Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 5556992, 12 pages
https://doi.org/10.1155/2021/5556992

Research Article

Cancer Classification with a Cost-Sensitive Naive Bayes Stacking Ensemble

Yueling Xiong,[1] Mingquan Ye,[1] and Changrong Wu[2]

[1]*School of Medical Information, Wannan Medical College, Wuhu 241002, China*
[2]*School of Computer and Information, Anhui Normal University, Wuhu 241002, China*

Correspondence should be addressed to Mingquan Ye; ymq@wnmc.edu.cn and Changrong Wu; wcr218@ahnu.edu.cn

Received 6 January 2021; Revised 17 March 2021; Accepted 15 April 2021; Published 26 April 2021

Academic Editor: Martti Juhola

Ensemble learning combines multiple learners to perform combinatorial learning, which has advantages of good flexibility and higher generalization performance. To achieve higher quality cancer classification, in this study, the fast correlation-based feature selection (FCBF) method was used to preprocess the data to eliminate irrelevant and redundant features. Then, the classification was carried out in the stacking ensemble learner. A library for support vector machine (LIBSVM), *K*-nearest neighbor (KNN), decision tree C4.5 (C4.5), and random forest (RF) were used as the primary learners of the stacking ensemble. Given the imbalanced characteristics of cancer gene expression data, the embedding cost-sensitive naive Bayes was used as the metalearner of the stacking ensemble, which was represented as CSNB stacking. The proposed CSNB stacking method was applied to nine cancer datasets to further verify the classification performance of the model. Compared with other classification methods, such as single classifier algorithms and ensemble algorithms, the experimental results showed the effectiveness and robustness of the proposed method in processing different types of cancer data. This method may therefore help guide cancer diagnosis and research.

1. Introduction

Cancer is a malignant tumor originating from epithelial tissues. It is a disease caused by the loss of normal regulation and the excessive proliferation of cells in the body. In recent years, cancer incidence and mortality have increased, thus posing severe risks to human health and life. In addition, because the occurrence and development of cancer are dynamic, most patients are diagnosed with cancer in late stages, thus making clinical diagnosis and treatment challenging [1, 2]. With the continual development of DNA microarray technology, gene expression profile data are gathered by synchronously tracking the expression of many genes. Consequently, early physiological information on cancer can be determined at the molecular level, and the type of cancer can be identified and used to guide biomedicine. However, many features are irrelevant and redundant for classification in gene expression profiles. Moreover, massive computational challenges, such as high dimensionality, small sample sizes, high noise, and unbalanced categories, introduce difficulties in the analysis and processing of cancer gene data. Therefore, various powerful methods have been proposed by researchers to address these problems [3].

At present, the application of machine learning methods to cancer classification is a significant research field in bioinformatics [4, 5]. Many traditional machine learning methods have been successfully applied to the classification analysis of gene expression data [6–9], such as RF, decision tree, KNN, and neural networks. However, with the increasing amounts and diversification of data, the traditional classification algorithm has been unable to meet the requirements of processing existing data and solving practical problems [10, 11]. Ensemble learning is a notable research direction in machine learning, in which multiple base learners are used for combined learning, and the combined classifier is often more accurate than its base classifier, thereby improving

performance in classification problems. Combined classification algorithms have consequently been widely applied in classification problems [12, 13].

Boosting [14] and bagging [15] are two popular combined classification methods. Among them, bagging is a typical representative of a parallel ensemble learning method [16]. In this method, new training subsets are generated by adopting random sampling with putting back on the training set; then, individual learners are trained with different training subsets, respectively, and finally, they are integrated as a whole [17]. In this sampling process, it is inevitable that some instances will be sampled multiple times while others will be ignored [18]. Therefore, for a specific subspace, the individual learner will have high classification accuracy, while for those neglected parts, the individual learner is difficult to correctly classify [19]. In addition, the classification performance of the bagging method depends on the stability of its base classifier. It has a good classification effect for unstable classification algorithms (such as decision tree, neural network, etc.), but it is not very ideal for stable classifier integration [20]. Different from bagging, boosting is an iterative algorithm that transforms weak learners into strong ones [21]. A new weak classifier is added to each round to produce a strong learner with superior performance by increasing the number of iterations. Although the algorithm improves the generalization performance of the combined classification algorithm, the algorithm will suffer from performance degradation and long training time due to the excessive tendency to some difficult samples and the fact that the update of each round of sample distribution depends on the accuracy of the previous round of classifiers [13, 19, 20].

Compared with the two ensemble classification algorithms of bagging and boosting, stacking [22] provides a novel idea for ensemble learning, by emphasizing the deviations of the classifier from the training set and then learning these deviations to enhance classification performance. Stacking improves flexibility in combining learners that provide category output. In addition, this algorithm uses multiple types of individual classifiers to form a two-layer combined classification model. The first layer adopts multiple base learners to train the datasets, and a metalearner is used in the second layer to learn the output of the base learners [16, 18]. Generally, to avoid the overfitting caused by directly using the training sets of the primary learner, cross-validation is usually used to generate the new secondary training set. In addition, how to choose the data type of the secondary training set and the best secondary learner are the two key points that the algorithm must solve [19]. In recent years, stacking ensemble learning methods have been successfully applied in many fields [21]. For example, Ekbal and Saha [23] proposed the extraction of biomedical entities with combined feature selection and a stacking ensemble. The feature selection technique based on genetic algorithms was used to determine the most relevant feature sets of the support vector machine and conditional random field classifiers. Kwon et al. [22] applied a stacking ensemble to breast cancer classification and achieved better classification performance by using a gradient boosting machine and generalized linear model as metaclassifiers. Wang et al. [24] proposed a decision tree ensemble method based on stacking for prostate cancer detection, which achieved good results in classification accuracy, sensitivity, and specificity.

To further explore the effects of ensemble learning applied to cancer gene expression data, we adopted a two-layer classification model using a stacking ensemble learning strategy in combination with feature selection technology to conduct a classification study on binary cancer datasets. First, the original gene expression dataset was standardized and transformed into data with a mean value of 0 and a standard deviation of 1. Then, we used FCBF to calculate the C-correlation value of each gene and category through symmetric uncertainty, and the irrelevant genes were eliminated. The F-correlation value between features was calculated to eliminate the redundant genes and obtain the candidate gene subset, so as to simplify the combined classification model. Second, in the multiclassifier combination method based on the stacking algorithm, LIBSVM, KNN, C4.5, and RF were used as primary learners. Given the problem of imbalanced cancer gene expression data, CSNB was used as the metalearner of the stacking ensemble to perform combinatorial learning, which was expressed as CSNB stacking. Nine cancer datasets were tested for experiments and then compared with other single classifier algorithms and ensemble algorithms: cost-sensitive KNN stacking (CSKNN stacking), cost-sensitive C4.5 stacking (CSC4.5 stacking), cost-sensitive LIBSVM stacking (CSLIBSVM stacking), bagging, AdaBoost, CSNB, NB, LIBSVM, KNN, RF, and C4.5. The experimental results demonstrated that the proposed method provided more accurate classification and was effective and robust in handling various cancer classification data.

The rest of the paper is organized as follows. Section 2 reviews the related work about cancer classification problem. Section 3 introduces the materials and methods of this study, and Section 4 exhibits and discusses the experimental results. In the end, we summarize the paper.

2. Related Work

The mature development of DNA microarray technology provides important guidance for cancer diagnosis and recognition. At present, many scholars have applied the machine learning method to cancer classification and thus designed various classification models and achieved satisfactory results. For example, Musheer et al. [25] used a naive Bayes classifier to classify and evaluate six microarray cancer datasets after feature reduction, which proved that the algorithm has certain significance. Ye et al. [1] applied the KNN classifier to evaluate the extracted information gene subset, which improved the classification accuracy. Besides these, there are also some classification models composed of hybrid methods. Ren et al. [26] proposed an integrated method named correntropy-induced loss-based sparse robust graph regularized extreme learning machine and applied it to the classification and recognition of cancer samples. Gao et al. [27] performed cancer classification based on SVM optimized by particle swarm optimization combined with artificial bee colony approaches, and the effectiveness of these methods was verified by the experimental results. In addition, with the

continuous development of machine learning, many studies have shown that the application of ensemble learning to classification problems is often better than traditional classification algorithms and single classifiers, and it can also solve the problem of increased data volume and data diversification [23, 26]. Therefore, a large number of classification models based on ensemble learning have been proposed. For example, Lee et al. [5] developed an ensemble model based on random forest and deep neural network for cancer classification and achieved an accuracy of 94%. ALzubi et al. [28] used the boosted weighted optimization neural network ensemble classification algorithm to classify cancer patients, thereby improving the accuracy of cancer diagnosis. Ghiasi and Zendehboudi [29] proposed a classification algorithm based on random forest and extreme random tree of decision tree, which was applied to the classification of breast cancer, and verified the diagnostic performance of the algorithm. Li and Luo [30] proposed a performance-weighted voting model for cancer classification. This ensemble model was composed of five weak classifiers: logistic region, SVM, RF, XGBoost, and neural networks, which achieved high accuracy of tumor diagnosis.

Stacking is widely used as a more flexible combination classification model in ensemble learning. However, in the process of constructing the combination model, how to choose the base classifier and give full play to their effectiveness and how to choose the best secondary learner are problems worthy of attention. In addition, the types and characteristics of experimental data will also affect the effectiveness of the classification model. Therefore, in view of the above problems, first of all, we used the FCBF algorithm to reduce the data dimension and achieve the purpose of simplifying the classification model. In addition, four base classifiers (LIBSVM, KNN, C4.5, and RF) were used as the primary learners of the ensemble model. Meanwhile, cost-sensitive learning idea was introduced as a secondary learner to solve the imbalance of microarray gene expression data, so as to overcome these problems and achieve high-quality classification results.

3. Materials and Methods

3.1. Cancer Datasets. For evaluation of the effectiveness of the proposed method, we used nine groups of cancer datasets of two classes derived from the Kent Ridge Biomedical Dataset database. These datasets included central nervous system embryonal tumors (NervSys), leukemia, three groups of diffuse large B-cell lymphoma (DLBCL), prostate cancer, ovarian cancer, and two groups of lung cancer. A detailed description of these datasets is shown in Table 1.

Among these sample data, DLBCL1 was derived from Stanford data, including a total of 62 samples of two subtypes. DLBCL2 and DLBCL3 were selected from two sets of data detected by Harvard. DLBCL2 included two types of patients: those with DLBCL and those with follicular lymphoma. DLBCL3 comprised the outcome prediction data, including cured patient samples and relapsed patient samples. Lung cancer1 was derived from the University of Michigan, including ten normal samples and 86 diseased samples. Lung cancer2 contained 181 samples comprising 31 cases of malignant pleural mesothelioma and 150 cases of adenocarcinoma, with 12533 genes detected. Notably, the NervSys and DLBCL3 data were outcome prediction data, whereas the ovarian cancer data were protein data, and the remaining samples were from two categories of data.

Table 1: Details of cancer datasets.

Datasets	Samples	No. of genes	Classes	Labels
NervSys	60	7129	2	Outcome prediction
Leukemia	72	7129	2	Two categories
DLBCL1	47	4026	2	Two categories
DLBCL2	77	7129	2	Two categories
DLBCL3	58	7129	2	Outcome prediction
Prostate	102	12600	2	Cancer or not
Ovarian	253	15154	2	Protein data
Lung1	96	7129	2	Cancer or not
Lung2	181	12533	2	Two categories

3.2. Stacking Ensemble Learning Algorithm. Stacking, also known as stacked generalization [31], is a technology involving heterogeneous classifier collections. By integrating multiple different types of base classifiers and combining them into a strong classifier, the generalization ability of the strong classifier can be improved. The stacking ensemble learning algorithm adopted a two-layer framework structure, as shown in Figure 1. The main idea of this algorithm was to train the dataset with multiple primary learners first. Then, the prediction results obtained by each base classifier were used as the input of the metaclassifier to perform training again. Finally, the training result of the metaclassifier was the final prediction result. The stacking ensemble algorithm took into account the learning ability of the primary classifier and metaclassifier, so that the final classification performance was significantly improved [32–34].

3.3. KNN. KNN [35] is a classification algorithm in supervised learning and also a lazy learning algorithm. The algorithm has the advantages of simple use, rapid calculation, and good predictive effects. However, when the sample distribution is uneven, the prediction error also increases.

The basic idea of the algorithm was that if a prediction sample has K-nearest neighbors in the feature space, the category of the prediction sample was usually determined by most of the categories of the K-nearest neighbors. The effects and performance were optimized by selecting the K value, distance measurement method, and classification decision rules.

3.4. C4.5. The three commonly used decision tree algorithms [36] are Iterative Dichotomiser (ID3), C4.5, and CART. Among them, decision tree C4.5 was an improvement on ID3. The C4.5 algorithm used the information gain ratio as the index to select the best split, which accommodated continuous variables and missing values, thereby addressing the disadvantage of ID3's tendency to select attributes with

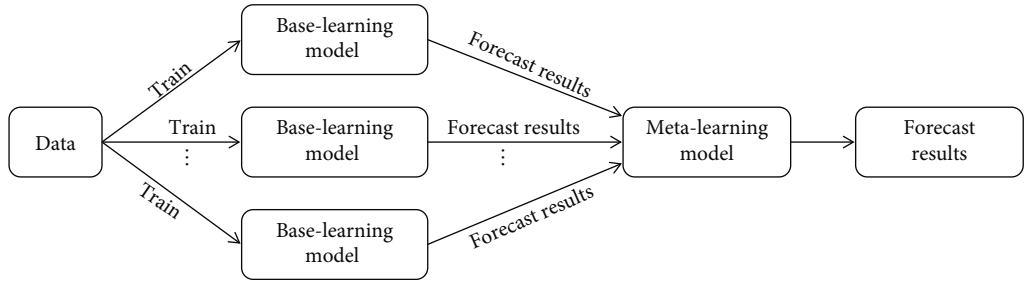


Figure 1: Ensemble learning method based on stacking.

many categories. Pruning could be performed during the construction of the tree to avoid overfitting.

3.5. RF. RF [37] is an ensemble algorithm that constructs a strong classifier by training multiple weak classifiers. The prediction results were determined by the average or voting of multiple base classifiers, thus giving the prediction model good accuracy and generalization ability. The decision tree was used as the base classifier in the RF. When making predictions, each decision tree in the forest participated in classification prediction. Finally, the classification with the highest number of votes was selected as the prediction value. The accuracy of RF depended on the strength of the base classifier and the dependence between them. Moreover, it was relatively robust to errors and outliers.

3.6. SVM. SVM [38] is a classic stability classifier. The SVM method was based on the VC dimension theory of statistical theory and the principle of minimum structural risk. The VC dimension represented the complexity of the problem. Normally, the higher the VC dimension, the more complex the function. SVM had advantages in handling nonlinear high-dimensional problems, and it is widely used in the classification and recognition fields [39].

For nonlinear samples, SVM used a kernel function to map the original data to a high-dimensional space, thus making the samples linearly separable. The optimal classification hyperplane was constructed to separate the samples correctly. Generally, different forms of kernel functions strongly affected the classification performance of SVM. Among them, the radial kernel function (RBF) [40] had fewer parameters and better performance and consequently was widely used in practical applications. The formula of the RBF kernel function can be expressed as follows:

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{\sigma^2}\right), \tag{1}$$

where x and x_i are two sample vectors, K is the value of the RBF kernel function, and σ is a free parameter.

3.7. CSNB. Cancer gene expression data are a type of unbalanced data [41]. Traditional classification algorithms often do not consider the factor of misclassification cost, thereby leading to classification results that tend to focus on the learning of large categories while ignoring the learning of small categories. In this experiment, the idea of cost sensitivity was introduced into a naive Bayes classification algorithm to make it sensitive to cost. In this way, the recognition rate of rare classes was improved, and the validity of classification was strengthened.

First, the definition of misclassification cost was given. Taking a binary classification dataset as an example, let the c_0 class be the minority class and the c_1 class be the majority class. The misclassification cost can usually be represented by a 2×2 cost matrix, and each element in the matrix represents the misclassification cost of samples [42].

$$C = \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix}, \tag{2}$$

where C_{ij} represents the cost of mistakenly classifying samples that are i classes into j classes. Generally, for the rationality condition of two classification problems [43], correct classification is not considered to bring losses; consequently, the cost of misclassification is 0. However, the cost of small classes being misclassified into large classes is much higher than that of large classes being misclassified into small classes. Therefore, we can derive the following relationship: $C_{00} = C_{11} = 0$ and $C_{01} > C_{10}$.

After the cost matrix is determined, the naive Bayes theory is used to construct the risk function [38]. When a sample x with an unknown category is classified by a classification algorithm, the sample x can be represented by a vector $(a_1, a_2, a_3, \cdots, a_n)$. The probability that it belongs to the category c_j is $P(c_j \mid x)$, which is expressed by the Bayesian formula as

$$P(c_j \mid x) = \frac{P(x \mid c_j)P(c_j)}{P(x)} = \frac{P(a_1, a_2, a_3, \cdots a_n \mid c_j)P(c_j)}{P(x)}. \tag{3}$$

When its category is determined to be c_j, the expected misclassification cost is

$$R(c_j \mid x) = \sum_i P(c_i \mid x)C_{ij}, \tag{4}$$

where $P(c_i \mid x)$ represents the posterior probability that sample x belongs to the c_i category. Its value is obtained by the naive Bayes formula. Furthermore, the corresponding

categories are determined by minimizing the posterior probability as follows:

$$c = \arg\min_{j=0,1} \{R(c_j|x)\}. \quad (5)$$

The sample x is finally predicted to be a certain category c_j that makes $R(c_j \mid x)$ have a minimum value, which can be expressed as

$$R(c_{j*} \mid x) = \min_j R(c_j \mid x), \quad (6)$$

where formula (6) is the CSNB formula.

3.8. FCBF. FCBF [39] is a supervised fast filtering feature selection algorithm. Its core idea is to define C-correlation and F-correlation, where C-correlation is the degree of correlation between features and categories and F-correlation is the degree of correlation between features. When a feature has high C-correlation with a category and low F-correlation with other selected features, the feature is marked as an important feature. In this algorithm, symmetric uncertainty (SU) was adopted as the standard to measure the degree of correlation. SU is defined as the standardized information gain:

$$\mathrm{SU}(X, Y) = 2\left[\frac{\mathrm{IG}(X|Y)}{H(X) + H(Y)}\right], \quad (7)$$

where X and Y represent two random variables, $\mathrm{IG}(X \mid Y)$ denotes the information gain, and $H(X)$ represents the information entropy.

3.9. The Proposed CSNB Stacking Algorithm. In this article, the FCBF algorithm was first used to reduce the dimensionality of the datasets. Then, LIBSVM, KNN, C4.5, and RF were used as primary learners. In addition, given the unbalanced classification of cancer gene expression data, the naive Bayes with embedded cost sensitivity was adopted as the metalearner of the stacking ensemble. In the experiment, the 5×10 -fold nested cross-validation method was used to divide the data to prevent overfitting. The experimental process is shown in Figure 2.

4. Results

4.1. Feature Selection. Cancer datasets contain many genes that are irrelevant and redundant for classification, thus leading to more complex classification tasks and inaccurate classification results. In fact, few informative genes are known to directly affect the classification results. Therefore, to save time and calculation costs and obtain better classification performance, we used the FCBF method to quickly and effectively reduce the dimensionality of the cancer datasets before classification, by discarding features making little or no contribution to classification. In this article, data were preprocessed first, and FCBF was used in WEKA for feature selection. The data information after reduction is provided in Table 2.

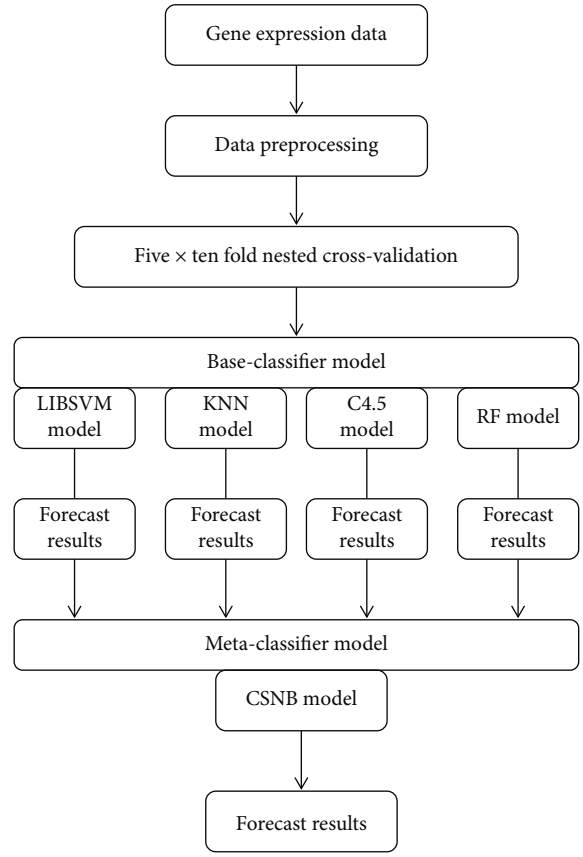


Figure 2: The experimental flow of the proposed method.

Table 2: Reduced attributes by FCBF.

Datasets	Original attributes	Reduced attributes
NervSys	7129	28
Leukemia	7129	51
DLBCL1	4026	60
DLBCL2	7129	73
DLBCL3	7129	27
Prostate	12600	77
Ovarian	15154	30
Lung1	7129	1
Lung2	12533	128

As shown in Table 2, the feature dimensions of the nine cancer datasets were greatly reduced after feature selection with the FCBF algorithm. Among them, ovarian cancer was reduced from the original 15154 attributes to 30, whereas the number of features of DLBCL2 was decreased from 7129 to 27. In addition, lung cancer ultimately retained only one important feature among 7129 original features. Hence, after the feature reduction by the FCBF algorithm, the subsequent classification task was greatly simplified.

4.2. Cancer Classification. In this article, the nine cancer datasets after feature reduction are used as the input to the proposed method CSNB stacking classifier for classification. To

Table 3: Classification accuracy (%) of different methods.

Datasets	NervSys	Leukemia	DLBCL1	DLBCL2	DLBCL3	Prostate	Ovarian	Lung1	Lung2
CSNB stacking	90.00	100	100	98.70	87.93	95.10	100	98.96	100
CSKNN stacking	83.33	100	100	98.70	84.48	92.16	100	98.96	100
CSC4.5 stacking	90.00	100	100	98.70	86.21	93.14	100	89.58	99.45
CSLIBSVM stacking	86.67	100	100	97.40	77.59	92.16	100	92.71	99.45
Bagging	83.33	100	100	97.40	84.48	93.14	100	98.96	100
AdaBoost	81.67	100	100	96.10	84.48	95.10	100	98.96	100
CSNB	76.67	100	100	96.10	86.21	92.16	100	98.96	100
NB	76.67	100	100	96.10	86.21	92.16	100	98.96	100
LIBSVM	86.67	100	100	98.70	82.76	94.12	100	98.96	99.45
KNN	86.67	100	95.74	93.51	70.69	89.22	100	98.96	100
RF	88.33	98.61	100	96.10	86.21	94.12	99.60	98.96	99.45
C4.5	75.00	81.94	78.72	84.42	72.41	88.24	98.02	98.96	95.58

test the quality of the proposed method, we compared this method with CSKNN stacking, CSC4.5 stacking, CSLIBSVM stacking, bagging, AdaBoost, CSNB, NB, LIBSVM, KNN, RF, andC4.5.

Among these methods, the primary learners with the CSKNN stacking, CSC4.5 stacking, and CSLIBSVM stacking were the same as with the proposed method, but the metalearners were different. CSKNN stacking adopts embedded cost-sensitive KNN as a metalearner for combinatorial learning. CSC4.5 stacking used embedded cost-sensitive C4.5 as a metalearner for classification, whereas the metalearner of CSLIBSVM stacking applies embedded cost-sensitive LIBSVM. Both bagging and AdaBoost ensemble methods used CSNB as the base classifier. The CSNB algorithm introduced cost-sensitive information into naive Bayes, thus making it sensitive to cost and emphasizing the learning of small samples. The above classification models all used cost-sensitive learning. In addition, several other comparison methods exist. Among them, the naive Bayes algorithm was simple and provided stable classification efficiency, but it was highly sensitive to the expression form of the input data; for example, if the training data error is large, the predictive effect will be poor. The LIBSVM classification algorithm had few parameters, flexible operation, and broad application capability. KNN was simple and effective and had low training costs, but it was computationally expensive. RF adopted an integrated algorithm with high accuracy and fast training speed, but the training required large amounts of time and space. Moreover, the RF model was prone to overfitting for sample sets with high noise. The classification rules generated by the C4.5 algorithm were easy to understand and have high accuracy; this algorithm handled discrete and continuous data, but its computational efficiency was low.

Furthermore, owing to the small training set, we evaluated the classification performance with 5×10-fold nested cross-validation to avoid overfitting. That is, the data were first divided by the 5-fold cross-validation method, and then, the datasets were divided according to the 10-fold cross-validation method before the primary classifier of stacking carries out the training data. The classification accuracy, recall rate, F-score, specificity, and receiver operating characteristic (ROC) curves were used to evaluate the effectiveness of the proposed method. The final classification accuracy results are shown in Table 3.

As shown in Table 3, the first row represented the classification results of the proposed CSNB stacking method, which always obtained the best value for the nine datasets. In these classification results, such as those with NervSys, the proposed CSNB stacking method achieved the same accuracy as CSC4.5 stacking (both 90%, a value higher than those of other methods). For the DLBCL3 and prostate cancer datasets, the method proposed in this article, compared with other classification methods, had the highest classification accuracy. In addition, the proposed classification model achieved 100% classification accuracy on the leukemia, DLBCL1, ovarian cancer, and lung cancer2 datasets. According to the classification results, the classification method based on a stacking ensemble was better than the single classification model. To visually demonstrate the classification effect of different models, we have converted the experimental results in Table 3 to a line graph, as shown in Figures 3 and 4. Because the classification results for the leukemia, DLBCL1, and ovarian datasets were all 100%, they were not shown in the figure. In Figure 4, the remaining classification methods were combined and compared with the proposed method. Similarly, because lung1 achieved 98% in these comparison methods, it was also not shown in the figure.

Figure 5 shows the values of the recall rates obtained by the various classification methods on nine cancer datasets. The results show that the proposed CSNB stacking method achieves the highest recall rate on all nine datasets, followed by the CSKNN stacking method. For DLBCL3 and prostate cancer, the recall rate of the proposed method reached 87.90% and 95.10%, respectively; these results were superior to those of other classification models.

Figures 6 and 7 illustrate the specificity values obtained by the various classification methods on nine cancer datasets. Figure 6 shows the comparison between the proposed method and the other five ensemble classifiers. Since the

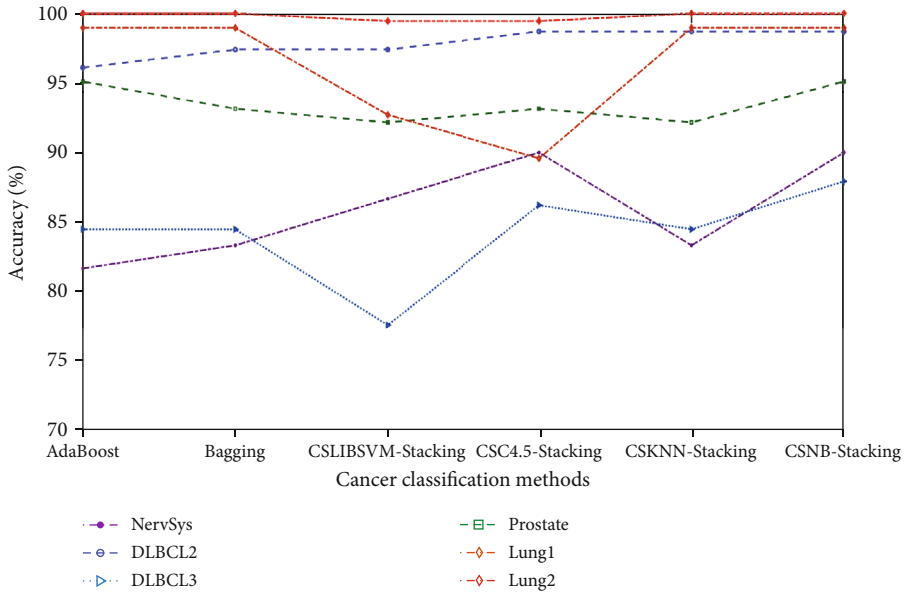


Figure 3: Cancer classification accuracy of different ensemble methods.

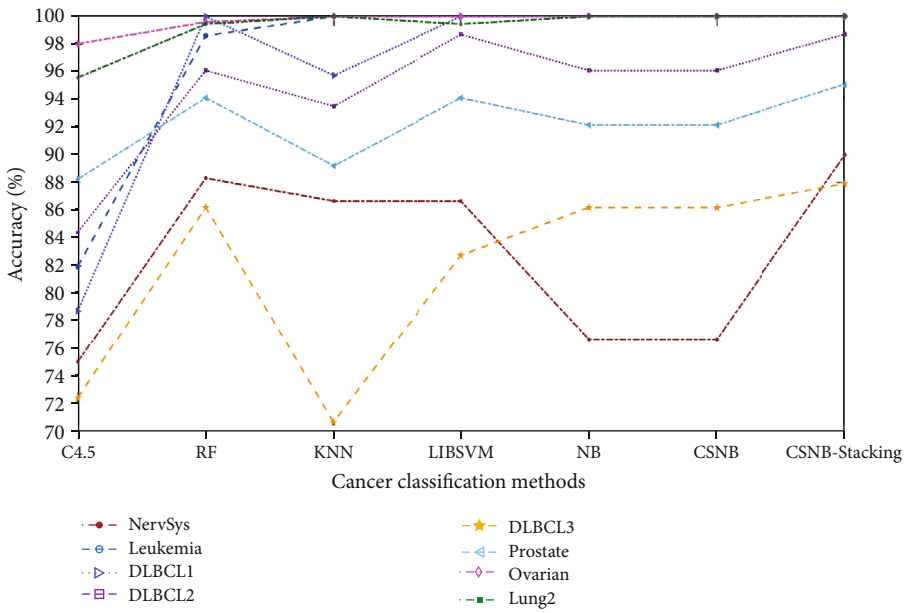


Figure 4: Cancer classification accuracy of different methods.

specificities of the six ensemble classifiers in the three datasets of leukemia, DLBCL1, and ovarian cancer were all 1, they were not shown in the figure. Figure 7 is a comparison of other remaining classification models and the proposed method. It can be seen from the figure that the proposed methods in this paper can obtain better specificity on nine datasets. In addition, compared with other ensemble classifiers, the proposed method had a better classification effect and achieves specificity of 0.9 and 1 on lung1 and lung2 datasets, respectively. However, CSC4.5 stacking performed poorly, especially for the lung1 dataset, with a specificity value of 0. Although C4.5 achieved a specificity of 1 on the lung1 dataset, the results on the remaining eight datasets were not good.

Table 4 lists the F-score of different classification methods on each cancer dataset. As can be seen from the table, the proposed method obtained the highest F-score on all nine datasets. Since the F-score represents the harmonic mean value of precision and recall rate, the experimental results further verified the effectiveness of the proposed method.

In order to display the ROC curves of different classification models more clearly and intuitively, we selected NervSys and prostate datasets from nine datasets for graphing. In addition, we divided all classifiers into two groups to show them more aesthetically and clearly. Figure 8 shows the ROC curves of the NervSys dataset on different classifiers. Figure 9 shows the ROC curves of the prostate dataset on

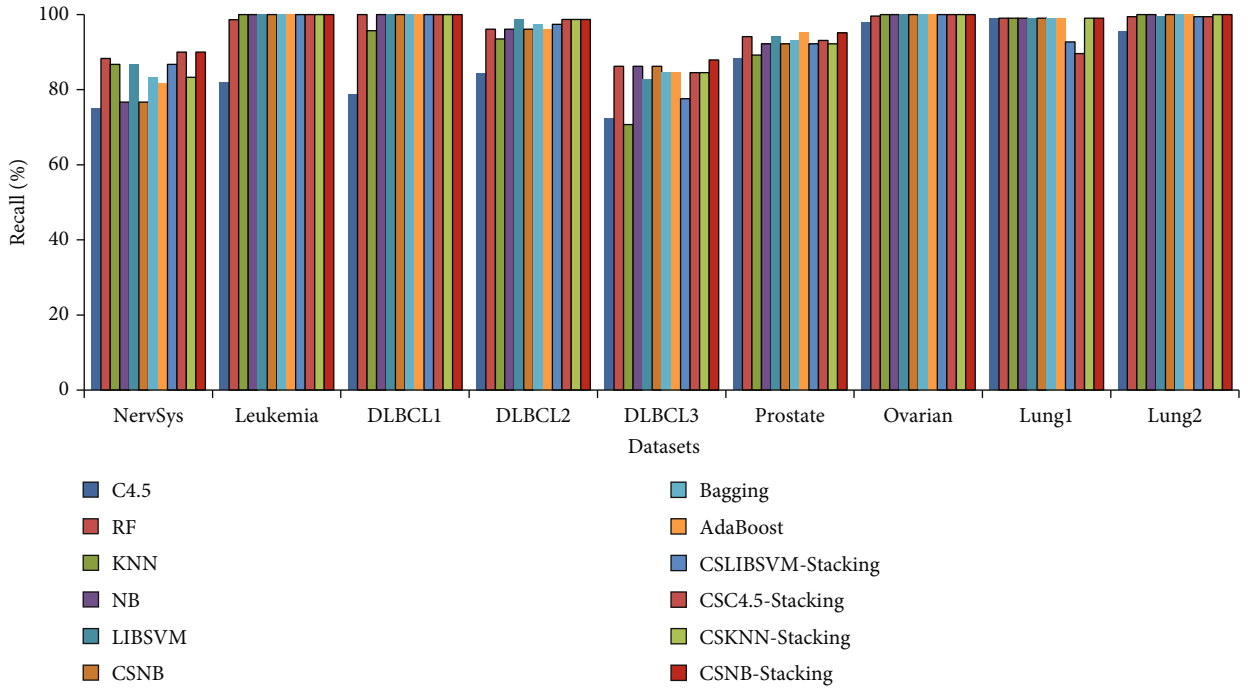


Figure 5: Recall rate of different methods.

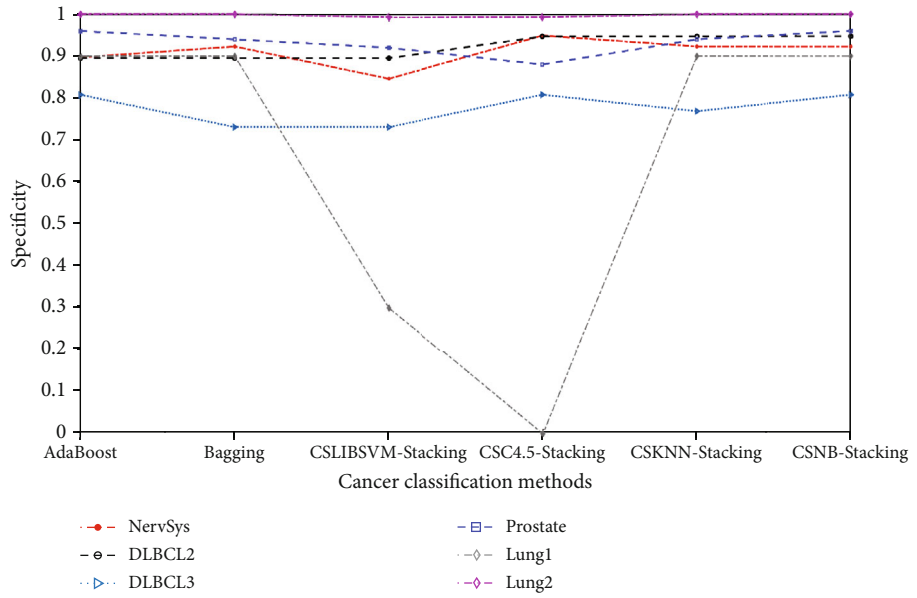


Figure 6: Specificity of different ensemble methods.

different classifiers. It can be seen from the figure that the proposed classification model in this study had a better performance, and the curve corresponding to CSNB stacking was closer to the upper left corner of the ROC chart, which further proved the great advantage of the proposed method in dealing with cancer datasets.

5. Discussion

As a combinatorial learning method, the stacking method is supported by less theoretical research than the boosting and bagging combination method, but it is welcomed by many researchers because of its strong flexibility and scalability in algorithms [44, 45]. However, the choice of the output data type of the base learner and the metalearner are longstanding problems in the stacking method. Some researchers favor using the output probability of the base learner as the input of the secondary learner. In addition, if some relatively simple learners are selected as primary learners and more complex learners are selected as metalearners, the stacking performance will be more robust in terms of classification accuracy [46, 47]. Therefore, herein, we propose the novel

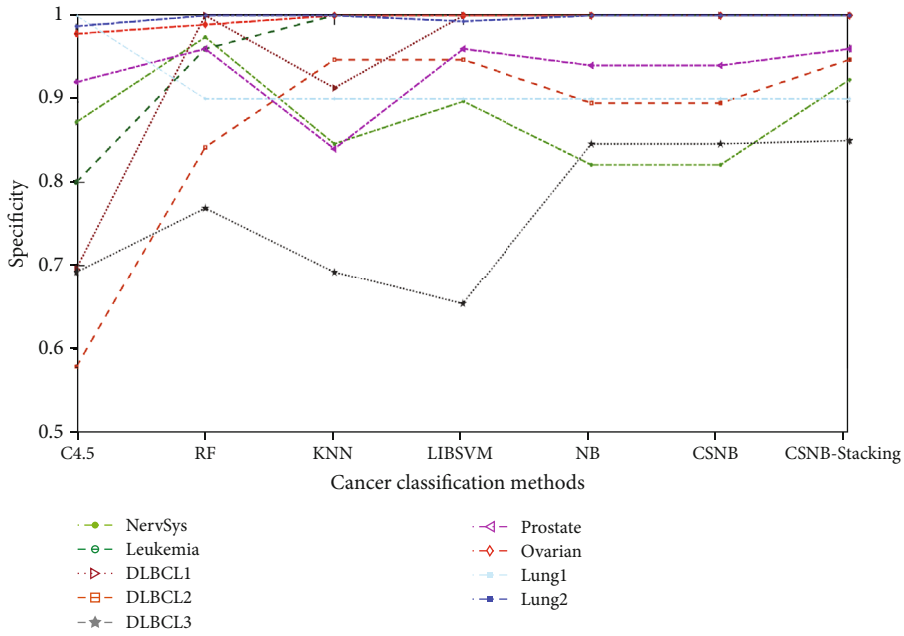


Figure 7: Specificity of different methods.

Table 4: *F*-score of different classification methods.

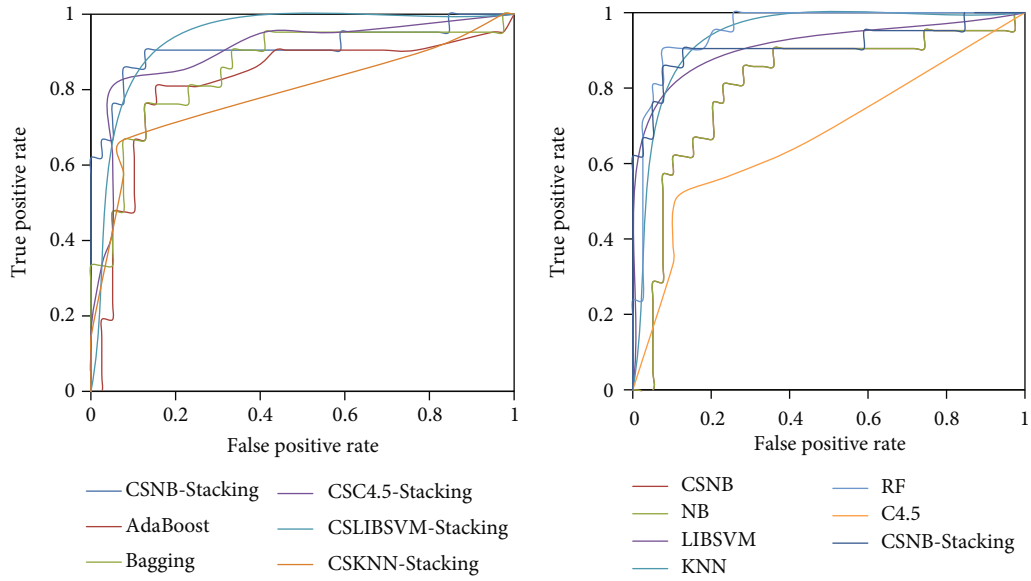
Datasets	NervSys	Leukemia	DLBCL1	DLBCL2	DLBCL3	Prostate	Ovarian	Lung1	Lung2
CSNB stacking	0.900	1.00	1.00	0.987	0.878	0.951	1.00	0.989	1.00
CSKNN stacking	0.829	1.00	1.00	0.987	0.844	0.922	1.00	0.989	1.00
CSC4.5 stacking	0.899	1.00	1.00	0.987	0.861	0.931	1.00	0.945	0.995
CSLIBSVM stacking	0.869	1.00	1.00	0.974	0.775	0.922	1.00	0.909	0.995
Bagging	0.829	1.00	1.00	0.974	0.842	0.931	1.00	0.989	1.00
AdaBoost	0.813	1.00	1.00	0.961	0.844	0.951	1.00	0.989	1.00
CSNB	0.767	1.00	1.00	0.961	0.862	0.922	1.00	0.989	1.00
NB	0.767	1.00	1.00	0.961	0.862	0.922	1.00	0.989	1.00
LIBSVM	0.867	1.00	1.00	0.987	0.821	0.941	1.00	0.989	0.995
KNN	0.869	1.00	0.957	0.937	0.707	0.892	1.00	0.989	1.00
RF	0.879	0.986	1.00	0.960	0.860	0.941	0.996	0.989	0.994
C4.5	0.741	0.822	0.785	0.838	0.724	0.882	0.980	0.952	0.955

combination classification model CSNB stacking and perform experimental verification.

The datasets selected in this experiment were diverse, including prognostic data, protein data, and two-category data. The experimental results showed that the proposed method achieved the best classification accuracy in different types of cancer datasets, thus reflecting the superiority of the CSNB stacking classification model in processing cancer datasets. On the other hand, through the experimental verification, the four base classifiers selected in this study constitute the primary learner of stacking, which is a satisfactory combination. At the same time, the cost-sensitive idea was introduced into the metalearner of stacking, which can well deal with the problem of unbalanced data. In addition, the proposed classification combination model was superior to the single classifier and ensemble algorithm in terms of stability and generalization ability, and it achieved better classification results. Furthermore, the proposed method provided a feasible reference scheme for the two-classification problem, and it may potentially aid in cancer prediction, identification, and classification.

6. Conclusion

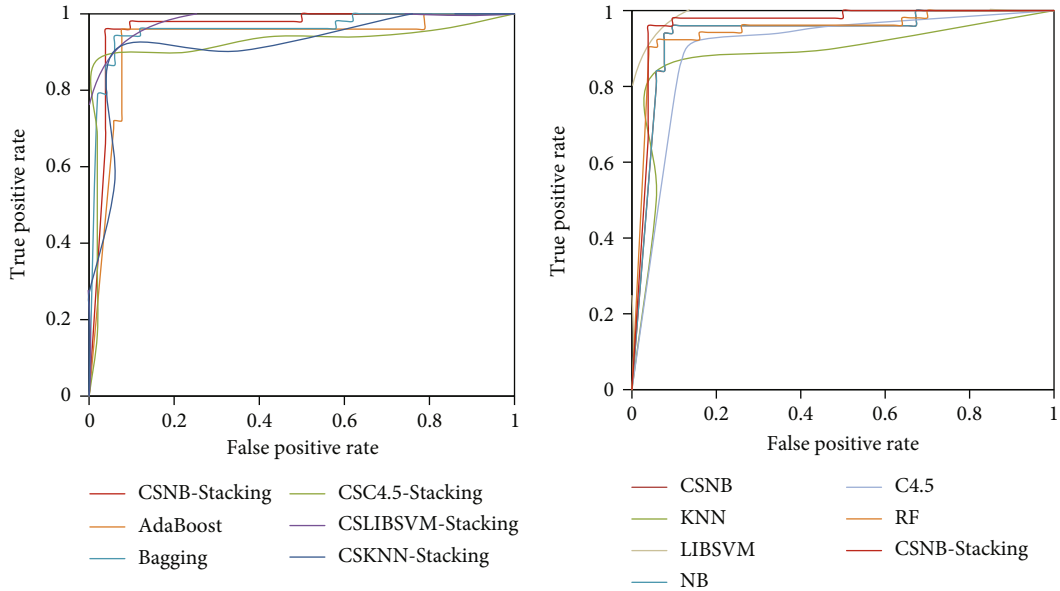
Owing to its high incidence and mortality, cancer has always been a threat to human health. Its complexity and variability make clinical diagnosis and treatment difficult. The emergence of cancer gene expression profiles, by synchronously tracking the expression levels of many genes, is important for early diagnosis of cancer at the molecular level. However, the analysis and processing of gene expression data are accompanied by problems such as high dimensionality, relatively small samples, high noise, and class imbalance. Therefore, to improve the quality of cancer classification and



(a) ROC curve of different ensemble methods

(b) ROC curve of different methods

Figure 8: ROC curves of NervSys by different classification methods.



(a) ROC curve of different ensemble methods

(b) ROC curve of different methods

Figure 9: ROC curves of the prostate dataset by different classification methods.

determine the genes that contribute to classification, we proposed a novel method for gene expression data classification: CSNB stacking, based on a CSNB stacking ensemble. This algorithm is based on the traditional stacking algorithm. In addition, because of the imbalanced characteristics of cancer gene expression data, we adopted the embedding CSNB as the metalearner of the stacking ensemble.

In the experiment, FCBF was first used to reduce the dimensionality of the data and discarded the irrelevant and redundant attributes. Then, the feature subset was input into the CSNB stacking classification model for classification. The experimental results for nine sets of cancer datasets demonstrated that the CSNB stacking method achieved the best classification performance among the tested methods. Simultaneously, this method had advantages in processing outcome prediction data, protein data, and two categories of data, and it therefore should have high guidance potential and clinical value for cancer classification and prognosis prediction.

In addition, through the comparison and analysis of the experimental results, the superiority of the proposed method in the binary classification problem was verified. Our next step will be to expand to the multiclassification task on the basis of the algorithm, to explore its effectiveness. Simultaneously, to

make the algorithm more conducive to solving practical problems, we will also focus on combining multiple types of data to achieve more detailed analysis in the future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We thank International Science Editing (http://www.internationalscienceediting.com) for polishing and examining the language of this manuscript. This article was supported by the National Natural Science Foundation of China (61672386), Humanities and Social Sciences Planning Project of Ministry of Education (16YJAZH071), and Anhui Provincial Natural Science Foundation of China (1708085MF142).

References

[1] M. Ye, W. Wang, C. Yao, R. Fan, and P. Wang, "Gene selection method for microarray data classification using particle swarm optimization and neighborhood rough set," *Current Bioinformatics*, vol. 14, no. 5, pp. 422–431, 2019.

[2] L. Gao, M. Ye, X. Lu, and D. Huang, "Hybrid method based on information gain and support vector machine for gene selection in cancer classification," *Genomics, Proteomics & Bioinformatics*, vol. 15, no. 6, pp. 389–395, 2017.

[3] Z.-Z. Xue, Y. Wu, Q.-Z. Gao, L. Zhao, and Y.-Y. Xu, "Automated classification of protein subcellular localization in immunohistochemistry images to reveal biomarkers in colon cancer," *BMC Bioinformatics*, vol. 21, no. 398, pp. 1–15, 2020.

[4] Yadavendra and S. Chand, "A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method," *Machine Vision and Applications*, vol. 31, no. 6, pp. 32–41, 2020.

[5] K. Lee, H.-o. Jeong, S. Lee, and W.-K. Jeong, "CPEM: accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network," *Scientific Reports*, vol. 9, no. 1, 2019.

[6] L. Jiang, C. M. T. Greenwood, W. Yao, and L. Li, "Bayesian hyper-LASSO classification for feature selection with application to endometrial cancer RNA-seq data," *Scientific Reports*, vol. 10, no. 1, pp. 9747–9749, 2020.

[7] L. Venkataramana, S. G. Jacob, and R. Ramadoss, "A parallel multilevel feature selection algorithm for improved cancer classification," *Journal of Parallel and Distributed Computing*, vol. 138, no. 1, pp. 78–98, 2020.

[8] D. Menaga and S. Revathi, "An empirical study of cancer classification techniques based on the neural networks," *Biomedical Engineering: Applications, Basis and Communications*, vol. 32, no. 2, 2020.

[9] P. García-Díaz, I. Sánchez-Berriel, J. A. Martínez-Rojas, and A. M. Diez-Pascual, "Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-seq data," *Genomics*, vol. 112, no. 2, pp. 1916–1925, 2020.

[10] B.-H. Kim, K. Yu, and P. C. W. Lee, "Cancer classification of single-cell gene expression data by neural network," *Bioinformatics*, vol. 36, no. 5, pp. 1360–1366, 2019.

[11] A. M. Mabu, R. Prasad, and R. Yadav, "Gene expression dataset classification using artificial neural network and clustering-based feature selection," *International Journal of Swarm Intelligence Research*, vol. 11, no. 1, pp. 65–86, 2020.

[12] S. Han, Y. Wang, W. Liao et al., "The distinguishing intrinsic brain circuitry in treatment-naive first-episode schizophrenia: ensemble learning classification," *Neurocomputing*, vol. 365, no. 1, pp. 44–53, 2019.

[13] D. Xu, X. Zhang, and H. Feng, "Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model," *International Journal of Finance & Economics*, vol. 24, no. 2, pp. 903–921, 2019.

[14] H. Gu, Y. F. Cui, L. Xu et al., "Bagging classification tree-based robust variable selection for radial basis function network modeling in metabonomics data analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 174, no. 1, pp. 76–84, 2018.

[15] J. Srimathi and V. Mayil, "Fuzzy gene optimized reweight boosting classification for energy efficient data gathering in WSN," *International Journal of Computer Networks & Communications*, vol. 11, no. 2, pp. 95–112, 2019.

[16] C. Li, W.-H. Zhang, R. Li, J.-Y. Wang, and J.-M. Lin, "Research on star/galaxy classification based on stacking ensemble learning," *Chinese Astronomy and Astrophysics*, vol. 44, no. 3, pp. 345–355, 2020.

[17] P. K. Mallick, S. Mishra, S. K. Satapathy, and A. R. Panda, "Emotion classification from EEG brain signal using weighted stacking of ensemble classifiers," *Indian Journal of Public Health Research & Development*, vol. 10, no. 11, pp. 4749–4752, 2019.

[18] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[19] M. M. Rahman, M. I. H. Bhuiyan, and A. B. Das, "Classification of focal and non-focal EEG signals in VMD-DWT domain using ensemble stacking," *Biomedical Signal Processing and Control*, vol. 50, no. 1, pp. 72–82, 2019.

[20] M. Ahmed, A. G. Rasool, H. Afzal, and I. Siddiqi, "Improving handwriting based gender classification using ensemble classifiers," *Expert Systems with Applications*, vol. 85, no. 1, pp. 158–168, 2017.

[21] S. Kang, S. Cho, and P. Kang, "Multi-class classification via heterogeneous ensemble of one-class classifiers," *Engineering Applications of Artificial Intelligence*, vol. 43, no. 1, pp. 35–43, 2015.

[22] H. Kwon, J. Park, and Y. Lee, "Stacking ensemble technique for classifying breast cancer," *Healthcare informatics research*, vol. 25, no. 4, pp. 283–288, 2019.

[23] A. Ekbal and S. Saha, "Stacked ensemble coupled with feature selection for biomedical entity extraction," *Knowledge-Based Systems*, vol. 46, no. 1, pp. 22–32, 2013.

[24] Y. Wang, D. Wang, N. Geng, Y. Wang, Y. Yin, and Y. Jin, "Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection," *Applied Soft Computing Journal*, vol. 77, no. 1, pp. 188–204, 2019.

[25] R. A. Musheer, C. K. Verma, and N. Srivastava, "Novel machine learning approach for classification of high5-dimensional microarray data," *Soft Computing*, vol. 23, no. 24, pp. 13409–13421, 2019.

[26] L.-R. Ren, Y.-L. Gao, J.-X. Liu, J. Shang, and C.-H. Zheng, "Correntropy induced loss based sparse robust graph regularized extreme learning machine for cancer classification," *BMC Bioinformatics*, vol. 21, no. 1, pp. 445–867, 2020.

[27] L. Gao, M. Ye, and C. Wu, "Cancer classification based on support vector machine optimized by particle swarm optimization and artificial bee colony," *Molecules*, vol. 22, no. 12, p. 2086, 2017.

[28] J. A. ALzubi, B. Bharathikannan, S. Tanwar, R. Manikandan, A. Khanna, and C. Thaventhiran, "Boosted neural network ensemble classification for lung cancer disease diagnosis," *Applied Soft Computing Journal*, vol. 80, pp. 579–591, 2019.

[29] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Computers in Biology and Medicine*, vol. 128, p. 104089, 2021.

[30] Y. Li and Y. Luo, "Performance-weighted-voting model: an ensemble machine learning method for cancer type classification using whole-exome sequencing mutation," *Quantitative Biology*, vol. 8, no. 4, pp. 347–358, 2020.

[31] H. Zhang and J. Dong, "Application of sample balance-based multi-perspective feature ensemble learning for prediction of user purchasing behaviors on mobile wireless network platforms," *EURASIP Journal on Wireless Communications and Networking*, vol. 2020, no. 1, 2020.

[32] P. Wu, X. Meng, and L. Song, "A novel ensemble learning method for crash prediction using road geometric alignments and traffic data," *Journal of Transportation Safety & Security*, vol. 12, no. 9, pp. 1128–1146, 2020.

[33] J. Zhang, Y. Wang, Y. Sun, and G. Li, "Strength of ensemble learning in multiclass classification of rockburst intensity," *International Journal for Numerical and Analytical Methods in Geomechanics*, vol. 44, no. 13, pp. 1833–1853, 2020.

[34] A. Sungheetha and R. R. Sharma, "Extreme learning machine and fuzzy *K*-nearest neighbour based hybrid gene selection technique for cancer classification," *Journal of Medical Imaging and Health Informatics*, vol. 6, no. 7, pp. 1652–1656, 2016.

[35] A. Nagpal and V. Singh, "Feature selection from high dimensional data based on iterative qualitative mutual information," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 6, pp. 5845–5856, 2019.

[36] M. Maktabi, H. Köhler, M. Ivanova et al., "Tissue classification of oncologic esophageal resectates based on hyperspectral data," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 10, pp. 1651–1661, 2019.

[37] R. Krishnamurthi, N. Aggrawal, L. Sharma, D. Srivastava, and S. Sharma, "Importance of feature selection and data visualization towards prediction of breast cancer," *Recent Patents on Computer Science*, vol. 12, no. 4, pp. 317–328, 2019.

[38] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[39] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[40] X. Tao, Q. Li, W. Guo et al., "Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification," *Information Sciences*, vol. 487, no. 1, pp. 31–56, 2019.

[41] J. Zhang and L. Chen, "Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis," *Computer Assisted Surgery*, vol. 24, no. sup2, pp. 62–72, 2019.

[42] G. Zhang, F. Porikli, H. Sun, Q. Sun, G. Xia, and Y. Zheng, "Cost-sensitive joint feature and dictionary learning for face recognition," *Neurocomputing*, vol. 391, no. 1, pp. 177–188, 2020.

[43] J. Howcroft, J. Kofman, and E. D. Lemaire, "Feature selection for elderly faller classification based on wearable sensors," *Journal of neuroengineering and rehabilitation*, vol. 14, no. 1, p. 47, 2017.

[44] B. A. Tama and K. H. Rhee, "Performance evaluation of intrusion detection system using classifier ensembles," *International Journal of Internet Protocol Technology*, vol. 10, no. 1, pp. 22–29, 2017.

[45] S. Malmasi and M. Dras, "Native language identification with classifier stacking and ensembles," *Computational Linguistics*, vol. 44, no. 3, pp. 403–446, 2018.

[46] N. Chaudhary, A. Abu-Odeh, I. Karaman, and R. Arróyave, "A data-driven machine learning approach to predicting stacking faulting energy in austenitic steels," *Journal of Materials Science*, vol. 52, no. 18, pp. 11048–11076, 2017.

[47] A. Gupta, R. U. Khan, V. K. Singh et al., "A novel approach for classification of mental tasks using multiview ensemble learning (MEL)," *Neurocomputing*, vol. 417, no. 1, pp. 558–584, 2020.