BMC Medical Informatics and Decision Making

RESEARCH

Open Access

DAPNet: multi-view graph contrastive network incorporating disease clinical and molecular associations for disease progression prediction

Haoyu Tian[1†], Xiong He[1†], Kuo Yang[1†], Xinyu Dai[1], Yiming Liu[1], Fengjin Zhang[2], Zixin Shu[1], Qiguang Zheng[1], Shihua Wang[3], Jianan Xia[1], Tiancai Wen[3], Baoyan Liu[4], Jian Yu[1] and Xuezhong Zhou[1*]

Abstract

Background Timely and accurate prediction of disease progress is crucial for facilitating early intervention and treatment for various chronic diseases. However, due to the complicated and longitudinal nature of disease progression, the capacity and completeness of clinical data required for training deep learning models remains a significant challenge. This study aims to explore a new method that reduces data dependency and achieves predictive performance comparable to existing research.

Methods This study proposed DAPNet, a deep learning-based disease progression prediction model that solely utilizes the comorbidity duration (without relying on multi-modal data or comprehensive medical records) and disease associations from biomedical knowledge graphs to deliver high-performance prediction. DAPNet is the first to apply multi-view graph contrastive learning to disease progression prediction tasks. Compared with other studies on comorbidities, DAPNet innovatively integrates molecular-level disease association information, combines disease co-occurrence and ICD10, and fully explores the associations between diseases;

Results This study validated DAPNet using a de-identified clinical dataset derived from medical claims, which includes 2,714 patients and 10,856 visits. Meanwhile, a kidney dataset (606 patients) based on MIMIC-IV has also been constructed to fully validate its performance. The results showed that DAPNet achieved state-of-the-art performance on the severe pneumonia dataset (F1=0.84, with an improvement of 8.7%), and outperformed the six baseline models on the kidney disease dataset (F1=0.80, with an improvement of 21.3%). Through case analysis, we elucidated the clinical and molecular associations identified by the DAPNet model, which facilitated a better understanding and explanation of potential disease association, thereby providing interpretability for the model.

Conclusions The proposed DAPNet, for the first time, utilizes comorbidity duration and disease associations network, enabling more accurate disease progression prediction based on a multi-view graph contrastive learning, which provides valuable insights for early diagnosis and treatment of patients. Based on disease association networks, our research has enhanced the interpretability of disease progression predictions.

†Haoyu Tian, Xiong He and Kuo Yang contributed equally to this work.

*Correspondence:
Xuezhong Zhou
xzzhou@bjtu.edu.cn
Full list of author information is available at the end of the article

Keywords Disease progression prediction, Disease association networks, Graph contrastive learning

Introduction

In the era of digital medicine, massive collection of clinical data provides a valuable opportunity for medical research and clinical decisions with the support of artificial intelligence (AI) approaches [1, 2]. Disease progression prediction, which usually predicts the progression of disease in the future by considering the patient's condition at multiple time points, such as using historical disease records to predict the progression of heart failure in the next six months, is a significant AI task for early disease intervention and outcome improvement [3, 4]. Modern medicine has shown that the occurrence and development of diseases are regular and phased, such as using oxygenation (PaO2) and elimination of carbon dioxide (PaCO2) as important indicators for dividing the progression of acute respiratory failure [5], which provides a theoretical basis for disease prediction. Early identification and more accurate prediction of these patients using better disease progression models may improve outcomes by facilitating early intervention of appropriate therapies, monitoring, and specialty referral [6]. In particular, it is important to predict and intervene in advance for patients who may be seriously ill in the future to effectively prevent the deterioration of their disease. Besides, current disease progression prediction research often relies heavily on phenotype data from patients, such as laboratory test results and imaging data [6–8]. However, this high dependency can lead to difficulties in data acquisition and increased costs. Therefore, this study proposed a novel predictive model to reduce dependence on phenotype data while still achieving predictive performance comparable to existing studies.

Early disease progression prediction methods mainly focused on current information, such as DeepPatient [7], HRFLM [9], and MedText [10], but these methods consider less medical information and rarely utilize past patient information. Subsequent models for disease progression prediction, on the other hand, typically analyze sequential historical diagnostic information of patients, such as those for chronic kidney disease progression prediction [11, 12] and COVID-19 progression prediction [6]. There are also some methods that combine deep learning algorithms, such as the RNN-based Doctor AI [8], the LSTM-based Hitanet [13], the attention-based heart failure prediction model GRAM [14] and tBNA-PR [15], and the common attention memory network CAMP [16]. Disease progression prediction models based on the knowledge graph also achieved good performance, such as GNDP [17] and Sherbet [18]. However, existing research often relies heavily on multi-modal patient data or comprehensive medical records, which implies a high demand for data richness in models. Due to the higher privacy concerns associated with medical data compared to other fields, obtaining complete medical records is challenging. Moreover, most methods fail to adequately consider potential relationships between diseases and lack sufficient utilization of domain knowledge, particularly at the molecular level.

It is worth noting that the emergence and development of diseases is a very complex physiological and pathological process [19–21]. Hidalgo et al. [22] provide suggestive evidence that patients develop diseases close in the phenotypic disease network to those already affecting them. Additionally, Lee et al. [23] discovered that the more connected a disease is to other diseases, the higher its prevalence and associated mortality rate. Meanwhile, molecular-level disease-gene relationships may also affect other cellular functions, leading to potential comorbidity effects [24]. Furthermore, these studies indicate the correlation between the occurrence of diseases and abnormal expression of certain genes, and we can use disease association networks to analyze their impact on other diseases. In recent years, the research on networks or graphs has been making excellent progress [25–28]. Inspired by the learning of network representation, these methods focus on the analysis of the potential relationship between diseases (such as disease co-occurrence relationships, shared genes, etc.), providing a new perspective for understanding the mechanisms of disease [29–31].

In this study, we proposed a novel disease progression prediction approach that leverages multi-source graph neural network fusion (Fig. 1). This study aims to explore the use of only patient diagnostic information for disease progression prediction, aiming to reduce data dependence and achieve predictive performance comparable to existing research. By combining the current popular Graph Contrastive Learning (GCL) [32–35] and the Convolutional Neural Network (CNN) [36–38], we designed a framework that integrates network embedding and network propagation methods, termed DAPNet. Around the DAPNet framework, we developed a novel method of patient representation calculation for the duration of the patient's illness and constructed a prediction dataset. At the same time, three disease association networks were constructed by integrating multi-source disease association data. This study offered several significant

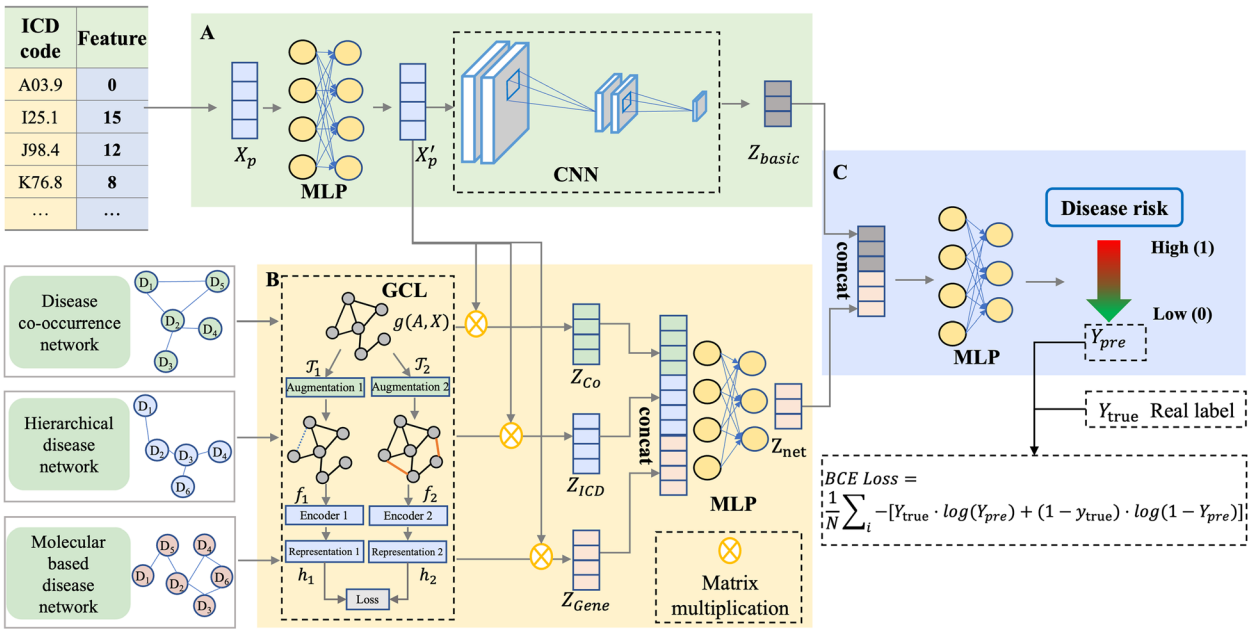


Fig. 1 Overview of disease progression prediction framework. **A** Clinical representation module based on clinical features. **B** Network representation module based on multi-source disease association networks. **C** Feature fusion and disease progression prediction module

contributions to the field of disease progression prediction, including the following aspects:

- DAPNet is the first to apply multi-view graph contrastive learning to disease progression prediction tasks. Compared with other studies, DAPNet integrates the molecular-level disease association network, the combines disease co-occurrence network and the ICD-10 network, and fully explores the associations among diseases.
- DAPNet reduces the dependence on clinical data, and can still achieve excellent performance only by relying on sequence diagnostic data (without relying on multi-modal data or comprehensive medical records). Even in this challenging task, DAPNet performs well, outperforming the baseline model on the severe pneumonia dataset (F1=0.84, with an improvement of 8.7%) and the kidney disease dataset (F1=0.80, with an improvement of 21.3%).
- This study constructs three disease association networks by integrating multi-source disease association data, with a total of 12,536 nodes and 1,763,872 edges, forming the foundation for the proposed method. These disease networks contribute to improving interpretability, which can better depict patient portraits to achieve more accurate prediction and provide valuable insights for early diagnosis and treatment of patients.

Materials and methods

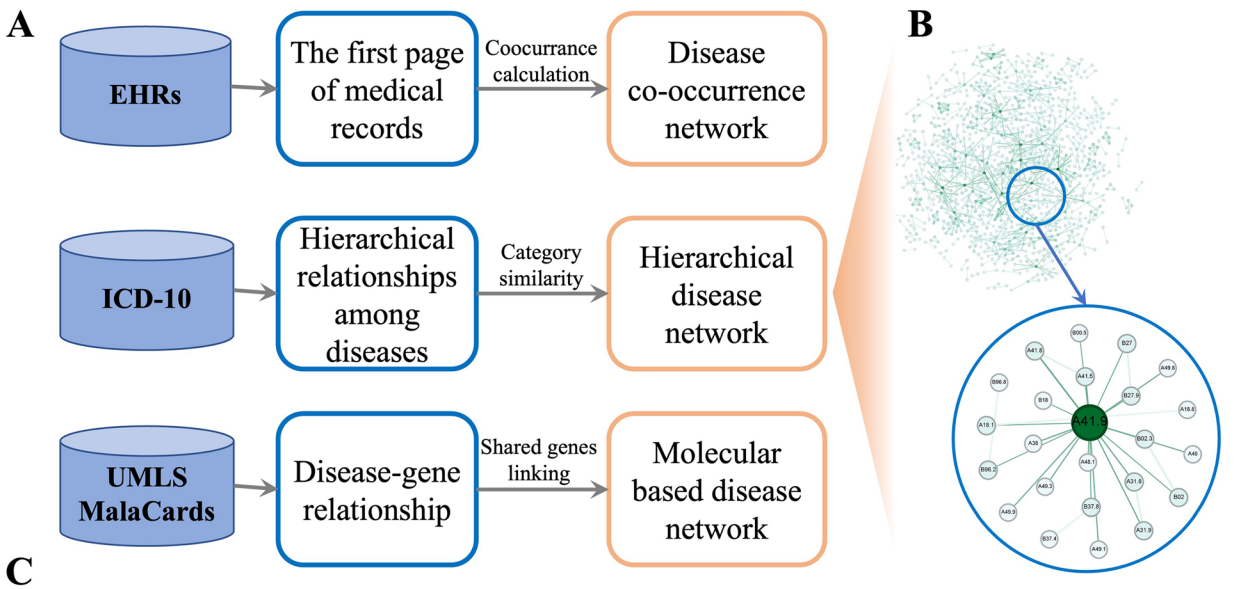
In this section, we first present the construction and scale of multi-source disease relationship networks. Subsequently, we introduce the primary process involved in constructing the dataset. Finally, we outline the framework of DAPNet for disease progression prediction.

Construction of multi-source disease association networks

As mentioned above, the existing disease prediction research lacks the consideration of the disease-gene relationship, but the association information at the molecular level is highly valuable [26, 28]. In order to build a more comprehensive disease association network, data collection encompassed not only disease data from electronic medical records (EMRs) but also disease association data at the molecular level sourced from MalaCards [39] and UMLS [40]. Based on the relationship data of diseases, this study sorted and constructed three disease networks from different perspectives (Fig. 2A and B), which were: disease co-occurrence network, hierarchical disease network, and molecular-based disease network.

Disease co-occurrence network

The disease co-occurrence network referred to patients suffering from multiple diseases at the same time. The network was constructed based on the co-occurrence of the patients' diseases. Each node in the network represented a disease, and each edge represents the co-occurrence associations between the corresponding diseases [41]. The



C

Visit ID (α)	Visit weight ($2^{\alpha-1}$)	D_1	D_2	D_3	D_4	D_5	D_6*	⋯	D_n
1	1	√							
2	2	√	√						
3	4	√	√	√					
4	8	√		√	√	√			
5	-	√		√		√	√		
Score ($\sum_{\alpha} 2^{\alpha-1}$)		**15**	**6**	**12**	**8**	**8**			

Fig. 2 Construction of disease prediction dataset. **A** Construction of disease association network based on three kinds of disease association data. **B** Partial node display of heterogeneous network. **C** Calculation method of the illness duration scores

weight was the number of patients with two diseases at the same time. Through pre-processing, the co-occurrence of diseases on the first page of the medical record data was counted, and the co-occurrence association data between diseases was obtained. To improve the quality of the network, the edges with weights greater than 5 were removed to preserve high-frequency edges.

Hierarchical disease network

The tree structure of the ICD-10 code reflected the medical knowledge contained in the ICD-10 code at the beginning of its design. The hierarchical disease network was constructed based on the medical knowledge behind the ICD-10 code. The nodes in the network represented a disease, and the edges were represented by the similarity between the diseases corresponding to the ICD-10 codes. Based on the tree structure characteristics of the ICD-10 code, the similarity between disease codes was calculated. This similarity could be calculated by the ontology similarity formula [42]. The similarity score was used as the weight of the edge in the network. The similarity formulas were as follows:

$$sim(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{1}$$

$$IC(c) = -\log\left(\frac{freq(c)}{freq(root)}\right) \tag{2}$$

where $lcs(c_1, c_2)$represented the nearest common root node found according to the ICD tree structure. The $freq(c)$ indicated the number of child nodes of node c. If c was a leaf node, it corresponded to 1.

Molecular based disease network

By collecting the "disease-gene" association data from MalaCards and UMLS, the genes in these "disease-gene" associations were used as bridges to establish links, thereby building associations between diseases. The purpose of selecting "shared genes linking" was to retain as much information as possible about the "disease-gene" associations at the molecular level.

Construction of benchmark dataset using medical claims data

As mentioned above, the data usually used by the existing disease progression prediction models encompasses modal data such as text, image, and physical and chemical indicators [28, 43]. However, due to the privacy of medical data, the more data formats required by the disease prediction model, the more difficult it is to obtain data [4]. To reduce the disease progression prediction model's dependence on data, this study utilizes past disease diagnosis information to predict the patient's disease course based on disease association knowledge. Meanwhile, most existing open-source datasets cannot meet our expectations (multiple visits), so this study collects real-world clinical cases to construct disease prediction datasets and focuses on considering the impact of disease duration, which other prediction models lack consideration [4, 17, 18, 28, 43]. Therefore, this section provides an overview of a new method for constructing datasets based on disease duration, using the prediction of comorbidities of severe lung diseases as an example.

Clinical data normalization

To ensure the authenticity of the data, we collected part of EMRs from a tertiary A-level hospital, paying particular attention to records of multiple visits, such as visit ID, patient ID, and disease diagnosis results, which were typically recorded using ICD-10 codes. The ICD-10 classification system, as mentioned, divides diseases into multiple levels based on their characteristics, forming a tree structure in which similar diseases share a common parent node [44]. Initially, we used the first page of the medical records as the core data and filtered the disease codes within it [45]. Given the privacy concerns surrounding medical data, it is often difficult to obtain medical data, so it's necessary to rely solely on the patient's diagnostic results for prediction. In this scenario, for each patient, only the diagnostic information within each visit is retained and sorted by visit time. Subsequently, we removed patient records containing aberrant codes (no ICD-10 codes). Furthermore, patient records with one diagnosis were excluded from the analysis, because only one record could not obtain the patient's disease progression. Under the guidance of clinical experts, four digits were reserved for each disease code to reduce data redundancy (e.g., from "J96.02" to "J96.0"). As a result of these criteria, the longitudinal diagnosis records of more than 470k records were ultimately retained for further analysis.

The illness duration scores calculation

To better represent the phenotypes of patients, this study proposed a new method for calculating illness duration to reflect the impact of disease duration. By comparing the results of two consecutive diagnoses, the difference between the latter diagnosis and the previous diagnosis is obtained, which is the new disease of the patient. Specifically, when the newly identified disease is the disease to be predicted, the focus is on the recent diagnosis records and their corresponding diagnosis times. To better depict the patient's disease progression, this study designed a new patient representation calculation method, namely the illness duration scores.

The scoring formula was designed to satisfy two main criteria:

- The score should reflect the criticality of a disease in relation to its temporal proximity to the target disease, with diseases occurring closer in time to the target disease being assigned higher scores.
- The cumulative impact of historical diagnosis results on the score should be reduced, such that recent diagnostic results have a greater impact than the cumulative impact of earlier diagnostic results.

The reason for choosing the exponential function is that we assume that the influence of the recent diagnosis is greater than the sum of influences of the past multiple diagnoses, and the characteristics of the exponential function align with our hypothesis. The final calculation formula is designed as follows:

$$score = \sum_{\alpha} 2^{\alpha - 1} \tag{3}$$

Figure 2C depicts the visit ID using the symbol α, and each row in the table corresponds to a specific diagnosis for the patient. Assuming that the patient had the disease to be predicted at the fifth diagnosis, the earlier diagnosis should be considered. The illness duration scores for these earlier diseases are determined based on the number of visits α, and are calculated using the specific scoring formula (3). Using the data in Fig. 2C as an example, the diagnosis ID corresponding to all diseases in the first diagnosis was the same ($\alpha = 1$). The score of the disease in this diagnosis can be calculated as 1 by substituting it into the formula. The scores for subsequent diseases are calculated as 2, 4, and 8 for the second, third, and

fourth diagnoses, respectively. These individual disease scores are then summed to generate the final illness duration score, as shown in the last row of the table. Therefore, the construction of the dataset focuses on the duration of the disease.

Construction of positive and negative samples in dataset

Modern medicine has shown that severe pneumonia affects a large portion of the population, and is associated with high-risk clinical conditions and even mortality [46]. Considering the occurrence of diseases in the collected medical records and the severity of COVID-19, this research focuses on predicting the risk of severe lung disease comorbidity, which comprised unspecified sepsis (A41.9), acute respiratory failure (J96.0), and unspecified heart failure (I50.9). To achieve this, the normalization method presented in the previous section is utilized to generate the training dataset, which is based on whether there was severe lung disease. If the new disease is classified as one of the severe lung diseases, it is considered a positive case. MIMIC-IV, a publicly available database, is sourced from the electronic health record of the Beth Israel Deaconess Medical Center [47]. In addition, this study also constructed an additional dataset based on the MIMIC-IV dataset for the acute kidney injury (AKI, ICD-10: N179) disease prediction task.

Disease progression prediction model based on multi-source graph neural network fusion

To better utilize disease association knowledge, this study proposed the disease progression prediction model DAPNet, which relies on patients' disease progression information and disease association networks to predict disease in the future. DAPNet consists of three modules (Fig. 1): Module A (Fig. 1A) aims to aggregate patients' clinical representations using CNN based on clinical features. Module B (Fig. 1B) utilizes GCL to learn the network representation based on multi-source disease association networks [48]. Finally, module C (Fig. 1C) combines the patient embedding learned from modules A and B, and uses a multi-layer perception (MLP) to predict the patient's future disease progression.

Module A (Fig. 1A) is mainly responsible for learning the clinical embedding of patients. This module mainly primarily comprises linear transformation operations and CNN-based embedding extraction. Initially, patient embeddings are constructed based on their disease conditions. To mitigate data sparsity, a large amount of 0 in the matrix is filled with values with the help of a fully connected network, which simultaneously learns the implicit relationships between diseases. Additionally, CNN is used to extract disease embedding, taking advantage of the ICD-10 tree structureal code information. Since input data corresponds to diseases sorted by disease codes, the tree structure of ICD-10 coding results in similar or consecutive disease codes corresponding to leaf nodes under the same broad category [18]. Compared to neural network models like MLP, CNN can effectively convolve patients' disease characteristic data, fuse the disease embeddings with those of similar diseases, and extract high-quality clinical diagnostic embeddings. In the DAPNet model, the patient embedding is first denoted as $X_p^{'}$, which is then passed through the MLP network for activation. Subsequently, the patient embedding is extracted using a two-layer CNN model. The convolutional operation is denoted by *Conv* in formula (5).

$$X_p' = X_p W^{(0)} + b^{(0)} \tag{4}$$

$$Z_{basic} = Conv\Big(Conv(X_p')\Big) \tag{5}$$

Module B (Fig. 1B) was mainly responsible for learning the characteristics of multi-source graph neural networks. It was mainly composed of GCL and MLP. The core of module B ware three disease association networks, whose adjacency matrix dimensions were the same. GCL has been a hot topic in graph representation in recent years [32–35]. In this study, we first introduced GCL into disease network representation learning and used the classical method GCN as a comparison. In each network, GCL was used to learn the link relationship of the disease association network and three node characteristic matrices were obtained. Since our network nodes correspond to diseases, we obtain the word vectors of diseases by weighted summation as the initial features of the disease network nodes. This study used Glove vectors and learned in advance from a large number of TCM books and EMRs in advance [49, 50]. Then it multiplied the activated patient's disease embedding and the three groups of embedding matrices respectively to obtain three groups of patient embedding based on the combined disease association network.

Usually, contrastive learning generates multiple views for each instance through various data extensions, maximizing the agreement of two jointly sampled positive pairs [18]. Given a graph $g(\boldsymbol{A}, \boldsymbol{X})$ (Fig. 1), K different transformations $\mathcal{T}_K$ can be applied to obtain multiple views $\{(\mathbf{A}_k, \mathbf{X}_k)\}_{k=1}^{K}$, defined as

$$\mathbf{A}_k, \mathbf{X}_k = \mathcal{T}_k(\mathbf{A}, \mathbf{X}), k = 1, 2, \ldots, K \tag{6}$$

Different graph encoders f_k can be used to generate different representations $\mathbf{h}_k$, defined as

$$\mathbf{h}_k = f_k(\mathbf{A}_i, \mathbf{X}_i), k = 1, 2, \ldots, K \tag{7}$$

The goal of contrastive learning is to maximize the mutual information between two views from the same instance

$$\max \sum_{\mathrm{i}} \sum_{\mathrm{j} \neq \mathrm{i}} \alpha_{\mathrm{i,j}} \mathcal{MI}(\mathbf{h}_{\mathrm{i}}, \mathbf{h}_{\mathrm{j}}) \tag{8}$$

where $i, j \in \{1, 2, \ldots, K\}$, $\{\mathbf{h}_i\}_{i=1}^{K}$ are representations generated from $g(\boldsymbol{A}, \boldsymbol{X})$, which are taken as positive samples. $\mathcal{MI}(\mathbf{h}_i, \mathbf{h}_j)$ are the mutual information between two representations $\mathbf{h}_i$ and $\mathbf{h}_j$. Besides, $\alpha_{i,j} \in \{0, 1\}$ can be used to determine if it comes from the same instance

To overcome the limitation that the graph enhancement in most GCL methods is not sufficient to filter out noise, MA-GCL proposes a new paradigm called GCL model enhancement, which focuses on perturbing the architecture of GNN encoders rather than graph inputs or model parameters [32]. MA-GCL presents three effective model augmentation tricks for GCL, namely asymmetric, random, and shuffling, which can respectively help alleviate high-frequency noises, enrich training instances, and bring safer augmentations. With the help of MA-GCL, we learned three groups of node embedding on three disease networks, defined as $GCLE_{ICD}$, $GCLE_{Co}$, and $GCLE_{Gene}$.

$$Z_{ICD} = GCLE_{ICD} X'_p \tag{9}$$

$$Z_{\mathbb{C}o} = GCLE_{Go} X'_p \tag{10}$$

$$Z_{Gene} = GCLE_{Gene} X'_p \tag{11}$$

In formula (12), splice the three groups of embedding and use a fully connected network to achieve the output size of module B consistent with that of module A.

$$Z_{\mathrm{net}} = Concat[Z_{ICD}, Z_{Co}, Z_{Gene}] W^{(1)} + b^{(1)} \tag{12}$$

Module C (Fig. 1C) is mainly responsible for disease progression prediction. After obtaining the high-dimensional patient embedding from module A and module B, these embeddings are concatenated together to form a single patient embedding to calculate the risk of patients suffering from severe lung disease in the future. The two groups of embeddings represent different aspects of patient information, and they are combined to form the final patient disease embedding. This step helps combine the patient's clinical characteristics and the characteristics learned from the disease network to obtain a comprehensive patient embedding representation, which is then used for disease progression prediction. The classification model is then used to predict the future disease progression of patients. To optimize the model parameters, we calculated the cross-entropy loss between the predicted and actual labels of the training data through iterative processes until the optimal model parameters are achieved. The formula (13) illustrates how the output results from modules A and B are integrated. The probability is then calculated using an MLP and SoftMax function.

$$Y_{pre} = softmax\Big(Concat[Z_{basic}, Z_{\mathrm{net}}] W^{(2)} + b^{(2)}\Big) \tag{13}$$

Experimental settings

Disease association networks

The sizes of the multi-source disease association networks are shown in Table 1. Due to different sources of data, the sizes and densities of the three disease association networks were inconsistent. Among them, the hierarchical disease network has the largest network size, with 11,249 nodes and 1,048,575 edges. The molecular-based disease network has the smallest size, with 1,048 nodes and 129,362 edges. Figure 2B showed related nodes with adjacency relationships in the ICD network of "A41.9" nodes, with typical cases of specified sepsis (A41.8), mycobacterial infection (A31.9), and bacterial infection (A49.9), which was consistent with the correlation between sepsis syndrome and these disease in clinical research.

Severe pneumonia dataset and kidney dataset

After the above segments, the severe pneumonia dataset contains 2,714 samples, and the kidney dataset contains 606 samples. The proportion of positive and negative samples was 1:1. To validate the DAPNet, we divided the datasets into training and testing sets, maintaining a ratio of 4:1. Leveraging the interpretability of the Random

Table 1 Statistical of disease association network scale

Network	Number of nodes	Number of edges	Average degree
Hierarchical disease network	11,249	1,048,575	186.43
Disease co-occurrence network	7,104	633,153	178.25
Molecular-based disease network	1,048	129,362	246.87

Table 2 Top 10 of comorbidity ranking

Index	ICD code	Feature importance
1	Atherosclerotic heart disease of native coronary artery (I25.1)	0.0341
2	Essential (primary) hypertension (I10)	0.0239
3	Other specified diseases of liver (K76.8)	0.0229
4	Other disorders of lung (J98.4)	0.0168
5	Type 2 diabetes mellitus without complications (E11.9)	0.0168
6	Cerebral infarction, unspecified (I63.9)	0.0155
7	Hyperlipidemia, unspecified (E78.5)	0.0128
8	Sequelae of cerebral infarction (I69.3)	0.0120
9	Anemia, unspecified (D64.9)	0.0111
10	Other forms of angina pectoris (I20.8)	0.0110

Forest model, we evaluated the importance of comorbidity in the datasets. Taking the severe pneumonia dataset as an example, Table 2 below lists the top ten comorbidities associated with severe pneumonia.

The ranking results indicate that the top three diseases most commonly associated with severe pneumonia are arteriosclerotic heart disease (I25.1), idiopathic (primary) hypertension (I10), and other specific liver diseases (K76.8). This suggests that patients with severe pneumonia often have comorbidities such as heart disease and hypertension, and that individuals with underlying pneumonia along with heart disease or hypertension are at an increased risk of developing severe pneumonia in the future. The top ten diseases identified in the ranking primarily focuses on heart disease, diabetes, and lung disease, indicating that these comorbidities were directly or indirectly related to severe lung disease. When patients are diagnosed with pneumonia along with any of these aforementioned diseases, it is recommended that doctors provide individualized treatment at an early stage to prevent the progression of severe lung disease or other life-threatening conditions (Fig. 3).

In addition, we also utilized SHAP to analyze the comorbidities, and the results indicate that the most relevant features associated with severe pneumonia are atherosclerotic heart disease (I25.1), other disorders of lung (J98.4), and chronic kidney disease (N18.9). Kurth et al. [51] have found that atherosclerotic heart disease is one of the typical diseases among lung disease deaths, which indirectly supports our analysis results.

Hyperparameters

Given that DAPNet is a classification model, we evaluated its performance using commonly used metrics, including accuracy, recall, and F1-score. Furthermore, we calculated the area under the receiver operating characteristic curve (AUC) to assist with the evaluation. The datasets were split into training and testing sets at a ratio of 4:1. The learning rate for the model was set to 0.00002, with a maximum of 10,000 iterations. The early stopping mechanism was implemented to save the optimal model.

Baselines

Because our study focuses on disease progression in patients' time series records (only relying on ICD codes), it is difficult to directly compare it with other similar disease prediction models. We designed a comprehensive set of comparative and ablation experiments to make up for this shortcoming. The baseline methods in this study include traditional machine learning methods, such as Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR), and MLP. For deep learning methods, this study compared the Hitanet [13] and the simplified version of tBNA-PR [15]. Since DAPNet only uses the disease information of patients and lacks the comprehensive data required for the full tBNA-PR model, the tBNA-PR model compared in this study has been simplified. Additionally, to validate the efficacy of graph contrastive learning, we conducted comparative experiments between the GCN and graph contrastive learning algorithms, including GRACE [35], NCLA [33], and MA-GCL [32]. At the same time, we designed a wealth of comparative experiments and ablation studies to verify the sensitivity and interpretability of the model.

Results

To evaluate the contribution of each module, in this section, we first assess the construction of multi-source disease networks and the distribution of disease datasets. At the same time, to better evaluate the performance of DAPNet, we designed multiple comparative experiments, including disease-associated network comparisons, overall comparison, module ablation experiments,

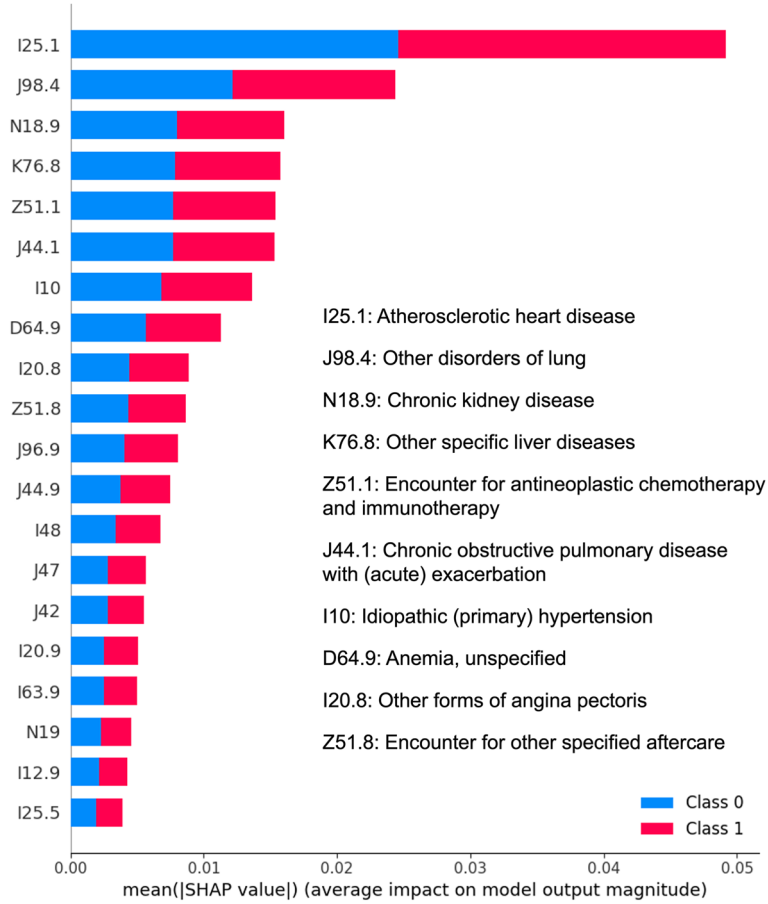


Fig. 3 Comorbidity assessment with SHAP. Class 1 corresponds to positive samples, Class 0 corresponds to negative samples

and parameter influence experiments. Finally, we invited clinicians to validate the results of the model prediction.

Performance of DAPNet

To better compare the performance of the DAPNet model, we conducted comparative experiments on the same dataset to verify the performance of the model. Notably, our research only relies on ICD code, which limits the direct comparability with existing similar studies. Therefore, we use machine learning methods as the baseline method. Meanwhile, this study aims to conduct as many ablation experiments as feasible to validate our model. The results of the performance comparison among models using severe pneumonia datasets are presented in Fig. 4A and Table 3. The DAPNet had the best overall performance, and its F1-score (0.8379) and AUC value (0.9172) were higher than other models. The RF model, as the best ML baseline model, also achieved good F1-score (0.7708) and AUC values (0.8528). Conversely, the LR and SVM models exhibited inferior performance in this experiment, lagging far behind DAPNet's exemplary performance. The tBNA-PR and HiTANet models, belonging to the deep learning group, performed not well, with their F1-score trailing behind those of the RF model.

The performance comparison of models with kidney datasets is shown in Fig. 4B and Table 4. The DAPNet demonstrated the best overall performance, and its F1-score (0.8018) and AUC value (0.838) being higher than those of other models. The HiTANet model, serving as the best baseline, also achieved good F1-score (0.661) and AUC values (0.6624). Although the tBNA-PR model achieved the highest Precision (0.959) and AUC (0.9523), it exhibited instability, characterized by a very low Recall (0.4528), indicating that the model did not fully capture the underlying knowledge.

Ablation experiments

This study compared different GNN models, and the results are shown in Table 5. It is noteworthy that for different GNN algorithms, the gap among AUC results is small. However, for the F1-score, all the GNN algorithms achieved results exceeding 0.8, with MA-GCL attaining the best performance (0.8379). The DAPNet model

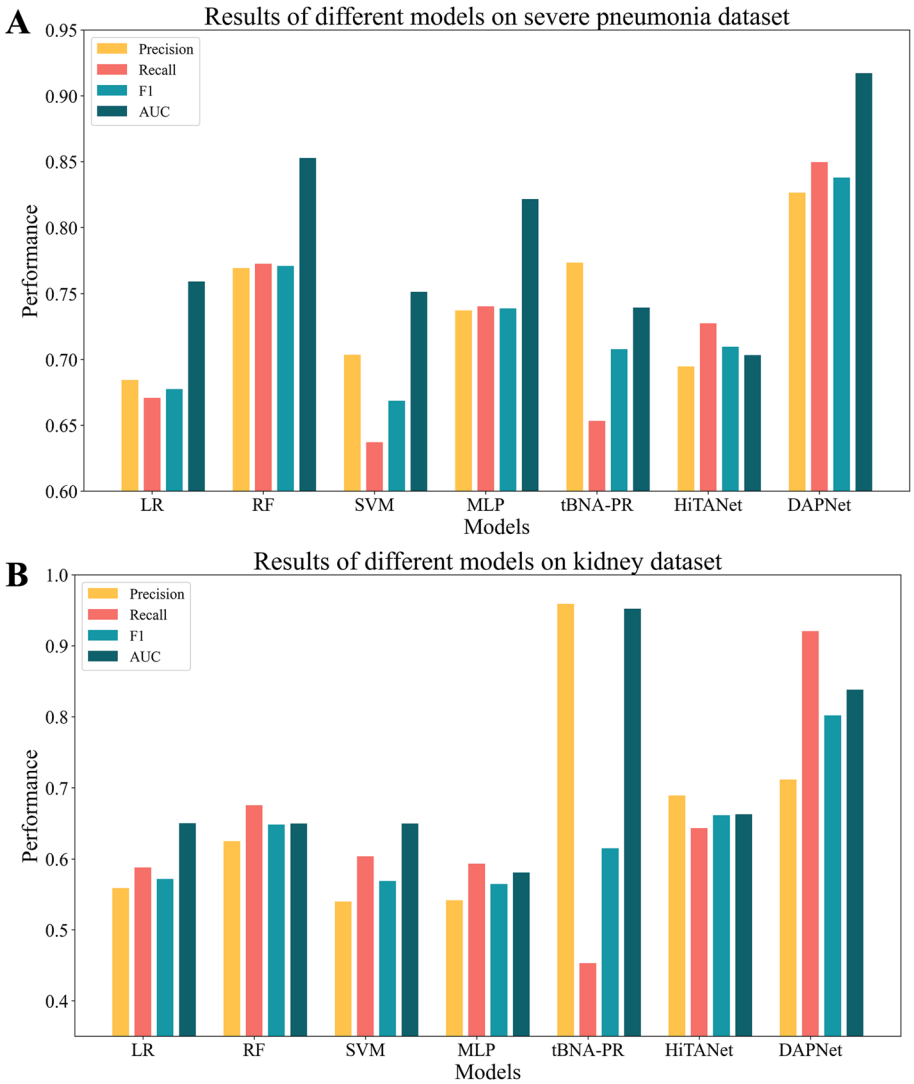


Fig. 4 Performance comparison between $DAPNet_{MA-GCL}$ and baseline. **A** Comparison of different models on severe pneumonia dataset. **B** Comparison of different models on kidney dataset

Table 3 Results of different models on severe pneumonia dataset

Model	Precision	Recall	F1	AUC
LR	0.6843±0.0118	0.6708±0.0097	0.6774±0.0086	0.759±0.0057
RF	0.7691±0.0096	0.7725±0.0111	0.7708±0.008	0.8528±0.0069
SVM	0.7035±0.008	0.637±0.0083	0.6685±0.0071	0.7514±0.0064
MLP	0.7373±0.0082	0.7402±0.0105	0.7387±0.007	0.8215±0.0068
tBNA-PR	0.7732±0.032	0.6532±0.0272	0.7077±0.024	0.7395±0.0244
HiTANet	0.6946±0.0178	0.7274±0.0387	0.7096±0.0119	0.7032±0.0072
DAPNet	**0.8265±0.0073**	**0.8497±0.0066**	**0.8379±0.0062**	**0.9172±0.0029**

Table 4 Results of different models on kidney dataset

Model	Precision	Recall	F1	AUC
LR	0.559±0.0523	0.5879±0.057	0.5717±0.047	0.6501±0.0392
RF	0.6249±0.0341	0.6751±0.0275	0.6482±0.0223	0.6499±0.0358
SVM	0.5399±0.0452	0.6035±0.0491	0.5688±0.0396	0.65±0.0341
MLP	0.5416±0.0421	0.5933±0.0381	0.5646±0.0267	0.5807±0.0365
tBNA-PR	**0.959±0.0068**	0.4528±0.0154	0.615±0.0145	**0.9523±0.0035**
HiTANet	0.6891±0.0434	0.6432±0.0838	0.661±0.041	0.6624±0.0312
DAPNet	0.7116±0.0492	**0.9207±0.022**	**0.8018±0.0332**	0.838±0.0236

Table 5 Results of different GNN algorithms

Model	Precision	Recall	F1	AUC
$DAPNet_{GCN}$	0.8066±0.0059	**0.8545±0.0065**	0.8298±0.0037	0.9106±0.0041
$DAPNet_{GRACE}$	**0.8282±0.0063**	0.8337±0.0076	0.8309±0.0033	**0.9191±0.0021**
$DAPNet_{NCLA}$	0.8213±0.0057	0.831±0.0054	0.8261±0.0041	0.9168±0.0021
DAPNet$_{MA-GCL}$	0.8265±0.0073	0.8497±0.0066	**0.8379±0.0062**	0.9172±0.0029

demonstrates a small variance, indicating its greater stability. Additionally, we also supplemented the t-test results in the appendix, which showed that the DAPNet model significantly outperforms the baseline model.

To ascertain the robustness of DAPNet, ablation experiments were conducted on multiple modules of DAPNet. Furthermore, a series of comparative experiments were also conducted on the scale of networks, as well as the principal parameters of CNN and GCN, across several groups with varying parameters.

The ablation experiment (Fig. 5A1-A4) demonstrated different combinations (A, B, C) and input forms (D, E, F) for the DAPNet model. The comparison of different modules primarily focused on the splicing of various submodules within DAPNet. The different input forms, on the other hand, centered around whether the input data was sorted based on the ICD code or underwent linear transformation. Specifically, "sort (input)" denoted the order of features, "random (input)" indicated a random ordering, and "without pretreatment" indicated no linear transformation calculation. Based on the MLP model, the experimental result of "A+C" combined with CNN model was better than that of "B+C" combined with GCN, reflecting that the performance of CNN was superior to that of disease network feature extraction. Furthermore, the input sorted according to the ICD code yielded better results than the unsorted input and was also superior to the input without linear transformation.

The experimental results with different disease networks are presented in Fig. 5B1-B4. Generally, the performance of the method employing multiple networks outperformed that of utilizing a single network. Notably, compared with other networks, especially molecular-based disease networks, DAPNet with disease co-occurrence networks exhibited the best performance (highest F1: 0.8366). This could be attributed to the lower quality of information in the "disease-gene" relationship data. Figure 5C1-C4 displays the experimental outcomes when employing varying proportions of network data under the same network. The results demonstrate that the model performance improved as the proportion of network structures used increased, indicating the availability of more edge information. Furthermore, it was observed that, for the hierarchical disease network, using more than 80% of edge data yielded results similar to those when all edge data was utilized.

Sensitivity analysis

The CNN module is primarily responsible for extracting the patient's clinical disease feature information and generating the high-dimensional features of the patient's clinical disease. In this set of comparative tests, we focused on comparing the size of the convolution kernel of CNN (Fig. 5D1-D4), represented by F, where the cases of 4, 8, 16, and 32 were compared. The results indicate that the optimal F1-score was achieved when F=16, and the optimal AUC result was obtained when F=8. The overall pattern of these two results was similar. However, when F was relatively large, such as F=32, the performance exhibited a decline. The GCN module was primarily responsible for extracting disease network node features and aiding in constructing patient disease features based on combined disease networks. In this set of comparative experiments, we mainly focused on comparing the

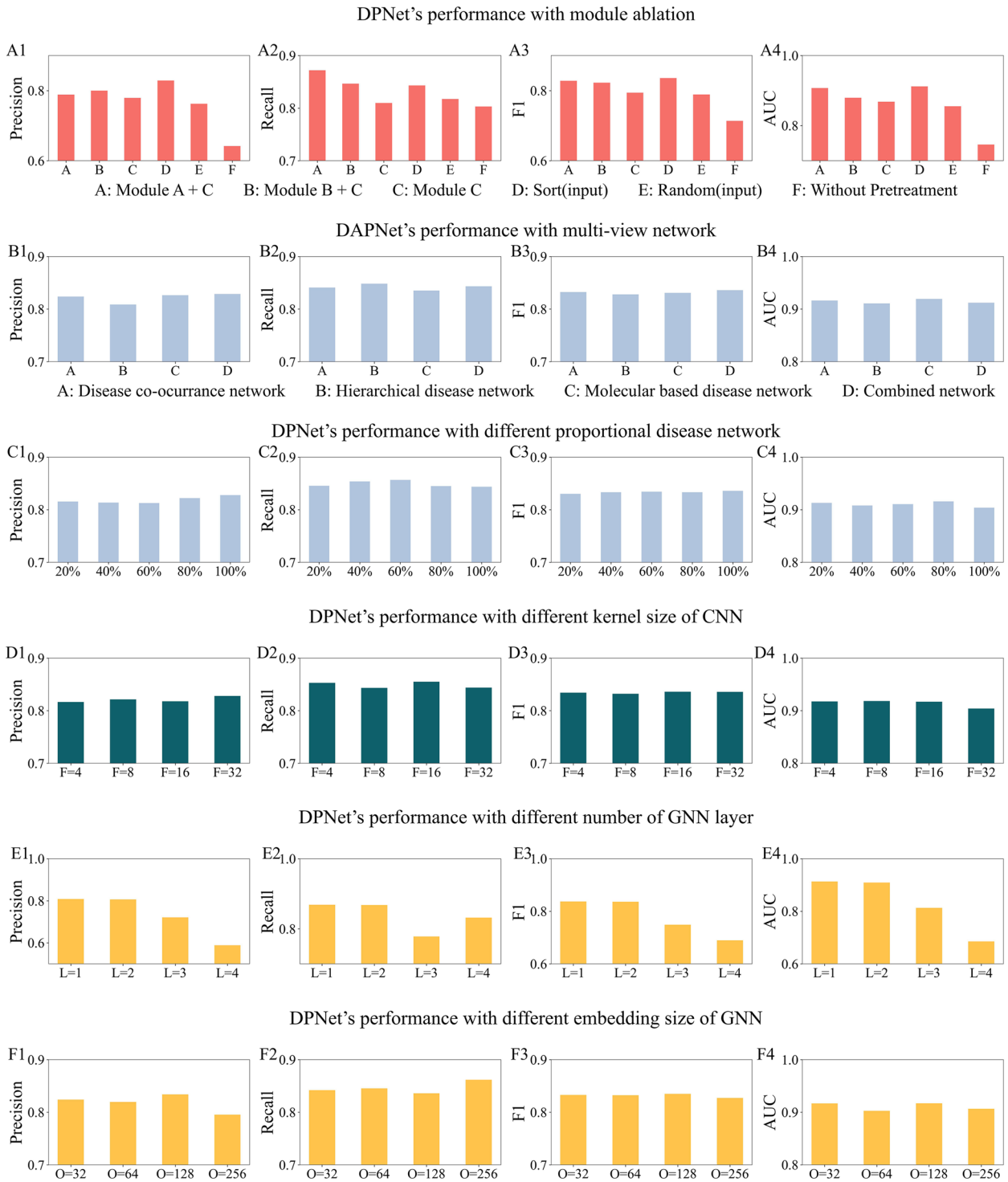


Fig. 5 Comparison of DAPNet core parameters. A1-A4 Performance comparison of different forms of DAPNet. B1-B4 Performance comparison of DAPNet with different networks. C1-C4 Performance comparison of DAPNet with different heterogeneous networks data scale. D1-D4 Performance comparison of DAPNet with different kernel size of CNN. E1-E4 Performance comparison of DAPNet with different number of GNN layer. F1-F4 Performance comparison of DAPNet with different embedding size of GNN

GCN module's layers and feature dimensions (Fig. 5E1-E4). The number of layers was represented by the letter L, while the output feature dimension was represented by the letter O. The results indicate that the performance of the GCN module was better when the number of layers was small. Generally, with an increase in the number of layers, the model's complexity increases, and the model's effectiveness exhibits a downward trend. In summary, the overall performance of the GCN module did not improve significantly with increasing layers. The model performed relatively well with a simple two-layer structure and an output feature dimension of 64. Figure 5F1-F4 presents the results of the experiment on the output dimension of the GNN module. The results indicate that the performance gap was relatively small and the performance was similar when the output feature dimension was less than 32. However, when the output feature dimension was increased to 256, there was a significant decline in performance. These findings suggest that a lower output feature dimension is preferable for the GNN module.

Indeed, the robustness of DAPNet could be attributed to the vital role played by the various components in the model's overall performance. The ablation experiments on different combinations of modules and input forms revealed that the performance of DAPNet was dependent on the specific combination of modules and the pre-processing of input data. Furthermore, experiments on disease networks demonstrated that incorporating multiple networks, especially disease co-occurrence networks, could lead to better results. Lastly, experiments on the CNN and GNN modules reveal that the size of the convolution kernel and the number of layers and output feature dimensions in the GNN module significantly impacted the overall performance of the model. Therefore, the robustness of DAPNet was attributed to the optimized combination of the different components, which allows for the effective extraction and integration of various types of data to generate high-quality disease prediction results.

Case study and interpretability

To evaluate the effectiveness of the proposed DAPNet model, we extracted actual patient cases to demonstrate the prediction results of the DAPNet model. Figure 6 showed the case study of disease progression prediction for severe pneumonia. The DAPNet model made accurate predictions for three random samples, with the predicted labels matching the actual labels. In addition, we examined the clinical significance of the ICD codes for Sample 53, which revealed that the patient suffered from serious lung diseases such as "J98.4, J15.9, C34.1, J42". The patient presented with "I50.9" in the subsequent diagnostic results, indicating that the coexistence of cancer and pneumonia increases the risk of severe conditions like heart failure [52–54]. For Sample 268, DAPNet predicted a score of 0.6, while the true label was negative. Although predicting this patient's outcome was relatively difficult, the model still provided an effective prediction. In Sample 354, there was no discernible connection between the patient's illness and pneumonia, and the model's prediction was consistent with the true label of low risk. These findings demonstrate that the DAPNet model could provide reliable predictions, thereby aiding healthcare professionals in making faster and more accurate diagnoses and preventing disease progression.

This study also explored the clinically relevant diseases and molecular associations identified by DAPNet. We analyzed the gene association information behind each patient's disease based on the disease information from the case study and our disease relationship network. Meanwhile, based on the MalaCards database, the relevant genes for each disease were searched and the number of matching genes was calculated, as shown in Fig. 7. The horizontal axis represents high-frequency co-occurrence genes, and the vertical axis represents their frequency. For sample 268, the high-frequency gene "CRP" is associated with disease "I70.9, E78.5, M81, I25.9, I25.1, M19.9". This finding is consistent with the study on atherosclerosis [55] and provides insights into the underlying molecular mechanisms of disease progression.

Discussion and conclusion

In this study, we proposed a novel disease progression prediction framework, DAPNet, which solely utilized the comorbidity, its duration, and disease association networks to deliver high-performance prediction. Leveraging three disease association networks, DAPNet was the first work of disease progression prediction only based on historical disease progression (only ICD data), thereby reducing data dependence while still maintaining good performance. The results showed that DAPNet achieved state-of-the-art performance on the severe pneumonia dataset (F1=0.84, with an improvement of 8.7%), and outperformed the six baseline models on the kidney disease dataset (F1=0.80, with an improvement of 21.3%). Even in this challenging task (only by historical disease progression), compared to other disease prediction models, DAPNet achieved better results on both datasets and reached the level of a clinically applicable model.

Unlike other disease progression prediction research often heavily relies on multi-modal patient data or comprehensive medical records [6–8], DAPNet only utilizes the patient's ICD information and still achieves excellent results, reducing the dependence on data. DAPNet aims to demonstrate that effective disease progression prediction can be achieved even with limited data, such as using

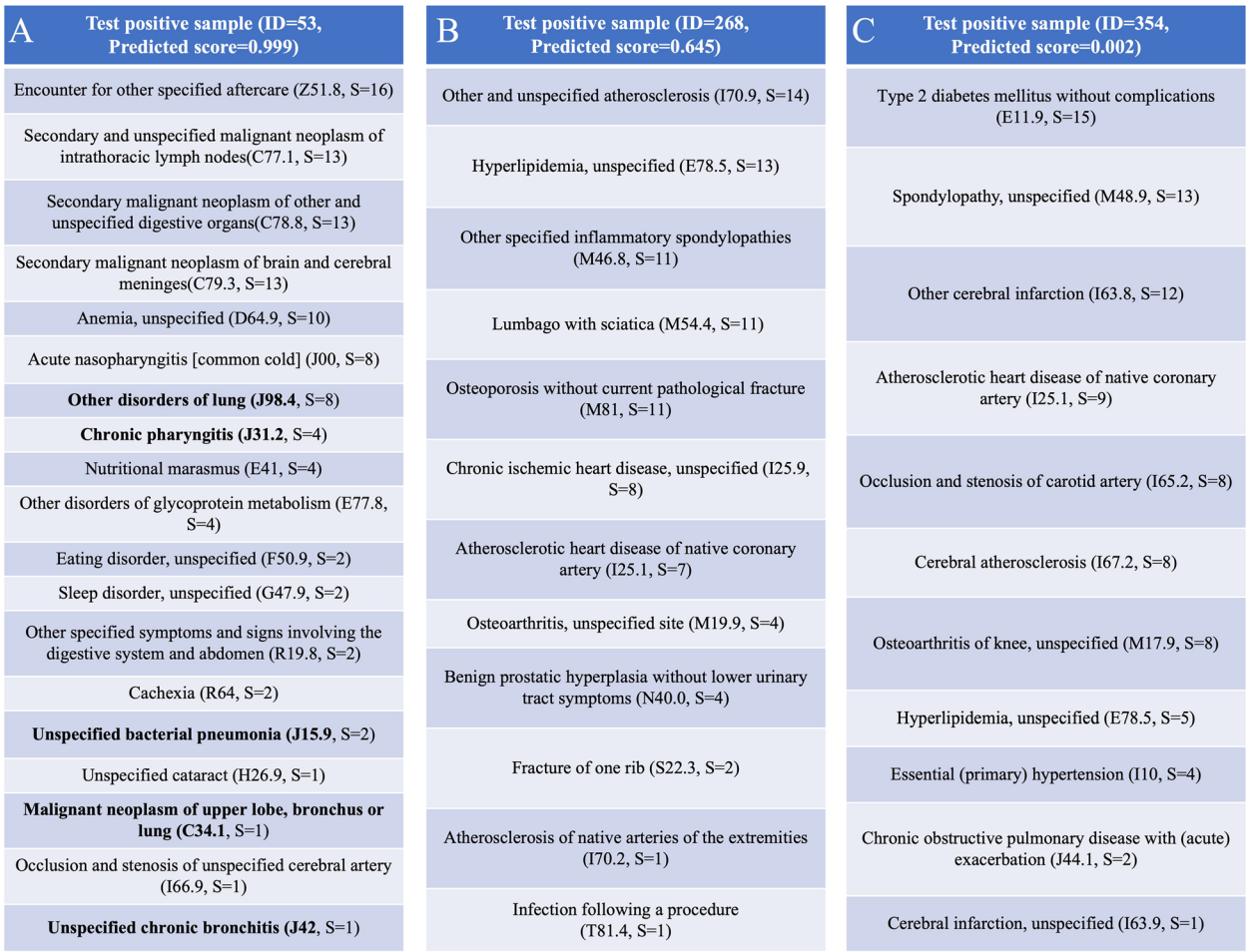


Fig. 6 Diagnoses of sample patients. Sample patients including one positive case (**A**), one neutral case (**B**), and one negative case (**C**)

medical claims data, thereby providing a more straightforward and reliable tool for clinical practice. It is worth noting that DAPNet, compared to other comorbidity studies, additionally integrates molecular-level disease association information, combining disease co-occurrence with the disease relationship information behind ICD-10 codes, and thoroughly explores the associations between diseases. Furthermore, DAPNet employs a multi-view GCL that fuses phenotype and molecular information, which is the first application in disease prediction problems. By leveraging the implicit disease relationships across three disease networks, DAPNet generated a more comprehensive patient representation and enhanced prediction accuracy. Compared to other disease progression prediction models, our DAPNet also showed reduced reliance on clinical data, using only historical disease progression for disease progression prediction while still achieving strong performance [6, 10, 12, 37]. Notably, we introduced a method for patient embedding construction that integrates historical disease progression to capture the duration and impact of prior illnesses

We acknowledge several limitations in the current research, which highlight opportunities for future improvements. Currently, our disease predictions cover only two types of diseases, and there is a lack of experimental validation for others [4]. In the future, we plan to continue collecting more clinical data and implement manual review and annotation procedures to refine the disease association data. Additionally, we intend to construct datasets for other diseases to broaden the range of disease categories, thereby increasing the dataset's scale. It is also important to note that disease association data is continually evolving. Hence, we will strive to collect more data in future work to enhance the scale and quality of the disease association networks [20, 26]. Besides, recognizing the importance of prospective research, we plan to design a prospective study in the future to address potential biases inherent in retrospective research and explore the design of diagnostic and treatment models

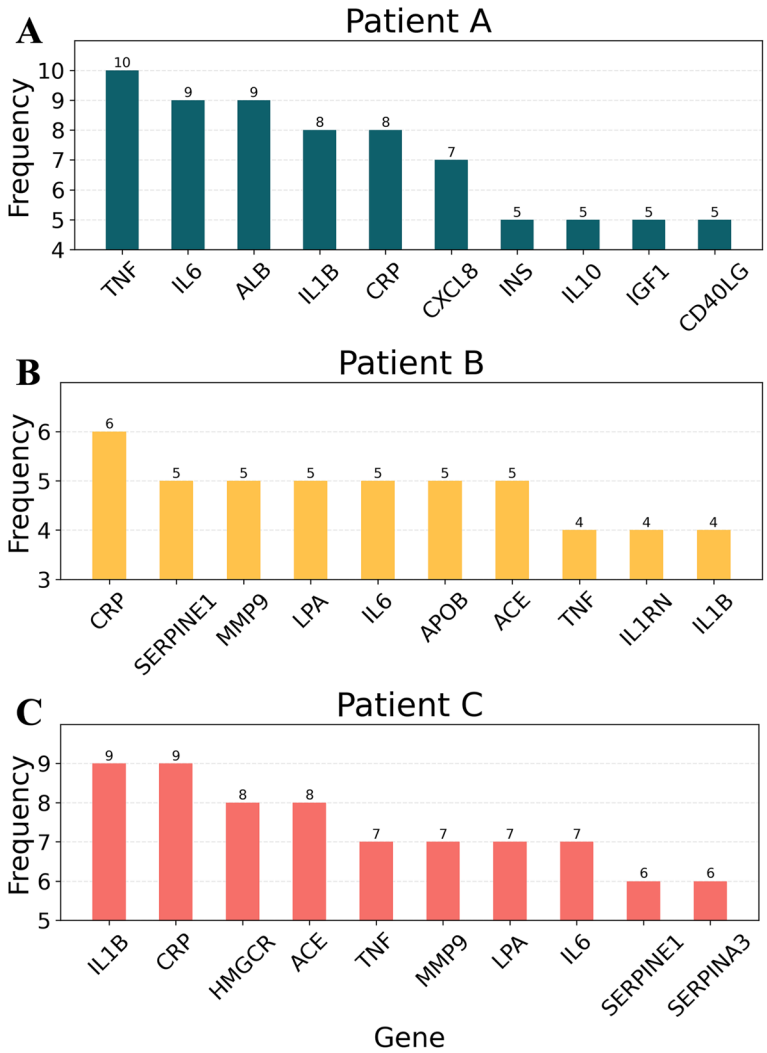


Fig. 7 Statistical analysis of disease-gene association

that integrate outcome information in subsequent studies [46, 56]. We plan to collect a new dataset covering a prolonged disease course. Using temporal information, patients' multiple visit records will be divided into two segments: the earlier segment will be used to train our model for predicting disease progression, while the later segment will be utilized for a retrospective study to validate the accuracy of the model's predictions. By integrating these two experimental approaches, we aim to uncover underlying patterns in the disease more comprehensively. In addition, we aim to explore the integration of clinical heterogeneous knowledge graphs to improve the model's performance and interpretability, ultimately enabling more effective and personalized early treatment strategies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02756-0.

Supplementary Material 1.

Acknowledgements
Not applicable.

Authors' contributions
Haoyu Tian, Kuo Yang and Xiong He wrote the main manuscript text. Fengjin Zhang, Shihua Wang, and Tiancai Wen collected and organized data. Xinyu Dai, Yiming Liu, Zixin Shu, Qiguang Zheng, and Jianan Xia assisted in the experiment and prepared figures. Xuezhong Zhou, Baoyan Liu, and Jian Yu provided guidance on the research. All authors have reviewed the manuscript.

Funding
This work is partially supported by the National Key Research and Development Program (2023YFC3502604), the National Natural Science Foundation

of China (U23B2062), the Natural Science Foundation of Beijing (L232033), the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0505700), and Key R&D Program Project of Ningxia Hui Autonomous Region(2022BEG02036). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability

The data that support the findings of this study were used under license for the current study. Preprocessing code has been released on Github (https://github.com/thy0621/DAPNet) to help readers have a clearer understanding of our data processing process. The datasets and materials used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

The process was reviewed and approved by the Research Ethics Committee of China Academy of Chinese Medical Science (2016NO.11-01). No potentially identifiable human images or data was presented in this study. This study obtained informed consent from all participants in the study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100063, Beijing, China. [2]Department of Nephrology, Third Hospital of Hebei Medical University, China Academy of Chinese Medical Sciences, Shijiazhuang 050051, Hebei, China. [3]Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, Beijing, China. [4]Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, Beijing, China.

Received: 21 February 2024 Accepted: 7 November 2024
Published online: 19 November 2024

References

1. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med. 2023;388(13):1201–8. https://doi.org/10.1056/NEJMra2302038.
2. Beam AL, Drazen JM, Kohane IS, Leong TY, Manrai AK, Rubin EJ. Artificial intelligence in medicine. N Engl J Med. 2023;388(13):1220–1. https://doi.org/10.1056/NEJMe2206291.
3. The Emerging Risk Factors Collaboration. C-Reactive Protein, Fibrinogen, and Cardiovascular Disease Prediction. N Engl J Med. 2012;367(14):1310–20. https://doi.org/10.1056/NEJMoa1107477.
4. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform. 2018;22(5):1589–604. https://doi.org/10.1109/JBHI.2017.2767063.
5. Delerme S, Ray P. Acute Respiratory Failure in the Elderly: Diagnosis and Prognosis. Age Ageing. 2008;37(3):251–7. https://doi.org/10.1093/ageing/afn060.
6. Feng Z, Yu Q, Yao S, Luo L, Zhou W, Mao X, et al. Early Prediction of Disease Progression in COVID-19 Pneumonia Patients with Chest CT and Clinical Characteristics. Nat Commun. 2020;11(1):4968. https://doi.org/10.1038/s41467-020-18786-x.
7. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep. 2016;6(1):26094. https://doi.org/10.1038/srep26094.
8. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In: Doshi-Velez F, Fackler J, Kale D, Wallace B, Wiens J, editors. Proceedings of the 1st Machine Learning for Healthcare Conference. vol. 56 of Proceedings of Machine Learning Research. Northeastern University, Boston: PMLR; 2016. pp. 301–318.
9. Mohan S, Thirumalai C, Srivastava G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access. 2019;7:81542–54. https://doi.org/10.1109/ACCESS.2019.2923707.
10. Lu Q, Nguyen TH, Dou D. Predicting Patient Readmission Risk from Medical Text via Knowledge Graph Enhanced Multiview Graph Convolution. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event Canada: ACM; 2021. pp. 1990–1994. https://doi.org/10.1145/3404835.3463062.
11. Cheng Y, Wang F, Zhang P, Hu J. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In: Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics; 2016. pp. 432–440. https://doi.org/10.1137/1.9781611974348.49.
12. Xiao J, Ding R, Xu X, Guan H, Feng X, Sun T, et al. Comparison and Development of Machine Learning Tools in the Prediction of Chronic Kidney Disease Progression. J Transl Med. 2019;17(1):119. https://doi.org/10.1186/s12967-019-1860-0.
13. Luo J, Ye M, Xiao C, Ma F. HiTANet: Hierarchical Time-Aware Attention Networks for Risk Prediction on Electronic Health Records. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Virtual Event: ACM; 2020. pp. 647–656. https://doi.org/10.1145/3394486.3403107.
14. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: Graph-Based Attention Model for Healthcare Representation Learning. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17. New York: Association for Computing Machinery; 2017. pp. 787–795. https://doi.org/10.1145/3097983.3098126.
15. Liang Y, Guo C. Heart Failure Disease Prediction and Stratification with Temporal Electronic Health Records Data Using Patient Representation. Biocybernetics Biomed Eng. 2023;43(1):124–41. https://doi.org/10.1016/j.bbe.2022.12.008.
16. Gao J, Wang X, Wang Y, Yang Z, Gao J, Wang J, et al. CAMP: Co-Attention Memory Networks for Diagnosis Prediction in Healthcare. In: 2019 IEEE International Conference on Data Mining (ICDM), 2019. pp. 1036–1041. https://doi.org/10.1109/ICDM.2019.00120.
17. Demeniconi C. Knowledge Guided Diagnosis Prediction via Graph Spatial-Temporal Network. Philadelphia: Society for Industrial and Applied Mathematics; 2020. https://doi.org/10.1137/1.9781611976236.
18. Lu C, Reddy CK, Ning Y. Self-Supervised Graph Learning With Hyperbolic Embedding for Temporal Health Event Prediction. IEEE Trans Cybern. 2021:1–13. https://doi.org/10.1109/TCYB.2021.3109881.
19. Degroot V, Beckerman H, Lankhorst G, Bouter L. How to Measure Comorbiditya Critical Review of Available Methods. J Clin Epidemiol. 2003;56(3):221–9. https://doi.org/10.1016/S0895-4356(02)00585-1.
20. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The Human Disease Network. Proc Natl Acad Sci. 2007;104(21):8685–90. https://doi.org/10.1073/pnas.0701361104.
21. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease Networks. Uncovering Disease-Disease Relationships through the Incomplete Interactome. Science (New York, NY). 2015;347(6224):1257601. https://doi.org/10.1126/science.1257601.
22. Hidalgo CA, Blumm N, Barabási AL, Christakis NA. A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Comput Biol. 2009;5(4): e1000353. https://doi.org/10.1371/journal.pcbi.1000353.
23. Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási AL. The Implications of Human Metabolic Network Topology for Disease Comorbidity. Proc Natl Acad Sci. 2008;105(29):9880–5. https://doi.org/10.1073/pnas.0802208105.
24. Park J, Lee DS, Christakis NA, Barabási AL. The Impact of Cellular Networks on Disease Comorbidity. Mol Syst Biol. 2009;5(1):262. https://doi.org/10.1038/msb.2009.16.
25. Lu H, Uddin S. A Weighted Patient Network-Based Framework for Predicting Chronic Diseases Using Graph Neural Networks. Sci Rep. 2021;11(1):22607. https://doi.org/10.1038/s41598-021-01964-2.
26. Lu H, Uddin S, Hajati F, Moni MA, Khushi M. A Patient Network-Based Machine Learning Model for Disease Prediction: The Case of Type 2

Diabetes Mellitus. Appl Intell. 2022;52(3):2411–22. https://doi.org/10.1007/s10489-021-02533-w.
27. Sieranoja S, Fränti P. Adapting K-Means for Graph Clustering. Knowl Inf Syst. 2022;64(1):115–42. https://doi.org/10.1007/s10115-021-01623-y.
28. Choudhary GI, Fränti P. Predicting Onset of Disease Progression Using Temporal Disease Occurrence Networks. Int J Med Inform. 2023;175: 105068. https://doi.org/10.1016/j.ijmedinf.2023.105068.
29. Peng J, Yang K, Tian H, Lin Y, Hou M, Gao Y, et al. The Mechanisms of Qizhu Tangshen Formula in the Treatment of Diabetic Kidney Disease: Network Pharmacology, Machine Learning. Molecular Docking and Experimental Assessment Phytomedicine. 2023;108: 154525. https://doi.org/10.1016/j.phymed.2022.154525.
30. Vidal M, Cusick ME, Barabási AL. Interactome Networks and Human Disease. Cell. 2011;144(6):986–98. https://doi.org/10.1016/j.cell.2011.02.016.
31. Yang Y, Yang K, Hao T, Zhu G, Ling R, Zhou X, et al. Prediction of Molecular Mechanisms for LianXia NingXin Formula: A Network Pharmacology Study. Front Physiol. 2018;9:489. https://doi.org/10.3389/fphys.2018.00489.
32. Gong X, Yang C, Shi C. MA-GCL: Model Augmentation Tricks for Graph Contrastive Learning. Proceedings of the AAAI Conference on Artificial Intelligence. 2023;37(4):4284–92. https://doi.org/10.1609/aaai.v37i4.25547.
33. Shen X, Sun D, Pan S, Zhou X, Yang LT. Neighbor Contrastive Learning on Learnable Graph Augmentation. Proceedings of the AAAI Conference on Artificial Intelligence. 2023;37(8):9782–91. https://doi.org/10.1609/aaai.v37i8.26168.
34. Wu L, Lin H, Tan C, Gao Z, Li SZ. Self-Supervised Learning on Graphs: Contrastive, Generative, or Predictive. IEEE Trans Knowl Data Eng. 2023;35(4):4216–35. https://doi.org/10.1109/TKDE.2021.3131584.
35. Zhu Y, Xu Y, Yu F, Liu Q, Wu S, Wang L. Deep Graph Contrastive Representation Learning. arXiv preprint arXiv:2006.04131. 2020. https://doi.org/10.48550/arXiv.2006.04131.
36. Li Z, Liu F, Yang W, Peng S, Zhou J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Trans Neural Netw Learn Syst. 2022;33(12):6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827.
37. Yuan Q, Chen K, Yu Y, Le NQK, Chua MCH. Prediction of Anticancer Peptides Based on an Ensemble Model of Deep Learning and Machine Learning Using Ordinal Positional Encoding. Brief Bioinform. 2023;24(1):bbac630. https://doi.org/10.1093/bib/bbac630.
38. Kha QH, Ho QT, Le NQK. Identifying SNARE Proteins Using an Alignment-Free Method Based on Multiscan Convolutional Neural Network and PSSM Profiles. J Chem Inf Model. 2022;62(19):4820–6. https://doi.org/10.1021/acs.jcim.2c01034.
39. Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, et al. MalaCards: An Amalgamated Human Disease Compendium with Diverse Clinical and Genetic Annotation and Structured Search. Nucleic Acids Res. 2017;45(D1):D877–87. https://doi.org/10.1093/nar/gkw1012.
40. Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Res. 2004;32(90001):267D – 270. https://doi.org/10.1093/nar/gkh061.
41. Zhou X, Menche J, Barabási AL, Sharma A. Human Symptoms-Disease Network Nat Commun. 2014;5(1):4212. https://doi.org/10.1038/ncomms5212.
42. Gan M, Dou X, Jiang R. From Ontology to Semantic Similarity: Calculation of Ontology-Based Semantic Similarity. Sci World J. 2013;2013:1–11. https://doi.org/10.1155/2013/793091.
43. Wang Y, Stroh JN, Hripcsak G, Low Wang CC, Bennett TD, Wrobel J, et al. A Methodology of Phenotyping ICU Patients from EHR Data: High-fidelity, Personalized, and Interpretable Phenotypes Estimation. J Biomed Inform. 2023;148:104547. https://doi.org/10.1016/j.jbi.2023.104547.
44. Manchikanti L, Falco FJE, Hirsch JA. Ready or Not! Here Comes ICD-10. J NeuroInterventional Surg. 2013;5(1):86–91. https://doi.org/10.1136/neurintsurg-2011-010155.
45. Jiadong Xie, Kongfa Hu, Peipei Fang, Guozheng Li, Liu B. Design and Implementation of the Platform for Collection and Analysis of the Inpatient Medical Record Home Page of Traditional Chinese Medicine. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Shenzhen: IEEE; 2016. pp. 1399–1402. https://doi.org/10.1109/BIBM.2016.7822726.
46. Chan KW, Shu Z, Chang K, Liu B, Zhou X, Li X. Add-on Chinese Medicine for Coronavirus Disease 2019 (COVID-19): A Retrospective Cohort. Eur J Integr Med. 2021;48:101903. https://doi.org/10.1016/j.eujim.2021.101903.
47. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a Freely Accessible Electronic Health Record Dataset. Sci Data. 2023;10(1):1. https://doi.org/10.1038/s41597-022-01899-x.
48. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. arXiv preprint arXiv:1609.02907. 2016. https://doi.org/10.48550/arXiv.1609.02907.
49. Pennington J, Socher R, Manning CD. Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL; 2014. p. 1532–1543. https://doi.org/10.3115/v1/D14-1162.
50. Zou Q, Yang K, Chang K, Zhang X, Li X, Zhou X. Phenonizer: A Fine-Grained Phenotypic Named Entity Recognizer for Chinese Clinical Texts. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2021. pp. 3963–3970.
51. Kurth L, Halldin C, Laney AS, Blackley DJ. Causes of Death among Federal Black Lung Benefits Program Beneficiaries Enrolled in Medicare, 1999–2016. Am J Ind Med. 2020;63(11):973–9. https://doi.org/10.1002/ajim.23176.
52. Gabet A, Juillière Y, Lamarche-Vadel A, Vernay M, Olié V. National Trends in Rate of Patients Hospitalized for Heart Failure and Heart Failure Mortality in France, 2000–2012: Trends in Hospitalization Rates for HF and HF Mortality in France. Eur J Heart Fail. 2015;17(6):583–90. https://doi.org/10.1002/ejhf.284.
53. Mikkelsen L, Iburg KM, Adair T, Fürst T, Hegnauer M, Von Der Lippe E, et al. Assessing the Quality of Cause of Death Data in Six High-Income Countries: Australia, Canada, Denmark, Germany, Japan and Switzerland. Int J Public Health. 2020;65(1):17–28. https://doi.org/10.1007/s00038-019-01325-x.
54. Villela PB, Santos SC, De Oliveira GMM. Heart Failure Quantified by Underlying Cause and Multiple Cause of Death in Brazil between 2006 and 2016. BMC Public Health. 2021;21(1):2100. https://doi.org/10.1186/s12889-021-12173-x.
55. Del Rincón I, Williams K, Stern MP, Freeman GL, O'Leary DH, Escalante A. Association between Carotid Atherosclerosis and Markers of Inflammation in Rheumatoid Arthritis Patients and Healthy Subjects. Arthritis Rheum. 2003;48(7):1833–40. https://doi.org/10.1002/art.11078.
56. Burrows B, Earle RH. Course and prognosis of chronic obstructive lung disease: a prospective study of 200 patients. N Engl J Med. 1969;280(8):397–404. https://doi.org/10.1056/NEJM196902202800801.

Publisher's Note