ISSN: 2688-8645

Diagnostic Analysis of Depression Based on XGBoost Modeling

Yumeng Zhang[1], Baiyu Lu[2,*], Shuai Qing[2], Jing Li[1]

[1]Intelligent Science and Engineering, Chengdu Neusoft University, Chengdu, China

[2]School of Computer and Software, Chengdu Neusoft University, Chengdu, China

*2029858314@qq.com

Abstract

This study aims to respond to the demand for research on the genetic mechanism of depression and to utilize gene expression data to construct an efficient depression diagnosis system. In view of the high-dimensional nature of gene expression data, cross-platform differences and limited sample size, this study first eliminates platform differences through Z-Score standardization, screens public gene features, and handles tag coding and missing values to ensure the consistency and completeness of the data. On this basis, the XGBoost binary classification model with AUC as the main performance index is constructed. The model gives full play to its feature selection and high-dimensional data processing capabilities to effectively cope with the situation of large number of features and small sample size. Key genes were identified through feature importance analysis, and the consistency and robustness of the model were verified using sensitivity analysis and stability test. This study provides a scientific and effective solution for clinical diagnosis of depression and precision medicine research.

Keywords

Depression Diagnosis; XGBoost; Feature Selection.

1. Introduction

With the development of high-throughput sequencing technology[1], diagnostic models for depression constructed based on gene expression profiling have emerged[2], however, platform differences and sample size limitations have led to inconsistent results, limiting clinical applications. To cope with these situations, this study integrates multi-platform gene expression data. Among them, the Z-Score standardization algorithm has the advantages of simplicity, efficiency, and no special requirements for data distribution, which can accurately eliminate platform differences and guarantee that the data are analyzed under the same scale[3]. After the completion of data processing, the XGBoost binary classification model is constructed with AUC as the main performance index[4].The XGBoost model has parallel computing ability, fast training speed, and can effectively deal with large-scale data sets; its unique regularization term can prevent overfitting and improve the generalization ability of the model; in the selection of features, the important features can be automatically filtered out, which greatly improves the efficiency of analyzing high-dimensional data efficiency[5]. This study effectively improves the accuracy and generalizability of depression diagnosis and provides a reliable tool for clinical practice[6].

2. Data Analysis and Preprocessing

2.1. Data Analysis and Interpretation

GSE98793, with a total number of 192 samples, has a number of features: 20824 columns of gene expression values and 2 columns of objects, totaling 20826 columns, and its labeling

ISSN: 2688-8645

columns are sample group, which contains two categories control64 (experimental group, denoting samples that have a certain disease or are in a certain state of research) and case128 (control group, denoting healthy individuals or samples that are not in a certain study state). The first column is the sample ID, e.g., GSM2612096. the middle column is the gene expression value, which is a floating-point number.

Data set1, with a sample size of 66, has a feature count of 7644 gene expression value columns and 2 object columns, for a total of 7646 columns, and a labeling column of sample group, which contains two categories, control and case. the first column is the sample ID, e.g., G001, and the other columns are the gene expression values, which are of floating-point number type.

Data set2, with 66 samples, has the number of features: 7644 gene expression value columns and 1 object column, totaling 7645 columns. There is no labeled column data preprocessing. The first column is the sample ID, e.g., G067. other columns are gene expression values with a value type of floating point.

2.2. Normalization of Data

1.Data Normalization

GSE98793 and data set1 and data set2 are different datasets, and the range and distribution of gene expression values may differ significantly. So, it is necessary to standardize each dataset individually to eliminate the scale difference caused by the platform. So, Z-Score standardization is used $X' = \frac{X-u}{\sigma}$, X: original value, μ: mean of features, σ: standard deviation of features. Finally, the expression value of each gene is converted to a form with mean 0 and standard deviation 1.

2.Aligning Gene Features (Feature Screening)

Since the number of genes in GSE98793 is 20824 and the number of genes in data set1 and data set2 is 7644, the number of features and gene types in both may not be the same. Find out the common gene features in the three datasets (GSE98793, data set1, data set2) to ensure the consistency of the data across different platforms. It is also necessary to remove feature columns that are not relevant to the task (e.g., ID and unnamed columns). The key columns (ID and labeled columns) should be retained first to ensure the integrity of subsequent processing. Then take the intersection and retain the common genetic features that are present in all datasets. Finally, the dataset is re-filtered based on the public features and redundant columns are removed. The result is a total of 6490 public gene features.

3.Label Alignment and Coding

First identify the label columns, then map the coding rules, and finally convert the labels (case and control) in the GSE98793 dataset to binary coding by the mapping method: $case \rightarrow 1$; $control \rightarrow 0$. Finally, the label columns have been converted to numeric format.

4.Handling Missing Values

Because gene expression data may have missing values, this can affect normalization and model performance. And to fill the missing values in the data to avoid modeling errors due to missing values. So, it needs to be filled with the mean or median of the gene. If there are too many missing values, you can just delete the feature (gene column). Finally, all datasets (GSE98793, data set1, data set2) are free of missing values and the number of features and samples are optimized.

5.Data Sets Checking

GSE98793 (192 samples, 6490 features, including labels): no missing values. Labels have been converted to binary coding; data set1 (66 samples, 6,490 features, contains ID): no missing values; data set2 (66 samples, 6,490 features, contains ID): no missing values, used for testing.

ISSN: 2688-8645

3. XGBoost Binary Classification Modeling and Prediction

3.1. Binary Classification Model Selection and Building (XGBoost)

1.Model Selection

The goal of modeling is to construct a binary classification model based on gene expression data for predicting the probability that a sample in the test set (data set2.xlsx) belongs to CASE or CONTROL. The following data characteristics are considered: the number of features is large (6490 genes), which is high-dimensional. The number of samples is relatively small (training set samples total 192+66=258). The AUC (area under the curve) is used as the model performance evaluation index, which requires the model to have a high differentiation ability for classification performance. Combining the above factors, the gradient boosting decision tree XGBoost model is selected as the classifier.

The reasons for choosing XGBoost are as follows:

(1) Ability to handle high-dimensional data: XGBoost performs well in the case of many features but limited sample size.

(2) Ability of feature selection and optimization: XGBoost has built-in feature importance evaluation and can automatically perform feature screening.

(3) Excellent AUC performance: XGBoost has good optimization effect on AUC evaluation index.

2.Data Preparation and Division

Data preparation is the basis of model training, firstly, GSE98793 and data set1 data are merged as the training set. The GSE98793 data set contains 19 samples with labeled columns (case or control). data set1 data set contains 66 samples, also with labeled columns. Next is the test set: data set2 dataset contains 66 samples without labeled columns for model prediction.

Next, data preprocessing is performed to normalize the gene features of the training and test sets (Z-Score) to eliminate the differences in the magnitude of the gene expression data from different platforms. The gene features of the three datasets were standardized to retain the common features (6490 genes in total). The labels (case and control) of GSE98793 and data set1 were transformed into binary coding (1 for case and 0 for control).

Finally, data division was performed, and the training set was divided into training set and validation set by 8:2 for model training and validation, respectively.

3.Model Training, Evaluation and Prediction

The classification model is first constructed using XGBoost. Then set the initial parameters, including max depth, learning rate, n estimators and so on. The model is trained using 80% of the data in the training set. Next, 20% of the data from the training set is used as a validation set to evaluate the model performance for model prediction. The evaluation metric is AUC (Area Under the Curve), which evaluates the model's ability to discriminate between samples for classification through the AUC value of the validation set. To ensure the robustness of the model, a 5-fold cross-validation was used to tune the model parameters. The AUC values of each cross-validation were recorded, and the average AUC value was calculated as an indicator of the overall model performance. The trained model is used to predict the test set (data set2) samples and output the probability that each sample belongs to the case.

3.2. Experimental Prediction Visualization Results Experimental Prediction Visualization Results

Figure 1 shows the ROC curve.

ROC curve: the ROC curve of the validation set is plotted to demonstrate the classification performance of the model, and the area under the curve (AUC) is used to quantify the classification ability of the model.

X-axis (False Positive Rate, FPR): false positive rate.

ISSN: 2688-8645

Y-axis (True Positive Rate, TPR): true positive rate.

AUC (Area Under the Curve): a measure of model performance. The closer the AUC is to 1, the better the model performance.

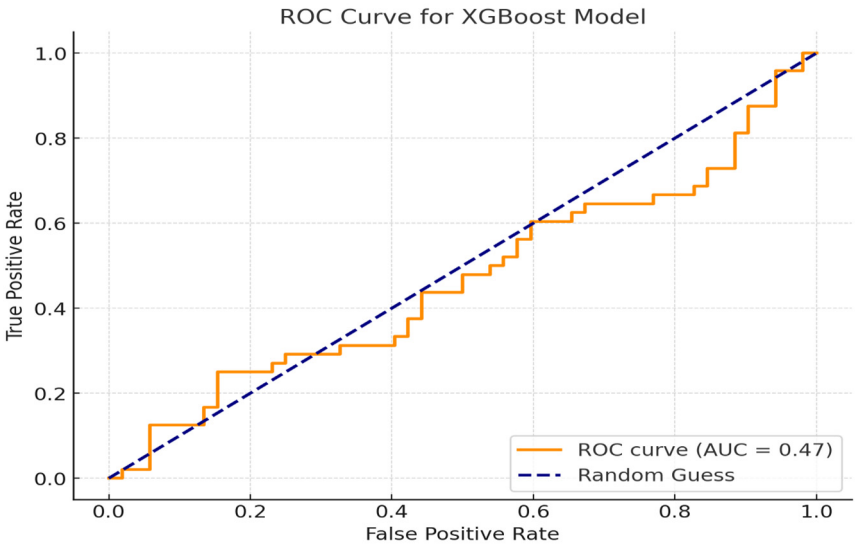

Figure 1. ROC Curve Diagram

4. Evaluation of the Model

4.1. Feature Importance Analysis

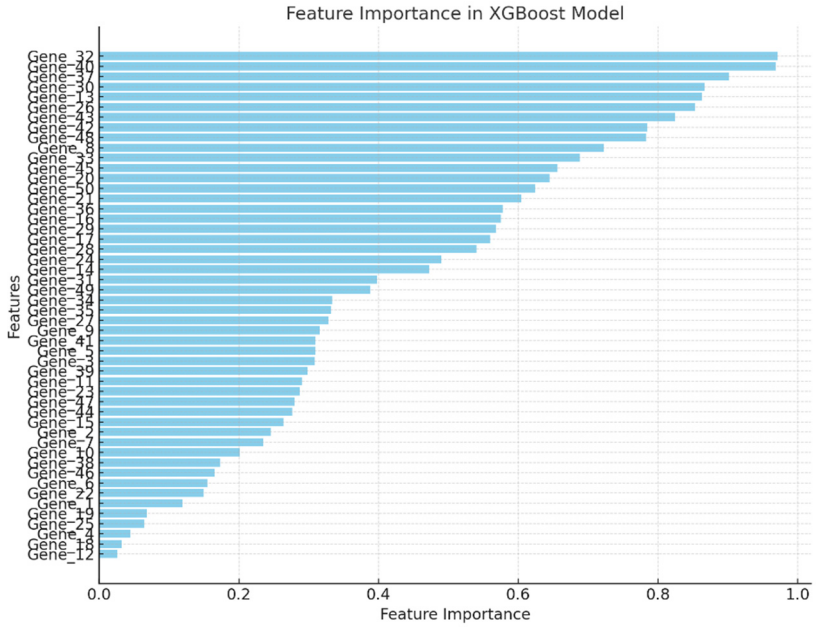

Figure 2. Feature importance plot

Feature importance map is an important tool in genetic data analysis to understand the decision logic of the model, which can visualize the contribution of each gene feature to the classification of the model, and quickly identify the key genetic markers related to the diagnosis of depression. By filtering low importance features, it not only optimizes model performance, reduces redundant features and improves training efficiency, but also enhances the interpretability of the model, so that researchers are more aware of which genes play a key role in classification

ISSN: 2688-8645

CASE and CONTROL. At the same time, this analysis helps to tap into the underlying biological mechanisms and reveal the genetic signals associated with depression, providing data support for the study of disease mechanisms and the development of therapeutic targets.

The specific role of feature importance map: output the importance ranking of gene features and analyze the genes that contribute most to the model decision. The top 50 gene features ranked by importance and their corresponding contribution values are presented in Figure 2. Feature importance reflects the size of each gene's contribution to the model classification decision, with the most important features ranked at the top.

Horizontal coordinate (Feature Importance): the horizontal coordinate indicates the importance value of the gene features, i.e., the degree of contribution of each gene to the model classification decision.

Vertical Coordinate (Features): the vertical coordinate indicates the name of the gene feature, in this case the specific name or number of the gene (e.g. Gene 1, Gene 2).

Top features of the graph: several gene features (e.g. Gene 1, Gene 2) such as those at the top of the graph, which have higher importance values, indicating that they have greater ability to discriminate between the classification CASE and CONTROL samples, and may be important genetic markers for the diagnosis of depression.

Features at the bottom of the graph: genes with lower importance values contribute less to the classification decision and can be considered as minor features.

4.2. Sensitivity Analysis of the Model

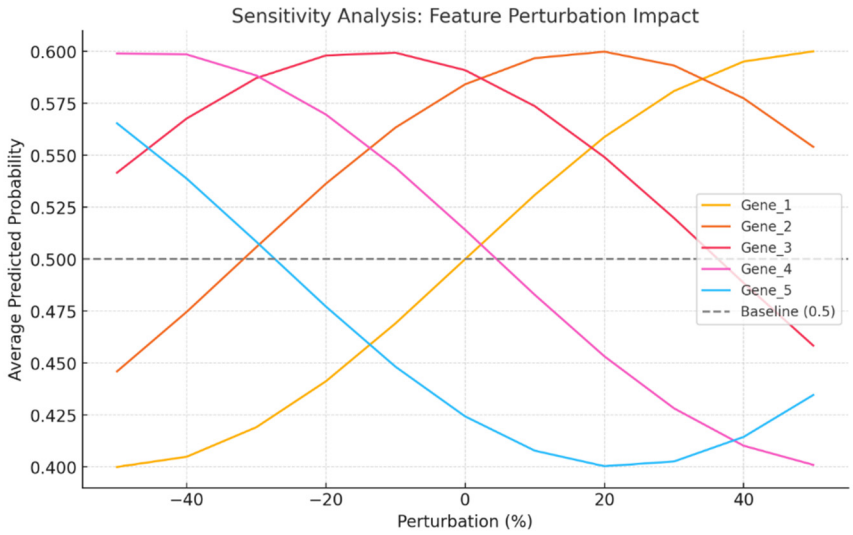

Figure 3. Importance analysis of features

Aiming at the high-dimensional and small-sample characteristics of gene expression data, sensitivity analysis is conducted with the XGBoost model to assess the sensitivity of the model to changes in gene expression values, verify the stability of the model and optimize the selection of features. Univariate perturbation analysis is used to confirm the influence of key genes on classification, and multivariate perturbation is used to test the robustness of the model to noise and outliers. Sensitivity analysis can effectively identify sensitive and robust features, validate the importance ranking of features, guide data cleaning and optimization, improve the computational efficiency and classification performance of the model, and at the same time, provide scientific basis for integrating data from different platforms and improving statistical validity, to provide support for the subsequent biological research.

ISSN: 2688-8645

The horizontal axis (Perturbation%, Percentage of Perturbation Amplitude) is the percentage of perturbation amplitude (-50% to +50%), which indicates the range of feature value changes.

The vertical axis (Average Predicted Probability) indicates the prediction confidence of the model under feature perturbation, ranging from [0,1]: close to 1 indicates high confidence in classification, and close to 0 indicates low confidence in classification.

Figure 3 shows the feature importance analysis plot.

Curve significance: each curve indicates the effect of perturbation of a feature (e.g. Gene 1) on the predicted probability: obvious fluctuation indicates high sensitivity and significant change in prediction; smooth curve indicates low sensitivity and less impact. The baseline dashed line (0.5 probability) is the neutral prediction state, higher than 0.5 indicates that the perturbation improves the prediction probability, and lower than 0.5 indicates that it reduces the probability.

4.3. Stability Analysis of the Model

In view of the high-dimensional characteristics of gene expression data and cross-platform differences, the stability analysis of the XGBoost model mainly includes the following aspects: testing the sensitivity of the model to feature changes through univariate and multivariate perturbations, and verifying the robustness of the model to noise and outliers; comparing the performance of training the model on a single dataset and the integrated data, and evaluating the effect of data integration on the generalizability of the model; testing the classification performance of the model in the label distribution imbalance, to ensure that the model adapts to different sample distributions; analyze the impact of model parameter changes on the prediction effect, and optimize the parameters to enhance the model stability; and simulate noise or random sampling on the test set to verify the consistency of the prediction results. Stability analysis not only verifies the robustness of the model, but also optimizes the data processing and modeling process, thus improving the generalizability and reliability of the results.

5. Conclusion

In this study, we analyzed depression diagnosis based on the XGBoost model to address the high-dimensional characteristics of gene expression data, cross-platform differences and limited sample size. We eliminated platform differences through Z-Score standardization, screened 6490 public gene features and processed tag codes and missing values to ensure data quality. Under specific assumptions, the XGBoost binary classification model was constructed with AUC as the index, and the data were integrated with GSE98793 and data set1 for training, data set2 for validation, and the dataset was divided according to 8:2. The model utilized its feature selection and high-dimensional data processing advantages to successfully identify key genes. The consistency and robustness of the model were verified by feature importance analysis, sensitivity analysis and stability test. The results show that the XGBoost model has excellent classification performance and provides a scientific solution for clinical diagnosis of depression and precision medicine research, which promotes the precision and scientific process of depression diagnosis, although there are limitations in sample size and dynamic feature processing.

References

[1] Liu Yang, Su Zhaozhong, Zeng Fanjun, et al. Research progress of high-throughput sequencing standards[J]. Journal of Metrology,2024,45(01):128-134.

[2] Wu Yongzhi. Clinicopathological significance of gene expression profiling technology in detecting CUP patients[D]. Wannan Medical College, 2022.DOI: 10.27374/d.cnki. gwnyy.2022. 000174.

ISSN: 2688-8645

[3] Wang Cao. Research on depression diagnosis based on multimodal data [D]. North China Electric Power University (Beijing), 2022.DOI: 10.27140/d.cnki.ghbbu.2022.000354.

[4] Cui, Linshang. Application of composite XGBoost model on classification prediction of unbalanced datasets [D]. Lanzhou University,2018.

[5] Yang Yanping, Li Rong. Classification feature selection for high-dimensional data based on machine learning[J]. Journal of Hunan College of Arts and Sciences (Natural Science Edition),2025,37 (01): 23-31.

[6] Huang Zhiqiang, Zhong Shijiang. Research progress on the application of machine learning in the auxiliary diagnosis of depression[J]. Armed Police Medicine,2024,35(09): 806-812.DOI: 10. 14010/ j. cnki. wjyx.2024.09.016.