Research Article

Improved Multiclassification of Schizophrenia Based on Xgboost and Information Fusion for Small Datasets

Wenjing Zhu,[1,2] **Shoufeng Shen,**[3] **and Zhijun Zhang** [1,2]

[1]*Department of Neurology, Affiliated Zhongda Hospital, Research Institution of Neuropsychiatry, School of Medicine, Southeast University, Nanjing 210009, China*
[2]*Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, Zhejiang University School of Medicine, Hangzhou 310000, China*
[3]*College of Science, Zhejiang University of Technology, Hangzhou, Zhejiang 310023, China*

Correspondence should be addressed to Zhijun Zhang; janemengzhang@vip.163.com

To improve the performance in multiclass classification for small datasets, a new approach for schizophrenic classification is proposed in the present study. Firstly, the Xgboost classifier is introduced to discriminate the two subtypes of schizophrenia from health controls by analyzing the functional magnetic resonance imaging (fMRI) data, while the gray matter volume (GMV) and amplitude of low-frequency fluctuations (ALFF) are extracted as the features of classifiers. Then, the D-S combination rule of evidence is used to achieve fusion to determine the basic probability assignment based on the output of different classifiers. Finally, the algorithm is applied to classify 38 healthy controls, 16 deficit schizophrenic patients, and 31 nondeficit schizophrenic patients. 10-folds cross-validation method is used to assess classification performance. The results show the proposed algorithm with a sensitivity of 73.89%, which is higher than other classification algorithms, such as supported vector machine (SVM), logistic regression (LR), K-nearest neighbor (KNN) algorithm, random forest (RF), BP neural network (NN), classification and regression tree (CART), naive Bayes classifier (NB), extreme gradient boosting (Xgboost), and deep neural network (DNN). The accuracy of the fusion algorithm is higher than that of classifier based on the GMV or ALFF in the small datasets. The accuracy rate of the improved multiclassification method based on Xgboost and fusion algorithm is higher than that of other machine learning methods, which can further assist the diagnosis of clinical schizophrenia.

1. Introduction

Schizophrenia (SZ) is a serious mental illness that interferes with a person's ability to think clearly, manage emotions, make decisions, and relate to others [1]. The positive symptoms, such as hallucinations and delusions, can lead to suicidal or aggressive behavior, while negative symptoms and cognitive impairment lead to a decline in quality of life and social function; all these symptoms will cause tremendous human suffering and economic burden [2]. However, SZ is diagnosed on the basis of clinical evaluation of symptoms and functional, no objective diagnostic biomarker set. In addition, there are two main types of schizophrenia, which are called deficit schizophrenia (DS) and nondeficit schizophrenia (NDS) [3]. Deficit syndrome of schizophrenia, also called negative symptoms of schizophrenia, includes social withdrawal, loss of motivation, poverty of speech, and blunting of affect. Compared with NDS, DS has greater cognitive impairment, worse long-term prognosis, and lower recovery rates, which persist or are found even during psychotic remissions [4, 5]. Therefore, it is quite important to diagnose SZ accurately and discriminate the two subtypes of SZ from healthy control (HC), particularly the discrimination between the DS and NDS.

Classification is a machine learning algorithm where we get the labeled data as input and we need to predict the output into a class [6]. If there are two classes, then it is called binary classification [7]. If there are more than two classes,

then it is called multiclass classification. Nowadays, classification algorithm is widely used in the medical diagnosis, especially in the field of mental disorders. Early published research applied support vector machines with radial basis function kernel method to classify 15 schizophrenic patients and 15 HCs based on the structural image of the hippocampal complex only with 63% classification accuracy [8]. With the rapid development of computational psychiatry, a growing body of classification approaches are applied to discriminate SZ in recent years, such as logistic regression (LR), support vector machine (SVM), neural network (NN), random forest (RF), extreme gradient boosting (XGBoost), and deep learning [9]. In the early studies, many researchers focused on the binary classification problem in SZ. Greenstein et al. proposed the logistic regress classifier to discriminate 99 SZ patients from 99 HCs with 73.7% accuracy [10]. Nieuwenhuis et al. proposed the SVM classifier to discriminate 128 SZ patients from 111 HCs and achieved an accuracy of 71.4% [11]. Thereby, in the past years, feature reduction approaches were discussed and applied to improve the performance of classification. Ershad and Hashemi proposed the dispelling reduction approaches [12], Juneja and his colleagues obtained the discriminative features by using SVD model and a novel multivariate feature selection algorithm [13]. However, the accuracy was not high enough based on the classical classification algorithms, which was usually less than 75%. Then, many new classifiers and feature selection approaches are proposed to improve the classification performance. Up to 2018, Wang et al. developed the SVM model to discriminate SZ from HCs and achieved an accuracy of 92.4% [14]. In 2020, Kim et al. proposed the feature reduction method when there are redundant or correlated features based on the FDR value and achieved an accuracy of 96.2% [15]. Patel et al. proposed a classification algorithm to discriminate SZ versus HCs by busing deep learning in fMRI, and the accuracy was 92% [16]. Nowadays, the binary classification is not a hard work in the field of mental disorders. However, there is not only one type of SZ, such as DS and NDS, which is more difficult to discriminate each other or from HCs. In order to solve this problem, multiple classification methods for schizophrenic subtypes are necessary.

Few previous literatures have reported multiclass classifications for different types of psychiatric patients, and most of these classifications used traditional machine learning methods such as SVM and LR. For examples, Zhu et al. proposed the SVM model to classify first-episode, drug-naive SZ, ultrahigh risk for psychosis and HC with the global balanced accuracy only 73.37% and the sensitivity only 68.42%, using the fivefold cross-validation method [17]. Soon afterwards, multiple classification methods have been explored, such as three SVM models to classify the SZ, bipolar disorder, and HC [18] and classify depression, bipolar disorder, and HC [19]; an SVM combined with recursive feature elimination was used to classify first-episode SZ, chronic SZ, and HC [20]. Unfortunately, almost all the accuracy rate of the multiclass classification is less than 70%. In addition, because of the poor coordination in psychiatric patients, the amount of imaging data is generally small, which leads to an accuracy far below 70%.

In recent years, deep learning is widely used in the pattern recognition. Zeng et al. proposed a deep discriminant autoencoder network to learn imaging site-shared functional connectivity features to discriminate SZ from normal subjects. In their work, the accuracy of 85% is achieved [21]. Oh et al. collected 873 structural MRI datasets and discriminate the SZ from normal subjects by using a deep convolutional neural network [22]. Srinivasagopalan et al. proposed a deep learning algorithm for diagnosing SZ [23]. These deep learning algorithms are based on the original image data, such as CNN and DNN. However, the original image data is usually quite difficult to obtain. Many deep learning algorithms on small dataset is usually overfitting. Therefore, the machine learning algorithms are more suitable to improve the performance of the classical classifiers.

To the best of our knowledge, there is no study to implement the multiclass classification of DS, NDS, and HC based on multimodal imaging data of schizophrenia. Therefore, in order to achieve multiclass classification of schizophrenia and obtain higher classification results in small data, a new classification algorithm is proposed in this paper. In this algorithm, GMV and ALFF are selected as the features to construct multiclassifier based on Xgboost, respectively. Then, the fused model is built to improve the accuracy for the small datasets. The D-S fusion model is used to combine the output from different classifiers to determine the probability assignment for different subtypes and HC. The rest of this paper is structured as follows. In Section 2, the Xgboost classifier is proposed to discriminate DS and NDS from HC and the fusion model is introduced to combine information of output. The results obtained by applying our model are shown in Section 3. In Section 4, the main contribution of this paper is summarized.

2. The Fused Classification Algorithm Based on Xgboost for Three-Class Classification

In this section, a fused classification algorithm is proposed to improve the accuracy for the small datasets. This algorithm is applied to discriminate the subtypes of SZ from HC. There are three labels (DS, NDS, and HC) that should be assigned to each collected subject. To solve this problem, an improved multiclassification algorithm is introduced. Firstly, Xgboost algorithm is applied to classify them, which is one of the most widely used machine learning algorithms in classification problems [24]. The classifiers are constructed based on the features of GMV or ALFF, which is extracted from fMRI data. Then, the fusion model is used to combine the output information of the different classifier to determine the probability assignment of each class. Finally, the test subject will be classified into the class with the maximum probability. The flow chart of the proposed algorithm is shown in Figure 1.

2.1. Classifier Based on the Xgboost Algorithm. The Xgboost algorithm is composed of many weak classification and regression trees (CARTs). Taking the ith dataset (x_i, y_i) as an example, x_i is the input variable with several attributes of fMRI data and y_i is the real value of the given subject.
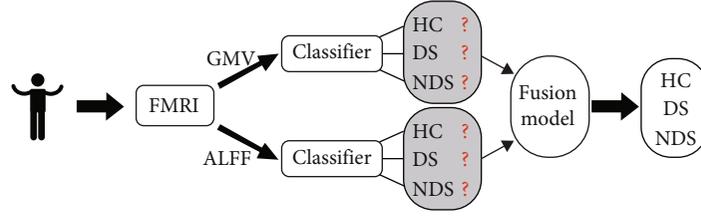
FIGURE 1: Flow chat of improved multiclassification algorithm.

For example, $y_i = (1, 0, 0)$ means the ith subject is HC, $y_i = (0, 1, 0)$ means the ith subject is DS, and $y_i = (0, 0, 1)$ means the ith subject is NDS.

Then, an Xgboost model can be mathematically expressed in the following form [25]:

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(x_i), \tag{1}$$

where K is the number of the CARTs, f_k is the predicted value of each independent CART, and $\widehat{y}_i$ is the predicted value with respect to input x_i.

The additive training model of Xgboost can be expressed as

$$\begin{cases} \widehat{y}_i^{(0)} = 0, \\ \widehat{y}_i^{(k)} = \sum_{t=1}^{k} f_t(x_i) = \widehat{y}_k^{(k-1)} + f_k(x_i), \end{cases} \tag{2}$$

where $\widehat{y}_i^{(k)}$ is the predicted value of the kth CART.

The objective function of Xgboost includes a loss function and regularization term, which is expressed as

$$\text{obj} = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3}$$

where $l(y_i, \widehat{y}_i)$ can be used to measure the error between predicted value and real value, n is the number of the subjects, and Ω is the regularization item to avoid overfitting.

The specific form of $\Omega(f_k)$ of the kth CART is given as

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{4}$$

where γ and λ present the penalty coefficients, T is the number of leaf nodes, and w is the weight of the leaf nodes.

Then, the objective function of the tth step $\text{obj}^{(t)}$ can be calculated by Equation (5) based on the previous step $\text{obj}^{(t-1)}$ based on the Equations (2) and (3).

$$\text{obj}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{k=1}^{t-1} \Omega(f_k) + \Omega(f_t). \tag{5}$$

By applying the second-order Taylor expansion to above

equation, the objective function can be transformed into

$$\text{obj}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i) + \Omega(f_t) + \text{Const.} \tag{6}$$

In the above expression, Const is a constant term at the step t; the parameters pair g_i and h_i can be calculated as

$$\begin{cases} g_i = \dfrac{\partial l\left(y_i, \widehat{y}_i^{(t-1)}\right)}{\partial \widehat{y}_i^{(t-1)}}, \\[4mm] h_i = \dfrac{\partial^2 l\left(y_i, \widehat{y}_i^{(t-1)}\right)}{\left(\partial \widehat{y}_i^{(t-1)}\right)^2}. \end{cases} \tag{7}$$

Because $l(y_i, \widehat{y}_i^{(t-1)})$ is the constant item, the objective function can be rewritten as

$$\text{obj}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \text{const}', \tag{8}$$

where const' is a new constant item at step t.

According to the definition of f_k, f_k can be written in the following form as

$$f_k(x) = w_{q(x)}. \tag{9}$$

Then, Equation (8) can be rewritten in the following form as

$$\text{obj}^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T + \text{const}', \tag{10}$$

denoted by $\sum_{i \in I_j} g_i = G_j$ and $\sum_{i \in I_j} h_i = H_j$; then, the objective function is expressed as

$$\text{obj}^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}(H_j + \lambda)w_j^2\right] + \gamma T + \text{const}'. \tag{11}$$

The leaf nodes of the tth CART are each independent; G_j

Require: I, instance set of current node; d, feature dimension
Ensure: Split with max score
1. $Gain \longleftarrow 0, G \longleftarrow \sum_{i \in I_j} g_i, H \longleftarrow \sum_{i \in I_j} h_i$
2. **for** $k = 1$ to m **do**
3. $G_L \longleftarrow 0, H_L \longleftarrow 0$
4.   **for all** i in sorted I by x_{jk} **do**
5. $G_L \longleftarrow G_L + g_j, H_L \longleftarrow H_L + h_j$
6. $G_R \longleftarrow G - G_L, H_R \longleftarrow H - H_L$
7. $score \longleftarrow \max (score, (G_L^2/H_L + \lambda) + (G_R^2/H_R + \lambda) - G^2/H + \lambda)$
8.   **end for**
9. **end for**

ALGORITHM 1: Greedy algorithm for split finding.

and H_j are the determined items. Then, minimizing the function equation (11), the optimal parameter w_j can be calculated as

$$w_j^* = -\frac{G_j}{H_j + \lambda}. \tag{12}$$

Therefore, the final objective function is shown in the following form as

$$obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T + const'. \tag{13}$$

The splitting algorithm [26] based on the above function is used to find the best split in Xgboost by

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma. \tag{14}$$

The gain function has four terms: the first two terms are the profits of left and right parts of a node, where G_L and G_R are the left and right parts of G_j and H_L and H_R are the left and right parts of H_j; and the third item is the total profit of that node. The last item is the regularization item for preventing overfitting. The greedy algorithm determines whether a node obtains the maximum gain. Thus far, the optimal tree structure that maximizes the gain can be generated.

The above description leads to the split finding algorithm for Xgboost presented as Algorithm 1.

2.2. The Fusion Model Based on the D-S Evidence Theory. Analyses of the amplitude of low-frequency fluctuations (ALFF) and gray matter volume (GMV) are two important methods used in fMRI studies. Selecting GMV as the feature to construct the classifier based on the Xgboost algorithm, the predicted value of the ith subject can be obtained and denoted as $\widehat{y}_i^{(1)}$, while the other predicted value $\widehat{y}_i^{(2)}$ can be calculated by selecting ALFF as the feature, $\widehat{y}_i^{(1)} \neq \widehat{y}_i^{(2)}$ usually. Taking the ith subject as an example, the output information of the classifiers is $\widehat{y}_i^{(1)} = (0.7, 0.3, 0.9)$ and

$\widehat{y}_i^{(2)} = (0.7, 0.5, 0.6)$. According to the output of classifier based on the feature of GMV, this subject should be classified into NDS, while the classifier based on the feature of ALFF will classify this subject into HC.

To overcome the conflict of different classifiers, the D-S evidence model is used to fuse the information of $\widehat{y}_i^{(1)}$ and $\widehat{y}_i^{(2)}$ [27]. The softmax function [28] is used in many machine learning applications for multiclass classifications to assign probability for each subject. The softmax function is expressed in the following form as Equation (15), which is used to calculate the probability assignment of each class.

$$\begin{cases} \widehat{y}_{ij}^{(1*)} = \dfrac{e^{\widehat{y}_{ij}^{(1)}}}{\sum_{j=1}^{3} e^{\widehat{y}_{ij}^{(1)}}}, \\[4mm] \widehat{y}_{ij}^{(2*)} = \dfrac{e^{\widehat{y}_{ij}^{(2)}}}{\sum_{j=1}^{3} e^{\widehat{y}_{ij}^{(2)}}}, \end{cases} \tag{15}$$

where $\widehat{y}_{ij}^{(1)}$ is the element in the set $\widehat{y}_i^{(1)}$ and $j = 1, 2, 3$ represents HC, DS, and NDS, respectively. For example, $\widehat{y}_{i1}^{(1*)}$ is the probability that the ith subject is HC according to the GMV feature, and $\widehat{y}_{i1}^{(2*)}$ is the probability that the ith subject is HC according to the ALFF feature.

Though softmax function, the probability assignment of the ith subject can be obtained as

$$\begin{aligned} \widehat{y}_i^{(1*)} &= (0.3458, 0.2318, 0.4224), \\ \widehat{y}_i^{(2*)} &= (0.3672, 0.3006, 0.3322). \end{aligned} \tag{16}$$

Then, the combined strategy based on the D-S evidence is expressed as

$$\begin{cases} \widehat{y}_i^{(1*)} \oplus \widehat{y}_i^{(2*)} (HC) = \dfrac{1}{K} \times \widehat{y}_{i1}^{(1*)} \times \widehat{y}_{i1}^{(2*)}, \\[3mm] \widehat{y}_i^{(1*)} \oplus \widehat{y}_i^{(2*)} (DS) = \dfrac{1}{K} \times \widehat{y}_{i1}^{(1*)} \times \widehat{y}_{i1}^{(2*)}, \\[3mm] \widehat{y}_i^{(1*)} \oplus \widehat{y}_i^{(2*)} (NDS) = \dfrac{1}{K} \times \widehat{y}_{i1}^{(1*)} \times \widehat{y}_{i1}^{(2*)}. \end{cases} \tag{17}$$
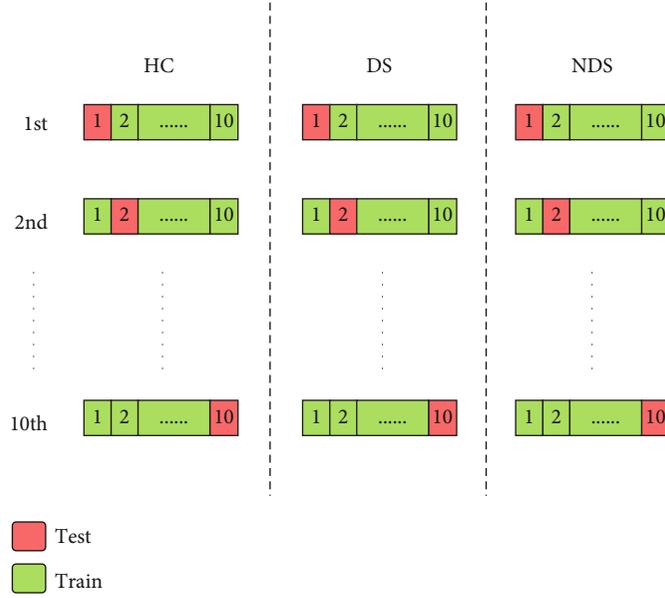
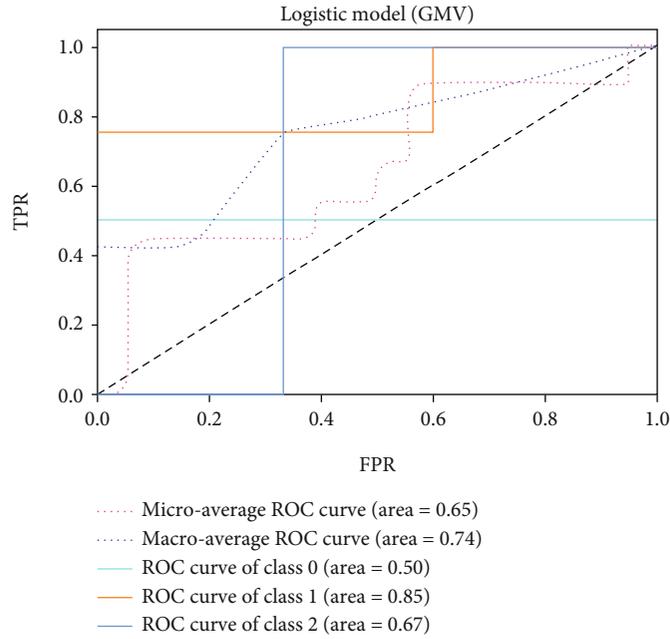FIGURE 2: Flow chat of 10-fold cross-validation.

FIGURE 3: Results of logistic classifier (GMV).

The probability of subject into each group $\hat{y}_i^{(1*)} \oplus \hat{y}_i^{(2*)} \in [0, 1]$, where K reflects the conflict level of evidences and can be represented as

$$K = \hat{y}_{i1}^{(1*)} \times \hat{y}_{i1}^{(2*)} + \hat{y}_{i1}^{(1*)} \times \hat{y}_{i1}^{(2*)} + \hat{y}_{i1}^{(1*)} \times \hat{y}_{i1}^{(2*)}. \qquad (18)$$

The result of the above example is $\hat{y}_i^{(1*)} \oplus \hat{y}_i^{(2*)} = ($ 0.3768, 0.2068, 0.4164$)$ by Equation (17). According to the output of the fusion information, this subject will be classified into the class of NDS.

3. Experiments and Results

To evaluate performance of the proposed classification method, we scan 246 brain regions of 85 subjects by fMRI to extract the features of GMV and ALFF, including 16 DS, 31 NDS, and 38 HC. Obviously, it is a hard task to classify them in these small datasets. We apply the proposed algorithm to classify all the subjects.

The present study classifies all subjects into three classes by applying linear regression (LR), supported vector classifier (SVC), K-nearest Neighbor (KNN), neural network (NN), naive Bayes (NB), classification and regression tree
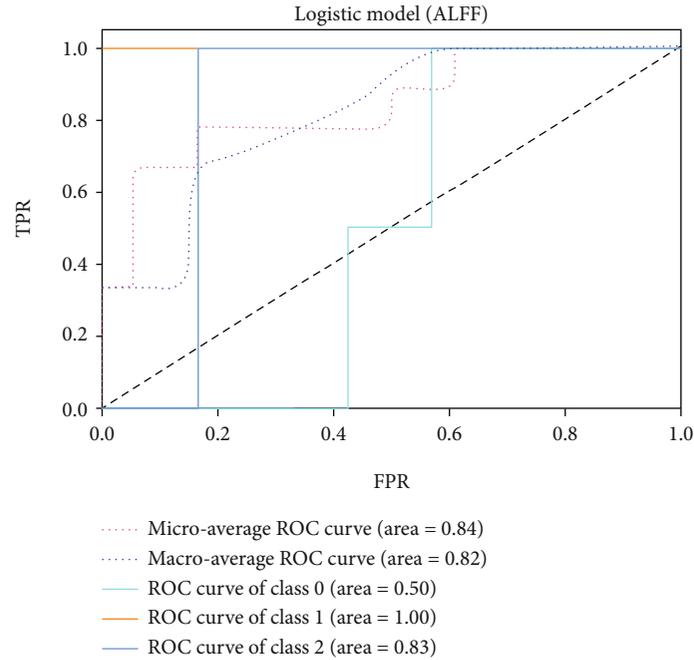
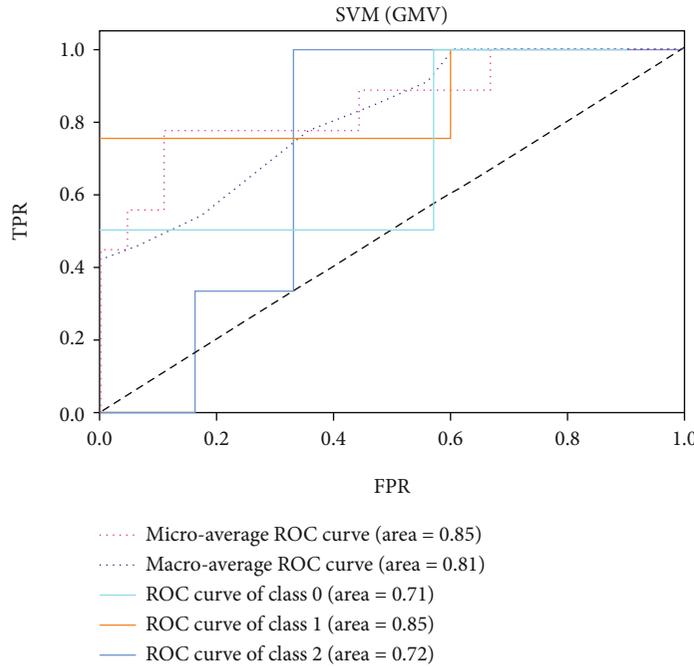Figure 4: Results of logistic classifier (ALFF).

Figure 5: Results of SVM with kernel of RBF (GMV).

(CART), random forest (RF), extreme gradient boosting (Xgboost), deep neural network (DNN), and the proposed fusion algorithm. The flow of 10-fold cross-validation is shown in Figure 2. The results of 10-fold cross-validation of different classifiers are shown as below.

The receiver operating characteristic (ROC) curve is considered to evaluate the performance of classifiers. For different classification thresholds, the true-positive rate (TPR) (Equation (19)) is plotted against the false-positive rate (FPR) (Equation (20)). The area under the ROC curve (AUC) indicates the classifier's ability to distinguish between classes. The value of the AUC is in the range [0,1]. AUC is 1 for a perfect classifier. In this work, the ROC curve is plotted for each class, as this is a multiclass problem. The microaverage and macroaverage are also computed by summing the individual values for true positive (TP), true negative (TN),
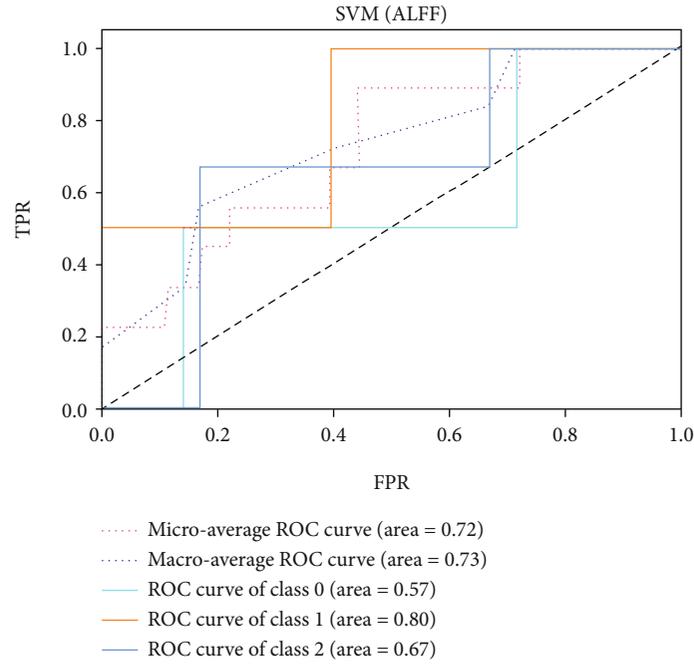
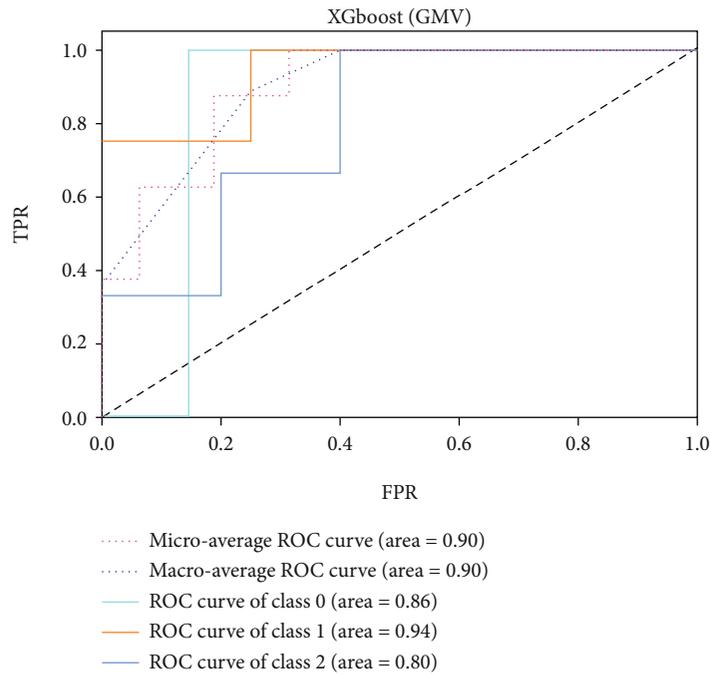FIGURE 6: Results of SVM with kernel of RBF (ALFF).

FIGURE 7: Results of Xgboost (GMV).

false positive (FP), and false negative (FN). Then, the accuracy (Equation (21)), recall (Equation (22)), precision (Equation (23)), and F1-score (Equation (24)) are selected as the important metrics to evaluate the performance of different classifiers [29].

$$TPR = \frac{TP}{TP + FN}, \tag{19}$$

$$FPR = \frac{FP}{FP + TN}, \tag{20}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{21}$$

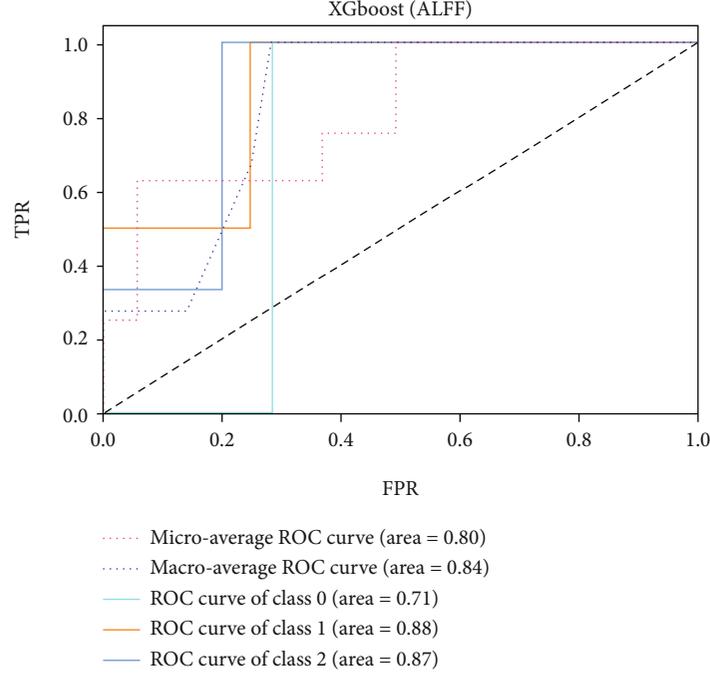$$Precision = \frac{TP}{TP + FP}, \tag{22}$$

FIGURE 8: Results of Xgboost (ALFF).

TABLE 1: Classification metrics of different methods.

Methods	Accuracy	Precision	Recall	F1-scores	AUC
LR (GMV)	64.8611%	65.0509%	64.8611%	0.6237	0.8160
LR (ALFF)	65.9722%	65.5394%	65.9722%	0.6388	0.7743
SVC (GMV)	61.5278%	51.5602%	61.5277%	0.5508	0.7403
SVC (ALFF)	67.0833%	59.4147%	67.0833%	0.6114	0.6114
KNN (GMV)	63.4722%	61.4259%	63.4722%	0.5899	0.7174
KNN (ALFF)	58.6111%	67.3175%	58.6111%	0.5601	0.6731
NN (GMV)	55.4167%	56.4749%	47.2222%	0.5425	0.8066
NN (ALFF)	64.4444%	67.0764%	60.8333%	0.6266	0.7618
NB (GMV)	53.8889%	55.8657%	53.8889%	0.5216	0.7389
NB (ALFF)	68.8889%	74.0139%	68.8889%	0.6781	0.8308
CART (GMV)	45.8333%	47.7870%	51.9444%	0.4721	0.6375
CART (ALFF)	60%	57.6075%	56.6667%	0.5667	0.6542
RF (GMV)	62.6389%	60.0883%	62.6389%	0.5874	0.7722
RF (ALFF)	65.6944%	57.1042%	65.6944%	0.5943	0.7969
Xgboost (GMV)	62.6389%	60.4445%	62.6389%	0.5888	0.7958
Xgboost (ALFF)	68.0556%	62.9610%	68.0556%	0.6364	0.8351
DNN	62.7083%	40.3380%	62.7083%	0.5152	0.6723
Our classifier (fusion)	73.8889%	65.4242%	73.8889%	0.6746	0.8524

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (23)$$

$$\text{F1-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (24)$$

The ROC of the above classifiers are shown in Figures 3–8. In these figures, DS is the class 0, HC is the class 1, and NDS is the class 2. The key classification metrics are extracted from the results listed in Table 1.

Due to the limited number of the DS, the AUC of class 0 is lower than other classes in the above figures. Therefore, it is hard to discriminate DS from NDS and HC. Taking Figure 5 as an example, the macroaverage AUC score is 0.81 and the microaverage AUC score is 0.85 based on the feature of GMV by using the SVM classifier, which are better
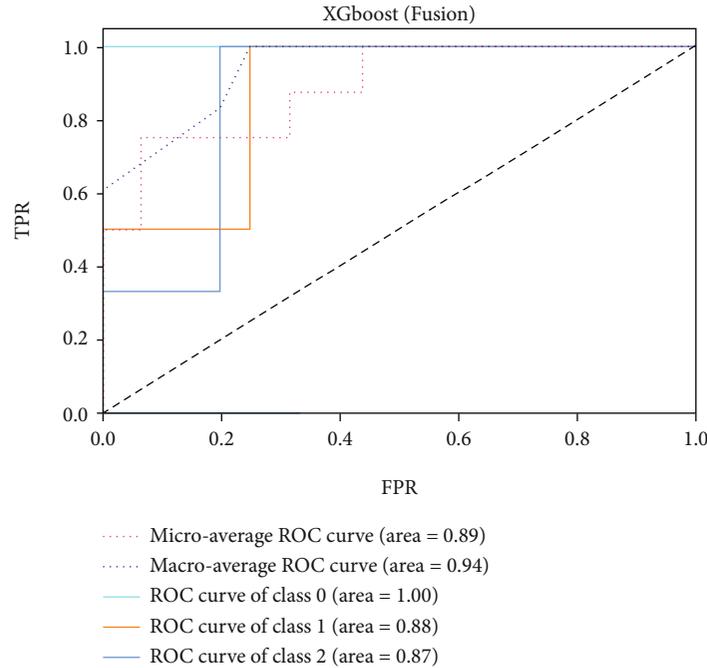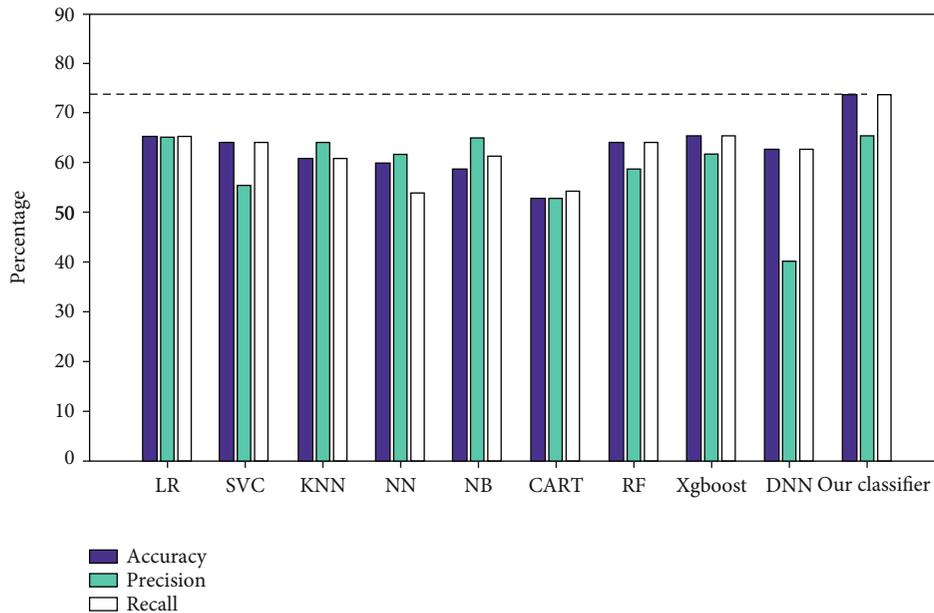
FIGURE 9: Results of Xgboost (fusion).

FIGURE 10: Accuracy of different classifiers.

than the performance based on the feature of ALFF. The AUC score of HC is 0.85 based on the feature of GMV, which is better than the AUC scores of DS and NDS. The Xgboost classifier showed better performance than the SVM and logistic classifier. The microaverage AUC score is 0.90 and 0.80, respectively, on the feature of GMV and ALFF. The macroaverage AUC score is 0.90 and 0.84, respectively, on the feature of GMV and ALFF. The ROC of the fusion algorithm is shown in Figure 9. In this figure, we find the AUC score of the class 0 is 1, which means it

is a perfect classifier. This classifier combines the advantages of the classifier based on feature of GMV and ALFF. Therefore, the performance of this fusion classifier is much better than others.

From Table 1, 10-fold cross-validation showed our algorithm with an accuracy of 73.89%, which is higher than other classifiers. Many metrics of the proposed classifier are better than other classifiers. In many present studies, the accuracy of the present classifier is usually less than 70% when the datasets are small. In this paper, the proposed

fusion classifier will improve the performance effectively for the small datasets by combining the advantages of each feature. The accuracy of different classifiers is shown in Figure 10.

4. Conclusion

In this paper, a new multiple classifier method was proposed for the small datasets and applied to discriminate the two subtypes of schizophrenia and health controls based on the fMRI data. Due to the limitation data and indexes, this study constructed the Xgboost algorithm based on the different features. To improve the accuracy, the fusion model was used to combine the information from different classifiers. Finally, the subject would be classified into the class with the maximum probability. This method was applied to classify 38 healthy controls, 16 deficit schizophrenic patients, and 31 nondeficit schizophrenic patients. 10-fold cross-validation showed our algorithm with a sensitivity of 73.89%, which was much higher than other classification algorithms when the datasets were small. In addition, the proposed algorithm can be used to discriminate different classes for the large datasets. The performance of the proposed algorithm would be effective than other algorithms when the datasets are small. It will bring the better performance in diagnosing subtypes of schizophrenia. Although the findings in our study are rigorous, there are some limitations: (1) relatively small sample size; (2) interference caused by antipsychotic drugs during experiment; and (3) limitations of the algorithm itself. In the future work, more subjects will be collected in the project, including different subtypes and HC. The original image data should be obtained in the processing of experiments, and more deep learning approaches will be proposed to solve this multiclass classification problem.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' Contributions

Wenjing Zhu was responsible for the conceptualization, investigation, methodology, and writing of original draft and undertook the data analysis. Shoufeng Shen was responsible for the software, conceptualization, and methodology. Zhijun Zhang supervised the data analysis, provided infrastructure, and was responsible for writing, reviewing, and editing. All authors contributed to and have approved the final manuscript.

References

[1] K. Kusudo, R. Ochi, S. Nakajima et al., "Decision tree classification of cognitive functions with D2 receptor occupancy and illness severity in late-life schizophrenia," *Schizophrenia Research*, vol. 241, p. 113, 2022.

[2] L. A. Antonucci, G. Pergola, A. Pigoni et al., "A pattern of cognitive deficits stratified for genetic and environmental risk reliably classifies patients with schizophrenia from healthy control subjects," *Biological Psychiatry*, vol. 87, no. 8, pp. 697–707, 2020.

[3] L. Fan, M. Yu, A. Pinkham et al., "Aberrant large-scale brain modules in deficit and non-deficit schizophrenia," *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, vol. 113, article 110461, 2022.

[4] G. P. Strauss, M. Harrow, L. S. Grossman, and C. Rosen, "Periods of recovery in deficit syndrome schizophrenia: a 20-year multi-follow-up longitudinal study," *Schizophrenia Bulletin*, vol. 36, no. 4, pp. 788–799, 2010.

[5] B. Kirkpatrick, O. Gurbuz Oflezer, M. Delice Arslan, G. Hack, and E. Fernandez-Egea, "An early developmental marker of deficit versus nondeficit schizophrenia," *Schizophrenia Bulletin*, vol. 45, no. 6, pp. 1331–1335, 2019.

[6] S. K. Khare and V. Bajaj, "A self-learned decomposition and classification model for schizophrenia diagnosis," *Computer Methods and Programs in Biomedicine*, vol. 211, article 106450, 2021.

[7] K. Dadi, M. Rahim, A. Abraham et al., "Benchmarking functional connectome-based predictive models for resting-state fMRI," *NeuroImage*, vol. 192, pp. 115–134, 2019.

[8] P. Golland, W. E. L. Grimson, M. E. Shenton, and R. Kikinis, "Small sample size learning for shape analysis of anatomical structures," *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, vol. 1935, pp. 72–82, 2000.

[9] J. W. Lai, C. K. E. Ang, U. R. Acharya, and K. H. Cheong, "Schizophrenia: a survey of artificial intelligence techniques applied to detection and classification," *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, p. 6099, 2021.

[10] D. Greenstein, J. D. Malley, B. Weisinger, L. Clasen, and N. Gogtay, "Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls," *Frontiers in Psychiatry*, vol. 3, p. 53, 2012.

[11] M. Nieuwenhuis, N. E. M. van Haren, H. E. Hulshoff Pol, W. Cahn, R. S. Kahn, and H. G. Schnack, "Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples," *NeuroImage*, vol. 61, no. 3, pp. 606–612, 2012.

[12] S. F. Ershad and S. Hashemi, "To increase quality of feature reduction approaches based on processing input datasets," in *2011 IEEE 3rd International Conference on Communication Software and Networks*, pp. 367–371, Xi'an, China, 2011.

[13] A. Juneja, B. Rana, and R. K. Agrawal, "A combination of singular value decomposition and multivariate feature selection method for diagnosis of schizophrenia using fMRI," *Biomedical Signal Processing and Control*, vol. 27, pp. 122–133, 2016.

[14] S. Wang, Y. Zhang, L. Lv et al., "Abnormal regional homogeneity as a potential imaging biomarker for adolescent-onset schizophrenia: a resting-state fMRI study and support vector machine analysis," *Schizophrenia Research*, vol. 192, pp. 179–184, 2018.

[15] J. Kim, M. Y. Kim, H. Kwon et al., "Feature optimization method for machine learning-based diagnosis of schizophrenia using magnetoencephalography," *Journal of Neuroscience Methods*, vol. 338, article 108688, 2020.

[16] P. Patel, P. Aggarwal, and A. Gupta, "Classification of schizophrenia versus normal subjects using deep learning," in *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 1–6, Guwahati, Assam, India, 2016.

[17] F. Zhu, Y. Liu, F. Liu et al., "Functional asymmetry of thalamocortical networks in subjects at ultra-high risk for psychosis and first-episode schizophrenia," *European Neuropsychopharmacology*, vol. 29, no. 4, pp. 519–528, 2019.

[18] H. G. Schnack, M. Nieuwenhuis, N. E. M. van Haren et al., "Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects," *NeuroImage*, vol. 84, pp. 299–306, 2014.

[19] S. Ranlund, M. J. Rosa, S. de Jong et al., "Associations between polygenic risk scores for four psychiatric illnesses and brain structure using multivariate pattern recognition," *NeuroImage: Clinical*, vol. 20, pp. 1026–1036, 2018.

[20] J. Liu, X. Wang, X. Zhang, Y. Pan, X. Wang, and J. Wang, "MMM: classification of schizophrenia using multi-modality multi-atlas feature representation and multi-kernel learning," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29651–29667, 2018.

[21] L. L. Zeng, H. Wang, P. Hu et al., "Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI," *eBioMedicine*, vol. 30, pp. 74–85, 2018.

[22] J. Oh, B. L. Oh, K. U. Lee, J. H. Chae, and K. Yun, "Identifying schizophrenia using structural MRI with a deep learning algorithm," *Frontiers in Psychiatry*, vol. 11, no. 11, p. 16, 2020.

[23] S. Srinivasagopalan, J. Barry, V. Gurupur, and S. Thankachan, "A deep learning approach for diagnosing schizophrenic patients," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 31, no. 6, pp. 803–816, 2019.

[24] F. Wang, J. Yu, Z. Liu, M. Kong, and Y. Wu, "Study on offshore seabed sediment classification based on particle size parameters using XGBoost algorithm," *Computers & Geosciences*, vol. 149, article 104713, 2021.

[25] B. Ullah, M. Kamran, and Y. Rui, "Predictive modeling of short-term rockburst for the stability of subsurface structures using machine learning approaches: t-SNE, K-means clustering and XGBoost," *Mathematics*, vol. 10, no. 3, p. 449, 2022.

[26] P. M. L. Ching, X. Zou, D. Wu, R. H. Y. So, and G. H. Chen, "Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine," *Environmental Research*, vol. 210, article 112953, 2022.

[27] Y. Dan, J. Hongbing, and G. Yongchan, "A robust D–S fusion algorithm for multi-target multi-sensor with higher reliability," *Information Fusion*, vol. 47, pp. 32–44, 2019.

[28] F. Gao, B. Li, L. Chen, Z. Shang, X. Wei, and C. He, "A softmax classifier for high-precision classification of ultrasonic similar signals," *Ultrasonics*, vol. 112, article 106344, 2021.

[29] A. Prabha, J. Yadav, A. Rani, and V. Singh, "Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier," *Computers in Biology and Medicine*, vol. 136, article 104664, 2021.