scientific reports

OPEN

A novel liver cancer diagnosis method based on patient similarity network and DenseGCN

Ge Zhang[1], Zhen Peng[1], Chaokun Yan[1], Jianlin Wang[1], Junwei Luo[2] & Huimin Luo[1]✉

Liver cancer is the main malignancy in terms of mortality rate, accurate diagnosis can help the treatment outcome of liver cancer. Patient similarity network is an important information which helps in cancer diagnosis. However, recent works rarely take patient similarity into consideration. To address this issue, we constructed patient similarity network using three liver cancer omics data, and proposed a novel liver cancer diagnosis method consisted of similarity network fusion, denoising autoencoder and dense graph convolutional neural network to capitalize on patient similarity network and multi omics data. We compared our proposed method with other state-of-the-art methods and machine learning methods on TCGA-LIHC dataset to evaluate its performance. The results confirmed that our proposed method surpasses these comparison methods in terms of all the metrics. Especially, our proposed method has attained an accuracy up to 0.9857.

Liver cancer is the main malignancy worldwide, and its incidence is still increasing annually[1]. According to GLOBOCAN 2020, liver cancer causes about 830,000 deaths, ranking third leading cause of cancer deaths in 2020[2]. Studies have shown that early diagnosis of cancer can help improve survival rates[3]. However, the symptoms of liver cancer in early stage are not obvious[4], most liver cancer patients are already in the middle and late stages when they are diagnosed, and treatment options are limited[5]. These factors causes that the liver cancer has a poor prognosis[6]. Therefore, it is of great practical importance to design a method that can effectively perform early diagnosis and help improve the treatment outcome of liver cancer.

With the emergence of gene sequencing technology, the amount of biological data has exploded[7,8], which has provided researchers with plenty of omics data from different aspects, such as proteomics, transcriptomics, epigenomics, and genomics. Analyze and utilize these omics data for cancer diagnosis is a hot issue[9–12]. The cancer diagnosis methods can be normally categorized into two kinds, machine learning methods and deep learning methods. Sun et al.[13] developed an improved feature selection method, called I-RELIEF, to extract hybrid features from breast cancer microarray data and clinical data. The extracted features were using to construct a breast cancer diagnostic model based on linear discriminant analysis (LDA). The excellent performance of the cancer diagnostic model was verified by comparing with several benchmark methods. Akay et al.[14] proposed a breast cancer diagnosis method using SVM and F-score[15]. They first ranked the features by F-score, and then carried out grid search method to find parameters for SVM model which can get the best performance. Final experiment results indicated that this method had a better performance compared with previous works. Tsai et al.[16] developed an artificial bee colony algorithm (ABC) combined with SVM for cancer diagnosis. They applied ABC to screen relatively important genes in gene expression data for cancer stage diagnosis, and identified some genes that could be used as biomarkers for further study. To address the problem that the success rate of liver cancer diagnosis is not satisfactory, Zhang et al.[17] introduced a hybrid cancer diagnosis method, which is based on SVM, incremental feature selection (IFS) and max-relevance and min-redundancy (mRMR). Firstly, mRMR was used to screen the gene expression data, then IFS was used for further selection of the screened features, and finally the obtained genes were input to SVM for liver cancer diagnosis. However, Machine learning methods have difficulty processing raw data directly, they usually transformed the raw data into appropriate feature vectors. This may bring additional computational cost[18].

Recent years, deep learning, which has the ability to capture intricate structures from raw data, started to gain attention in bioinformatics field and many cancer diagnosis methods based on deep learning method have been proposed[19]. Fakoor et al.[20] reduced the dimensionality of gene expression data by principal component analysis (PCA). Then sparse autoencoder (SAE) was used for further feature extraction and finally softmax was

[1]School of Computer and Information Engineering, Henan University, Kaifeng, China. [2]Present address: Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng, China. ✉email: luohuimin@henu.edu.cn

used for cancer diagnosis. lyu and Haque[21] transformed the gene expression data into 2-D images, then input the 2-D images into convolutional neural network to classify cancer of 33 tumor types. Gao et al.[22] introduced a novel cancer diagnosis method (DeepCC). DeepCC performs gene enrichment analysis to transform the gene expression data into functional spectra. Then the resulting functional spectra are input into a multilayer neural network for subsequent training. For both colorectal and breast cancer, DeepCC outperforms random forest (RF) and SVM for cancer subtype classification. However, previous deep learning-based models mainly use single omics data, which is limited to describe all the features of cancer[23]. It limits the performance of deep learning in cancer diagnosis.

Accordingly, cancer diagnosis methods based on multiple omics data are increasingly adopted[24–26]. Sun et al.[27] proposed a deep learning method which is based on model fusion, named MDNNMD, for breast cancer prognosis. They used two types of omics data, gene expression data and copy number variation (CNV), as well as clinical data, and constructed three deep neural network (DNN) models for the three types of data, and finally fused the prediction scores of the three independent models as the final prediction result. Zhang et al.[28] used variational autoencoder (VAE) to integrate methylation data and gene expression data to diagnose cancer. They used ten-fold cross-validation on 33 types of cancers to evaluate their method, and the final accuracy obtained by their method is 97.49%. Copy number variation, gene expression, and methylation data were used in these researches on cancer diagnosis. This indicated that copy number variation, gene expression, and methylation data bring useful information to cancer diagnosis. Thus, all these three omics data were selected in this work.

Previous studies have often only used genomics data. Interpretability is particularly required in genomics because of relatively smaller sample sizes and to better understand the molecular causes of disease so that targeted therapies can be designed[29]. Patient similarity network (PSN) can address these problems and specializes in integrating multi-omics data and generating interpretable models[30]. However, previous works rarely took the patient similarity into account. To address this issue, we integrated three omics data of liver cancer and calculated the similarity between patients. As the similarity network is none-Euclidean data, previous neural networks like CNNs, are hard to handle this data[31,32]. Thus, graph convolutional network (GCN), which has the advantages in processing non-Euclidean data is used in this work. Meanwhile, since omics data have small sample size, we need a deeper network to fit the data and thus avoid the disadvantages associated with the small sample size[33]. But the number of GCN layers is rarely more than four because of the vanishing gradient problem[34]. To deal with this challenge, we selected the dense graph convolutional neural network (DenseGCN)[35]. DenseGCN improves information flow in the network by densely connecting different layers. DenseGCN is able to overcome the vanishing gradient problem and make the GCN architecture deeper, thus enabling better utilization of patient similarity network and multi-omics data for cancer diagnosis. To the best of our knowledge, this is the first effort to employ DenseGCN in cancer diagnosis field.

In this work, A novel liver cancer diagnosis method (pDenseGCN) based on patient similarity network and DenseGCN is proposed. We first used similarity network fusion (SNF) to construct the patient similarity network using three liver cancer omics data. Then, we extracted latent embedding representation of omics data by using denoising autoencoder (DAE). This can provide a more precise representation of liver cancer. Finally, we adopted DenseGCN for liver cancer diagnosis based on the patient similarity network and latent representation of omics data. By incorporating the supplemental information PSN into the model, we got a more comprehensive view of cancer and finally obtained better performance on liver cancer diagnosis. According to the reliable experiments, our method pDenseGCN gained an accuracy score of 0.9857, and performed better compared with five state-of-the-art methods and machine learning methods.

The main contributions of this paper are as follows.

- A novel deep learning method, named pDenseGCN, is proposed for effectively liver cancer diagnosis.
- pDenseGCN utilizes SNF to construct a patient similarity network based on multi-omics, thus captures the similarity information between patients, which helps in liver cancer diagnosis.
- pDenseGCN adopts DenseGCN as the classifier. DenseGCN connects different layers densely to improve information flow in the network, which can overcome vanishing gradient problem. This brings better results in liver cancer diagnosis.

Methods

Proposed method.

There are three components in the proposed method pDenseGCN. The first component is generating patient similarity network by omics datasets. Three omics datasets were applied as the input of similarity network fusion method to produce patient similarity network. The second component is extracting feature by denoising autoencoder. In this step, RNA-Seq, DNA Methylation and CNV were put into denoising autoencoder respectively to obtain low-dimensional features. The next component is to input the obtained patient similarity network and feature matrix into dense graph convolutional network (DenseGCN) for classified training and prediction, and a cancer prediction framework was finally built. Figure 1 describes the overall workflow of our proposed method pDenseGCN.

Construction of patient similarity network.

In order to construct the patient similarity network (PSN), we employed a method named Similarity network fusion (SNF), which can make full use of multi-omics[36]. SNF is applied to combine RNA-seq, DNA methylation and CNV data to generated a patient similarity network. Assuming that there are n patients, each of them has m type data (such as RNA-Seq and DNA methylation). We represent the PSN as a graph G=(V,E), where V represents the set of patients $\{x_1, x_2, x_3..., x_n\}$ and the edges E correspond to the similarity between vertices $v \in V$. The weights between edges are represented by an $n \times n$ similarity matrix W which is computed by Eq. (1).

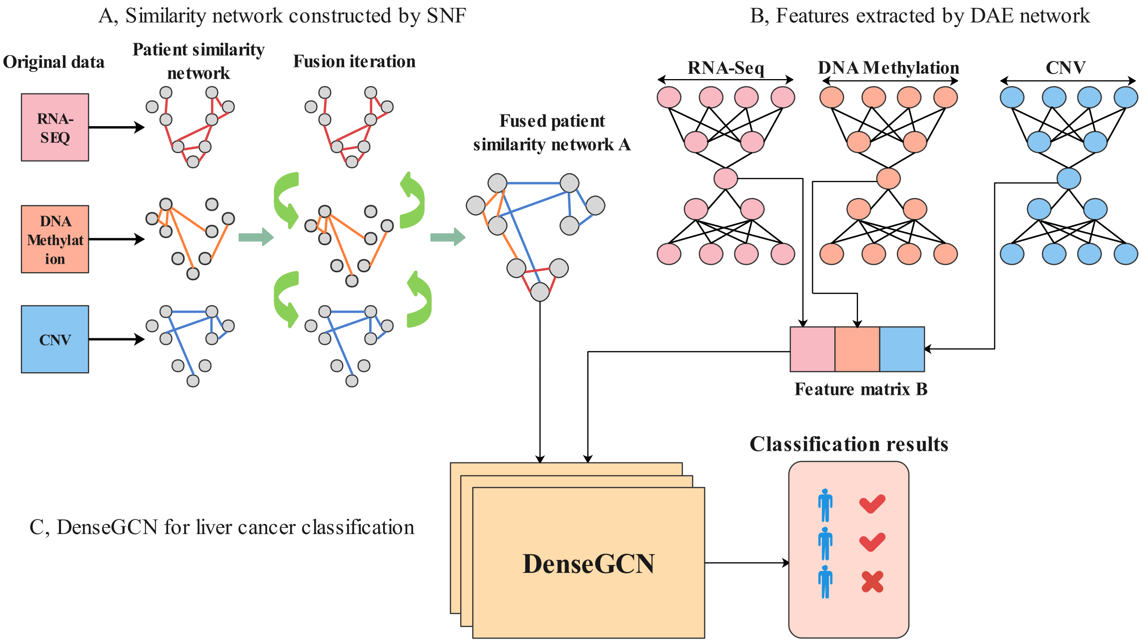
A, Similarity network constructed by SNF
B, Features extracted by DAE network
Original data
Patient similarity network
Fusion iteration
RNA-SEQ
DNA Methylation
CNV
Fused patient similarity network A
RNA-Seq
DNA Methylation
CNV
Feature matrix B
Classification results
C, DenseGCN for liver cancer classification
DenseGCN

Figure 1. The overall workflow of pDenseGCN. (**A**) Similarity network constructed by SNF. (**B**) Features extracted by DAE network. (**C**) DenseGCN for cancer diagnosis.

$$W_{i,j} = \exp(-\frac{\phi^2(x_i, x_j)}{\alpha\gamma_{i,j}}) \tag{1}$$

where α is a hyperparameter, ϕ (x_i, x_j) is the Euclidean distance between patients x_i and x_j and $\gamma_{i,j}$ is used to eliminate the scaling problem. In order to compute the fused matrix from multiple types of data, the similarity matrix is normalized as Eq. (2).

$$P_{i,j} = \begin{cases} \frac{W_{i,j}}{2\sum_{k\neq i} W_{i,k}} & j \neq i \\ \frac{1}{2} & j = i \end{cases} \tag{2}$$

Assuming N_i is a set of x_i's neighbors. Then local affinity matrix S is calculated by Eq. (3).

$$S_{i,j} = \begin{cases} \frac{W_{i,j}}{\sum_{k\in N_i} W_{j,k}} & j \in N_i \\ 0 & otherwise \end{cases} \tag{3}$$

Let $P_t{}^{(h)}$ represent normalized similarity matrix of h-th type data ($1 \le h \le m$) in the t-th iteration, $P_t{}^{(h)}$ is updated according to Eq. (4).

$$P_{t+1}^{(h)} = S^{(h)}(\frac{\sum_{k\neq h} P_t^{(k)}}{m-1})(S^{(h)})^T \tag{4}$$

where the $S^{(h)}$ represents local affinity matrix of h-th type data. Through this process of continuous iterative fusion, a patient similarity network which contains complementary information from three omics dataset is finally obtained. The fused network can be used for classification or clustering, and in this work the fused similarity network is taken as the input of DenseGCN for cancer diagnosis.

Feature extraction by denoising autoencoder. To reduce the noise in the row omics data and the computational cost, we constructed three independent denoising autoencoders to extract latent embedding representation from the omics datasets, respectively. The autoencoder (AE) is a neural network which typically contains two networks: an encoder network and a decoder network. The encoder network takes a feature vector $x \in \Re^d$ as input and encodes it into a low-dimensional representation $y \in \Re^q$, define as f_e: x → y. The decoder network maps the low-dimensional representation y back to the input space, define as f_d: y → z. The autoencoder is optimised by minimizing the reconstruction loss L between original input x and reconstructed input z as Eq. (5).

$$\arg\min_{f_e,f_d} L(x,z) \tag{5}$$

where f_e, f_d represent the parameters of the encoder network and the decoder network, respectively.

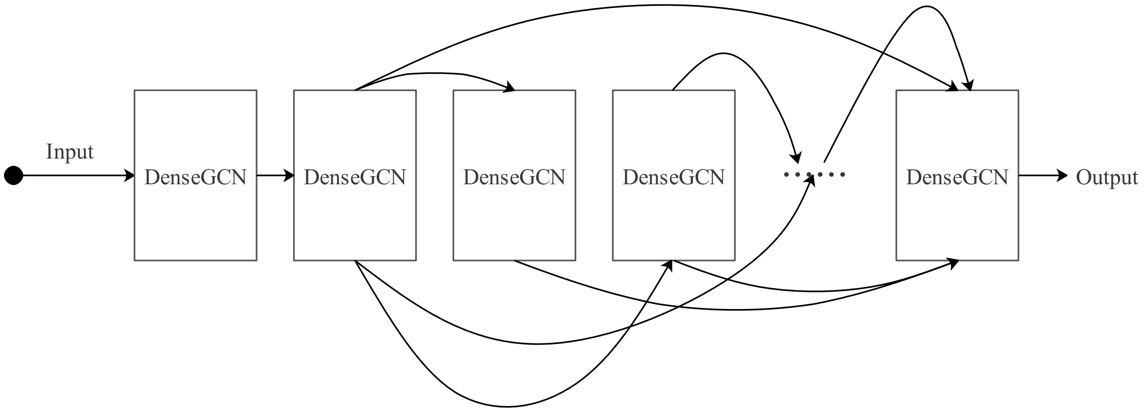

Figure 2. The structure of DenseGCN.

In this work a denoising autoencoder (DAE)[37] is applied to extract latent embedding representation. The architecture of DAE is the same as AE, but the way to train network is different. DAE first corrupted the input data by adding noise, then the corrupted input data x_noise is fed to the autoencoder. By recovering the damaged input data, DAE extracts robust latent embedding representation. We use the loss function Mean Squared Error to train DAE. The latent embedding representations extracted by three independent DAE are connected and then fed to the further work together with patient similarity network.

DenseGCN.

The patient similarity network constructed by SNF is non-Euclidean data that CNNs fail to handle[32], so GCN is considered in this work because of their advantages in processing non-Euclidean data[38]. However, original GCN model is usually very shallow due to the vanishing gradient problem, this limits the ability of GCN to fit the data[35]. So an improved GCN model named DenseGCN is used in this work.

GCN takes a feature matrix X which describes every node in the graph and an adjacency matrix A which illustrate the structure of the graph as input and generates a node-level matrix Z. The layer-wise propagation rule of GCN can be formulated as Eq. (6).

$$H(N) = f(H(N-1), A) = \sigma(AH(N-1)W(N)) \tag{6}$$

where H(N) is the output of the N layer, and W(N-1) is a weight matrix of the N-1 layer. $f(\cdot)$ represents graph convolution operation. $\sigma(\cdot)$ is an activation function which is usually non-linear. This rule is valid but still has some limitations. Frist the feature vectors of all neighboring nodes are taken into consideration, but the node itself is ignored. This limitation can be fixed by adding self-connections to the adjacency matrix A, define as $\hat{A}$ = A+E, where E represents the identity matrix. The second limitation is that A is usually not normalized, this means that the scale of the feature vectors will change when multiplying with A. To get rid of this limitation, symmetric normalization, defining as $D^{-1/2}$AD$^{-1/2}$, is applied to standardize A, where D is the diagonal node degree matrix. Thus, propagation rule is reformulated as Eq. (7).

$$H(N) = f(H(N-1)) = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H(N-1)W(N)) \tag{7}$$

Theoretically, deeper networks are able to learn more abstract representations and require less data for training than shallow neural networks[33,39], and at the same time, omics data are characterized by high dimensionality and few samples. This indicates that deep networks are more applicable to omics data. However, GCN is usually very shallow because of the vanishing gradient problem[35], and most state-of-the-art GCNs are less than 4 layers[34]. Inspired by the dense connectivity of DenseNet[40], a similar idea is adapted to GCN to improve information flow in the network and avoid gradient vanishing problem[35]. This dense model, named DenseGCN, has a new propagation rule which is define as Eq. (8).

$$H(N) = T(f(H(N-1), A), f(H(N-2), A), \ldots, f(H(0), A)) \tag{8}$$

where H(0) is the input feature matrix X, $T(\cdot)$ represents a vertex-wise concatenation function. The structure of DenseGCN is shown in Fig. 2.

In summary, the original GCN is limited by the gradient disappearance problem, which makes it difficult to have a deep network architecture. In contrast, DenseGCN improves the flow of information by connecting layers densely to solve the gradient vanishing problem, and is able to have a deeper network architecture compared with original GCN. Thus, DenseGCN is more suitable for omics datasets.

Results

A series of experiments were conducted to evaluate the performance of proposed method pDenseGCN. First, pDenseGCN was compared with five state-of-the art methods, namely ASVM[41], Xgboost-AD[42], MGRFE-GaRFE[43], ET-SVM[44], XOmiVAE[45], and four machine learning methods, namely Linear Discriminant Analysis (LDA), Naïve bayes (NB), Random Forest (RF), and Decision Tree (DT). Then we investigated the influence

Omics type	Number of samples	Number of features
RNA-Seq	424	20,530
DNA methylation	429	20,421
CNV	760	24,924

Table 1. The details of three omics datasets.

of patient similarity network and different omics data. Finally, we discussed the impact of different number of DenseGCN layers and features selected by DAE.

Datasets and data preprocessing.

We performed our proposed method pDenseGCN on Liver Hepatocellular Carcinoma (LIHC) omics datasets acquired from TCGA portal (https://www.cancer.gov/tcga). A R package named TCGA-assembler[46] was used to obtain DNA methylation, RNA-seq and CNV data of LIHC. The detail of above three datasets is described in Table 1.

Similar to the previous literature[47], these three datasets are preprocessed by following steps. The first step is outlier removal. We delete these features which have more than 20% missing values. Similarly, these sample which have moved than 20% features have been removed. 404 common samples remained in this step. The next step is missing-data imputation. We use the mean of remaining features to impute the missing values based on the python package sklearn[48]. Finally, these three datasets are normalized according to Eq. (1).

$$X_{nor} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{9}$$

where X is any column in the omics dataset, X_{nor} is the corresponding columns after normalization, X_{max} is the maximum values in X and X_{min} represent the minimum values in X.

Evaluation metrics.

To fully evaluate different methods, accuracy, precision, recall, F1-score[49], and AUC[50] were used as the metrics. All of them are defined as follows.

Accuracy: The ratio of correctly predictions. Accuracy can be calculated as Eq. (10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Precision: The ratio of samples categorized as positive to those which are actually positive. The formula of precision is Eq. (11).

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

Recall: The ratio of true positive samples divided into positive samples. It is defined as Eq. (12).

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

F1-score: The harmonic means of recall and precision. It can be calculated as Eq. (13).

$$F1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{13}$$

AUC: The area under the receiver operating characteristics curve.

Experiment and parameter settings.

For these omics dataset, 60% of the data was randomly selected to train models and 20% of the data was randomly selected as the validation set. The remaining 20% data was used for testing. To reduce the deviation, we repeated the experiments five times and the average result of the five experiments was taken as the ultimate result of the experiment. All of our models were implemented using Pytorch. The experiments were executed on a PC with an Intel core i7-10700 processor of 2.90 GHz and 32.0 GB RAM. The relevant parameters of the used methods are listed in this part. For pDenseGCN, we determined the optimal learning rate (Lr) and the batch size according to the grid search method. For the comparison algorithm, the parameters given in its original paper were slightly modified to make it more suitable for our dataset. Table 2 describes the detailed parameters.

Comparison with other methods.

To validate the performance of our proposed method pDenseGCN, we compared it with five state-of-the-art methods and four machine learning methods. We replicated them according to their publications or using publicly available programs. The details of these five state-of-the-art methods are described below.

Methods	Parameters
pDenseGCN	Lr(DAE) = 0.01, epochs(DAE)=50, batch size(DAE)=8, Lr(DenseGCN)=0.01, epoch(DenseGCN)=500
ASVM	m=4, n=8, q=5, numGlobal=30, numLocal=20
Xgboost-AE	Lr(AE)=1.0, batch size(AE)=16, epoch(AE)=100
MGRFE-GaRFE	global_bestsize = 120, layer_bestsize = 100 , total_layer = 2
ET-SVM	C=0.004, kernel='linear', decision_function_shape='ovo', gama=1
XOmiVAE	learning_rate=0.01, dropout=0.5, epoch=100
LDA	solver='svd'
NB	var_smoothing=1e-09
RF	n_estimators=10
DT	splitter='best', min_samples_split=2, min_samples_leaf=1

Table 2. Parameter settings.

	Precision	Recall	F1-Score	Accuracy	AUC
pDenseGCN	**0.9865**	**0.9865**	**0.9865**	**0.9857**	**0.9856**
ASVM	0.937	0.9744	0.9553	0.9208	0.8531
XGBoost-AD	0.9736	0.9729	0.9732	0.9726	0.9759
MGRFE-GaRFE	0.9689	0.9397	0.9183	0.954	0.8306
ET-SVM	0.96	0.6316	0.7619	0.7945	0.8015
XOmiVAE	0.946	0.8974	0.9211	0.8537	0.8718
LDA	0.7262	0.8133	0.7673	0.7466	0.7447
RF	0.9605	0.9125	0.9359	0.937	0.9848
NB	0.8977	0.7914	0.8412	0.8452	0.8492
DT	0.9254	0.8267	0.8732	0.8767	0.8781

Table 3. Results of comparison methods and proposed method. Significant values are in bold.

- ASVM[41] is a novel multilayer recursive feature elimination algorithm based on embedded variable length encoding genetic algorithm aiming at cancer classification. It utilizes the Shuffled Frog Leaping algorithm to adaptively adjust the parameters of the Support Vector Machine based on data attributes to classify early stage cancers.
- Xgboost-AD[42] is a novel cancer classification method. It integrates multi-omics data by autoencoder and utilizes extreme gradient boosting to accurately diagnostic classify cancer.
- MGRFE-GaRFE[43] is aiming to use fewer genes for better cancer classification results. It applies a multilayer recursive feature elimination method based on an embedded genetic algorithm to get a better feature subset for cancer classification.
- ET-SVM[44] adopts extra trees and variance threshold to select features from gene expression data, and uses these important features to diagnostic classify cancer based on SVM.
- XOmiVAE[45] is an interpretable deep learning model for cancer diagnosis based on variational autoencoder. It uses variational autoencoder to extract low-dimensional expressions from genomics data, which are then fed into a multilayer perceptron for cancer classification.

The results are displayed in Table 3. As seen in Table 3, pDenseGCN has a better performance compared with other methods among all the metrics in LIHC dataset. In terms of accuracy, pDenseGCN achieves 98.57% accuracy, which is 1.31% better than the best remaining method XGBoost-AD and up to 23.9% better than other comparison methods. As for the other four metrics, pDenseGCN gains a best performance which are up to 26.03%, 35.49%, 22.46%, 24.09% better than other methods in terms of precision, recall, f1-score, and AUC. It proves that by introducing the patient similarity network, our proposed method is more advantageous in cancer diagnosis and more applicable to the LIHC dataset.

The influence of patient similarity network.

Constructing patient similarity network is one important component of pDenseGCN, since the patient similarity network allows DenseGCN to gain information from the neighboring patients. To investigate the influence of patient similarity network on cancer diagnosis, we designed two experiments. One experiment took patient similarity network as the input and the other one took an identity matrix as the input. The results are presented in Fig. 3. As Fig. 1 shows, the model trained with patient similarity

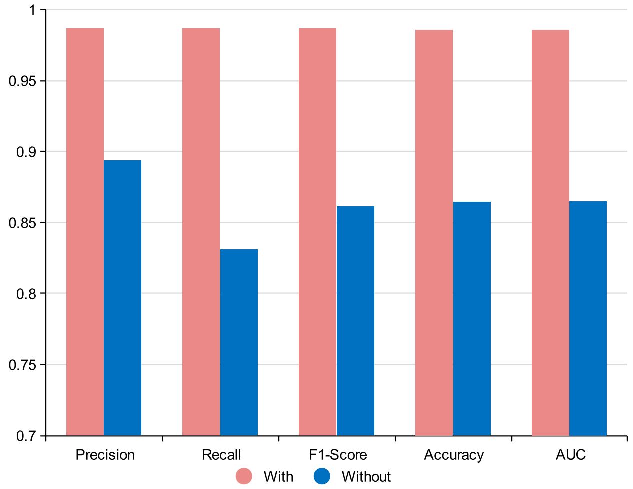

Figure 3. The influence of the patient similarity network.

	Precision	Recall	F1-score	Accuracy	AUC
RNA-Seq	0.8197	0.6757	0.7407	0.75	0.7545
DNA methylation	0.931	0.7297	0.8182	0.8286	0.8346
CNV	0.9574	0.6081	0.7438	0.7785	0.7889
RNA-Seq+DNAMethy	0.9855	0.9189	0.951	0.95	0.9519
RNASeq+CNV	0.8375	0.9853	0.9054	0.9	0.9024
DNAMethy+CNV	0.9589	0.9459	0.9524	0.95	0.9502
Multi-omics	**0.9865**	**0.9865**	**0.9865**	**0.9857**	**0.9856**

Table 4. Results of different omics data. Significant values are in bold.

network performs prior to the model trained without patient similarity network. In the case of precision, recall, F1-score, accuracy, and AUC, the model trained with patient similarity network is 9.29%, 15.6%, 12.6%, 12.2%, 12.1% higher than the model trained without patient similarity network. This demonstrates that by introducing a patient similarity network, our proposed method pDenseGCN takes the information from neighboring patients into consideration when predicting the label of a patient. This effectively improves the classification results.

Effectiveness of different omics data. We carried out experiments with varied type of data to confirm the effectiveness of varied omics data and the effect of multi-omics data combination. The results are displayed in Table 4. In Table 4, RNA-seq, DNA Methylation and CNV represent three single omics data, respectively. RNA-Seq+DNAMethy, RNASeq+CNV, and DNAMethy+CNV represent three omics data pairwise combinations, respectively. Multi-Omics represents our proposed method with three omics data. We can see from Table 4 that the performance of our proposed method rises over time as the type of data used increases. These models trained with single omics data have an accuracy of up to 0.8286, however, when the model was trained with two omics data, the lowest accuracy was 0.9. The optimal performance is attained when the model is trained with three kinds of omics data with an accuracy value of 0.9857. It confirms that multiple omics do outperform single omics, and that the performance improves progressively as the number of omics data increases. This indicate that different omics data contain complementary information, which provides a comprehensive view of cancer and improves the result of cancer diagnosis. Besides, the model trained with DNA Methylation performs better in the three single omics data, this may indicate that DNA Methylation contains more information that facilitates cancer diagnosis.

The effect of DenseGCN layer numbers. In order to explore the effect of different DenseGCN layer numbers on the final result, we designed several models with various number of layers. The results of different models are shown in Table 5. As seen in Table 5, unlike the conventional GCN models, pDenseGCN still performs well even if the number of layers is more than three. Meanwhile, excluding 7-layers and 8-layers, the performance of pDenseGCN increases gradually with the increase of layers. This illustrates that deep network is more suitable to fit omics data and can gain a better performance than shallow network in cancer diagnosis. When the number of pDenseGCN layers reaches 10, multiple metrics such as Recall, F1-Score, Accuracy and AUC perform best. However, as the number of layers keeps increasing, these scores of metrics gradually decline. This is probably because although DenseGCN overcomes gradient vanishing by densely connecting layers to

	Precision	Recall	F1-Score	Accuracy	AUC
3-layers	1	0.6892	0.816	0.8357	0.8446
4-layers	0.8024	0.8784	0.8387	0.8214	0.8179
5-layers	1	0.9324	0.965	0.9643	0.9662
6-layers	0.9125	0.9865	0.9481	0.9429	0.9402
7-layers	1	0.7297	0.8437	0.8571	0.8649
8-layers	0.9667	0.7838	0.8657	0.8714	0.8767
9-layers	1	0.9459	0.9722	0.9714	0.973
10-layers	0.9865	**0.9865**	**0.9865**	**0.9857**	**0.9856**
15-layers	0.925	1	0.961	0.9571	0.9545

Table 5. Results of different DenseGCN layer numbers. Significant values are in bold.

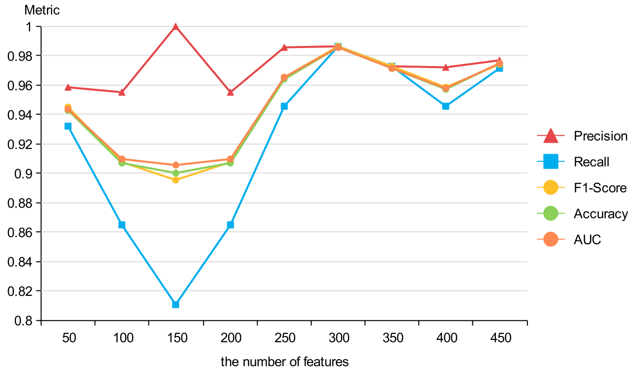

Figure 4. Results of different number of features.

some extent, the number of DenseGCN layers is not the more the better. Therefore, the number of layers of pDenseGCN in this work is set to 10.

The effect of feature numbers extracted by DAE.

To investigate the effect of feature numbers on model performance, we conducted several experiments with different number of features extracted by DAE to explore the changes in experimental results. The results are displayed in Fig. 4.

As can be seen from Fig. 4, when the quantity of features ranges from 100 to 300, three metrics F1-Score, Accuracy, AUC exhibit an obvious rising trend. This may be because that useful information that the model can learn gradually increases as the amount of features increases. The proposed method reaches best performance when the number of features is set to 300, with an accuracy value of 0.9857. As the number of features continues to grow, the performance of our proposed method begins to gradually decrease instead. This indicates that when the number of features is large, irrelevant or redundant information may be incorporated, which does harm to the performance of model. Thus, 300 is selected as the number of features in this work.

Conclusion

Liver cancer is one of the common malignant tumors worldwide with a poor prognosis. Since effective diagnosis helps to improve the cure of liver cancer, there is an urgent need for a method that can accurately perform diagnosis of liver cancer. In this work, we establish a novel method pDenseGCN which consists of similarity network fusion, denoising autoencoder, and dense graph convolutional network for liver cancer diagnosis. The pDenseGCN takes multi-omics data to construct a patient similarity network, which brings more patient information for cancer diagnosis. We explore the differences in the results of pDenseGCN trained with and without patient similarity network. The results indicate that the similarity information does contribute to cancer diagnosis. In addition, since the patient similarity network is non-Euclidean data, and the omics data is characterized by high dimensionality and few samples, pDenseGCN utilizes densely connected graph convolutional neural network to fit them better. Compared with state-of-the-art methods, pDenseGCN achieves better results in terms of the final prediction performance metrics. It demonstrates that our proposed pDenseGCN is a promising method for liver cancer diagnosis. In our future work, we are committed to extend our proposed method to multi-classification tasks, such as cancer subtype classification as well as pan-cancer classification.

Received: 14 October 2021; Accepted: 5 April 2022
Published online: 26 April 2022

References

1. Alqahtani, A. *et al.* Hepatocellular carcinoma: Molecular mechanisms and targeted therapies. *Medicina* **55**, 526. https://doi.org/10.3390/medicina55090526 (2019).
2. Sung, H. et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249. https://doi.org/10.3322/caac.21660 (2021).
3. McPhail, S., Johnson, S., Greenberg, D., Peake, M. & Rous, B. Stage at diagnosis and early mortality from cancer in england. *Br. J. Cancer* **112**, S108–S115. https://doi.org/10.1038/bjc.2015.49 (2015).
4. Arciero, C. A. & Sigurdson, E. R. Liver-directed therapies for hepatocellular carcinoma. *J. Natl. Compr. Canc. Netw.* **4**, 768–774. https://doi.org/10.6004/jnccn.2006.0067 (2006).
5. Tabrizian, P., Jibara, G., Shrager, B., Schwartz, M. & Roayaie, S. Recurrence of hepatocellular cancer after resection: Patterns, treatments, and prognosis. *Ann. Surg.* **261**, 947–955. https://doi.org/10.1097/SLA.0000000000000710 (2015).
6. Anwanwan, D., Singh, S. K., Singh, S., Saikam, V. & Singh, R. Challenges in liver cancer and possible treatment approaches. *Biochim. Biophys. Acta (BBA)-Rev. Cancer* **1873**, 188314. https://doi.org/10.1016/j.bbcan.2019.188314 (2020).
7. Zhang, C., Cai, H., Huang, J. & Song, Y. nbcnv: A multi-constrained optimization model for discovering copy number variants in single-cell sequencing data. *BMC Bioinf.* **17**, 1–10. https://doi.org/10.1186/s12859-016-1239-7 (2016).
8. Zhang, G., Hou, J., Wang, J., Yan, C. & Luo, J. Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm. *Interdiscip. Sci. Comput. Life Sci.* **12**, 288–301. https://doi.org/10.1007/s12539-020-00372-w (2020).
9. Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nat. Rev. Genet.* **14**, 333–346. https://doi.org/10.1038/nrg3433 (2013).
10. Lin, E. & Lane, H.-Y. Machine learning and systems genomics approaches for multi-omics data. *Biomark. Res.* **5**, 1–6. https://doi.org/10.1186/s40364-017-0082-y (2017).
11. Zhang, W. *et al.* Computational resources associating diseases with genotypes, phenotypes and exposures. *Brief. Bioinform.* **20**, 2098–2115. https://doi.org/10.1093/bib/bby071 (2019).
12. Mahmud, M., Kaiser, M. S., McGinnity, T. M. & Hussain, A. Deep learning in mining biological data. *Cogn. Comput.* **13**, 1–33. https://doi.org/10.1007/s12559-020-09773-x (2021).
13. Sun, Y., Goodison, S., Li, J., Liu, L. & Farmerie, W. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* **23**, 30–37. https://doi.org/10.1093/bioinformatics/btl543 (2007).
14. Akay, M. F. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst. Appl.* **36**, 3240–3247. https://doi.org/10.1016/j.eswa.2008.01.009 (2009).
15. Chen, Y.-W. & Lin, C.-J. Combining svms with various feature selection strategies. In *booktitleFeature extraction*, 315–324. https://doi.org/10.1007/978-3-540-35488-8_13 (Springer, 2006).
16. Tsai, M.-H., Chen, M.-Y., Huang, S. G., Hung, Y.-C. & Wang, H.-C. A bio-inspired computing model for ovarian carcinoma classification and oncogene detection. *Bioinformatics* **31**, 1102–1110. https://doi.org/10.1093/bioinformatics/btu782 (2015).
17. Zhang, Z.-M. *et al.* Early diagnosis of hepatocellular carcinoma using machine learning method. *Front. Bioeng. Biotechnol.* **8**, 254. https://doi.org/10.3389/fbioe.2020.00254 (2020).
18. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444. https://doi.org/10.1038/nature14539 (2015).
19. Munir, K., Elahi, H., Ayub, A., Frezza, F. & Rizzi, A. Cancer diagnosis using deep learning: A bibliographic review. *Cancers* **11**, 1235. https://doi.org/10.3390/cancers11091235 (2019).
20. Fakoor, R., Ladhak, F., Nazi, A. & Huber, M. Using deep learning to enhance cancer diagnosis and classification. In *booktitleProceedings of the international conference on machine learning*, vol. 28 (ACM, New York, USA, 2013).
21. Lyu, B. & Haque, A. Deep learning based tumor type classification using gene expression data. In *booktitleProceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 89–96. https://doi.org/10.1145/3233547.3233588 (2018).
22. Gao, F. *et al.* Deepcc: A novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* **8**, 1–12. https://doi.org/10.1038/s41389-019-0157-8 (2019).
23. Lemsara, A., Ouadfel, S. & Fröhlich, H. Pathme: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. *BMC Bioinformatics* **21**, 1–20. https://doi.org/10.1186/s12859-020-3465-2 (2020).
24. Cheerla, A. & Gevaert, O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **35**, i446–i454. https://doi.org/10.1093/bioinformatics/btz342 (2019).
25. Lin, X. *et al.* The robust classification model based on combinatorial features. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **16**, 650–657. https://doi.org/10.1109/tcbb.2017.2779512 (2017).
26. Su, R., Zhang, J., Liu, X. & Wei, L. Identification of expression signatures for non-small-cell lung carcinoma subtype classification. *Bioinformatics* **36**, 339–346. https://doi.org/10.1093/bioinformatics/btz557 (2020).
27. Sun, D., Wang, M. & Li, A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **16**, 841–850. https://doi.org/10.1109/TCBB.2018.2806438 (2018).
28. Zhang, X. et al. Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In *booktitle2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 765–769. https://doi.org/10.1109/BIBM47256.2019.8983228 (IEEE, 2019).
29. Pai, S. *et al.* netdx: Interpretable patient classification using integrated patient similarity networks. *Mol. Syst. Biol.* **15**, e8497 (2019).
30. Pai, S. & Bader, G. D. Patient similarity networks for precision medicine. *J. Mol. Biol.* **430**, 2924–2938. https://doi.org/10.1016/j.jmb.2018.05.037 (2018).
31. Henaff, M., Bruna, J. & LeCun, Y. Deep convolutional networks on graph-structured data. *arXiv preprint*. arXiv:1506.05163 (2015).
32. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *arXiv preprint*arXiv:1606.09375 *(2016).*
33. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828. https://doi.org/10.1109/TPAMI.2013.50 (2013).
34. Zhou, J. *et al.* Graph neural networks: A review of methods and applications. *AI Open* **1**, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001 (2020).
35. Li, G., Muller, M., Thabet, A. & Ghanem, B. Deepgcns: Can gcns go as deep as cnns? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9267–9276 (2019).
36. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333. https://doi.org/10.1038/nmeth.2810 (2014).
37. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, 1096–1103. https://doi.org/10.1145/1390156.1390294 (2008).
38. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint*: arXiv:1609.02907 (2016).
39. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint*arXiv:1611.03530 *(2016).*

40. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
41. Liu, L., Chen, X. & Wong, K.-C. Early cancer detection from genome-wide cell-free dna fragmentation via shuffled frog leaping algorithm and support vector machine. *Bioinformatics*https://doi.org/10.1093/bioinformatics/btab236 *(2021).*
42. Ma, B. *et al.* Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Comput. Biol. Med.* **121**, 103761. https://doi.org/10.1016/j.compbiomed.2020.103761 (2020).
43. Peng, C., Wu, X., Yuan, W., Zhang, X. & Li, Y. Mgrfe: multilayer recursive feature elimination based on an embedded genetic algorithm for cancer classification. *IEEE/ACM Trans. Comput. Biol. Bioinf.* (2019).
44. Hsu, Y.-H. & Si, D. Cancer type prediction and classification based on rna-sequencing data. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 5374–5377. https://doi.org/10.1109/EMBC.2018.8513521 (IEEE, 2018).
45. Withnell, E., Zhang, X., Sun, K. & Guo, Y. Xomivae: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Brief. Bioinf.*https://doi.org/10.1093/bib/bbab315 *(2021).*
46. Wei, L. *et al.* Tcga-assembler 2: software pipeline for retrieval and processing of tcga/cptac data. *Bioinformatics* **34**, 1615–1617. https://doi.org/10.1093/bioinformatics/btx812 (2018).
47. Ding, Z., Zu, S. & Gu, J. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics* **32**, 2891–2895. https://doi.org/10.1093/bioinformatics/btw344 (2016).
48. Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830. https://hal.inria.fr/hal-00650905 (2011).
49. Goutte, C. & Gaussier, E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, 345–359. https://doi.org/10.1007/978-3-540-31865-1_25 Springer, (2005).
50. Huang, J. & Ling, C. X. Using auc and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* **17**, 299–310. https://doi.org/10.1109/TKDE.2005.50 (2005).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 61802113, 61802114, 61972134). Science and Technology Development Plan Project of Henan Province, (No. 202102210173, No. 212102210091). China Postdoctoral Science Foundation (No. 2020M672212). Henan Province Postdoctoral Research Project Founding.

Author contributions

G.Z. and Z.P. conceived and designed the approach. Z.P. performed the experiments. C.Y. analyzed the data. C.Y. and H.L. wrote the manuscript. J.W. and J.L. supervised the whole study process and revised the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.