ORIGINAL RESEARCH
published: 25 March 2021
doi: 10.3389/fgene.2021.644378

A Novel Biomarker Identification Approach for Gastric Cancer Using Gene Expression and DNA Methylation Dataset

Ge Zhang, Zijing Xue, Chaokun Yan, Jianlin Wang* and Huimin Luo*

School of Computer and Information Engineering, Henan University, Kaifeng, China

As one type of complex disease, gastric cancer has high mortality rate, and there are few effective treatments for patients in advanced stage. With the development of biological technology, a large amount of multiple-omics data of gastric cancer are generated, which enables computational method to discover potential biomarkers of gastric cancer. That will be very important to detect gastric cancer at earlier stages and thus assist in providing timely treatment. However, most of biological data have the characteristics of high dimension and low sample size. It is hard to process directly without feature selection. Besides, only using some omic data, such as gene expression data, provides limited evidence to investigate gastric cancer associated biomarkers. In this research, gene expression data and DNA methylation data are integrated to analyze gastric cancer, and a feature selection approach is proposed to identify the possible biomarkers of gastric cancer. After the original data are pre-processed, the mutual information (MI) is applied to select some top genes. Then, fold change (FC) and T-test are adopted to identify differentially expressed genes (DEG). In particular, false discover rate (FDR) is introduced to revise p_value to further screen genes. For chosen genes, a deep neural network (DNN) model is utilized as the classifier to measure the quality of classification. The experimental results show that the approach can achieve superior performance in terms of accuracy and other metrics. Biological analysis for chosen genes further validates the effectiveness of the approach.

Keywords: gastric cancer, omics data, biomarkers, feature selection, deep neural network, machine learning

OPEN ACCESS

Edited by:
Wang Guohua,
Harbin Institute of Technology, China

Reviewed by:
Wei Lan,
Guangxi University, China
Xiulong Liu,
Tianjin University, China

****Correspondence:***
Chaokun Yan
ckyan@henu.edu.cn
Jianlin Wang
jlwang@henu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *21 December 2020*
Accepted: *16 February 2021*
Published: *25 March 2021*

Citation:
Zhang G, Xue Z, Yan C, Wang J and Luo H (2021) A Novel Biomarker Identification Approach for Gastric Cancer Using Gene Expression and DNA Methylation Dataset. Front. Genet. 12:644378. doi: 10.3389/fgene.2021.644378

1. INTRODUCTION

Gastric cancer is one of the most common malignant tumors of the digestive system (Nogueira et al., 2017). The pathogenesis is mainly relevant to helicobacter pylori infection, diet, environment, and genetic factors. It remains one of the most deadly cancers worldwide, especially among older males (Siegel et al., 2020). Generally speaking, early detection of cancer is crucial for increasing the chances for successful treatment and prolonging the patient's life. The 5-year survival rate of early-stage gastric cancer can reach more than 95% (Song et al., 2017). However, the early stage of gastric cancer is hard to monitor because of rare symptoms and some potential patients' cancer may be advanced when they are first diagnosed. Therefore, early targeting and treatment are very important in clinical practice of gastric cancer (Wang et al., 2020). In recent years, with the development of sequencing technology, the genome data of cancer patients can be obtained easily. These genomic data have been used to study the association between genetic changes and diseases and contribute to diagnosis and prognosis. However, these data always have the characteristics of high dimensions and low sample size (HDLSS) (Han et al., 2019). It is hard to process these data directly (Yan et al., 2018). Therefore, feature selection technology is usually adopted to assist in analyzing the possible cancer-causing genes, also called biomarkers, from massive cancer data. The biomarkers can facilitate us to understand the pathogenesis of diseases at a detailed molecular level and play an auxiliary role in clinical diagnosis.

Till now, many researchers have applied the feature selection methods to the field of gene expression data analysis (Ding and Peng, 2005; Lu et al., 2017; Zhao et al., 2020). However, it is incomprehensive to analyze cancer only using gene expression data. The rapid accumulation of omics data can provide disparate, partially independent, and complementary information about the entire genome (Zhang et al., 2016). The multi-omic data can lay an important foundation for mining informative biomarkers for cancer (Ruffalo et al., 2015). Among these omics data, DNA methylation is an important epigenetic event that affects gene expression during the development in various diseases such as cancer (Bird, 1986; Wang et al., 2018). In general, DNA methylation status is more reliable than gene expression (Paziewska et al., 2014). The combination of DNA methylation data and gene expression data is more beneficial to explain the pathogenesis of gastric cancer. Therefore, these two kinds of data are utilized to identify the biomarkers of gastric cancer in our study.

In this paper, we propose a novel gastric cancer biomarker identification approach, referred to GCBMI, to discover the possible biomarkers of gastric cancer. First, the gene expression data and DNA methylation data of gastric cancer are collected and processed. Then, fold change, statistical test, and mutual information are utilized to identify the differentially expressed genes of gastric cancer and the selected genes can serve as guidelines to reduce the dimension of omics data. At last, the DNN model is adopted as the classifier to measure the quality of classification. Experimental results indicate that GCBMI can obtain more favorable performance than other state-of-art methods.

The main contributions of this study are summarized as follows:

- For gastric cancer, a novel feature selection approach is proposed to identify the potential biomarkers. Here, DNA methylation data is integrated with the gene expression data effectively to obtain a comprehensive analysis to discover the relationship between gastric cancer and potential biomarkers.
- Besides *T*-test and FC, mutual information is introduced as a preliminary screening method to filter out redundant genes and FDR is adopted to revise *p*_value to further screen genes.
- The experimental results suggest that our approach can achieve improvement in different evaluation indicators than other state-of-art methods. In addition to evaluating accuracy, GO analysis, heatmap, and literature review are executed. The above biological validation is able to demonstrate that the genes selected by our approach are associated with gastric cancer.

The remainder of this paper is organized as follows: In section 2, we review related works of feature selection methods. The proposed approach is introduced in section 3. section 4 introduces the experimental design. Experimental results and biological analysis are described in section 5. Finally, we summarize the paper and make a vision for the future in section 6.

2. RELATED WORK

With the development of sequencing technology, massive amounts of cancer genome data have been accumulated at an accelerated speed. A number of feature selection methods have been extensively applied to cancer data. Traditional feature selection methods can be divided into two categories: filter methods and wrapper methods. Among them, the filter method has the advantage of low time consumption. So far, some filter methods had been well-applied to gene expression data.

Principal Component Analysis (PCA) is an effective dimensionality reduction method (Wold et al., 1987). Ding et al. combined feature extraction with feature selection in gene expression data (Ding et al., 2009). The relief was utilized to feature selection, and PCA was used to extract features. Then, they used the support vector machines (SVM) for classification. Experimental results illustrated that their method is effective to reduce the classification error rate in eight cancer datasets. But such methods cannot guarantee that the features still remain the corresponding biological significance. For example, the dimensionality reduction of features by PCA is equivalent to mapping the new features on the original features, and the features obtained after PCA are different from the original genes (Shen and Huang, 2008). Thus, it is often difficult to interpret the results.

Hsu et al. used extremely randomized trees (ET) to calculate the weight of the features (Hsu and Si, 2018). Feature selection was achieved by selecting features with high weight. Then, the linear SVM was combined to achieve about 95% accuracy on TCGA datasets. Lee et al. developed a novel filter method to identify the biomarkers of lung cancer and confirmed seven possible biomarkers (Lee et al., 2011).

In addition to filter methods, the wrapper methods utilize classification accuracy as a measurement standard for evaluation and find the optimal feature subset by iteration of meta-heuristic algorithms (Rodrigues et al., 2014). A lot of meta-heuristic algorithms had been well-applied to wrapper methods for feature selection of cancer such as bat algorithm (BA), recursive memetic algorithm (RMA), binary krill herd algorithm (MBKH), and so on (Dashtban et al., 2018; Ghosh et al., 2019; Zhang et al., 2020).

Dashtban et al. proposed MOBBA-LS which utilized fisher criterion and BA (Dashtban et al., 2018). They tested their method on three microarray cancer datasets. The accuracy achieved 100, 97, and 100% on leukemia, prostate, and SRBCT datasets, respectively. Ghosh et al. developed a recursive memetic algorithm (RMA) model for feature selection (Ghosh et al., 2019), and Zhang et al. proposed a pre-screening method of feature ranking, IG-MBKH, which is based on information gain (IG) and an improved binary krill herd (MBKH) (Zhang et al., 2020). The above methods can obtain favorable classification accuracy on microarray data of cancer.

Multiple-omics data can enable to provide a more comprehensive analysis of the entire genome. Among them, DNA methylation is one of the important epigenetic regulatory mechanisms (Luo et al., 2020). Especially, it is considered as a molecular factor that controls and regulates gene expression levels near the CpG sites. Its status is closely associated with diverse diseases and is generally more stable than gene expression (Ding et al., 2019). Therefore, the function of DNA methylation data was widely recognized. Increasing feature selection methods, which are based on gene expression data and DNA methylation data, were proposed.

For Alzheimer's disease, Park et al. proposed a biomarker prediction model, which integrated multi-omic data (Park et al., 2020). They used the Limma package to select possible biomarkers. Experimental results showed that their method can achieve better accuracy than using single data, and some chosen genes were reported in AlzGene database.

Mallik et al. proposed a method to identify biomarkers of cancer based on omics data (Mallik et al., 2017). The maximal relevance and minimal redundancy (mRMR) and parameter test like T-test were used to select the genes. The results suggested that their method had stable performance on different classifiers and classification accuracy can achieve about 95 and 90% in gene expression data and DNA methylation data, respectively.

Wang et al. proposed a feature selection method based on gene expression data and DNA methylation data of the six types of cancer (Wang et al., 2020). Their method can be divided into three steps. First, the correlation between gene expression profile and methylation profile of each gene was calculated to screen genes initially. Then, the genes were further filtered by T-test and FDR value. Finally, the genes selected in first two steps are filtered by Elastic Net. Finally, support vector machine was utilized as the classifier. The accuracy can be as high as 98% for the training set and 97% for the independent test set.

3. THE PROPOSED APPROACH

In this section, the proposed approach GCBMI is introduced. The overall workflow of GCBMI is shown in **Figure 1**. GCBMI consists of three stages: data pre-processing, selection of DEG and data combination, and using deep neural network as the classifier.

3.1. Data Pre-processing

In this section, we regularize the gene expression data, and then merge the individual gene expression data files. In addition, on the basis of annotation file of the gene chip, the column (feature) name of each sample is converted to the gene name, and the label column is added. In the annotation file of the gene chip, the gene name corresponding to each probe is stored. If a gene corresponds to multiple probes, we take the median of expression value as new expression value of the gene. After that, the genes with null values are further removed. In order to eliminate the influence of outliers, the dataset is standardized by z-score according to the following formula (Zhang et al., 2014). Finally, the datasets are divided into training set and test set in our experiment.

$$x' = \frac{x - \overline{x}}{\sigma} \tag{1}$$

where x and $x^{'}$ represent a column of data before and after standardization. $\overline{x}$ and σ represent the mean and standard deviation of a column of data in training set.

Likewise, DNA methylation data are also processed accordingly to eliminate the influence of outliers.

3.2. Selection of Differentially Expressed Genes and Data Combination

In this section, how to identify DEG in our approach is introduced. For gene expression data, the characteristics of high dimension and low sample size make it hard to construct a prediction model directly and may lead to the over-fitting (Ma and Zhang, 2019). For this issue, an appropriate method is required to reduce the size of feature space and the risk of over-fitting.

In GCBMI, the DEG and the differentially methylated positions (DMP) are utilized to train the model. The overall process contains three steps as follows.

First, MI (Liu H. et al., 2009) is applied to select TopN genes for gene expression data and DNA methylation data, respectively. It is a classic filter method of feature selection, which has been successfully applied to many feature selection problems (Peng and Fan, 2017). In order to avoid redundancy, the MI is adopted to filter out irrelevant genes. N is set to 3,000 through the subsequent experiments.

Second, FC and T-test are adopted to do identify DEG and DMP. What is more, the FDR is applied to revise the p_value. Taking DEG as an example, FC value for each selected genes in the first step is calculated. Since the data obey the normally distributed by Z-score standardization. Parametric statistics like T-test can work well on this kind of data. Then, Levene-test (Ankarali et al., 2009) is applied to verify whether the samples with variance homogeneity or not. If they have variance homogeneity, performing the standard T-test (Gauvreau and Pagano, 1993) to calculate p_value. Otherwise, the Welch's T-test (Algina et al., 1994) is executed to calculate the p_value. After that, the FC value and significant p_value for each gene are obtained. Finally, FDR is utilized to revise p_value to further screen candidate genes. A suitable threshold for FC value, p_value, and FDR are set to filter genes. And then we can obtain DEG. Similarly, DMP can be obtained. As shown in **Figure 1**, in gene expression data, the $|FC| \ > 2.1$ and p_value $\ < 0.05$.The $|FC| \ > 1.8$ and $p \ < 0.05$ in DNA methylation data. The FDR threshold value of both experimental datasets is set as 0.01. A hypothesis is made that if the gene is differentially expressed and occur hypermethylated and hypomethylated in different samples. This gene may have a potential relationship with gastric cancer. So the overlapping genes in DEG and DMP are the possible biomarkers of gastric cancer.

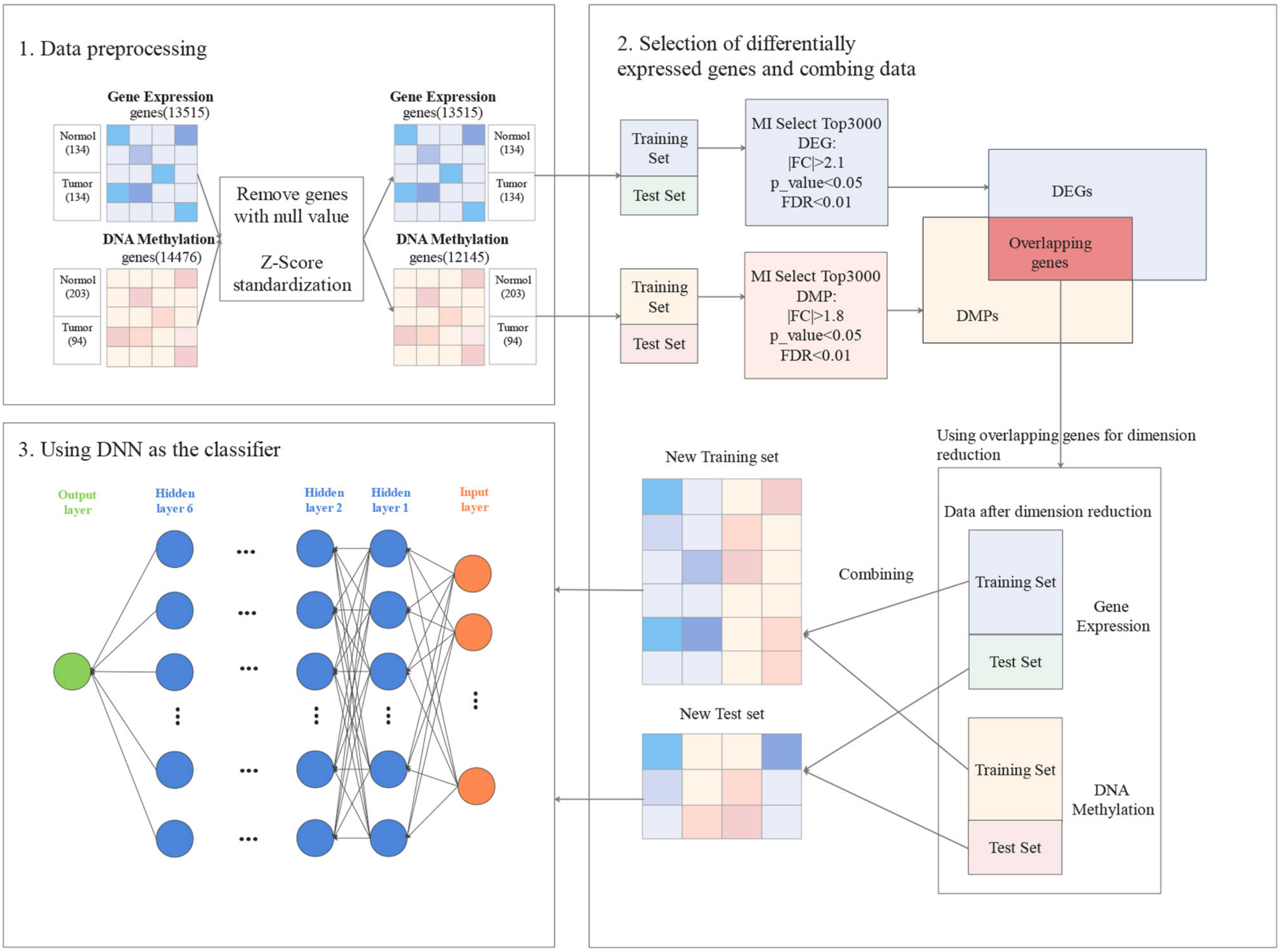

FIGURE 1 | The workflow of gastric cancer biomarker identification approach (GCBMI).

Finally, in order to extend training samples, all possible pairs of gene expression data and DNA methylation data for tumor and normal samples are utilized to merge into a new dataset. As shown in **Figure 2**, Cartesian product (Emelyanov and Ponomaryov, 2017) is performed on the gene expression data and DNA methylation data. The gene expression data and methylation data that labeled as tumor are combined into new tumor samples, and which labeled as normal are combined into new normal samples. In this way, the gene expression matrix and DNA methylation matrix are combined into a new expression matrix. This matrix has a large sample size. For example, in one of the cross-validation, the training set of gene expression data has 214 samples, which contains 112 tumor samples and 102 normal samples. DNA methylation data have 237 samples, which contains 160 tumor samples and 77 normal samples. After the combination, we will obtain 17,920 tumor samples and 7,854 normal samples. Taking them as new tumor samples and normal samples, so the new training set contains 25,774 samples, including 17,920 tumor samples and 7,854 normal samples.

3.3. Using Deep Neural Network as the Classifier

DNN model has excellent classification performance compared with traditional classifiers in previous studies, such as (Chen et al., 2020; Singh and Yamada, 2020). Here, the DNN also adopted as the classifier and the parameters of the DNN are determined through experiments.

In this section, the structure of the network is introduced. Our DNN model consists of three parts: input layer, hidden layer, and output layer. The input layer consists of two parts, corresponding to gene expression data and DNA methylation data, respectively. Then we add six hidden layers that applied ReLU as the activation function. Each layer contains 100 nodes and a additional bias nodes. The dropout is added for each hidden layer to avoid overfitting, which refers to drop some neurons randomly according to a certain probability during the learning iteration. It is equivalent to train a sparser network than the original network. Each of iterations is training a different network model to prevent overfitting. Finally, since our data only have two categories, the output layer with one node is sufficient. Sigmoid

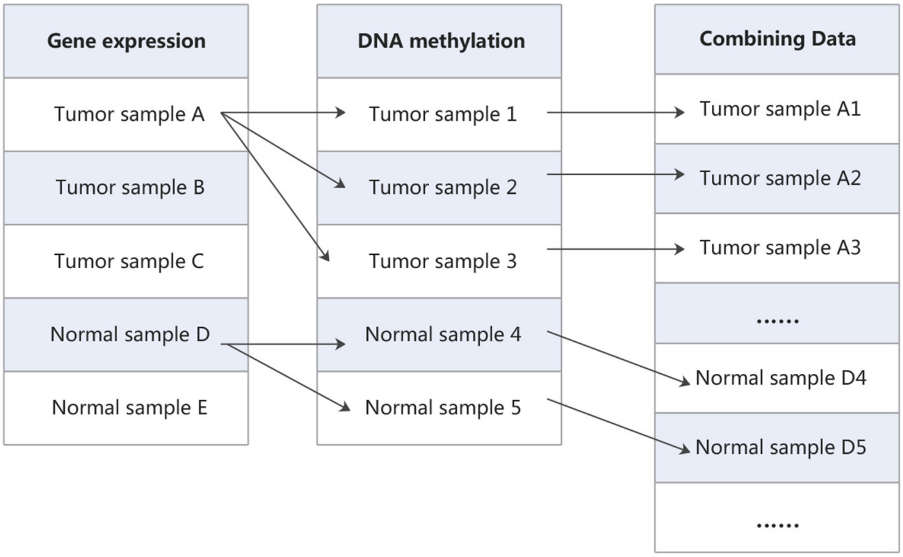

FIGURE 2 | The process of combining data.

function is adopted as the activation function of the output layer to make the output value between 0 and 1.

In the DNN model, the loss function is binary cross entropy and cost function is the reduced average value of cross entropy. Adam algorithm is applied to optimize the parameters of the network model. The formula of the loss function and cost function are as follows:

$$L(\hat{y}, y) = -ylog(\hat{y}) - (1-y)log(1-\hat{y}) \tag{2}$$

$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m}(-y^i log(\hat{y}^i) - (1-y^i)log(1-\hat{y}^i) \tag{3}$$

where y and $\hat{y}$ represent the true value and the predicted value of a sample. $\hat{y}$ is the result of sigmoid regression. m is the total number of samples and i represents the index of the sample. w and b represent weights and biases, respectively.

4. EXPERIMENTAL SETTING

The experiments can be divided into two parts. First, we compare GCBMI with other state-of-art methods. The ET (Hsu and Si, 2018), Elastic Net (Wang et al., 2020), IG-MBKH (Zhang et al., 2020), and MOBAA-LS (Dashtban et al., 2018) are selected as the baselines. A detailed description of the comparison methods is as follows:

- ET was proposed by Hsu et al. They used ET to calculate the weight of the features and select features with high weight. SVM was combined to evaluate the feature subsets. This method achieved about 95% accuracy on TCGA datasets.
- Elastic Net was a novel method that integrates the Pearson correlation coefficient, T-test, and FDR. The data are based on gene expression data and DNA methylation data. In six types of omics-data, the accuracy can up to about 98% by combing with SVM.
- IG-MBKH was presented and applied to feature selection for high-dimensional datasets. This method combined IG and krill herd algorithm and they used K-Nearest Neighbor (KNN) classifier to evaluate the classification accuracy. The accuracy of classification on nine different cancer datasets was more than 90%.
- MOBAA-LS is based on fisher criterion and BA. The accuracy achieved 100, 97, and 100% on leukemia, prostate, and SRBCT datasets, respectively.

Second, we investigate the prediction performance of DNN in biomarker identification for gastric cancer and how our method using different classifiers can affect the classification accuracy. We undertake experiments to compare our method using DNN classifier compared with using the traditional classifiers, such as KNN (Tahir et al., 2007), SVM (Vieira et al., 2013), and Naive Bayesian (NB) (Bielza and Larrañaga, 2014).

4.1. Dataset

We select the GEO database, which is an authoritative database of cancer applied in many previous studies (Zouridis et al., 2012; Wang et al., 2013) as the benchmark database. And the gene expression data GSE29272 (Li et al., 2014) and DNA methylation data GSE30601 (Lei et al., 2013; Kurashige et al., 2016) of gastric cancer are downloaded to construct our experiment dataset. As shown in **Table 1**, there are 268 samples of gene expression data including 134 tumor samples, 134 normal samples, and 13,515 features. And DNA methylation data contains 203 tumor samples, 94 normal samples, and 14,476 features.

4.2. Parameter Setting

The experiments are conducted on Intel Dual Core CPU, 8 GB RAM, Windows 7 operating system. The procedure

TABLE 1 | Benchmark dataset.

Dataset	Gene expression	DNA methylation
GEO ID	GSE29272	GSE30601
Normal samples	134	203
Tumor samples	134	94
Features	13515	14476

TABLE 2 | Parameter setting.

Methods	Parameter setting
GCBMI	MI: n = 3,000; Gene expression: $\|FC\| > 2$, $p < 0.05$, FDR < 0.01; DNA methylation: $\|FC\| > 1.8$, $p < 0.05$, FDR < 0.01
ET	Default parameters
IG-MBKH	N = 20; Iterations = 400; TopM = 80; Nmax = 4; Vf = 0.02; Dmax = 0.005
Elastic Net	$p < 0.05$, FDR < 0.01, ElasticNetCV (cv = 10)
MOBBA-LS	opN = 500, Population = 20, iteration = 300, alpha = 0.9, sigma = 0.7, injRate = 0.01, extRate = 0.01

is implemented under the programming environment Python version 3.6. The feature selection algorithms, statistical detection methods, and classifiers are provided by the Scikit-learn package and scipy package and the DNN is built by Keras package. Related parameters are given as follows: DNN is set as described in the Section 3.3; SVM: degree = 3, gamma = auto, kernel = "rbf," cache_size = 200; KNN: K = 5. The parameters of methods are set according to the original literature (Dashtban et al., 2018; Hsu and Si, 2018; Wang et al., 2020; Zhang et al., 2020). The specific settings are shown in **Table 2**.

According to Park et al. (2020), all experiments use five-fold cross validation. The dataset is divided into five parts, and one part is taken as the test set in order and the rest parts are taken as the training set in each cross validation. After the Cartesian product is executed, there are average 8,053 normal samples, 17,400 tumor samples as training set, and 496 normal samples, 1,079 tumor samples as test set. The accuracy, precision, recall, F1-score and Area Under Curve (AUC) are utilized to evaluate the classification results of the model (Tanzi et al., 2020). These evaluation indicators are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$Prediction = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - Score = \frac{2 \cdot Prediction \cdot Recall}{Prediction + Recall} \tag{7}$$

The positive samples are tumor samples and the negative samples are normal samples. True positive (TP) indicates the number of tumor samples that have been correctly classified, false positive (FP) indicates the number of normal samples which are misclassified as tumor samples, true negative (TN) indicates the number of correctly classified normal samples, and false negative (FN) indicates the number of tumor samples, which are misclassified as normal samples.

TABLE 3 | Performance comparison on different metrics (the accuracy, precision, recall, F1-score, and AUC value are average).

Methods	Accuracy	Precision	Recall	F1-score	AUC
GCBMI + DNN	**0.9870**	**0.9971**	0.9836	**0.9903**	**0.9891**
ET + SVM	0.9259	0.8571	**1.0**	0.9230	0.9333
Elastic Net + SVM	0.8922	0.9003	0.9433	0.9210	0.8598
IG-MBKH + KNN	0.9518	0.9730	0.9166	0.9437	0.9483
MOBBA-LS + SVM	0.94	0.9477	0.9327	0.9401	0.9412

The bold values represent the highest value of each metrics.

5. RESULTS AND DISCUSSION

5.1. Comparison of Other State-of-Art Methods

In this section, GCBMI is compared with other state-of-art methods, and the experimental results are shown in **Table 3**. The accuracy of GCMBI achieved is 98.7%. The Elastic net also applies omics data, but the accuracy of GCBMI is 9% higher than the Elastic net. The performance of two wrapper methods IG-MBKH and MOBBA-LS are similar in our experiment. In terms of accuracy, these two methods are about 5% lower than our approach. The accuracy of extremely randomized trees achieved is 93%. What is more, in terms of precision and recall, GCBMI also has the highest precision and the second highest recall. This indicates FP and FN appear less frequently and the classification performance of GCBMI is superior to other state-of-art methods.

F1-score and AUC value are often applied to evaluate the stability and robustness of models. The two indicators of GCBMI can achieve about 99%. It is 5–7% higher than other state-of-art methods. In order to display the advantages of our method more intuitively, the histogram of experimental results is plotted in **Figure 3**.

Overall, GCBMI can get better performance on different evaluation indicators than other feature selection methods, which indicates that the genes identified by GCBMI have more sufficient capacity to classify gastric cancer. The high F1-score and AUC value also illustrate that our model has better stability. The experimental results suggest that combined omics data are meaningful, and it may reveal some causal relationships between different biological layers.

5.2. The Impact of Classifiers on Performance

In this section, the impact of different classifiers is evaluated on our feature selection method. **Table 4** displays the experimental results, which indicates that DNN model compared with the other three classifiers has better performance in different evaluation indicators. The performance of KNN is similar to

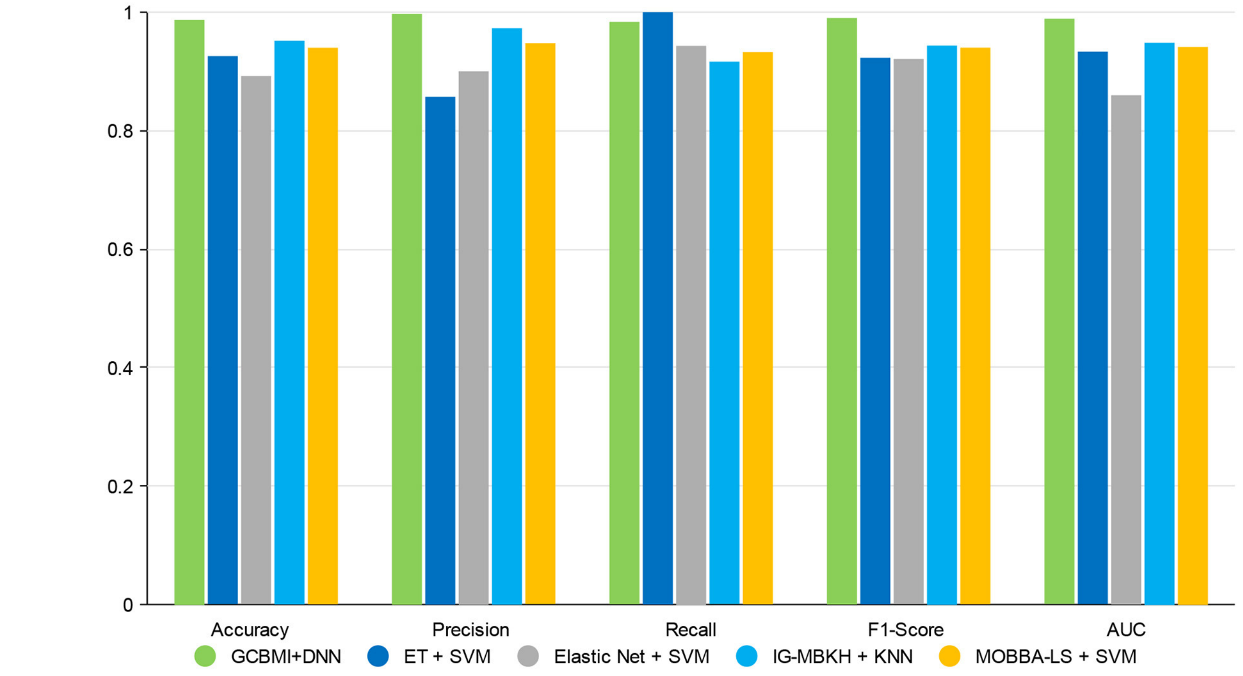

FIGURE 3 | The experimental results of gastric cancer biomarker identification approach (GCBMI) compared with other methods.

TABLE 4 | Results with different classifiers (the accuracy, precision, recall, F1-score, and AUC value are average).

Classifiers	Accuracy	Precision	Recall	F1-score	AUC
DNN	**0.9870**	**0.9971**	**0.9836**	**0.9903**	**0.9891**
KNN	0.9776	0.9934	0.9729	0.9830	0.9795
SVM	0.9819	0.9878	0.9826	0.9862	0.9803
NB	0.9651	0.9698	0.9777	0.9737	0.9557

The bold values represent the highest value of each metrics.

SVM and NB is worst but still reaches 96%. The performance of our method is stable in different classifiers. GCBMI integrates gene expression data and DNA methylation data and expands the number of samples. In this way, the DNN model can be trained better and achieves superior results than other classifiers.

On the whole, when compared with the KNN, SVM, and NB, our deep neural network model has better performance in different metrics, which indicates the validity of our feature selection approach. All the experimental results indicate that DNN model is a more appropriate classifier to feature selection in our approach. **Figure 4** shows the histogram of the average accuracy, F1 score, and AUC value of GCBMI with different classifiers, respectively. The classification advantage of DNN model has been shown in it, which has demonstrated the effectiveness of GCBMI.

5.3. Gene Analysis

In our experiment, the overlapped genes are recorded, which are shown in **Table 5**. In each fold of cross-validation, about 20 genes are selected. These genes are the intersections of DEG and DMP. Among them, eight genes appear in each intersection and they are thought to be biomarkers of gastric cancer. In this section, the selected genes are further analyzed to understand the biological relevance.

Through literature retrieving, we can find the coding protein of PGC is a digestive enzyme produced by the stomach and it is the main component of the gastric mucosa. Polymorphism of this gene is associated with gastric cancer susceptibility. Serum levels of this enzyme are used as the biomarker for certain stomach diseases, including *Helicobacter pylori* associated gastritis (Sun et al., 2009). Moreover, Liu et al. discovered PGC was positively expressed in normal gastric mucosa (100%), and the expression rate was 6.45% in gastric cancer (Liu D. et al., 2009). The results suggested that PGC has important application value in the diagnosis of gastric cancer.

For gene PSCA, relevant research demonstrated that proteins encoded by PSCA play an important role in cell proliferation. In addition to being highly expressed in the prostate, it is also expressed in differentiating gastric epithelial cells. This gene includes a polymorphism that results in an upstream start codon in some individuals; this polymorphism is thought to be associated with a risk for gastric cancers (Bahrenberg et al., 2000; Sakamoto et al., 2008).

Except for PGC and PSCA, gene PDGFD as a member of PDGF family (Huang et al., 2014), its signaling pathway has been considered as a new target for the treatment of gastric cancer (Wang et al., 2009). Besides, gene KCNE2 is expressed mainly in the cytoplasm of parietal cells. Kuwahara et al. discovered that the loss of KCNE2 expression could cause gastric adenocancer (Kuwahara et al., 2013).

For these eight genes identified, in order to observe their expression level, gene expression heatmap is constructed. As

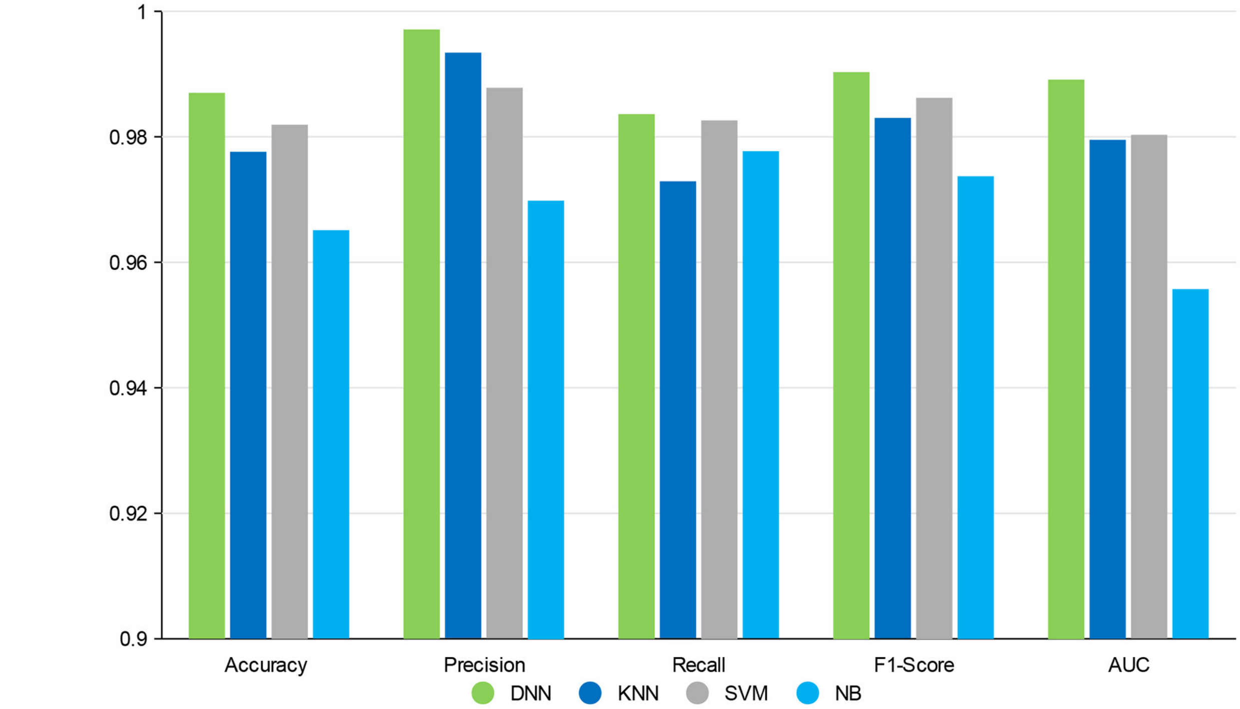

FIGURE 4 | The experimental results of gastric cancer biomarker identification approach (GCBMI) with different classifiers.

TABLE 5 | Selected genes from integrating gene expression and DNA methylation dataset.

K-fold	Number of overlapping genes	Selected genes
K = 1	17	FAHD2A,PGC,FIGF,PPAP2B,FOXA1,IFITM2,HOXC10, GPRC5C,CLEC3B,FBN1,LIF,C5,PSCA,PDGFD,KCNE2, RORC,C3
K = 2	19	PGC,FIGF,NID2,PPAP2B,IFITM2,RAB31,RORC,GPRC5C,FSCN1,TEAD4,CLEC3B,RAB17,IGFALS,C5,PSCA,PDGFD,KCNE2,COL4A1,C3
K = 3	17	FAHD2A,PGC,PPAP2B,FOXA1,IFITM2,IGFALS,GPRC5C, TEAD4,DNM1,ORM1,PTPRN2,FBN1,PSCA,PDGFD, KCNE2,RORC,C3
K = 4	24	PGC,FIGF,PDGFRB,PSMA7,TEAD4,C5,RORC,ADA, IFITM1,FAHD2A,PPAP2B,IGFALS,SLC1A2,GPRC5C, CLEC3B,CAPN9,KCNE2,PSCA,IFITM2,FSCN1,RPRM, PDGFD,SERPINA4,FBN1
K = 5	17	IFITM1,PGC,FIGF,PPAP2B,KCNE2,IFITM2,HOXC10, GPRC5C,CAPN9,FBN1,HRAS,C5,PSCA,PDGFD, SERPINA4,RORC,C3
Overlapped genes in 5-CV	8	PGC,RORC,GPRC5C,PDGFD,KCNE2,PSCA,IFITM2, PPAP2B

shown in **Figure 5**, the expression levels of these eight genes in all samples are demonstrated. The first half of the heatmap are normal samples, and others are tumor samples. Basically, the result shows that these genes have different expression in normal and tumor samples. Some of these genes differed significantly between the two classes and may have some relationship with gastric cancer.

What is more, the enrichment analysis is conducted by DAVID database for selected genes. As shown in **Table 6**, biological significance of the genes are reported through Gene Ontology (GO). "GO:0008284 positive regulation of cell proliferation," "GO:0046597 negative regulation of viral entry into host cell," "GO:0030335 positive regulation of cell migration" are common biological activities in human cancer (Dyrskjøt et al., 2009). Among them, there have some items about platelet, some studies have suggested that gastric cancer may lead to changes in platelet count and morphology (Matowicka-Karna et al., 2013). In addition, some studies also have been pointed out that interferon (Ferrantini et al., 2007) and other related factors may have relationship with the occurrence of cancer.

6. CONCLUSION

In this work, we propose a novel feature selection approach, GCBMI, which uses gene expression and DNA methylation data for identifying the biomarkers of gastric cancer. GCBMI consists of three main parts, namely data pre-processing, selection of differentially expressed genes and data combination, and

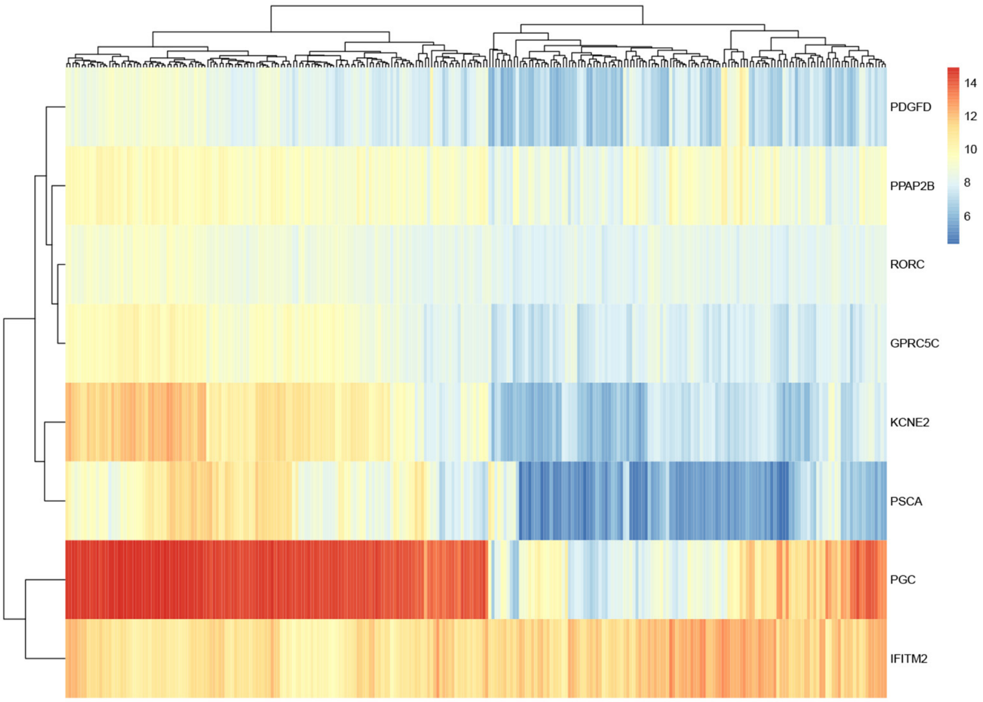

FIGURE 5 | Heatmap of eight overlapped genes.

TABLE 6 | GO analysis of selected genes.

Category	Term	*p*-value	Gene
GOTERM_BP_DIRECT	GO:0071560 cellular response to transforming growth factor beta stimulus	0.003912643	CLEC3B,FBN1, PDGFD
GOTERM_BP_DIRECT	GO:0043406 positive regulation of MAP kinase activity	0.005625548	HRAS,PDGFRB, PDGFD
GOTERM_BP_DIRECT	GO:0008284 positive regulation of cell proliferation	0.01138237	LIF,HOXC10,HRAS, PDGFRB,PDGFD
GOTERM_BP_DIRECT	GO:0002576 platelet degranulation	0.016395992	ORM1,CLEC3B, SERPINA4
GOTERM_BP_DIRECT	GO:0035456 response to interferon-beta	0.017024892	IFITM1,IFITM2
GOTERM_BP_DIRECT	GO:0035455 response to interferon-alpha	0.018899122	IFITM1,IFITM2
GOTERM_MF_DIRECT	GO:0048407 platelet-derived growth factor binding	0.020021643	COL4A1,PDGFRB
GOTERM_MF_DIRECT	GO:0005102 receptor binding	0.026443684	LIF,C3,C5,PDGFRB
GOTERM_MF_DIRECT	GO:0005161 platelet-derived growth factor receptor binding	0.02720561	PDGFRB,PDGFD
GOTERM_BP_DIRECT	GO:0036120 cellular response to platelet-derived growth factor stimulus	0.033768846	PDGFRB,PDGFD
GOTERM_BP_DIRECT	GO:0046597 negative regulation of viral entry into host cell	0.033768846	IFITM1,IFITM2
GOTERM_BP_DIRECT	GO:0030335 positive regulation of cell migration	0.047784333	HRAS,PDGFRB, PDGFD
GOTERM_BP_DIRECT	GO:0048008 platelet-derived growth factor receptor signaling pathway	0.053858697	PDGFRB, PDGFD

deep neural network as the classifier. Differential expression analysis, statistical test, and MI are integrated to obtain comprehensive view to implement the biomarkers identification after data pre-processing. MI is introduced to filter out irrelevant gene, and FC and *T*-test are utilized to select differentially expressed genes. In particular, FDR is applied to revise the *p*_value to further screen genes. After that, Cartesian product is performed to expand samples. Moreover, GCBMI adopts DNN as the classifier to evaluate the classification ability of selected genes. Experimental results on GEO dataset indicate that the proposed approach outperforms other state-of-the-art feature methods. The results of biological relevant verification indicate the status of the selected gene as the biomarkers of gastric cancer.

What is more, the performance of combined with omics data tends to be more superior than using a single omics data alone. In the future, some other omics data will be combined such as copy number variation (CNV) data to identify cancer biomarkers, and our methods will be applied to other fields as well (Liu et al., 2020). Besides, some measures will also be taken to improve our method so that its classification performance can be improved further.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CY and ZX conceived and designed the approach. ZX performed the experiments. HL analyzed the data. GZ and ZX wrote the manuscript. CY and JW supervised the whole study process and revised the manuscript. All authors have read and approved the final version of manuscript.

FUNDING

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

ACKNOWLEDGMENTS

This paper is recommended by the 5th Computational Bioinformatics Conference.

REFERENCES

Algina, J., Oshima, T., and Lin, W.-Y. (1994). Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. *J. Educ. Stat.* 19, 275–291. doi: 10.3102/10769986019003275

Ankarali, H., Yazici, A. C., and Ankarali, S. (2009). A bootstrap confidence interval for skewness and kurtosis and properties of t-test in small samples from normal distribution. *Med. J. Trakya Univ.* 26, 297–305. doi: 10.1620/tjem.219.337

Bahrenberg, G., Brauers, A., Joost, H.-G., and Jakse, G. (2000). Reduced expression of psca, a member of the ly-6 family of cell surface antigens, in bladder, esophagus, and stomach tumors. *Biochem. Biophys. Res. Commun.* 275, 783–788. doi: 10.1006/bbrc.2000.3393

Bielza, C., and Larrañaga, P. (2014). Discrete bayesian network classifiers: a survey. *ACM Comput. Surv.* 47, 1–43. doi: 10.1145/2576868

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* 321, 209–213. doi: 10.1038/321209a0

Chen, Z., Pang, M., Zhao, Z., Li, S., Miao, R., Zhang, Y., et al. (2020). Feature selection may improve deep neural networks for the bioinformatics problems. *Bioinformatics* 36, 1542–1552. doi: 10.1093/bioinformatics/btz763

Dashtban, M., Balafar, M., and Suravajhala, P. (2018). Gene selection for tumor classification using a novel bio-inspired multi-objective approach. *Genomics* 110, 10–17. doi: 10.1016/j.ygeno.2017.07.010

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* 3, 185–205. doi: 10.1142/S0219720005001004

Ding, W., Bu, H., Zheng, S., and Qian, F. (2009). "Tumor classification by using PCA with relief wrapper," in *2009 2nd IEEE International Conference on Computer Science and Information Technology* (Beijing: IEEE), 514–517. doi: 10.1109/ICCSIT.2009.5234895

Ding, W., Chen, G., and Shi, T. (2019). Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis. *Epigenetics* 14, 67–80. doi: 10.1080/15592294.2019.1568178

Dyrskjøt, L., Ostenfeld, M. S., Bramsen, J. B., Silahtaroglu, A. N., Lamy, P., Ramanathan, R., et al. (2009). Genomic profiling of microRNAs in bladder cancer: miR-129 is associated with poor outcome and promotes cell death *in vitro*. *Cancer Res.* 69, 4851–4860. doi: 10.1158/0008-5472.CAN-08-4043

Emelyanov, P., and Ponomaryov, D. (2017). "Cartesian decomposition in data analysis," in *2017 Siberian Symposium on Data Science and Engineering (SSDSE)* (Novosibirsk: IEEE), 55–60. doi: 10.1109/SSDSE.2017.8071964

Ferrantini, M., Capone, I., and Belardelli, F. (2007). Interferon-α and cancer: mechanisms of action and new perspectives of clinical use. *Biochimie* 89, 884–893. doi: 10.1016/j.biochi.2007.04.006

Gauvreau, K., and Pagano, M. (1993). Student's t test. *Nutrition* 9:386.

Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D., and Maulik, U. (2019). Recursive memetic algorithm for gene selection in microarray data. *Expert Syst. Appl.* 116, 172–185. doi: 10.1016/j.eswa.2018.06.057

Han, F., Tang, D., Cheng, Z., Jiang, J., and Li, Q.-W. (2019). A hybrid gene selection method based on gene scoring strategy and improved particle swarm optimization. *BMC Bioinformatics* 20:289. doi: 10.1186/s12859-019-2773-x

Hsu, Y.-H., and Si, D. (2018). "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)* (Honolulu, HI: IEEE), 5374–5377. doi: 10.1109/EMBC.2018.8513521

Huang, F., Wang, M., Yang, T., Cai, J., Zhang, Q., Sun, Z., et al. (2014). Gastric cancer-derived msc-secreted pdgf-dd promotes gastric cancer progression. *J. Cancer Res. Clin. Oncol.* 140, 1835–1848. doi: 10.1007/s00432-014-1723-2

Kurashige, J., Hasegawa, T., Niida, A., Sugimachi, K., Deng, N., Mima, K., et al. (2016). Integrated molecular profiling of human gastric cancer identifies ddr2 as a potential regulator of peritoneal dissemination. *Sci. Rep.* 6:22371. doi: 10.1038/srep22371

Kuwahara, N., Kitazawa, R., Fujiishi, K., Nagai, Y., Haraguchi, R., and Kitazawa, S. (2013). Gastric adenocarcinoma arising in gastritis cystica profunda presenting with selective loss of kcne2 expression. *World J. Gastroenterol.* 19:1314. doi: 10.3748/wjg.v19.i8.1314

Lee, I.-H., Lushington, G. H., and Visvanathan, M. (2011). A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *J. Clin. Bioinform.* 1:11. doi: 10.1186/2043-9113-1-11

Lei, Z., Tan, I. B., Das, K., Deng, N., Zouridis, H., Pattison, S., et al. (2013). Identification of molecular subtypes of gastric cancer with different responses to pi3-kinase inhibitors and 5-fluorouracil. *Gastroenterology* 145, 554–565. doi: 10.1053/j.gastro.2013.05.010

Li, W.-Q., Hu, N., Burton, V. H., Yang, H. H., Su, H., Conway, C. M., et al. (2014). PLCE1 mRNA and protein expression and survival of patients with esophageal squamous cell carcinoma and gastric adenocarcinoma. *Cancer Epidemiol. Prevent. Biomark.* 23, 1579–1588. doi: 10.1158/1055-9965.EPI-13-1329

Liu, D., Wu, J., and Wu, H.-X. (2009). Expression of MG7 and PGC in gastric cancer and precancerous lesion and its significance. *China Cancer*, 1.

Liu, H., Sun, J., Liu, L., and Zhang, H. (2009). Feature selection with dynamic mutual information. *Pattern Recogn.* 42, 1330–1339. doi: 10.1016/j.patcog.2008.10.028

Liu, X., Chen, S., Liu, J., Qu, W., Xiao, F., Liu, A. X., et al. (2020). Fast and accurate detection of unknown tags for RFID systems – hash collisions are desirable. *IEEE/ACM Trans. Network.* 28, 126–139. doi: 10.1109/TNET.2019.2957239

Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., and Gao, Z. (2017). A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing* 256, 56–62. doi: 10.1016/j.neucom.2016.07.080

Luo, X., Wang, F., Wang, G., and Zhao, Y. (2020). Identification of methylation states of DNA regions for Illumina methylation BeadChip. *BMC Genomics* 21:672. doi: 10.1186/s12864-019-6019-0

Ma, T., and Zhang, A. (2019). "Affinitynet: semi-supervised few-shot learning for disease type prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI), 1069–1076. doi: 10.1609/aaai.v33i01.33011069

Mallik, S., Bhadra, T., and Maulik, U. (2017). Identifying epigenetic biomarkers using maximal relevance and minimal redundancy based feature selection for multi-omics data. *IEEE Trans. Nanobiosci*. 16, 3–10. doi: 10.1109/TNB.2017.2650217

Matowicka-Karna, J., Kamocki, Z., Polińska, B., Osada, J., and Kemona, H. (2013). Platelets and inflammatory markers in patients with gastric cancer. *Clin. Dev. Immunol*. 2013:6. doi: 10.1155/2013/401623

Nogueira, C., Mota, M., Gradiz, R., Cipriano, M. A., Caramelo, F., Cruz, H., et al. (2017). Prevalence and characteristics of epstein-barr virus-associated gastric carcinomas in portugal. *Infect. Agents Cancer* 12:41. doi: 10.1186/s13027-017-0151-8

Park, C., Ha, J., and Park, S. (2020). Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. *Expert Syst. Appl*. 140:112873. doi: 10.1016/j.eswa.2019.112873

Paziewska, A., Dabrowska, M., Goryca, K., Antoniewicz, A., Dobruch, J., Mikula, M., et al. (2014). DNA methylation status is more reliable than gene expression at detecting cancer in prostate biopsy. *Brit. J. Cancer* 111, 781–789. doi: 10.1038/bjc.2014.337

Peng, H., and Fan, Y. (2017). Feature selection by optimizing a lower bound of conditional mutual information. *Informat. Sci*. 418, 652–667. doi: 10.1016/j.ins.2017.08.036

Rodrigues, D., Pereira, L. A., Nakamura, R. Y., Costa, K. A., Yang, X.-S., Souza, A. N., et al. (2014). A wrapper approach for feature selection based on bat algorithm and optimum-path forest. *Expert Syst. Appl*. 41, 2250–2258. doi: 10.1016/j.eswa.2013.09.023

Ruffalo, M., Koyutürk, M., and Sharan, R. (2015). Network-based integration of disparate omic data to identify "silent players" in cancer. *PLoS Comput. Biol*. 11:e1004595. doi: 10.1371/journal.pcbi.1004595

Sakamoto, H., Yoshimura, K., Saeki, N., Katai, H., Shimoda, T., Matsuno, Y., et al. (2008). Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nat. Genet*. 40, 730–40. doi: 10.1038/ng.152

Shen, H., and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal*. 99, 1015–1034. doi: 10.1016/j.jmva.2007.06.007

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. Ca A Cancer *J. Clin*. 70, 7–30. doi: 10.3322/caac.21590

Singh, D., and Yamada, M. (2020). FSNet: feature selection network on high-dimensional biological data. *arXiv [preprint] arXiv*:2001.08322.

Song, Z., Wu, Y., Yang, J., Yang, D., and Fang, X. (2017). Progress in the treatment of advanced gastric cancer. *Tumor Biol*. 39:1010428317714626. doi: 10.1177/1010428317714626

Sun, L.-P., Gong, Y.-H., Dong, N.-N., Wang, L., and Yuan, Y. (2009). Correlation of pepsinogen c (PGC) gene insertion/deletion polymorphism to PGC protein expression in gastric mucosa and serum. *Chin. J. Cancer* 28, 487–492.

Tahir, M. A., Bouridane, A., and Kurugollu, F. (2007). Simultaneous feature selection and feature weighting using hybrid tabu search/k-nearest neighbor classifier. *Pattern Recogn. Lett*. 28, 438–446. doi: 10.1016/j.patrec.2006.08.016

Tanzi, L., Vezzetti, E., Moreno, R., Aprato, A., Audisio, A., and Massé, A. (2020). Hierarchical fracture classification of proximal femur x-ray images using a multistage deep learning approach. *Eur. J. Radiol*. 133:109373. doi: 10.1016/j.ejrad.2020.109373

Vieira, S. M., Mendonça, L. F., Farinha, G. J., and Sousa, J. M. (2013). Modified binary pso for feature selection using svm applied to mortality prediction of septic patients. *Appl. Soft Comput*. 13, 3494–3504. doi: 10.1016/j.asoc.2013.03.021

Wang, G., Hu, N., Yang, H. H., Wang, L., Su, H., Wang, C., et al. (2013). Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in china. *PLoS ONE* 8:e63826. doi: 10.1371/journal.pone.0063826

Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., et al. (2018). MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res*. 46, D146–D151. doi: 10.1093/nar/gkx1096

Wang, X., Shang, W., Li, X., and Chang, Y. (2020). Methylation signature genes identification of cancers occurrence and pattern recognition. *Comput. Biol. Chem*. 85:107198. doi: 10.1016/j.compbiolchem.2019.107198

Wang, Z., Kong, D., Li, Y., and Sarkar, F. H. (2009). PDGF-D signaling: a novel target in cancer therapy. *Curr. Drug Targets* 10, 38–41. doi: 10.2174/138945009787122914

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometr. Intell. Lab. Syst*. 2, 37–52. doi: 10.1016/0169-7439(87)80084-9

Yan, C., Ma, J., Luo, H., and Wang, J. (2018). A hybrid algorithm based on binary chemical reaction optimization and tabu search for feature selection of high-dimensional biomedical data. *Tsinghua Sci. Technol*. 23, 733–743. doi: 10.26599/TST.2018.9010101

Zhang, C., Cai, H., Huang, J., and Song, Y. (2016). nbCNV: a multi-constrained optimization model for discovering copy number variants in single-cell sequencing data. *BMC Bioinformatics* 17:384. doi: 10.1186/s12859-016-1239-7

Zhang, G., Hou, J., Wang, J., Yan, C., and Luo, J. (2020). Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm. *Interdiscipl. Sci. Comput. Life Sci*. 12, 288–301. doi: 10.1007/s12539-020-00372-w

Zhang, Z., Cheng, Y., and Liu, N. C. (2014). Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of web of science subject categories. *Scientometrics* 101, 1679–1693. doi: 10.1007/s11192-014-1294-7

Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020). ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles. *BMC Bioinformatics* 21:43. doi: 10.1186/s12859-020-3388-y

Zouridis, H., Deng, N., Ivanova, T., Zhu, Y., Wong, B., Huang, D., et al. (2012). Methylation subtypes and large-scale epigenetic alterations in gastric cancer. *Sci. Transl. Med*. 4:156ra140. doi: 10.1126/scitranslmed.3004504

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.