Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 20(11): 19133–19151.
DOI: 10.3934/mbe.2023845
Received: 15 August 2023
Revised: 25 September 2023
Accepted: 26 September 2023
Published: 13 October 2023

Research article

Integrative approach for classifying male tumors based on DNA methylation 450K data

Ji-Ming Wu[1], Wang-Ren Qiu[1,*], Zi Liu[1], Zhao-Chun Xu[1] and Shou-Hua Zhang[2]

[1] Computer Department, Jing-De-Zhen Ceramic University, Jingdezhen 333403, China
[2] Department of General Surgery, Jiangxi Provincial Children's Hospital, Nanchang 330006, China

* **Correspondence:** Email: qiuone@163.com.

Abstract: Malignancies such as bladder urothelial carcinoma, colon adenocarcinoma, liver hepatocellular carcinoma, lung adenocarcinoma and prostate adenocarcinoma significantly impact men's well-being. Accurate cancer classification is vital in determining treatment strategies and improving patient prognosis. This study introduced an innovative method that utilizes gene selection from high-dimensional datasets to enhance the performance of the male tumor classification algorithm. The method assesses the reliability of DNA methylation data to distinguish the five most prevalent types of male cancers from normal tissues by employing DNA methylation 450K data obtained from The Cancer Genome Atlas (TCGA) database. First, the chi-square test is used for dimensionality reduction and second, L1 penalized logistic regression is used for feature selection. Furthermore, the stacking ensemble learning technique was employed to integrate seven common multiclassification models. Experimental results demonstrated that the ensemble learning model utilizing multiple classification models outperformed any base classification model. The proposed ensemble model achieved an astonishing overall accuracy (ACC) of 99.2% in independent testing data. Moreover, it may present novel ideas and pathways for the early detection and treatment of future diseases.

Keywords: cancer; methylation; multiclassification; ensemble learning; stacking

1. Introduction

Cancer is a serious disease that profoundly affects human physical and mental health. According to the International Agency for Research on Cancer (IARC) of the World Health Organization, approximately 19.3 million people worldwide were diagnosed with cancer in 2020 [1], with over half being male patients. Moreover, male-specific tumors [2] (such as prostate adenocarcinoma) have garnered significant attention due to their high incidence rates and their impact on men's health. There are notable differences in the global occurrence rates of male cancers [3]. Accurately classifying these cancers is one of the fundamental strategies to furnish clinical decision-making information and reduce the mortality rates of male cancers [4]. Among these, bladder urothelial carcinoma [5], colon adenocarcinoma [6], liver hepatocellular carcinoma [7], lung adenocarcinoma [8] and prostate adenocarcinoma [9] are prevalent cancers in males. The incidence of these prevalent cancers in men, among which prostate adenocarcinoma is one of the most common, increases with age. Bladder urothelial carcinoma and colon adenocarcinoma are usually associated with diet, lifestyle and genetic factors. On the other hand, liver hepatocellular carcinoma is primarily associated with hepatitis B and C virus infection, while lung adenocarcinoma is associated with smoking and exposure to airborne pollutants. These cancers' high incidence and mortality rates pose considerable threats to human health and life. Hence, cancer classification is crucial in selecting appropriate treatment strategies and improving the patient's prognosis. DNA methylation analysis has emerged as a promising tool for cancer classification [10], providing valuable insights into tumor biology and revealing potential therapeutic targets.

DNA methylation is the process of covalently modifying DNA by adding a methyl group to cytosine residues located in CpG dinucleotide contexts without altering the DNA sequence itself [11]. This process is critical in regulating gene expression, maintaining genomic stability and silencing transposable elements [12]. Increasing evidence suggests that abnormal DNA methylation patterns are associated with many diseases [13], especially cancer. Specifically, abnormal DNA methylation patterns in CpG island promoter regions [14] can lead to an increased loss of control of gene expression and genomic instability, thus promoting tumor initiation and progression. It is noteworthy that DNA methylation analysis has become an effective tool for cancer classification primarily because this technique can provide comprehensive information on the methylation status of individual CpG sites [15]. Consequently, it can accurately identify differential methylation patterns between normal and tumor tissues, making it an essential tool for cancer diagnosis and classification.

In recent years, high-throughput sequencing technology [16] has emerged as one of the most crucial tools in cancer research. DNA methylation data, which are closely associated with cancer development, are one of the types of data analyzed using this technology. With the continuous advancement of sequencing technology and computer processing capabilities, an increasing amount of large-scale DNA methylation data has been amassed. The challenge is now to extract useful information from these data and classify cancer, a critical issue in current cancer research. In addition to integrating multiple high-throughput sequencing data, artificial intelligence technology has also been widely used in cancer research. For instance, deep learning algorithms can be utilized to automate tasks and improve work efficiency in cancer diagnosis and treatment. For example, Mohammed et al. [17] used multiple One-Dimensional Convolutional Neural Network (1D-CNN) models stacked together to classify five types of cancer based on The Cancer Genome Atlas (TCGA) RNA-seq data. Jia et al. [18] proposed a method that combines variance selection with recursive feature elimination, successfully selecting 20 optimal features from over 480,000 dimensions of DNA methylation data. They compared the performance of four different estimators and five classifiers and achieved an accuracy of over 93%. Furthermore, Lin et al. [19] developed a new cancer prediction model, iCancer-Pred, utilizing deep neural networks. This model can classify seven different cancer datasets obtained from the TCGA Hub database on the University of California Santa Cruz (UCSC) XENA platform [19,20]. The authors compared this method with machine learning techniques such as support vector machines (SVM), logistic regression (LR) and random forest (RF). By means of 5-fold cross-validation, they achieved the highest accuracy of the model to be up to 97%. Although several existing studies have made significant progress in cancer classification using various models, there is still a need to overcome model limitations and improve the overall performance of cancer classification.

In this study, we propose an ensemble learning-based classification algorithm called Stacking for classifying male tumors. Specifically, we utilize the chi-square test and L1 regularity based logistic regression to select features highly associated with the characteristics of the cancer dataset. Subsequently, we devised an ensemble learning algorithm to distinguish the five most common cancers in males and their corresponding normal tissues. Stacking has been tested on DNA methylation 450K cancers data set, where the results demonstrated a significant advantage in the accurate classification of cancer. In addition, this study explores the relationship between potential genes and the survival rates of these five common cancers through gene ontology analysis, survival analysis, literature review and other related methods. Our findings suggest that the SRC gene is associated with bladder urothelial carcinoma survival, while RPS2, RPL23A, RPL22, RPL27 and SRC genes are related to liver hepatocellular carcinoma survival. Furthermore, KRAS gene is associated with lung adenocarcinoma survival, and SRC gene is associated with prostate adenocarcinoma survival. These discoveries may assist in the early identification and precise categorization of these cancer types, while also pinpointing potential treatment approaches to enhance the survival rates among high-risk males.

2. Materials and methods

2.1. Data collection and preprocessing

UCSC XENA is one of the websites derived from the TCGA database. The site stores several large public datasets on cancer, including TCGA, GETX and TARGET, among others with powerful and intuitive functionality.

The DNA methylation 450K data used in this study were downloaded exclusively from the UCSC XENA platform, which included datasets for bladder urothelial carcinoma (BLCA), colon adenocarcinoma (COAD), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD) and prostate adenocarcinoma (PRAD). A total of 2241 samples of both cancer and normal tissue were obtained by combining these five datasets, as shown in Table 1. The dataset was then divided into a training set and a testing set at a ratio of 9:1.

Table 1. Number of samples per type of cancer.

Cancer tumor	Number of samples	Training (≈90%)	Testing (≈10%)
BLCA	434	390	44
COAD	337	303	34
LIHC	429	386	43
LUAD	492	443	49
PRAD	549	494	55
Total	**2241**	**2016**	**225**

2.2. Feature selection method

Given the high-dimensional nature of the data in this study, with more than 480,000 dimensions, the sample size seems somewhat limited. However, it is essential to note that not all features hold equal importance for the classification model. Therefore, it is crucial to identify and select the most informative features to ensure accurate and effective cancer classification. This task is achieved through feature selection and dimensionality reduction, where representative features are selected. In the training dataset, features containing "NaN" were removed, and in the test dataset, they were substituted with 0.

To select features relevant to the five common tumor classifications, the chi-square test was initially employed for feature selection. The chi-square test [21] is a statistical method used to evaluate the independence between categorical variables. It is employed to assess the significance of each feature in predicting the target variable. We can determine the association between features and cancer classifications by utilizing the chi-square test, thereby selecting the crucial features. Specifically, we used the SelectKBest [22] function from the scikit-learn library [23] to filter out features with top chi-square scores. Based on this, through a cross-validation approach, we determined that the performance of the chi-square test was significantly improved at a feature count of 22,120. The features selected by the chi-square test are highly relevant to cancer classification tasks [24] and thus are of significant importance. Consequently, these features were utilized as input features for subsequent classifier training and testing. The formula used is as follows:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}. \tag{1}$$

O_{ij} denotes the observed value of the cross term in row i, column j; E_{ij} represents the expected value of the cross term in row i, column j; n denotes the number of rows; m denotes the number of columns.

Although approximately 400,000 CpG sites were removed from the cancer dataset using chi-square testing, which significantly reduced the number of sample features, it is still necessary to further reduce the number of features to construct a high-performance predictor. Feature selection can assist in reducing model complexity [25], which minimizes the risk of overfitting and enhances model interpretability and explainability. By selecting features with strong predictive power for the target variable, feature selection can improve the predictive performance of the model [26]. Additionally, feature selection can decrease data processing and modeling time and expenses.

To accomplish this aim, we employed a logistic regression model based on the L1 parametric penalty [27], and the SelectFromModel function of the scikit-learn library was used to filter features. "L1" refers to L1 regularization, which is a regularization technique used in machine learning models like linear regression and logistic regression [28]. This approach helps identify crucial features of the classification task by penalizing the model's complexity, thus preventing overfitting. These features not only enhance the model's performance and predictive power but also its interpretability and practical application value. In practical applications, we can perform more nuanced feature selection and optimization based on the significance and weight of these features to further improve the model's performance and application. The L1 regularization method naturally possesses feature selection properties because of its sparse solution characteristic.

The logistic regression model computes the probability of a data point belonging to a specific class based on a linear combination of input features [29]. The fundamental logistic regression formula without regularization is:

$$p(y = n|x) = \frac{1}{(1 + e^{-(w_0+w_1x_1+w_2x_2+\cdots+w_nx_n)})}. \tag{2}$$

Here, $p(y = n|x)$ denotes the probability of data point x belonging to class n, w_i represents the weights for each feature x_i and exp is the exponential function.

When L1 regularization is applied, the objective function is the sum of the log loss to be minimised and the L1 regularization term, which is the absolute sum of the weights. The L1 regularization term is added with a regularization strength parameter λ. The objective function is:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[-y\,log\big(p(y = n|x)\big) - (1 - y)\,log\big(1 - p(y = n|x)\big)\right] + \lambda\sum_{i=1}^{m}|w_i|, \tag{3}$$

where y is the true label of the data points, and the L1 regularized logistic regression model is derived by minimizing the objective function $J(\theta)$ with respect to the weight w_i.

3. Model architecture

3.1. Overall process

Stacking is an ensemble method for models [30], where the combination of multiple weaker models often yields better performance than a single strong model. This approach involves training several base learners and using their predictions as input to a meta-learner. The stacked ensemble algorithm offers superior performance, generalization capabilities and flexibility compared to individual algorithms by leveraging the advantages of multiple base learners to enhance model accuracy and robustness. In this study, we propose a framework combining a chi-square test, logistic regression with L1 penalty and stacking ensemble learning to construct a multiclass classifier for five types of cancer data. The overall flowchart of this study is presented in Figure 1.

The approach has trained seven base classification models: random forest (RF), support vector machine (SVM), bootstrap aggregated algorithm (Bagging), stochastic gradient descent (SGD), multilayer perceptron (MLP), logistic regression (LR) and LightGBM (LGBM) [31–34]. The reason for selecting these models is that they are based on different algorithms and can capture different data features. The LR, SVM and SGD models are linear models, the RF model captures nonlinear relationships and interactions, Bagging and LighTGBM capture nonlinear relationships by boosting weak learners and MLP can solve linearly inseparable problems. In this study, the integrated algorithm is designed to leverage the strengths of multiple models more effectively than a single algorithm. This approach aims to enhance performance robustness and accuracy.

All base learners were trained on the whole training set and then evaluated with the validation set. These predictions were used to train the meta-learner along with the true labels in the validation set. For the meta-learner, we chose the LGBM model. The specific prediction process is shown in Figure 2.

In summary, stacking is an effective method of ensemble learning that combines multiple models to achieve higher performance than any single model [35]. By leveraging different base learners and using a meta-learner to find the best way to combine them, stacking can produce better results compared to any single model. Our experimental performance demonstrated the effectiveness of the stacking approach for this classification task.

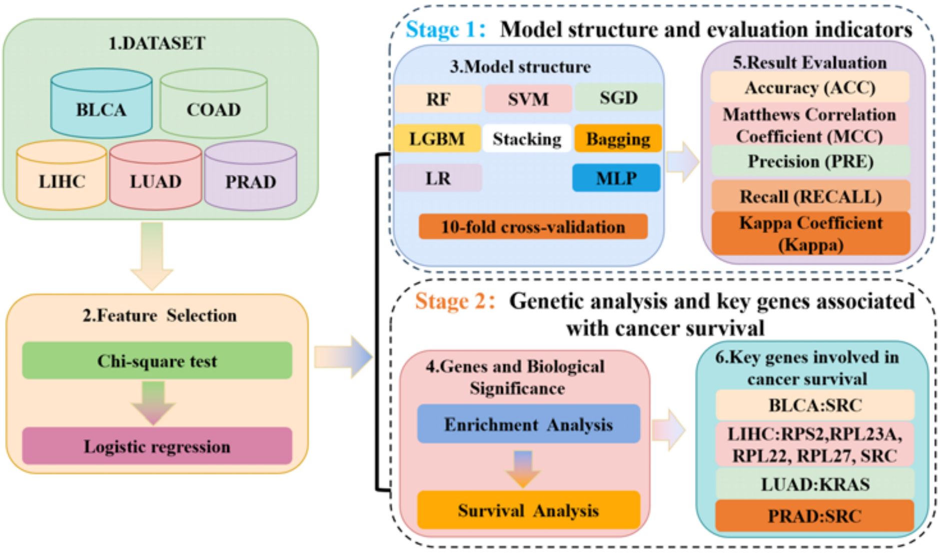
1.DATASET
BLCA
COAD
LIHC
LUAD
PRAD
2.Feature Selection
Chi-square test
Logistic regression
Stage 1: Model structure and evaluation indicators
3.Model structure
RF
SVM
SGD
LGBM
Stacking
Bagging
LR
MLP
10-fold cross-validation
5.Result Evaluation
Accuracy (ACC)
Matthews Correlation Coefficient (MCC)
Precision (PRE)
Recall (RECALL)
Kappa Coefficient (Kappa)
Stage 2: Genetic analysis and key genes associated with cancer survival
4.Genes and Biological Significance
Enrichment Analysis
Survival Analysis
6.Key genes involved in cancer survival
BLCA:SRC
LIHC:RPS2,RPL23A, RPL22, RPL27, SRC
LUAD:KRAS
PRAD:SRC

Figure 1. Overall workflow diagram.

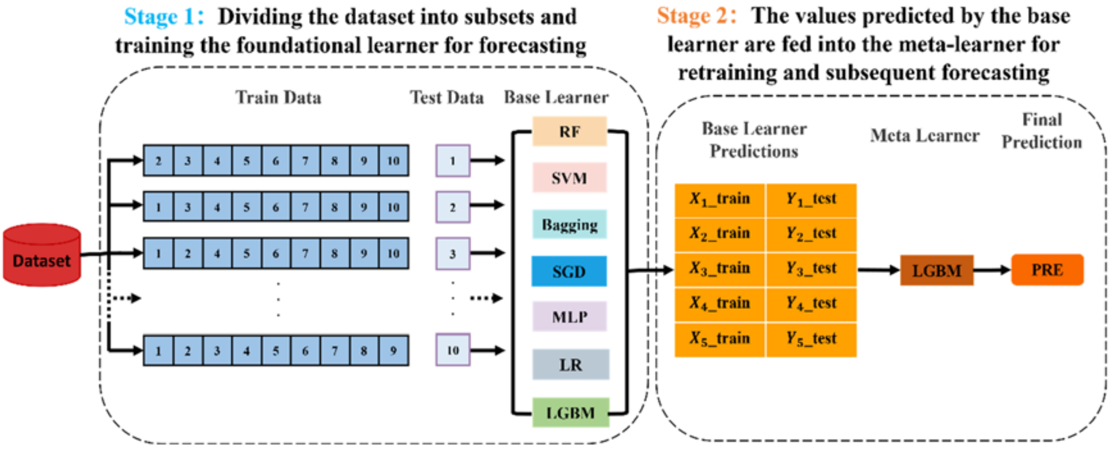

Figure 2. Stacking ensemble modeling framework.

3.2. Classifier performance evaluation

Scientific evaluation metrics are crucial to the performance metrics of a model. We usually use a variety of evaluation metrics to measure the performance of a model, such as accuracy and recall. These metrics can not only help us understand the predictive ability of the model but also help us optimize the parameters and structure of the model to improve its performance. In this study, the evaluation of model performance contains five metrics: Accuracy (ACC), Matthews Correlation Coefficient (MCC) [36], Precision (PRE), Geometric mean (Gmean), Recall (RECALL) and Kappa Coefficient (KAPPA) [37].

$$\begin{cases} ACC = \frac{1}{6}\sum_{i=0}^{5}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \\ MCC = \frac{1}{6}\sum_{i=0}^{5}\frac{TP_i \times TN_i - FP_i \times FN_i}{(TP_i + FP_i)(TP_i + +FN_i)(TN_i + FP_i)(TN_i + FN_i)} \\ PRE = \frac{1}{6}\sum_{i=0}^{5}\frac{TP_i}{TP_i + FP_i} \\ REC = \frac{1}{6}\sum_{i=0}^{5}\frac{TP_i}{TP_i + FN_i} \\ Gmean = \frac{1}{6}\sum_{i=0}^{5}\sqrt{\frac{TP_i \times TN_i}{(TP_i + FP_i)(TP_i + FN_i)}} \\ KAPPA = \frac{p_0 - p_e}{1 - p_e} \end{cases} . \quad (4)$$

In this context, TP (True Positive) represents the true positives, indicating the number of times the model predicted the positive class correctly; TN (True Negative) represents the true negatives, indicating the number of times the model predicted the negative class correctly; FP (False Positive) represents the false positives, indicating the number of times the model predicted the negative class as positive; FN (False Negative) represents the false negatives, indicating the number of times the model predicted the positive class as negative. The kappa coefficient is a measure of agreement between a classifier and human classification. It compares the observed classification accuracy with the chance agreement. p_o is the observed classification accuracy, and p_e is the expected classification accuracy by chance. They can be expressed as follows:

$$p_o = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}, \quad (5)$$

$$p_e = \sum_{i=0}^{5}\frac{(TP_i + FN_i)(TP_i + FP_i) + (FN_i + TN_i)(FP_i + TN_i)}{(TP_i + TN_i + FP_i + FN_i)^2}, \quad (6)$$

where TP_i, TN_i, FP_i, FN_i (i=0,1,2...,5) are TP, TN, FP and FN for each subset, respectively.

3.3. Results

During model training, we performed 10-fold cross-validation on the training set to optimize the parameters. In this, 10-fold cross-validation is performed on the training dataset. The original dataset is first divided into training and test sets. Then, we further divide the training set into ten equal-sized subsets for cross-validation. In each cross-validation, one subset is used as the validation set and the remaining nine subsets are used to train the model, ensuring that each subset has acted as a validation set.

Selecting an effective feature selection method improves the performance of predictive models and obtains better explanatory power. For this purpose, a comprehensive comparison of various feature selection methods was performed and recursive feature elimination (RFE) [38], elastic network (ENET) [39], and a combination of logistic regression with chi-square test based on L1 regularity were considered.

During the comparison process, we observed that the combined methods exhibit promising performance in the feature selection task. The results presented in Table 2 enable us to clearly compare the variations in performance among these methods for the prediction task. In the fourth row of Table 2, "99.22 ± 0.004" indicates that in the 10-fold cross-validation, the value of ACC is 99.22 and the variance is 0.004.

Building upon these findings, the chi-square test proves to be highly valuable in categorization problems as it allows us to identify features that are significantly associated with the target variable, potentially related to cancer in this study. As for the logistic regression method based on L1 regularization, it induces sparsity in the feature coefficients by applying L1 regularization, and this sparsity helps to filter out irrelevant or redundant features, thus improving the generalization ability of the model. Consequently, we opted for the combination method to screen features and have continued utilizing this approach in subsequent studies.

Table 2. 10-fold cross-validation results for different feature methods.

Method	ACC (%)	PRE (%)	RECALL (%)	MCC (%)	Gmean (%)	KAPPA (%)
RFE	98.36 ± 0.010	97.95 ± 0.013	97.74 ± 0.014	98.01 ± 0.012	98.01 ± 0.012	99.03 ± 0.006
ENET	97.97 ± 0.009	97.07 ± 0.010	96.18 ± 0.016	96.81 ± 0.011	96.79 ± 0.012	98.40 ± 0.006
Ours	**99.22** ± 0.004	**99.24** ± 0.004	**99.23** ± 0.004	**99.07** ± 0.005	**99.60** ± 0.002	**99.07** ± 0.005

Afterwards, we conducted a 10-fold cross-validation of all classification methods. The statistical results are shown in Tables 3 and 4. The performance of the stacking ensemble learning method was found to be superior to those of the base learners while also exhibiting good stability. Due to the instability of independent testing results at each training, we took the average of 10 results in the experiment. The data achieved very good results on the stacking ensemble learning model, with over 99% in all criteria except RECALL, as shown in Table 4 and Figure 3(b).

Table 3. 10-fold cross-validation results.

Method	ACC (%)	PRE (%)	RECALL (%)	MCC (%)	Gmean (%)	KAPPA (%)
MLP	97.52 ± 0.010	96.85 ± 0.014	97.01 ± 0.013	96.99 ± 0.013	98.54 ± 0.006	96.98 ± 0.013
Bagging	97.62 ± 0.015	96.81 ± 0.020	97.48 ± 0.017	97.13 ± 0.018	98.63 ± 0.009	97.10 ± 0.018
LR	98.01 ± 0.011	97.37 ± 0.015	97.78 ± 0.019	97.60 ± 0.013	98.85 ± 0.006	97.59 ± 0.013
SVM	97.87 ± 0.011	97.17 ± 0.016	97.67 ± 0.012	97.42 ± 0.014	98.76 ± 0.006	97.41 ± 0.014
RF	97.77 ± 0.012	97.03 ± 0.016	97.28 ± 0.014	97.30 ± 0.014	98.70 ± 0.007	97.28 ± 0.014
SGD	97.42 ± 0.013	96.87 ± 0.014	96.93 ± 0.021	96.91 ± 0.015	98.49 ± 0.008	96.86 ± 0.016
LGBM	98.11 ± 0.013	97.56 ± 0.017	97.62 ± 0.015	97.71 ± 0.016	98.89 ± 0.008	97.70 ± 0.016
Stacking	**99.22** ± 0.004	**99.24** ± 0.004	**99.23** ± 0.004	**99.07** ± 0.005	**99.60** ± 0.002	**99.07** ± 0.005

Table 4. Average results of 10 independent tests.

Method	ACC (%)	PRE (%)	RECALL (%)	MCC (%)	Gmean (%)	KAPPA (%)
MLP	98.22 ± 0.003	97.20 ± 0.009	97.25 ± 0.006	97.82 ± 0.004	98.94 ± 0.002	97.82 ± 0.004
Bagging	98.58 ± 0.009	96.89 ± 0.004	97.62 ± 0.001	98.26 ± 0.002	99.20 ± 0.001	98.25 ± 0.002
LR	98.22 ± 0.000	96.73 ± 0.000	97.32 ± 0.000	97.83 ± 0.000	98.97 ± 0.000	97.82 ± 0.000
SVM	98.67 ± 0.000	97.09 ± 0.000	97.69 ± 0.000	98.37 ± 0.000	99.25 ± 0.000	98.36 ± 0.000
RF	98.36 ± 0.005	96.98 ± 0.011	97.35 ± 0.004	97.99 ± 0.005	99.05 ± 0.002	97.98 ± 0.006
SGD	97.73 ± 0.008	97.10 ± 0.016	96.22 ± 0.012	97.24 ± 0.009	98.63 ± 0.005	97.22 ± 0.010
LGBM	99.11 ± 0.000	98.26 ± 0.000	98.07 ± 0.000	98.91 ± 0.000	99.48 ± 0.000	98.91 ± 0.000
Stacking	**99.29** ± 0.004	**99.21** ± 0.007	**98.20** ± 0.007	**99.13** ± 0.005	**99.56** ± 0.002	**99.13** ± 0.005

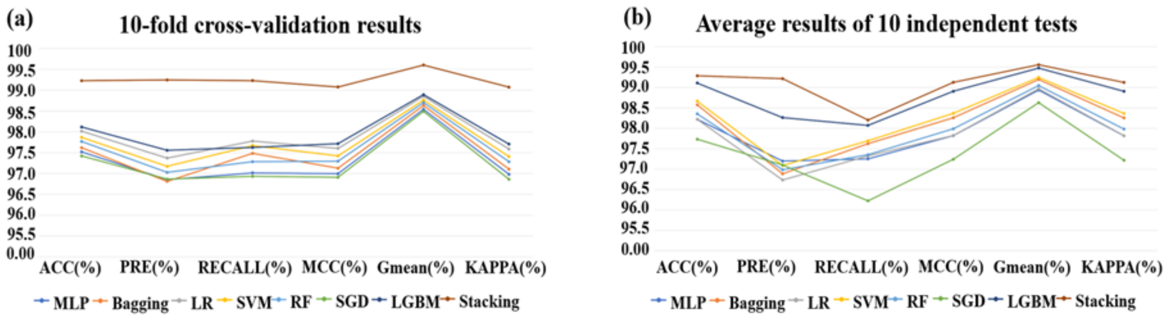

Figure 3. 10-fold cross-validation results and average results of 10 independent test line graphs.

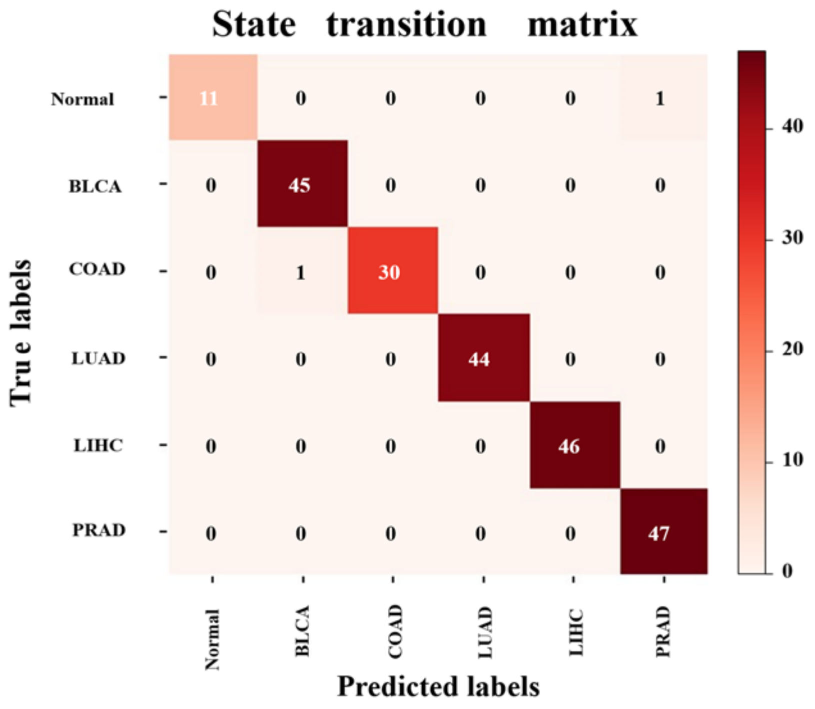

Figure 4. Confusion matrix for independent testing of multiclass predictors.

The confusion matrix [40] is shown in Figure 4, and it can be seen that the model performs well in distinguishing between the five types of cancer and normal tissue. It can also be observed that out of all the samples in the independent testing data, only two were misclassified, where one normal tissue sample was incorrectly predicted as a PRAD sample, and another COAD sample was incorrectly predicted as a BLCA sample.

This result is satisfactory, which indicates that the model has high accuracy and reliability and can be used in clinical practice. At the same time, although the misclassification rate is low, we still need to continue to optimize and improve the model to improve its accuracy and applicability in future research.

To validate the generalizability of the proposed model, we utilized the dataset and neural network employed by Lin et al. [19] to assess the performance of our classifier in this study. By comparing the performance of our model with their dataset and methods, our model consistently outperforms theirs, as demonstrated in Tables 5 and 6. These results indicate that our model excels not only on the original dataset but can also be successfully applied to other datasets with a degree of generality and replicability. This, in turn, enhances the reliability and stability of the model for practical medical applications.

Table 5. Comparison of results using our dataset with the iCancer-Pred approach.

Method	ACC (%)	PRE (%)	RECALL (%)	MCC (%)	Gmean (%)	KAPPA (%)
ICancer-Pred	83.56 ± 0.199	78.49 ± 0.229	81.81 ± 0.179	82.50 ± 0.194	80.16 ± 0.236	89.43 ± 0.135
Stacking	**99.29** ± 0.004	**99.29** ± 0.007	**98.20** ± 0.007	**99.13** ± 0.005	**99.56** ± 0.002	**99.13** ± 0.005

Table 6. Comparison of results using iCancer-Pred’s dataset with our approach.

Method	ACC (%)	PRE (%)	RECALL (%)	MCC (%)	Gmean (%)	KAPPA (%)
ICancer-Pred	97.27 ± 0.006	97.37 ± 0.005	96.99 ± 0.007	96.82 ± 0.007	—	96.81 ± 0.007
Stacking	**98.22** ± 0.006	**98.18** ± 0.007	**97.96** ± 0.009	**97.93** ± 0.007	**98.35** ± 0.002	**97.92** ± 0.007

3.4. Explaining model predictions using LIME

To achieve interpretable predictions and gain insights into feature contributions, we utilized the local interpretable model-agnostic explanations (LIME) [41] model. In this study, we used 9511 CpG sites as features for predicting cancer types. Figure 5 shows the LIME prediction results of a sample by using the stacking integrated learning model, which screens the top 6 predictive biomarkers most helpful in classifying normal tissue, bladder urothelial carcinoma (BLCA), colon adenocarcinoma (COAD), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD) and prostate adenocarcinoma (PRAD). The prediction probability table in the top-left corner of Figure 5 shows the model’s probability of predicting a given sample as one of these types of cancer. In this case, LIME assigns a feature weight 0.20 for cg11055493 feature values less than or equal to 0.39 (cg11055493 $\leq$ 0.39).

Additionally, we demonstrated the feature weights of other predicted features. We detailed each feature’s values and color codes in the “Feature-Value” table, which specifies whether a given feature contributes to the prediction. Specifically, normal tissue is displayed in blue, bladder urothelial carcinoma (BLCA) is color coded in orange, colon adenocarcinoma (COAD) is color-coded in green, liver hepatocellular carcinoma (LIHC) is color coded in purple, lung adenocarcinoma (LUAD) is color coded in red and prostate adenocarcinoma (PRAD) is color coded in brown.

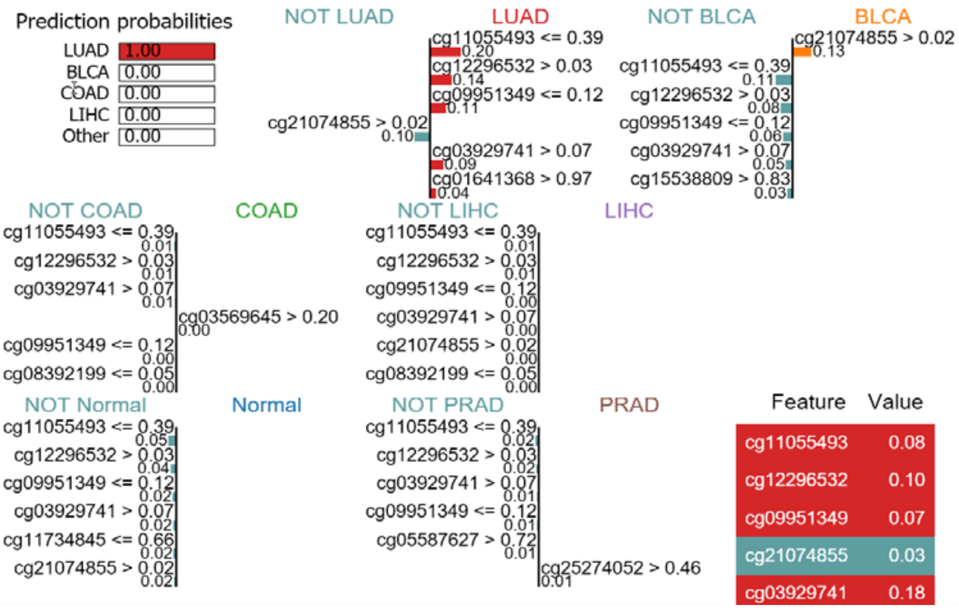

Figure 5. LIME results of 6 major biomarkers illustrated using stacking ensemble learning classifier for normal tissue (Normal), bladder urothelial carcinoma (BLCA), colon adenocarcinoma (COAD), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD) and prostate adenocarcinoma (PRAD); LIME: Local Interpretable Model-Agnostic Explanations.

4. Genes and biological significance

We obtained 9511 CpG loci by screening and annotating them into genes and finally obtained 8087 genes. By comparing these 8087 genes with published CpG biomarkers, we found that Ding et al.'s study [42] included data for the five types of cancer used in our study, as well as 3000 CpG biomarker genes. At the same time, 863 genes overlapped between the two studies, as shown in Figure 6.

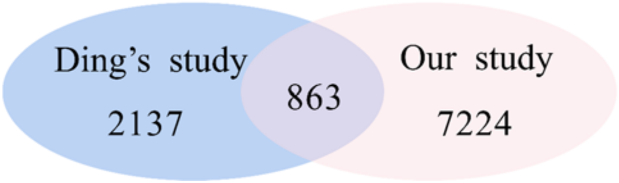

Figure 6. Venn diagram of overlapping genes between the two studies.

For the list of overlapping genes, we performed pathway and process enrichment analysis by using multiple ontology sources, including GO Biological Processes, GO Cellular Components, GO Molecular Functions and KEGG Pathway [43]. A series of criteria were applied to screen for biologically significant enrichment terms, including p values less than 0.01, minimum counts of 3 and enrichment factors greater than 1.5 (enrichment factor refers to the ratio between observed counts and randomly expected counts). Based on their membership similarity, we grouped the enriched terms into clusters and used Kappa scores as a similarity measure in the hierarchical clustering process, treating subtrees with a similarity greater than 0.3 as a cluster. Finally, in each cluster, the most significant enriched terms in terms of the above metrics (p-values, counts, etc.) were selected to represent its clusters. For example, for GO:0016570 we filtered the following metrics: count = 34, Log10(P) = -8.09, Log10(q) = -4.16. "Count" is the number of genes in the user-provided lists with membership in the given ontology term. "Log10(P)" is the p-value in log base 10. "Log10(q)" is the multi-test adjusted p-value in log base 10. The results of these enrichment analyses can help us to deeply understand the functions of these overlapping genes in different biological processes and pathways, providing important clues for further studies.

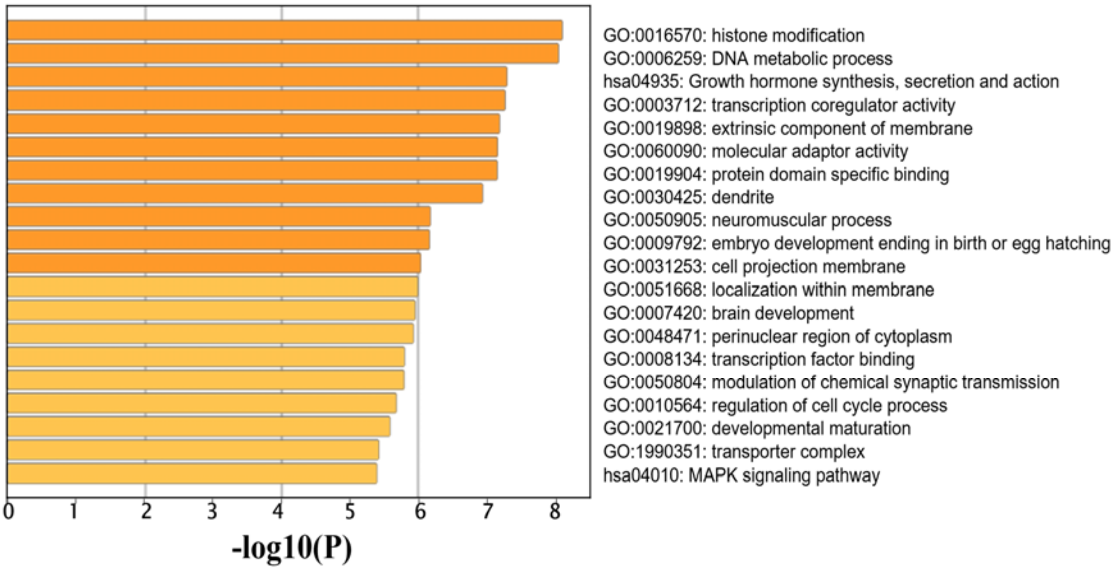

Figure 7. Heatmap of enriched terms for overlapping genes. The intensity of the color indicates the level of enrichment, with darker colors indicating higher levels of enrichment. On the right side, there is a wealth of information on terms from the Gene Ontology (GO) and KEGG Pathway that can be used to clarify the meaning and function of each enrichment term.

As shown in Figure 7, gene ontology analysis shows that overlapping genes are present in biological processes of histone modification (GO:0016570), DNA metabolic process (GO:0006259), neuromuscular process (GO:0050905), embryo development ending in birth or egg hatching (GO:0009792), localization within membrane (GO:0051668), brain development (GO:0007420), modulation of chemical synaptic transmission (GO:0050804), regulation of cell cycle process (GO:0010564), developmental maturation (GO:0021700) and other related genes; molecular functions exist in transcription coregulator activity (GO:0003712), molecular adaptor activity (GO:0060090), protein domain specific binding (GO:0019904) and transcription factor binding (GO:0008134); among cellular components, there is extrinsic component of membrane (GO:0019898), dendrite (GO:0030425), cell projection membrane (GO:0031253), perinuclear region of cytoplasm (GO:0048471) and transporter complex (GO:1990351) were enriched. In addition, growth hormone synthesis (hsa04935) secretion and action (hsa04935) and the MAPK signaling pathway (hsa04010) were identified in the KEGG pathway.

In this study, the STRING database was employed to search for potential interactions among encoded proteins and investigate their potential interactions. Through this step, we obtained a representation of the protein–protein interaction network, as shown in Figure 8. This network describes the relationships between genes and proteins, such as physical contacts and targeted regulation. Our goal was to elucidate the meaningful molecular regulatory networks in living organisms.

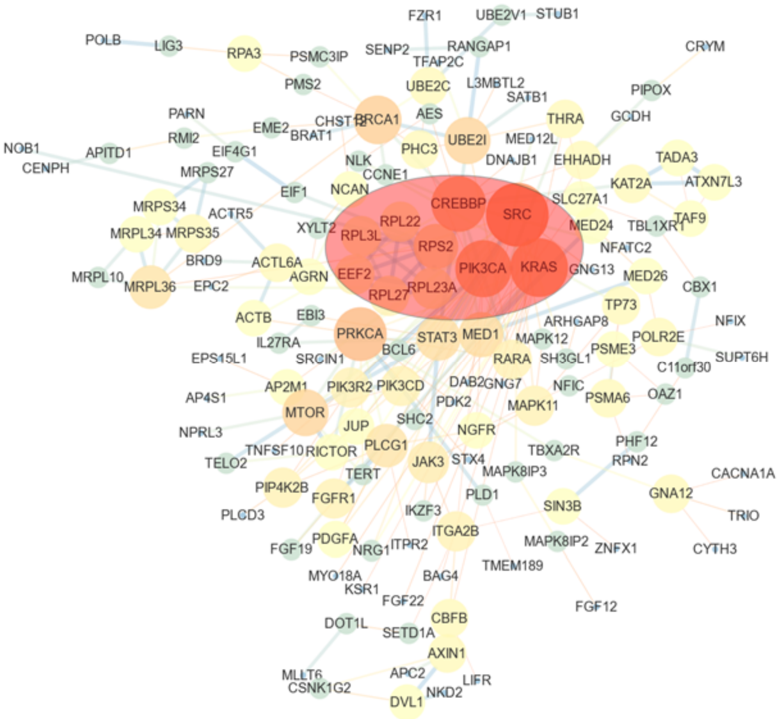

Figure 8. Protein–protein interaction networks generated by overlapping gene lists. The genes marked in red are the top 10 key genes scored by the MCC method (RPS2, RPL23A, RPL3L, RPL22, RPL27, EEF2, PIK3CA, SRC, KRAS and CREBBP).

Subsequently, the "cytoHubba" plugin in Cytoscape [44] software calculated the node scores of genes in the PPI network and identified the top 10 key genes: RPS2, RPL23A, RPL3L, RPL22, RPL27, EEF2, PIK3CA, SRC, KRAS and CREBBP (as shown in Figure 9).

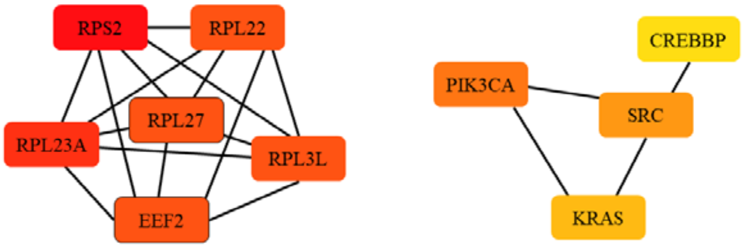

Figure 9. Key genes for scoring the top 10.

To investigate the effect of these genes on the survival of cancer patients, we performed a survival analysis of 10 potential biomarkers screened from the PPI network and used the TIMER database [45] to further draw Kaplan–Meier survival curves (Figure 10). The Kaplan-Meier survival curve was first proposed by Kaplan and Meier in 1958 [46]. It is a non-parametric method used for analyzing survival data, capable of estimating survival probabilities at different time points and visualizing the changes in survival curves. In the field of cancer research, the Kaplan-Meier survival curve is widely employed for analyzing patients' survival data [47,48]. Recent studies have shown that this method remains highly effective in predicting patient survival rates. For instance, Hamid Bakhtiari et al. [49] utilized this method to predict the survival rates of hypertensive patients with COVID-19. According to statistical significance, a gene was considered to be significantly associated with cancer survival when $P < 0.05$. Our analysis revealed that the SRC gene is significantly associated with the prognosis of bladder urothelial carcinoma, liver hepatocellular carcinoma and prostate adenocarcinoma. Additionally, the RPS2, RPL23A, RPL22, RPL27 and SRC genes are significantly associated with the prognosis of liver hepatocellular carcinoma. Furthermore, we found that the KRAS gene is significantly associated with the prognosis of lung adenocarcinoma. These results suggest that these genes have potential prognostic value. However, further clinical validation of these potential biomarkers is needed before they can be used.

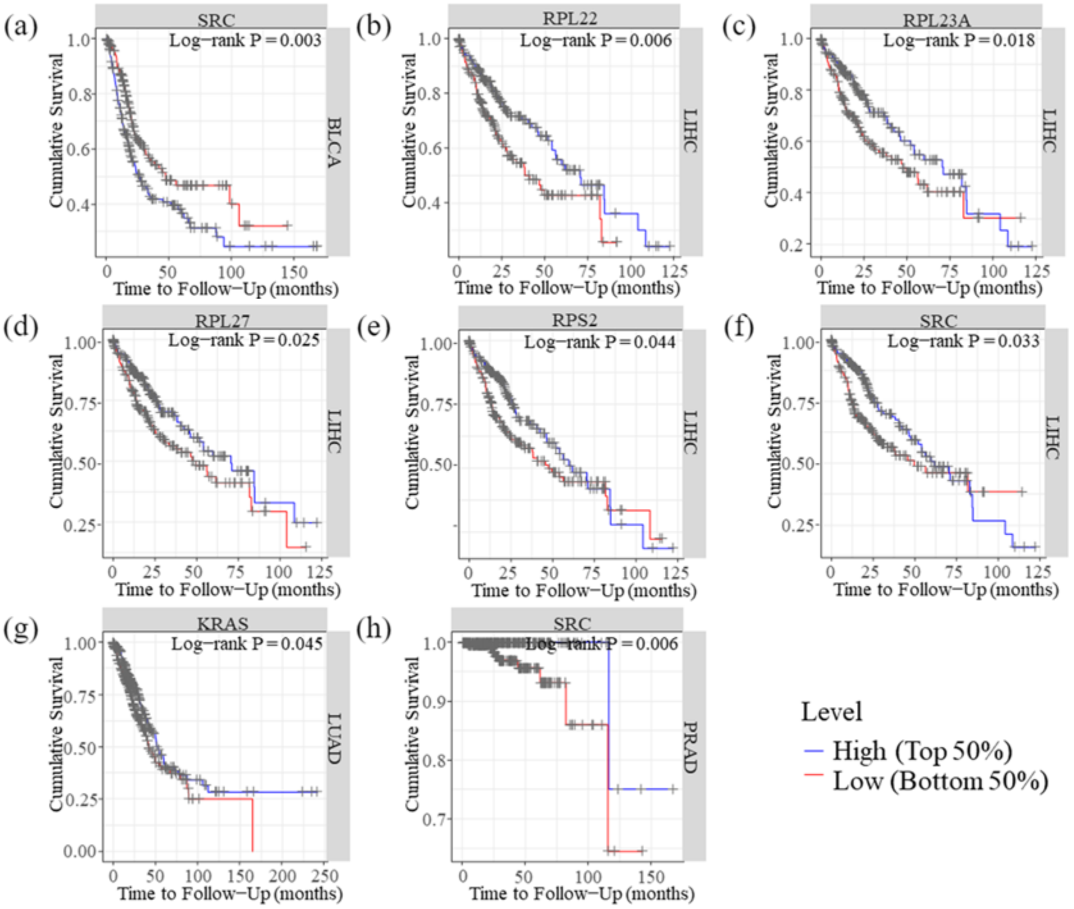

Figure 10. Kaplan–Meier curve plot showing the genes with p values less than 0.05.

5. Literature review

After conducting a comprehensive review of the literature, we found that many of the critical genes associated with cancer survival identified in this study have been previously reported in the literature. For example, Sree Karani Kondapuram et al. [50] identified SRC as a central autophagy gene associated with LIHC survival, making it a potential drug target. Kowalczyk et al. [51] discovered that RPS2 is overexpressed in mouse liver hepatocellular carcinoma samples and may impact the accuracy of mRNA translation related to aminoacyl-tRNA binding ribosomes, thus promoting cell proliferation. Additionally, Pan et al. [52] identified SRC as a potential therapeutic target for docetaxel-resistant prostate adenocarcinoma and an effective prognostic indicator that was significantly correlated with the immune score, ferroptosis, methylation and OCLR score. Wang et al. [53] found that high expression of ribosome-related genes RPL23A and RPL27 significantly reduced the survival rate of patients with liver hepatocellular carcinoma. Moreover, Xu et al. [54] identified SRC as a prognostic gene for BLCA through multivariate Cox regression analysis [55]. In lung cancer, KRAS gene mutation is most common in patients with lung adenocarcinoma, and approximately 33% of patients will have this mutation [56].

In summary, the comprehensive literature review confirms the importance of the key genes identified in this study in cancer survival. Our feature selection method has proven to be effective in extracting potential biomarkers. Furthermore, these findings provide additional evidence supporting the potential clinical relevance of our model and the importance of integrating machine learning methods into cancer research. Further research is needed to validate these findings and explore the underlying mechanisms of these genes in cancer development and progression.

6. Conclusions

In this paper, we present a novel model for predicting the types of five different cancers and their corresponding normal tissues in the DNA methylation 450K dataset. Our proposed model includes a stacked ensemble learning approach combined with a feature selection method based on a chi-square test and logistic regression with L1 regularization. This framework effectively addresses the challenges posed by the high-dimensional nature of the data. Specifically, we utilize a chi-square test for feature selection, followed by logistic regression with L1 regularization as the estimator for SelectFromModel to create an optimized feature set. These selected features are then employed in our stacking ensemble learning model for prediction. Additionally, we have taken steps to mitigate the issue of an unbalanced sample distribution between cancer samples and normal tissues by applying SMOTETomek integrated sampling to the training set.

Compared to existing methods, our proposed stacking ensemble learning model consistently performs better in classifying different cancer types. Our study establishes a robust multiclass predictor capable of identifying a patient’s cancer type. Furthermore, we have conducted survival analysis on essential genes to identify potential biomarkers associated with cancer survival, and we have performed comprehensive GO and KEGG pathway analyses to underscore the biological relevance of our findings. In conclusion, our model has great potential in the field of cancer diagnosis and treatment, highlighting the value of combining machine learning methods with DNA methylation data analysis. However, despite conducting relevant bioinformatics analyses, our study still has limitations and requires further validation and testing on a broader cancer dataset. In order to enhance the robustness of our approach, we plan to explore and integrate other types of cancer and related multi-omics data in future research efforts.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (No. 62162032, 62062043, 32270789), the Scientific Research Plan of the Department of Education of Jiangxi Province, China (GJJ2201004, GJJ2201038).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA Cancer J Clin.*, **71** (2021), 209–249. https://doi.org/10.3322/caac.21660
2. W. Wang, L. R. Meadows, J. M. den Haan, N. E. Sherman, Y. Chen, E. Blokland, et al., Human HY: a male-specific histocompatibility antigen derived from the SMCY protein, *Science*, **269** (1995), 1588–1590. https://doi.org/10.1126/science.7667640
3. K. Shibuya, C. D. Mathers, C. Boschi-Pinto, A. D. Lopez, C. J. Murray, Global and regional estimates of cancer mortality and incidence by site: II. Results for the global burden of disease 2000, *BMC Cancer*, **2** (2002), 37. https://doi.org/10.1186/1471-2407-2-37
4. A. Jemal, R. Siegel, J. Xu, E. Ward, Cancer statistics, 2010, *CA Cancer J. Clin.*, **60** (2010), 277–300. https://doi.org/10.3322/caac.20073
5. *Cancer Genome Atlas Research*, Comprehensive molecular characterization of urothelial bladder carcinoma, *Nature*, **507** (2014), 315–322. https://doi.org/10.1038/nature12965
6. J. Terzic, S. Grivennikov, E. Karin, M. Karin, Inflammation and colon cancer, *Gastroenterology*, **138** (2010), 2101–2114. https://doi.org/10.1053/j.gastro.2010.01.058
7. F. X. Bosch, J. Ribes, M. Diaz, R. Cleries, Primary liver cancer: worldwide incidence and trends, *Gastroenterology*, **127** (2004), S5–S16. https://doi.org/10.1053/j.gastro.2004.09.011
8. *Cancer Genome Atlas Research*, Comprehensive molecular profiling of lung adenocarcinoma, *Nature*, **511** (2014), 543–550. https://doi.org/10.1038/nature13385
9. P. Rawla, Epidemiology of prostate cancer, *World J. Oncol.*, **10** (2019), 63–89. https://doi.org/10.14740/wjon1191
10. P. Jurmeister, M. Leitheiser, P. Wolkenstein, F. Klauschen, D. Capper, L. Brcic, DNA methylation-based machine learning classification distinguishes pleural mesothelioma from chronic pleuritis, pleural carcinosis, and pleomorphic lung carcinomas, *Lung Cancer*, **170** (2022), 105–113. https://doi.org/10.1016/j.lungcan.2022.06.008
11. Z. D. Smith, A. Meissner, DNA methylation: roles in mammalian development, *Nat. Rev. Genet.*, **14** (2013), 204–220. https://doi.org/10.1038/nrg3354

12. P. A. Jones, Functions of DNA methylation: islands, start sites, gene bodies and beyond, *Nat. Rev. Genet.*, **13** (2012), 484–492. https://doi.org/10.1038/nrg3230

13. T. Bozic, C. C. Kuo, J. Hapala, J. Franzen, M. Eipel, U. Platzbecker, et al., Investigation of measurable residual disease in acute myeloid leukemia by DNA methylation patterns, *Leukemia*, **36** (2022), 80–89. https://doi.org/10.1038/s41375-021-01316-z

14. C. Stirzaker, D. S. Millar, C. L. Paul, P. M. Warnecke, J. Harrison, P. C. Vincent, et al., Extensive DNA methylation spanning the Rb promoter in retinoblastoma tumors, *Cancer Res.*, **57** (1997), 2229–2237.

15. I. Huh, X. Yang, T. Park, S. V. Yi, Bis-class: a new classification tool of methylation status using bayes classifier and local methylation information, *BMC Genomics*, **15** (2014), 608. https://doi.org/10.1186/1471-2164-15-608

16. J. Jo, J. Oh, C. Park, Microbial community analysis using high-throughput sequencing technology: a beginner's guide for microbiologists, *J. Microbiol.*, **58** (2020), 176–192. https://doi.org/10.1007/s12275-020-9525-5

17. M. Mohammed, H. Mwambi, I. B. Mboya, M. K. Elbashir, B. Omolo, A stacking ensemble deep learning approach to cancer type classification based on TCGA data, *Sci. Rep.*, **11** (2021), 15626. https://doi.org/10.1038/s41598-021-95128-x

18. S. Jia, Y. Zhang, Y. Mao, J. Gao, Y. Chen, Y. Jiang, et al., A new parsimonious method for classifying Cancer Tissue-of-Origin Based on DNA Methylation 450K data, preprint, arXiv:2101.00570. https://doi.org/10.48550/arXiv.2101.00570

19. W. Lin, S. Hu, Z. Wu, Z. Xu, Y. Zhong, Z. Lv, et al., iCancer-Pred: A tool for identifying cancer and its type using DNA methylation, *Genomics*, **114** (2022), 110486. https://doi.org/10.1016/j.ygeno.2022.110486

20. M. J. Goldman, B. Craft, M. Hastie, K. Repecka, F. McDade, A. Kamath, et al., Visualizing and interpreting cancer genomics data via the Xena platform, *Nat. Biotechnol.*, **38** (2020), 675–678. https://doi.org/10.1038/s41587-020-0546-8

21. N. Pandis, The chi-square test, *Am. J. Orthod. Dentofacial Orthop.*, **150** (2016), 898–899. https://doi.org/10.1016/j.ajodo.2016.08.009

22. T. Desyani, A. Saifudin, Y. Yulianti, Feature selection based on naive bayes for caesarean section prediction, *IOP Conf. Ser.: Mater. Sci. Eng.*, **879** (2020), 01209. https://doi.org/10.1088/1757-899X/879/1/012091

23. A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, et al., Machine learning for neuroimaging with scikit-learn, *Front. Neuroinf.*, **8** (2014), 14. https://doi.org/10.3389/fninf.2014.00014

24. M. Wimmer, G. Sluiter, D. Major, D. Lenis, A. Berg, T. Neubauer, et al., Multi-task fusion for improving mammography screening data classification, *IEEE Trans. Med. Imaging*, **41** (2022), 937–950. https://doi.org/10.1109/TMI.2021.3129068

25. P. Khumprom, D. Grewell, N. Yodo, Deep neural network feature selection approaches for data-driven prognostic model of aircraft engines, *Aerospace*, **7** (2020), 132. https://doi.org/10.3390/aerospace7090132

26. H. Kaneko, Examining variable selection methods for the predictive performance of regression models and the proportion of selected variables and selected random variables, *Heliyon*, **7** (2021), e07356. https://doi.org/10.1016/j.heliyon.2021.e07356

27. H. Gao, H. Zhao, Multilevel bioluminescence tomography based on radiative transfer equation Part 1: l1 regularization, *Opt. Express*, **18** (2010), 1854–1871. https://doi.org/10.1364/OE.18.001854
28. P. Ravikumar, M. J. Wainwright, J. D. Lafferty, High-dimensional Ising model selection using ℓ_1-regularized logistic regression, *Ann. Statist.*, **38** (2010), 1287–1319. https://doi.org/10.1214/09-aos691
29. K. Shah, H. Patel, D. Sanghvi, M. Shah, A comparative analysis of logistic regression, random forest and KNN models for the text classification, *Augment. Hum. Res.*, **5** (2020). https://doi.org/10.1007/s41133-020-00032-0
30. Y. Wang, D. Wang, D. Geng, Y. Wang, Y. Yin, Y. Jin, Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection, *Appl. Soft Comput.*, **77** (2019), 188–204. https://doi.org/10.1016/j.asoc.2019.01.015
31. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1007/978-1-4419-9890-3_12
32. C. J. C. Burges, K. Discovery, A tutorial on support vector machines for pattern recognition, *Data Min. Knowl. Discov*., **2** (1998), 121–167. https://doi.org/10.1023/A:1009715923555
33. R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, *IJCAI*, **7** (1995), 1137–1143. https://dl.acm.org/doi/10.5555/1643031.1643047
34. B. Recht, C. Re, S. Wright, F. Niu, Hogwild!: A lock-free approach to parallelizing stochastic gradient descent, *Adv. Neural Inf. Process. Syst.*, **24** (2011), 693–701. https://doi.org/10.48550/arXiv.1106.5730
35. S. Cui, Y. Yin, D. Wang, Z. Li, Y. Wang, A stacking-based ensemble learning method for earthquake casualty prediction, *Appl. Soft Comput.*, **101** (2021). https://doi.org/10.1016/j.asoc.2020.107038
36. S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric, *PLoS One*, **12** (2017), e0177678. https://doi.org/10.1371/journal.pone.0177678
37. T. S. Tsou, A robust likelihood approach to inference about the kappa coefficient for correlated binary data, *Stat. Methods Med. Res.*, **28** (2019), 1188–1202. https://doi.org/10.1177/0962280217751519
38. L. Li, W. K. Ching, Z. P. Liu, Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods, *Comput. Biol. Chem.*, **100** (2022), 107747. https://doi.org/10.1016/j.compbiolchem.2022.107747
39. H. Zou, T. Hastie, Regularization and variable selection via the elastic nets, *J. R. Stat. Soc. Series B Stat. Methodol.*, **67** (2015), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
40. T. P. Hettinger, J. F. Gent, L. E. Marks, M. E. Frank, A confusion matrix for the study of taste perception, *Percept. Psychophys.*, **61** (1999), 1510–1521. https://doi.org/10.3758/bf03213114
41. I. Palatnik de Sousa, M. Maria Bernardes Rebuzzi Vellasco, E. Costa da Silva, Local interpretable model-agnostic explanations for classification of lymph node metastases, *Sensors (Basel)*, **19** (2019). https://doi.org/10.3390/s19132969
42. S. Ding, H. Li, Y. H. Zhang, X. Zhou, K. Feng, Z. Li, et al., Identification of pan-cancer biomarkers based on the gene expression profiles of cancer cell lines, *Front. Cell Dev. Biol.*, **9** (2021), 781285. https://doi.org/10.3389/fcell.2021.781285
43. Y. H. Zhang, T. Zeng, L. Chen, T. Huang, Y. D. Cai, Determining protein–protein functional associations by functional rules based on gene ontology and KEGG pathway, *Biochim. Biophys. Acta Proteins Proteom.*, **1869** (2021), 140621. https://doi.org/10.1016/j.bbapap.2021.140621

44. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.*, **13** (2003), 2498–2504. https://doi.org/10.1101/gr.1239303
45. T. Li, J. Fan, B. Wang, N. Traugh, Q. Chen, J. S. Liu, et al., TIMER: A web server for comprehensive analysis of tumor-infiltrating immune cells, *Cancer Res.*, **77** (2017), e108–e110. https://doi.org/10.1158/0008-5472.CAN-17-0307
46. E. L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, *J. Am. Stat. Assoc.*, **53** (1958), 457–481. https://doi.org/10.1080/01621459.1958.10501452
47. K. J. Jager, P. C. van Dijk, C. Zoccali, F. W. Dekker, The analysis of survival data: the Kaplan-Meier method, *Kidney Int.*, **74** (2008), 560–565. https://doi.org/10.1038/ki.2008.217
48. P. Guyot, A. E. Ades, M. J. Ouwens, N. J. Welton, Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves, *BMC Med. Res. Methodol.*, **12** (2012), 9. https://doi.org/10.1186/1471-2288-12-9
49. A. Emami, F. Javanmardi, A. Akbari, J. Kojuri, H. Bakhtiari, T. Rezaei, et al., Survival rate in hypertensive patients with COVID-19, *Clin. Exp. Hypertens.*, **43** (2021), 77–80. https://doi.org/10.1080/10641963.2020.1812624
50. S. K. Kondapuram, M. S. Coumar, Pan-cancer gene expression analysis: Identification of deregulated autophagy genes and drugs to target them, *Gene*, **844** (2022), 146821. https://doi.org/10.1016/j.gene.2022.146821
51. P. Kowalczyk, M. Woszczynski, J. Ostrowski, Increased expression of ribosomal protein S2 in liver tumors, posthepactomized livers, and proliferating hepatocytes in vitro, *Acta Biochim. Pol.*, **49** (2002), 615–624. https://doi.org/10.18388/abp.2002_3770
52. K. H. Pan, L. L. Wan, M. Chen, Exploration and identification of potential therapeutic targets and biomarkers for docetaxel resistant prostate cancer, preprint, 2022. https://doi.org/10.21203/rs.3.rs-1172051/v2
53. C. Wang, S. Qin, W. Pan, X. Shi, H. Gao, P. Jin, et al., mRNAsi-related genes can effectively distinguish hepatocellular carcinoma into new molecular subtypes, *Comput. Struct. Biotechnol. J.*, **20** (2022), 2928–2941. https://doi.org/10.1016/j.csbj.2022.06.011
54. W. Xu, A. Anwaier, C. Ma, W. Liu, X. Tian, M. Palihati, et al., Multi-omics reveals novel prognostic implication of SRC protein expression in bladder cancer and its correlation with immunotherapy response, *Ann. Med.*, **53** (2021), 596–610. https://doi.org/10.1080/07853890.2021.1908588
55. K. A. Myers, J. A. Fuller, D. F. Scott, T. J. Devine, M. J. Denton, A. Chan, Multivariate Cox regression analysis of covariates for patency rates after femorodistal vein bypass grafting, *Ann. Vasc. Surg.*, **7** (1993), 262–269. https://doi.org/10.1007/BF02000252
56. S. A. Best, S. Ding, A. Kersbergen, X. Dong, J. Y. Song, Y. Xie, et al., Distinct initiating events underpin the immune and metabolic heterogeneity of KRAS-mutant lung adenocarcinoma, *Nat. Commun.*, **10** (2019), 4190. https://doi.org/10.1038/s41467-019-12164-y