electronic health records: a scoping review. npj Digit Med 2022;5(1):77. http://dx.doi.org/10.1038/s41746-022-00617-6.

[41] Wang ZY, Zhang HY. Rational drug repositioning by medical genetics. Nature Biotechnol 2013;31(12):1080–2. http://dx.doi.org/10.1038/nbt.2769.

[42] Lin Z, Lai J, Chen X, et al. Curriculum reinforcement learning based on K-fold cross validation. Entropy 2022;24(12):1787. http://dx.doi.org/10.3390/e24121787.

[43] Li Z, Li H, Braimah A, et al. A novel ontology-guided attribute partitioning ensemble learning model for early prediction of cognitive deficits using quantitative structural MRI in very preterm infants. Neuroimage 2022;260:119484. http://dx.doi.org/10.1016/j.neuroimage.2022.119484.