WILEY | Hindawi

Research Article

Optimized Naive-Bayes and Decision Tree Approaches for fMRI Smoking Cessation Classification

Amirhessam Tahmassebi [iD],[1] Amir H. Gandomi,[2] Mieke H. J. Schulte,[3,4] Anna E. Goudriaan,[3] Simon Y. Foo [iD],[5] and Anke Meyer-Baese [iD][1]

[1]*Department of Scientific Computing, Florida State University, Tallahassee, FL 32306-4120, USA*
[2]*School of Business, Stevens Institute of Technology, Hoboken, NJ 07030, USA*
[3]*Amsterdam Institute for Addiction Research, Department of Psychiatry, Academic Medical Center,*
University of Amsterdam, Amsterdam, Netherlands
[4]*Addiction, Development, and Psychopathology (ADAPT) Lab, Department of Psychology,*
University of Amsterdam, Amsterdam, Netherlands
[5]*Department of Electrical and Computer Engineering, FAMU-FSU College of Engineering,*
Tallahassee, FL 32310-6046, USA

Correspondence should be addressed to Amirhessam Tahmassebi; atahmassebi@fsu.edu

This paper aims at developing new theory-driven biomarkers by implementing and evaluating novel techniques from resting-state scans that can be used in relapse prediction for nicotine-dependent patients and future treatment efficacy. Two classes of patients were studied. One class took the drug N-acetylcysteine and the other class took a placebo. Then, the patients underwent a double-blind smoking cessation treatment and the resting-state fMRI scans of their brains before and after treatment were recorded. The scientific research goal of this study was to interpret the fMRI connectivity maps based on machine learning algorithms to predict the patient who will relapse and the one who will not. In this regard, the feature matrix was extracted from the image slices of brain employing voxel selection schemes and data reduction algorithms. Then, the feature matrix was fed into the machine learning classifiers including optimized CART decision tree and Naive-Bayes classifier with standard and optimized implementation employing 10-fold cross-validation. Out of all the data reduction techniques and the machine learning algorithms employed, the best accuracy was obtained using the singular value decomposition along with the optimized Naive-Bayes classifier. This gave an accuracy of 93% with sensitivity-specificity of 99% which suggests that the relapse in nicotine-dependent patients can be predicted based on the resting-state fMRI images. The use of these approaches may result in clinical applications in the future.

1. Introduction

Smoking cigarettes is the leading cause of preventable mortality in the United States, with around 50% of lifelong smokers dying from illnesses such as heart disease, stroke, and cancer [1]. In addition to this, insomnia, tremors and quivering, lightheadedness, high blood pressure, heart attack, and decreasing bone density are just a few symptoms that nicotine could cause. Nicotine has been shown to have addictive potential [2, 3]. Moreover, nicotine has been shown to activate the mesolimbic dopamine system, specifically, the ventral tegmental area, which reinforces the effects of nicotine [4]. Developing a cessation treatment with a compound that will reduce a patient's dependence on nicotine, as well as the effects of withdrawal, could help millions of people [5]. One of these new, potentially effective compounds is N-acetylcysteine (NAC) [6]. NAC ($C_5H_9NO_3S$, mw: 163.19) is a derivative of the amino acid cysteine prodrug which is approved as a mucolytic agent and an acetaminophen antidote. NAC restores the basal level of glutamate in the accumbens which may reduce the drug seeking behavior [7].

In a pilot study, a trend was observed for fewer withdrawal symptoms after smoking cessation for subjects taking NAC versus subjects taking the placebo [8]. The goal of this

paper is to show that NAC affects brain functions related to addiction. NAC has been shown to normalize glutamate levels in cocaine-dependent patients [6], which indicates that it could also affect smoking behavior by changing glutamate levels. Although there have been negative trials of NAC on cessation of cocaine use in cocaine dependence, some trials do show that in subjects who were already abstinent at the start of the trial, NAC may be useful in reducing relapse in these subjects [7, 9, 10]. Functional magnetic resonance imaging (fMRI) is a noninvasive technique used for functional brain mapping [11, 12]. Interpreting fMRI connectivity maps based on machine learning algorithms represents an important step towards a computational classification model [13–15]. Employing machine learning algorithms and statistical inference methods are commonly used to implement computational models to find relations between fMRI data and related tasks. Machine learning may also result in clinical applications in the future. For instance, in mental health disorders, no clinical factors are known that can predict whether patients will respond to a specific treatment. Being able to identify which patients will respond to a treatment, using baseline neuroimaging data, could result in improved treatment outcomes, for instance, by early identification through fMRI. Thus, personalized medicine could in the future result in targeted interventions [16–20].

Previously, Smitha et al. [21] have investigated the functional connectivity applied to region of interests (ROIs) extracted by Brodmann template using CONN toolbox. The classification accuracy was less than 50% using linear support vector machines (Linear SVM). Since Linear SVM results in low accuracy, it is important to employ new methods for extracting ROIs and multivariate machine learning algorithms for classifications. In addition, the use of decision trees results in even lower accuracy [22]. Moreover, the suitability of the random subspace ensemble method for fMRI classification has been investigated [23]. It has previously been shown that decision trees were successfully used for variable selection and classification in fMRI brain activities [24–27].

The aim of this study was to be able to develop a way to use fMRI data to predict which patients will relapse and which will not. This classification is normally done after 6 months of treatment or 12 months past the start of addiction treatment [28, 29]. It is based on self-reports by the addicted persons. When it is possible to classify responders to NAC and placebo based on the fMRI data using machine learning algorithms, there would be a tool which could be used in a future sample, to predict who will relapse after a mixed intervention and who needs additional treatment when trying to stop. This would be useful in clinical practice to make earlier decisions about when to start NAC treatment (based on their fMRI classification): at the start of treatment, instead of having to wait for 6 or 12 months to see whether NAC has effects.

In this study, four different machine learning classifiers along with features based on high activity areas in the brain extracted by a novel scheme for voxel selection were employed. Areas of high activity are defined to be those where more oxygen-rich blood is flowing and fMRI is able to map these areas. In addition to this, the accuracy of classification will rely heavily on how the data is reduced. Thus, three different data reduction techniques were employed to perform this study. Tahmassebi et al. [30] have recently shown that 10 components could be employed as the optimal number of features extracted using their novel voxel selection scheme in the fMRI smoking cessation study. First, validation of the voxel selection schemes as well as the machine learning classifiers is required. Therefore, the extracted features from fMRI brain scans were employed to predict who received the drug NAC and who received a placebo. Although, this prediction will not have any clinical utility, it would guarantee that the machine learning models are valid and ready to be applied for prediction relapse versus nonrelapse.

2. Data Acquisition

The main goal of this study was to determine whether or not the drug N-acetylcysteine (NAC) would decrease nicotine dependency. NAC may have an effect on relapse in smoking cessation [8, 31]. In this regard, 39 regular smokers participated in this treatment study at the Spinoza Center of the University of Amsterdam. 19 heavy smokers who wanted to quit took the drug NAC (class 1) and the other 20 subjects took a placebo (class 0) for two weeks. Anatomical and functional scans of their brains were taken at baseline and after two weeks of NAC treatment. Then, the relapse data were assessed at six months after NAC treatment [32]. The Spinoza Center of University of Amsterdam is equipped with a 3.0 T Intera MRI scanner (Philips Health care, Best, The Netherlands) with a 32-channel SENSE head coil to obtain MRI data. The subjects were asked to keep their eyes closed, stay relaxed, and stay awake during the scan (resting state). Two hundred 3-dimensional functional images of the subjects' brains of size $80 \times 80 \times 37$ with a voxel size of $3\,mm^3$ with 2.3 seconds as repetition time were taken due to the sensitivity to blood oxygen level-dependent (BOLD) contrast by the gradient-echo planar sequence. In addition to this, the 3-dimensional anatomical data of size $240 \times 240 \times 220$ with a voxel size of $1\,mm^3$ have been acquired. Figure 1 shows slices of the brain from one patient in all three axes for both the anatomical (Figure 1(a)) and the functional (Figure 1(b)) representations.

Magnetic fields of the nuclei in oxygen-rich blood are flipped due to the combination of a strong magnetic field and radio waves. This produces a detailed map of the regions where the ratio of flow of oxygen-rich blood to the brain is high which explains the high activity areas of the brain. This is called BOLD signal which is also studied in this paper. The BOLD signal is generally modeled as the convolution of the stimulus function with Hemodynamic Response Function (HRF) [33–36]. The energy due to an influx of oxygenated blood to a local area of neuronal activity produces the BOLD signal. Oxygenated hemoglobin has a diamagnetic effect. However, hemoglobin would show paramagnetic characteristics once it is deoxygenated. The MRI machine produces a magnetic field which aligns the randomly oriented atomic nuclei within the direction of the magnetic field [21, 37].

(a) Anatomical
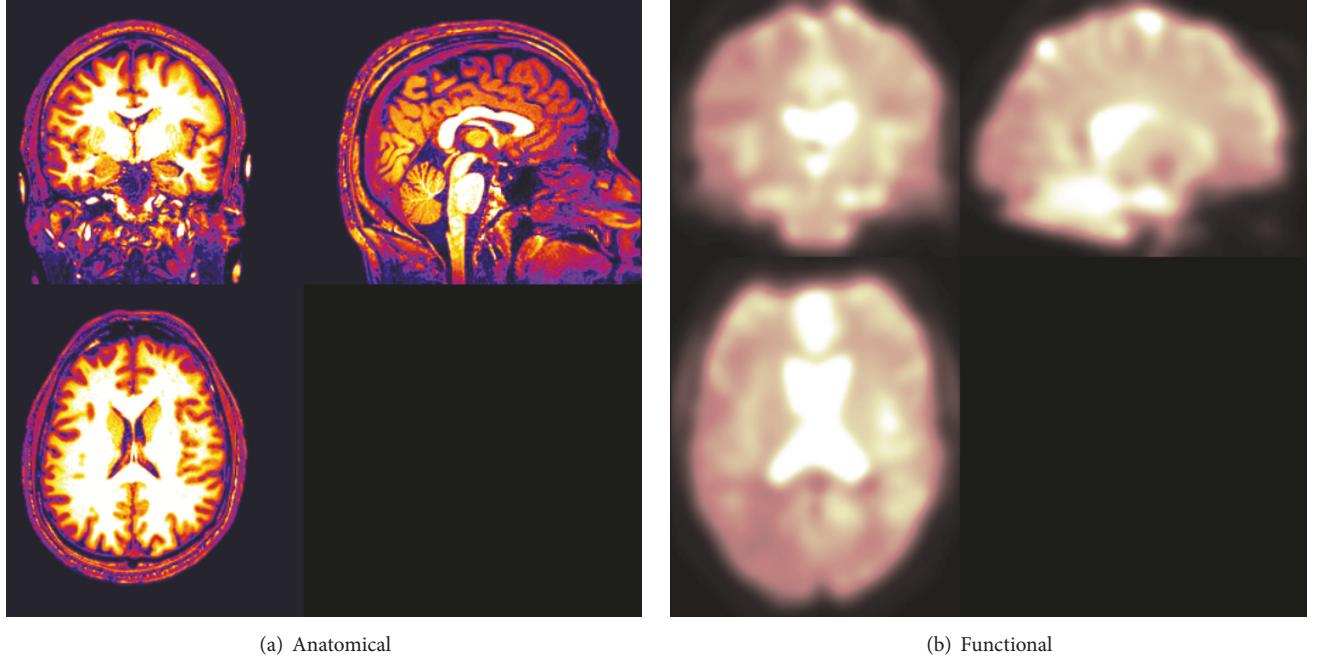
(b) Functional

FIGURE 1: Anatomical and functional slices of the brain.

3. Data Preprocessing

We were given the fMRI data in 4-dimensional spatiotemporal NIFTI (Neuroimaging Informatics Technology Initiative) format. The data contains subject-dependent artifacts due to the long process of the scans, possible movements of the subjects, and physiological noise [38]. The fMRI data analysis pipeline is made using a combination of Statistical Parametric Mapping (SPM12) and FMRIB Software Library (FSL) to increase the BOLD contrast to noise ratio.

The preprocessing stage [39] includes the following. (1) Motion correction to shift the voxels location which would not match the anatomical location. 1% variation of the signal due to the movement of the subject can be greater than the BOLD signal which are extracted as features. (2) Segmentation and realignment to correct in-plane rotations and translations of the head within the image by selecting the first image as the reference. (3) Temporal slice timing to correct and shift in order the interleaved fashion slices. (4) Smoothing to increase the signal-to-noise (SNR) ratio using a general linear model along with a Gaussian to approximate the haemodynamic response function (HRF) [40, 41]. (5) Normalization using a Gaussian full width half maximum (FWHM) kernel of 3 mm for each voxel. (6) Coregistration to map functional information into anatomical space using Talairach standard coordinates [42] and the Montreal Neurological Institute (MNI) [43] brain templates. Figure 2 demonstrates the correlation matrix of raw (Figure 2(a)) and preprocessed (Figure 2(b)) data. It should be noted that Figure 2 is reproduced from Tahmassebi et al. (2017) [44] [under the Creative Commons Attribution License/public domain].

To extract high activity features from the big data, a novel voxel selection scheme (mask) [30, 45] was applied on 94,720,000 features of the preprocessed data for each subject. A mask is a 3-dimensional array of 0 s and 1 s, where 1 signifies keeping the voxel in that position and 0 indicates ignoring the voxel in the data. This makes the problem more feasible to feed the machine learning algorithms with 94,720 features which has 0.1% share of the given big data.

4. Features Extraction

Dealing with a size of $39 \times 94,720,000$ as the feature matrix with available computational equipment would be computationally expensive and inaccurate. After applying voxel selection scheme, the size of the feature matrix was reduced to $39 \times 94,720$, which is still a large number for a feature vector for classification. Here, three robust algorithms were employed for data reduction to find the feature matrix.

4.1. Independent Component Analysis (ICA). ICA [46] is a technique to separate a multivariate signal into multiple independent non-Gaussian signals. It extracts the hidden spatiotemporal structure in neuroimaging. Maximum independence of underlying signals is the essential assumption of ICA. In principle, independence of two random variables would lead to uncorrelation between variables [47]. ICA demixes the recorded brain signals at each voxel and extracts the relevant components along with the artifacts [48]. There are some algorithms such as infomax, JADE, and FastICA to employ this approach [49]. In this paper, the FastICA algorithm was used. The FastICA is a hierarchical and symmetric approach which minimizes the mutual information by

(a) Raw   (b) Preprocessed
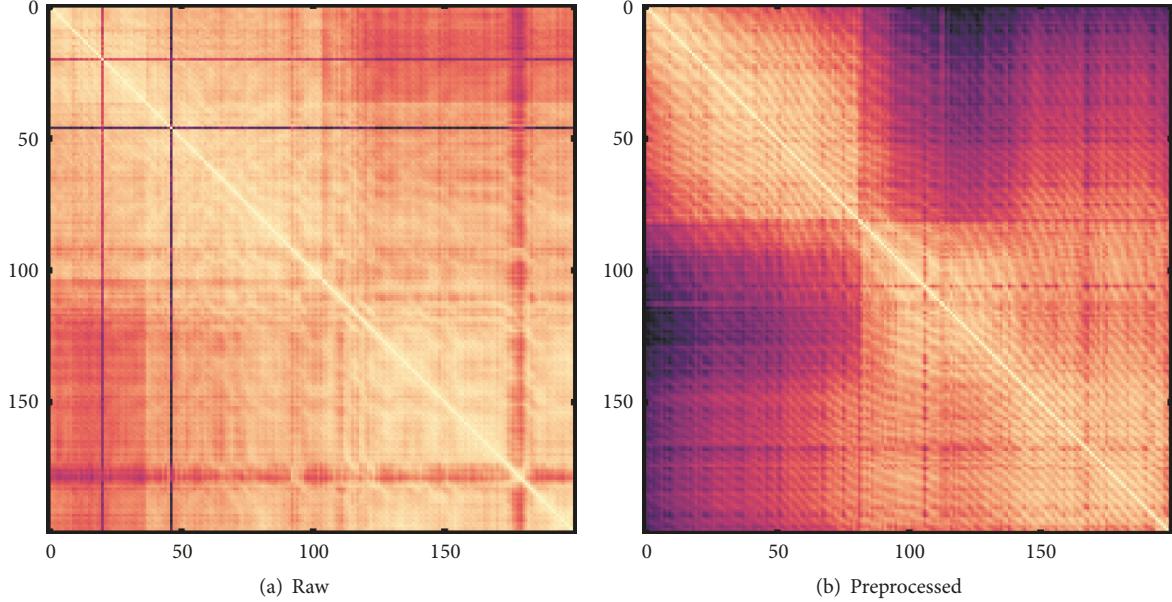
FIGURE 2: Raw and preprocessed fMRI data.

employing non-Gaussianity and measurement of negentropy [48]. ICA tries to extract a feature matrix like U from the full rank matrix A. With having N patients and M features for each patient, the feature matrix size would be $[A]_{N\times M}$. It is desired to find a good approximation with respect to the source, W, which is an unmixing matrix providing a linear decomposition of A. Thus, for extracting q features out of M features, we have

$$[U]_{N\times q} = [A]_{N\times M} [W]_{M\times q}. \tag{1}$$

Feeding vector $[U]_{39\times 10}$ into various classifiers would be the next level. We hope that ICA will tell us what regions of the brain share similar brain activities.

4.2. Principal Component Analysis (PCA). PCA transforms features of the original input data orthogonally to a new space to reduce redundancy and reach high information density. These variables in the new space are principal components which are linearly uncorrelated. PCA is also referred to as Karhunen-Loeve transformation or the Hotelling transform [50]. PCA normalizes the input data within the unit interval and chosen based variance. It was shown that better discriminatory results were found by choosing a larger variance. The first principal component has the largest variance within the data set and the last one has the smallest variance. Orthogonality of all principal components to each other gives us an orthogonal basis set. Let R be the correlation matrix and λ_i the corresponding ith eigenvalue of the matrix R and Q the eigenvector matrix with q_i columns. It should be noted that Q is an orthogonal matrix ($Q^T Q = I$). By spectral theorem we have

$$[R]_{m\times m} = [A]_{m\times n} \left[A^T \right]_{n\times m} = \sum_{i=1}^{m} \lambda_i q_i q_i^T. \tag{2}$$

We could create a new basis by choosing eigenvectors q_i to rewrite the original data again using C as the coefficient vector, which is the projection of A onto the principal directions:

$$A = \sum_{i=1}^{m} q_i c_i = QC = QQ^T A. \tag{3}$$

To demonstrate the importance of PCA, which is reducing the dimension of the data, we could have a rank k approximation A_r of the original data:

$$A_r = \sum_{i=1}^{k} q_i c_i = Q_r Q_r^T A. \tag{4}$$

Employing all the above-mentioned equations leads us to a strategy such as subspace decomposition to find the largest eigenvalue and project the original data orthogonally onto its subspace. Thus, the eigenvalues closer to zero which contain the redundant information will be discarded [50].

4.3. Singular Value Decomposition (SVD). SVD is the factorization of matrix $[A]_{m\times n}$ to the form $U\Sigma V^T$, where U is an $m \times m$ unitary matrix, Σ is an $m \times n$ diagonal matrix, and V is an $n \times n$ unitary matrix. The diagonal values of Σ are the singular values of the original matrix and the columns of U and V are the left and right singular values of the original matrix, respectively. In this paper, the diagonals of the matrix Σ are used since this tells us the properties of the matrix. It can be used to compare with the other matrices and reduces the dimension of the matrix

$$[A]_{m\times n} = [U]_{m\times m} [\Sigma]_{m\times n} [V]_{n\times n}^T. \tag{5}$$

The performance of different data reduction algorithms with different numbers of components has been previously

(a) ICA

(b) PCA

(c) SVD
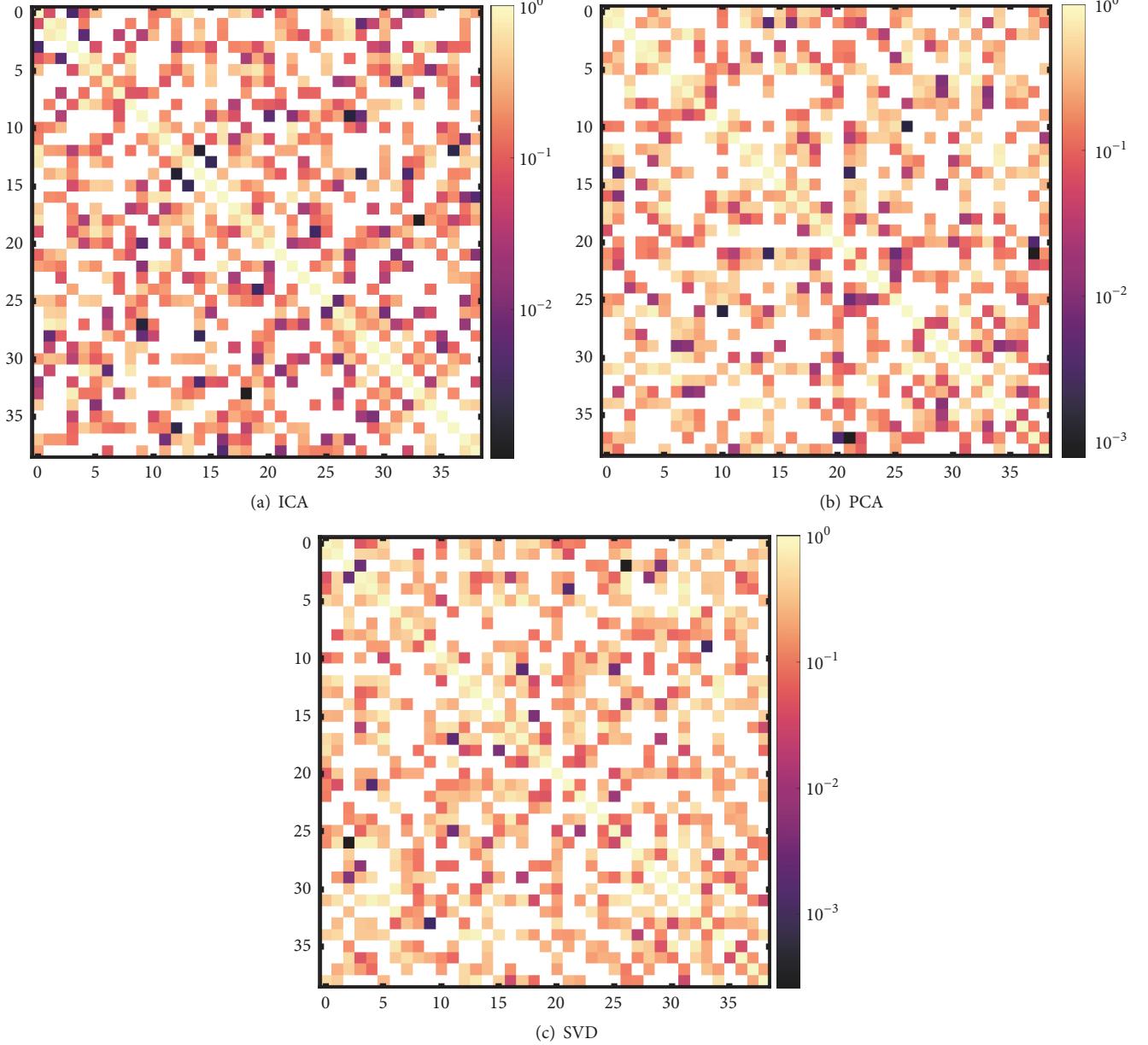
Figure 3: Illustration of correlation matrices with 10 components for ICA, PCA, and SVD data reduction techniques.

investigated, which suggests that 10 components would be ideal for this analysis [30, 44, 51]. Figure 3 depicts the correlation matrices produced by getting the correlation of the independent components (Figure 3(a)), principal components (Figure 3(b)) and singular values (Figure 3(c)).

5. Classification

There are two classification strategies for medical images. For the first strategy, measurements of a set of features from a region in an image would be employed as the feature vector. This is called region-based classification. For the second strategy, voxel-based classification is used, where contextual or noncontextual information about every single voxel is

used as feature vector to feed into the classifier [50]. In this paper, multivariate voxel-based analyses along with various classifiers such as decision tree and Gaussian Naive-Bayes with two implementations were employed [52].

5.1. Decision Tree Algorithm. Decision tree is a machine learning algorithm which partitions a set of input data recursively. A decision tree structure is made of a root node, which has no incoming edges and zero or more outgoing edges, internal nodes, each of which has one incoming edge and more outgoing, and leaf nodes, each of which has one incoming and no outgoing edges [53]. Based on the volume of the data and available memory resources, decision tree algorithms can be implemented in a serial fashion such as

CART (classification and regression tree) [54], C4.5 [55], and IDE3 (iterative dichotomizer 3) [56, 57] or parallel fashion such as SLIQ (supervised learning in ques) [58] and SPRINT (scalable parallelizable induction of decision trees) [59]. Dealing with a feature matrix of size 39×10 after data reduction and voxel selection scheme motivates us to employ a serial algorithm to implement the decision tree. In addition to this, decision trees are easy to interpret by Boolean logic and can also be visualized.

5.1.1. CART. The CART, by Breiman et al. [54], builds both classification and regression trees based on binary splitting of the features selected based on Langs et al. [60] index splitting measure. In principle, it follows Hunt et al.'s algorithm [61]. It yields the largest information gain at each node. For a given training data set X and label vector Y, CART partitions the space recursively such that the matched instances and labels are clustered together [62]. For the D amount of data at node k, CART partitions the data into D_{Left} and D_{Right} subsets based on the splitting threshold and feature. The impurity at node k is calculated through the impurity measure Gini as $\sum_i p_{ki}(1 - p_{ki})$ with p_{ki} as the proportion of class i observations in node k [62]. For a data set of size $\{N \text{ samples} \times M \text{ features}\}$, the run time cost order to build the tree is $O(NM \log(N))$ and the query time is $O(\log(N))$.

Based on the CART algorithm of how to build trees, if the splitting process continues to the point that there are few samples in each leaf of the tree, it is likely to overfit the data. On the other hand, a small tree also might not capture the important structural information about the sample space. This problem is known as the horizon effect. Therefore, the complexity of the tree in such a way that the estimated true error is low is desired. In this regard, a reduced error pruning algorithm, which is a bottom-up fashion pruning method, was employed. This improves predictive accuracy by starting at the leaves and replacing each node with its most popular class to reduce overfitting and increase the simplicity of the tree and speed of the process. This process continues until the prediction accuracy is not affected. The optimization part was repeated 51 times for each of the data reduction methods to reach the most efficient result.

5.2. Naive-Bayes Algorithm. Naive-Bayes is a classification technique based on Bayes Theorem with an assumption of independence among predictors to model probabilistic relationships between the feature matrix and the class labels [53]. In simple terms, a Naive-Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Bayes Theorem combines prior knowledge of the classes with new evidence gathered from training data [53]. First, the Naive-Bayes model builds the frequency table of the training data set. Then, it creates the likelihood table by calculating the probabilities. Finally, it calculates the posterior probability for each class and the class with maximum posterior probability is the result of the prediction. Naive-Bayes classifier is easy to implement, useful for big data problems, and known to outperform even highly sophisticated classifiers. In this paper, a standard algorithm

for Gaussian Naive-Bayes and an optimized version of Naive-Bayes were employed.

5.2.1. Gaussian Naive-Bayes (GNB). The essential principle in Bayes method is assuming a known a priori and then minimization of the classification error probability, respectively. The class-conditional density function could be known or estimated from the available training dataset. During Bayesian estimation, the training set conditioned density function is getting updated by the training set which acts as observations to allow the conversion of the a priori information into an a posteriori density [50].

A simple introduction can be given by considering two pattern classes: (1) NAC and (2) Placebo. To make the mathematical notations easier, we call them w_1 and w_2, respectively. Recalling the Bayes rule, we have

$$P(w_i \mid x) = \frac{p(x \mid w_i) P(x_i)}{\sum_{i=1}^{2} p(x \mid w_i) P(w_i)}. \tag{6}$$

For the two-class patterns w_1 and w_2, we have two Bayes classification rules as follows:

$$
\begin{aligned}
&\text{If } P(w_1 \mid x) > P(w_2 \mid x), \quad x \text{ is assigned to } w_1, \\
&\text{If } P(w_1 \mid x) < P(w_2 \mid x), \quad x \text{ is assigned to } w_2.
\end{aligned} \tag{7}
$$

Considering the Gaussian probability distribution function with μ_i as the mean value and Σ_i as the covariance matrix for discriminant functions makes it more feasible to be solved.

$$
\begin{aligned}
& p(x \mid w_i) \\
& = \frac{1}{(2\pi)^{n/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\right) \quad (8) \\
& \hspace{8cm} i = 1, 2.
\end{aligned}
$$

By choosing a monotonic logarithmic discriminant function we have

$$
\begin{aligned}
g_i(x) = &-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) + \ln P(w_i) \\
&-\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| \quad i = 1, 2.
\end{aligned} \tag{9}
$$

Now, by calculating the mean vector and covariance matrices of the discriminant function for each class from the training data, the data can be separated by a hyperplane (if they have an equal covariance matrix) or hyperquadrics (if they have an unequal covariance matrix).

5.2.2. Optimized Naive-Bayes (ONB). To optimize the Naive-Bayes standard algorithm, the bag-of-token model was employed [63]. In this way, the value of each feature k is calculated based on the nonnegative number of occurrences of token k in the observation. For estimated probability we have

$$p(\text{token} = k \mid \text{class} = l) = \frac{1 + \beta_1}{N + \beta_2}, \tag{10}$$

TABLE 1: Statistics of 51 runs of the reduced error pruning of the CART with 10-fold cross-validation with ICA, PCA, and SVD data reduction algorithms.

Algorithm	Type of tree	Min	Mean	Median	Max	STD
ICA	Original tree	0.435	0.435	0.435	0.435	0.0
	Pruned tree	0.307	0.472	0.487	0.487	0.035
PCA	Original tree	0.794	0.794	0.794	0.794	0.0
	Pruned tree	0.358	0.484	0.487	0.487	0.017
SVD	Original tree	0.615	0.615	0.615	0.615	0.0
	Pruned tree	0.410	0.482	0.487	0.487	0.018

where β_1 is the weighted number of occurrences of token k in class l, β_2 is the total weighted number of occurrences of all tokens in class l, and N is the number of instances in the training set. Then, the classifier predicts the class label for each observation based on the estimated posterior probability presented in (10). In this way, each observation is assigned to the class with the maximum posterior probability [64].

As discussed, the participants were indeed randomized at the first session to receive NAC or placebo. In addition to this, we first learned the two classes (relapse versus nonrelapse) and then tested them in a randomized way choosing randomly participants from the two groups. Moreover, due to the low numbers of participants, k-fold cross-validation was used to overcome overfitting. In principle, in an overfitted model, the underlying relationships of the entire database are not learned, which might give falsified information about the process. In this paper, 10-fold cross-validation was used to test the accuracy of the classification. In this procedure, our testing set is provided by leaving 4 subjects out of the data set and the training set is provided by aggregating the other 9 folds (35 subjects) to use in the predictive process. Each classifier's discriminant function was fitted on the training data set for each fold. This was done 10 times in order to cover all the subjects in the testing set. The classification accuracy in each fold is the number of times that the model predicted correctly divided by the number of subjects in the testing fold, (4). The total classification accuracy is the average of the classification accuracies in all 10-folds.

6. Results and Discussion

Two classification tasks with three major data reduction methods, ICA, PCA, and SVD, were developed in Python [62, 65] and MATLAB [64]. Based on the previous studies using this database, the data reduction was done by keeping 10 components [30, 44, 51]. The classification tasks were done in two different phases: (1) validation and (2) relapse. For the model validation as discussed in Section 6.1, the classification was done based on two classes NAC (class 1) and placebo (class 0). This will not have any clinical benefits, but it would guarantee that the voxel selection schemes which were employed to extract features from fMRI brain scans were valid. In addition to this, due to the equal number of subjects for each class (19 NAC, 20 placebo), the classification task is unbiased. This would present the performance of the proposed model without overfitting. Next, classification

was done to separate subjects into two groups of subjects (relapse (class 0) and nonrelapse (class 1)). In this case, class labels were made based on the number of cigarettes that each subject smoked after the treatment for six month. In this paper, people who smoked less than 10 cigarettes during the six months past the treatment were chosen for nonrelapse class (13 subjects) and the rest of the subjects were assigned to the relapse class (26 subjects). In this section, the results regarding the classification tasks for the model validation and the relapse are presented and discussed.

6.1. Validation. As discussed, the pruning process of the cart using the reduced error pruning algorithm was repeated 51 times per data reduction algorithm. Table 1 presents the statistical parameters of the pruning process. In addition to this, the optimization process of the CART misclassification error is shown in Figure 4 for different numbers of terminal nodes. For each of the reduction techniques, cross-validation (solid purple line) and resubstitution (dashed pink line) were done and the best choice (magenta circle) based on the cross-validation accuracy was reported. It is preferred to use a simpler tree if it is roughly as good as a more complex tree according to Occam's Razor Principle. Therefore, a cut-off value equal to the minimum cost plus one standard error (dotted black line) was employed.

For the learning phase as Figures 4(a), 4(b), and 4(c) picture, the resubstitution error decreased as the number of terminal nodes increased. It is assured that the classifier can predict the training dataset very well. The most complex tree was made by SVD with 7 terminal nodes. This pattern normally was expected to be repeated. However, for the cross-validation error in Figure 4(c) until the second terminal nodes of the tree, the classifier did not predict the testing samples very well. This might be the reason why the model finished with 7 terminal modes. The lowest misclassification error for 10-fold cross-validation was presented in Figure 4(a) and also in Table 1 with 0.307 for ICA.

To summarize the number of samples predicted correctly (true positive (TP) for positive class, and true negative (TN) for negative class) and incorrectly (false positive (FP) for positive class and false negative (FN) for negative class) they were calculated as a confusion matrix [53]. The fraction of positive instances predicted correctly defines true positive rate (TPR) or sensitivity. In addition to this, true negative rate or specificity is defined as the fraction of correct prediction of negative instances. Figure 5 shows the confusion matrices of

(a) ICA

(b) PCA

(c) SVD
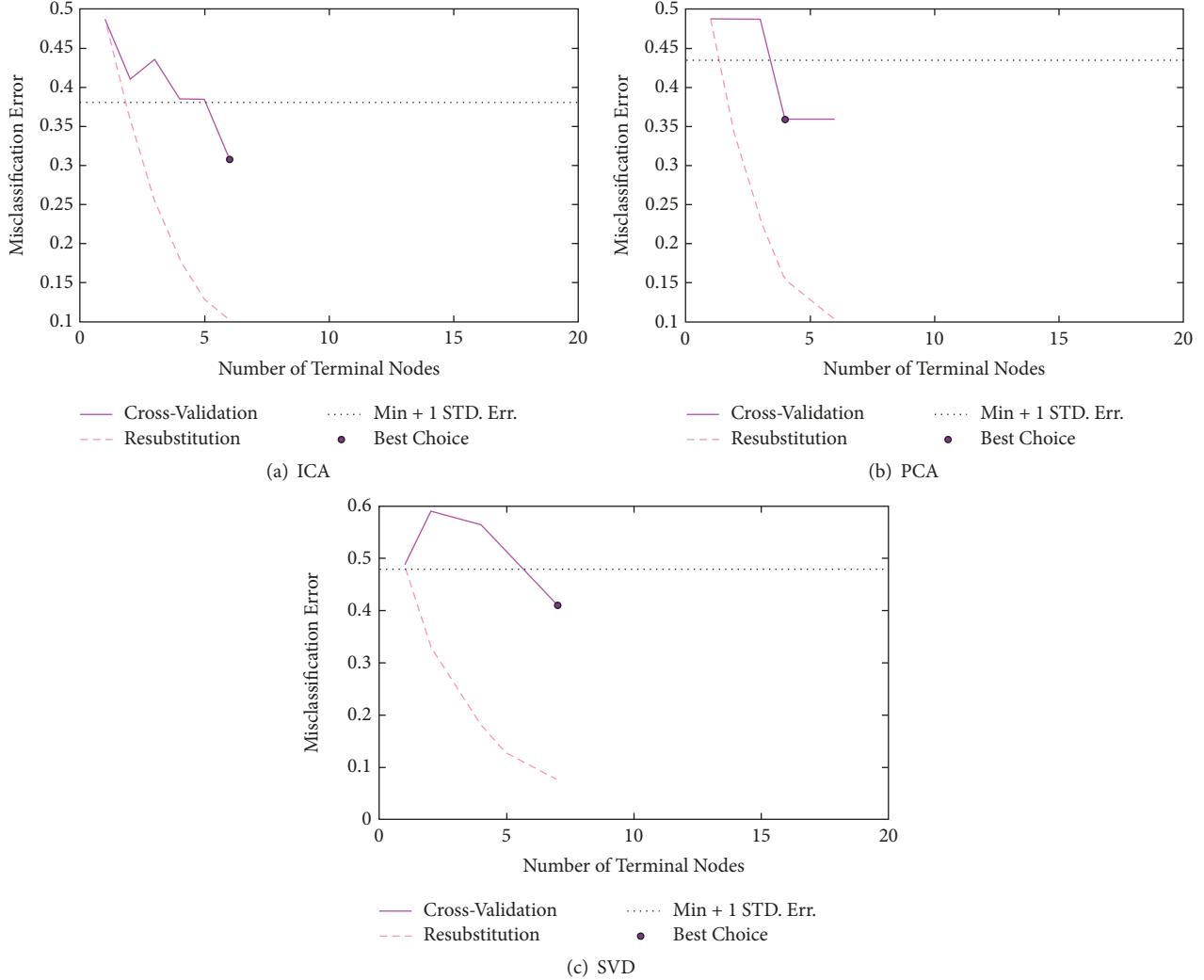
FIGURE 4: Misclassification error for the CART for different numbers of terminal nodes with ICA, PCA, and SVD data reduction techniques.

classification using the CART with different data reduction methods.

As shown in Figure 5, the rows of the confusion matrix plot correspond to the predicted class (Output Class), and the true class (Target Class) is shown as the columns of the confusion matrix plot. In addition to this, the prediction accuracy of the classes of observations by the trained classifier is depicted on the diagonal cells while the off diagonal cells depict where the classifier failed at predictions of the classes of observations. The column on the far right of the plot shows the accuracy for each predicted class, while the row at the bottom of the plot shows the accuracy for each true class. The overall accuracy is shown in the bottom right cell of the plot [64].

As seen in Figure 5, the best classification accuracy was reported for SVD (Figure 5(c)). ICA and SVD feature extraction schemes performed well on placebo predictions since the CART predicted all the 20 subjects in placebo class correctly (51.3%). Employing SVD along with the CART, out of 23 placebo predictions, 87.0% were correct and 13.0% were

incorrect. Out of 16 NAC predictions, 100% were correct, and out of 20 subjects in the placebo class, 100% were correctly predicted as placebo with zero error. Out of the 19 subjects in the NAC class, 78.9% were correctly classified as NAC and 21.1% were classified as placebo. Overall, 92.3% of the predictions were correct and 7.7% were incorrect classifications based on the cross-validation. The overall accuracy for ICA and PCA was 89.7% and 84.6%, respectively. It should also be noted that the model with SVD being more complex was already discussed. This complexity becomes more clear in Figure 6(c). As we compare the tree evolution of the models, it can be seen that the ICA model can improve with reduced complexity and high accuracy. Moreover, the contour plots of the probability surface for classification using the CART were presented in Figure 7. In a statistical-classification, a decision boundary or decision surface is a hypersurface that partitions the underlying feature space into two sets, one for each class. The classifier classifies all the points on one side of the decision boundary as belonging to one class and all those on the other side as belonging to the other class.

(a) ICA

(b) PCA

(c) SVD
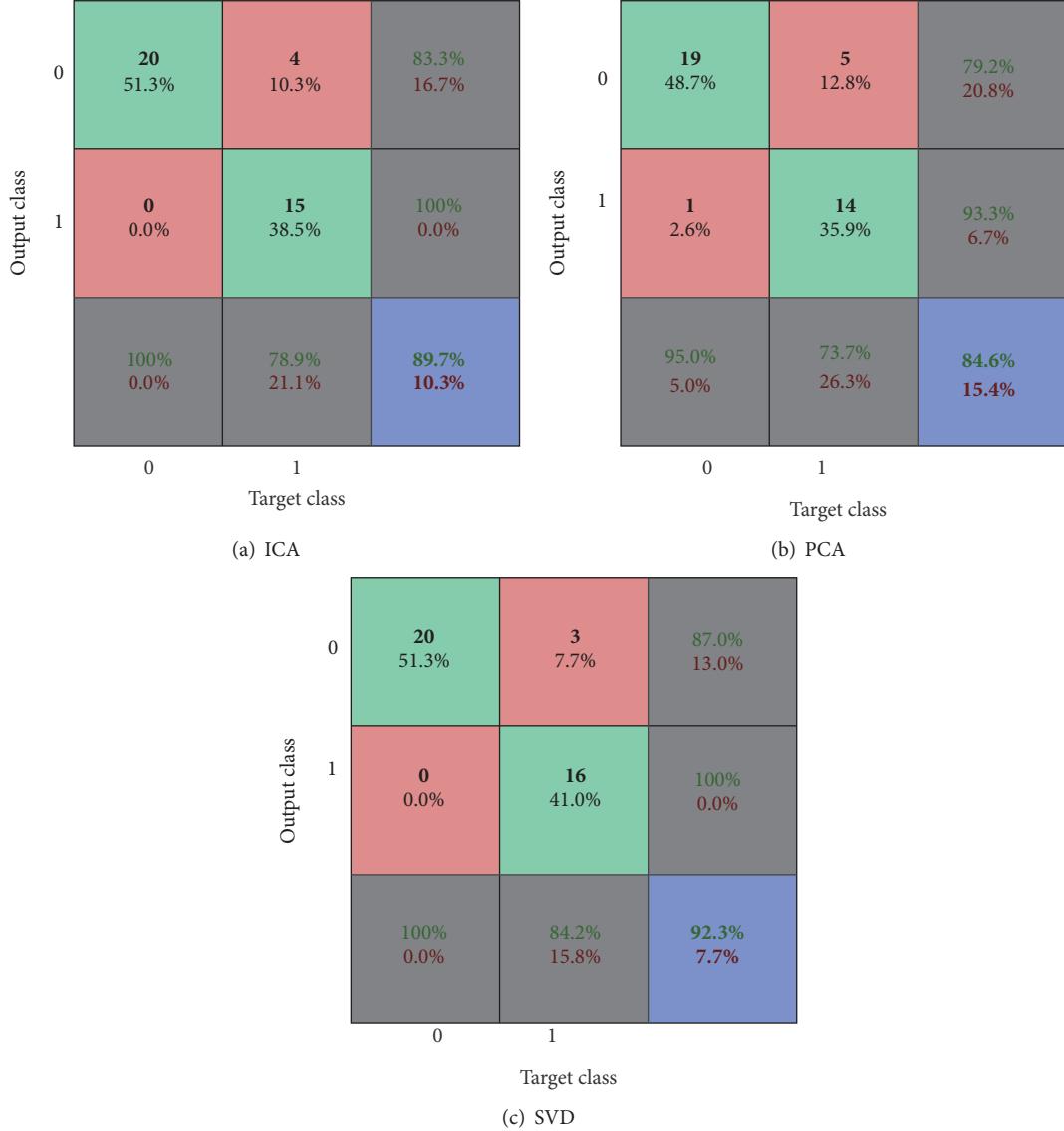
FIGURE 5: Confusion matrices of classification using CART for ICA, PCA, and SVD data reduction techniques.

Decision boundaries are not always clear-cut. In fact, they are the transition from one class in the feature space to another, which is not discontinuous, but gradual. In this plot, red color was chosen to demonstrate the highest probability and blue color represents the lowest probability as well. The black scatter points are the samples predicted by the classifier for each class. It can be seen that it is less probable to predict a subject in the surface area with blue color.

In classification with GNB, as shown in Figure 8, the confusion matrices contain high values for FP, which means the classifier failed at the prediction of the subjects in the class NAC. The low specificity can be seen clearly in Figure 8(b) where the GNB just predicted 4 subjects in the NAC class and did not detect the other 15 subjects in that class. That is the reason FP is 38.5%, which is relatively high for any classification task. Furthermore, this can also be seen clearly in Figure 9, where scatter black points can be found in blue

parts of the surface areas, specifically for class 1 prediction. This matches the previous results with confusion matrices and confirms that, overall, the GNB classifier could not predict the subjects in the NAC class for different feature matrices based on ICA, PCA, and SVD very well. This is more obvious in Figure 9 where the classifier has predicted the subjects in the NAC class with 30% probability and the subjects in the placebo class with a probability of more than 90%. Results in Figure 9 confirm the accuracies presented in Figure 8.

Similar to the pruning process of the CART, the ONB algorithm ran for 51 times as well, and the results are presented in Table 2. The performance of the optimization algorithm is obvious by comparing the confusion matrices in Figures 8 and 11. For example, the accuracy for GNB with ICA is improved by 30%. More importantly, the optimization algorithm improved the strength of the classifier in prediction

TABLE 2: Statistics of 51 runs of the optimized and Gaussian Naive-Bayes classifier with 10-fold cross-validation for ICA, PCA, and SVD data reduction algorithms.

Algorithm	GNB	ONB				
	Total	Min	Mean	Median	Max	STD
ICA	0.564	0.358	0.411	0.384	0.538	0.053
PCA	0.615	0.333	0.498	0.487	0.666	0.058
SVD	0.615	0.384	0.495	0.487	0.666	0.039

(a) ICA

(b) PCA

(c) SVD
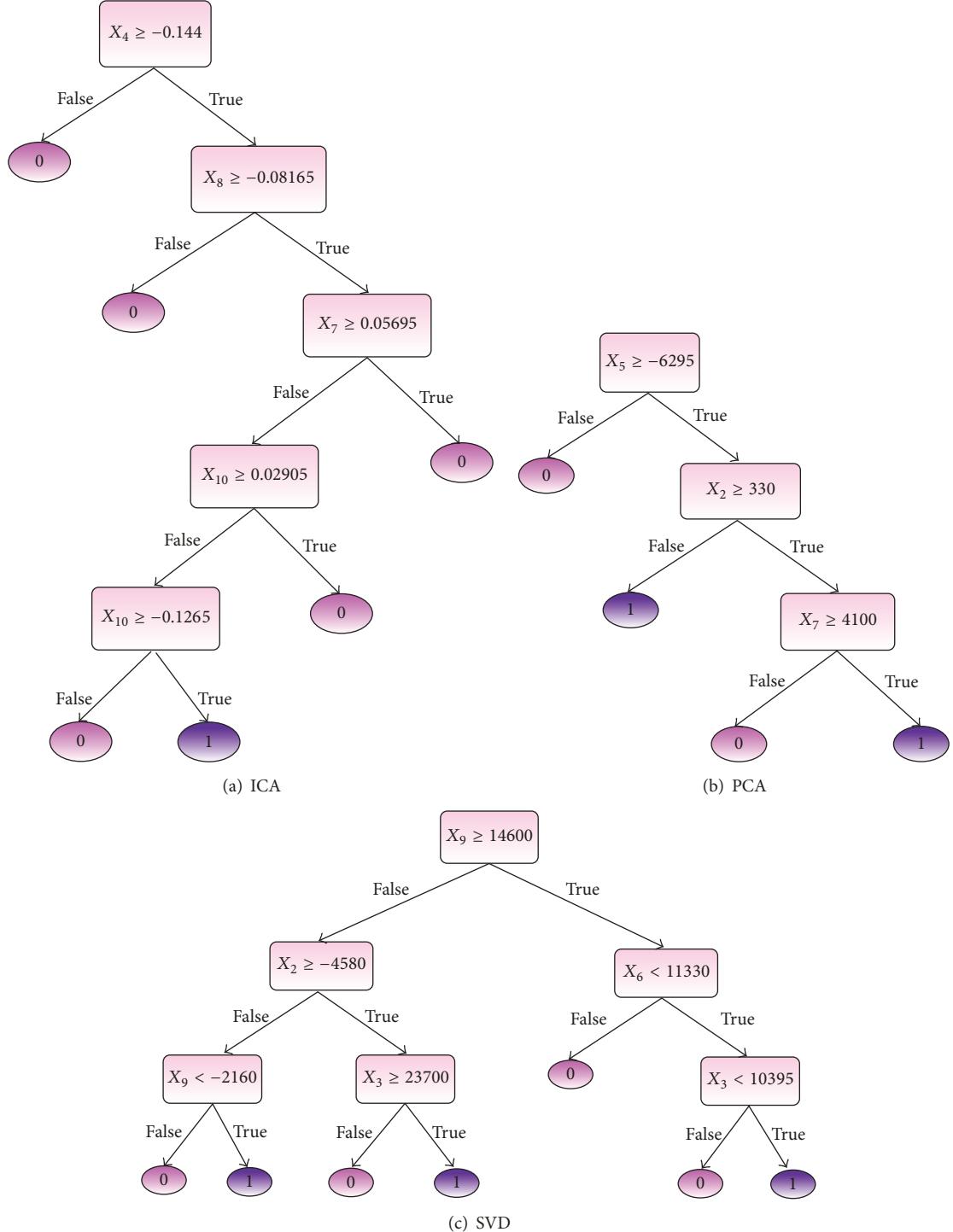
FIGURE 6: Illustration of tree evolution for the CART for ICA, PCA, and SVD data reduction techniques.

(a) ICA

(b) PCA

(c) SVD
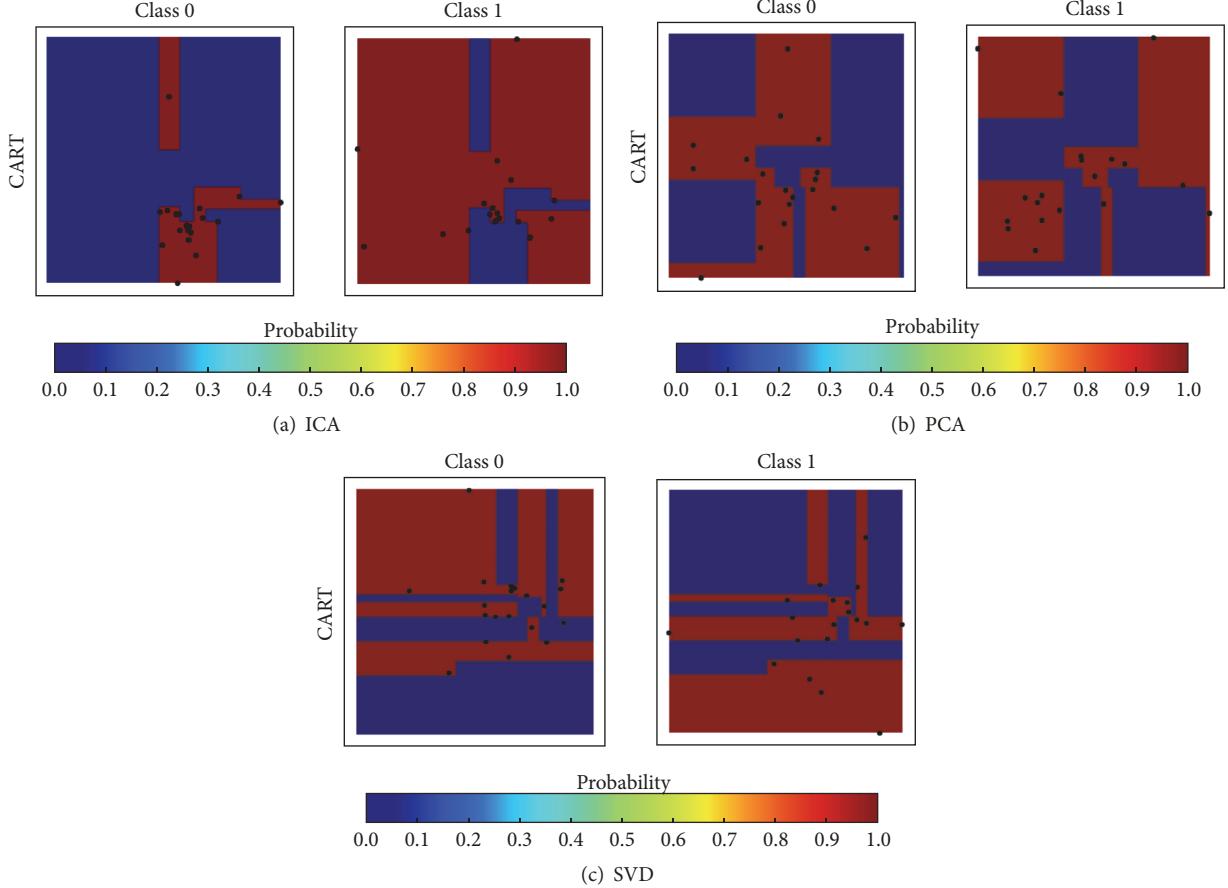
FIGURE 7: Contour plots of probability surface for classification using the CART for ICA, PCA, and SVD data reduction techniques.

of the subjects in the NAC class. As shown in Figure 8(a), GNB predicted 9, 4, and 5 subjects in the NAC class for ICA, PCA, and SVD algorithms, respectively. On the other hand, the correct number of predictions for the NAC class was almost doubled as shown in Figure 11(a). The best result, as presented in Table 2, was achieved using ICA. This is better depicted in Figure 10, which pictures the convergence of the minimum estimated and observed cross-validation error. As seen, the estimated error matches the observed error with ICA. On the other hand, the estimated error started fluctuating after 80 function evaluations for PCA and SVD.

To display the trade-off between sensitivity and specificity of the classifiers, Receiver Operating Curves (ROC) have been employed [53]. The closer the curve follows the left-hand border and the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. The ROC curve visualizes the detection performance of the classifier based on TPR and FPR. As the number of features increases, the classifier's performance increases until we reach the optimal number of features due to the curse of dimensionality. Here, 10 components were chosen as the optimal number. It should be taken into account that increasing the number of features decreases the performance of the classifier. The area under the ROC curve (AUC) for

each data reduction method was also calculated. An AUC of 1.0 confirms the ideal performance of the classifier and an AUC of 0.5 confirms that the performance of the classifier is not acceptable.

Figure 12 shows the ROC curves for the CART, the GNB, and the ONB classifiers with the three feature selection schemes: ICA, PCA, and SVD, respectively. The performance of the ICA as the feature selection scheme along with the ONB with an AUC of 0.99 and the CART with an AUC of 0.95 as classifiers is the best. Moreover, the CART classifier shows an acceptable performance along with all the employed data reduction algorithms. This gives the conclusion that the ICA extracted structural information to be classified by the CART.

6.2. Relapse. Next, the validated model (previously presented in Section 6.1) was employed to predict relapse in nicotine-dependent subjects. Similarly, the CART model underwent an optimization procedure using the reduced error pruning algorithm using 10-fold cross-validation with 51 runs to achieve the best estimation of the error. Table 3 presents the statistical parameters of the estimated error of the CART. As seen, both errors for the original tree and the pruned tree are presented and the best estimated error was achieved using PCA as the feature extraction scheme.

TABLE 3: Statistics of 51 runs of the reduced error pruning of the CART with 10-fold cross-validation with ICA, PCA, and SVD data reduction algorithms.

Algorithm	Type of tree	Min	Mean	Median	Max	STD
ICA	Original tree	0.512	0.512	0.512	0.512	0.0
	Pruned tree	0.333	0.333	0.333	0.333	0.0
PCA	Original tree	0.410	0.410	0.410	0.410	0.0
	Pruned tree	0.256	0.330	0.333	0.333	0.012
SVD	Original tree	0.512	0.512	0.512	0.512	0.0
	Pruned tree	0.307	0.332	0.333	0.333	0.003

(a) ICA

(b) PCA

(c) SVD

FIGURE 8: Confusion matrices of classification using the Gaussian Naive-Bayes for ICA, PCA, and SVD data reduction techniques.

Figure 13 demonstrates the bias-variance trade-off phenomenon [66]. The resubstitution error is the training error of the built tree for prediction of the subjects in terms of relapse and nonrelapse. As shown, the substitution error decreased through the number of terminal nodes for each of the employed data reduction algorithms. The tree was made using 4, 5, and 7 terminal nodes for ICA, PCA, and SVD, respectively. Based on Table 3 and Figure 13(b), it can be concluded that the PCA algorithm along with the CART showed the best performance. The cross-validation estimate of the true error rate decreases as we increase the number of terminal nodes.
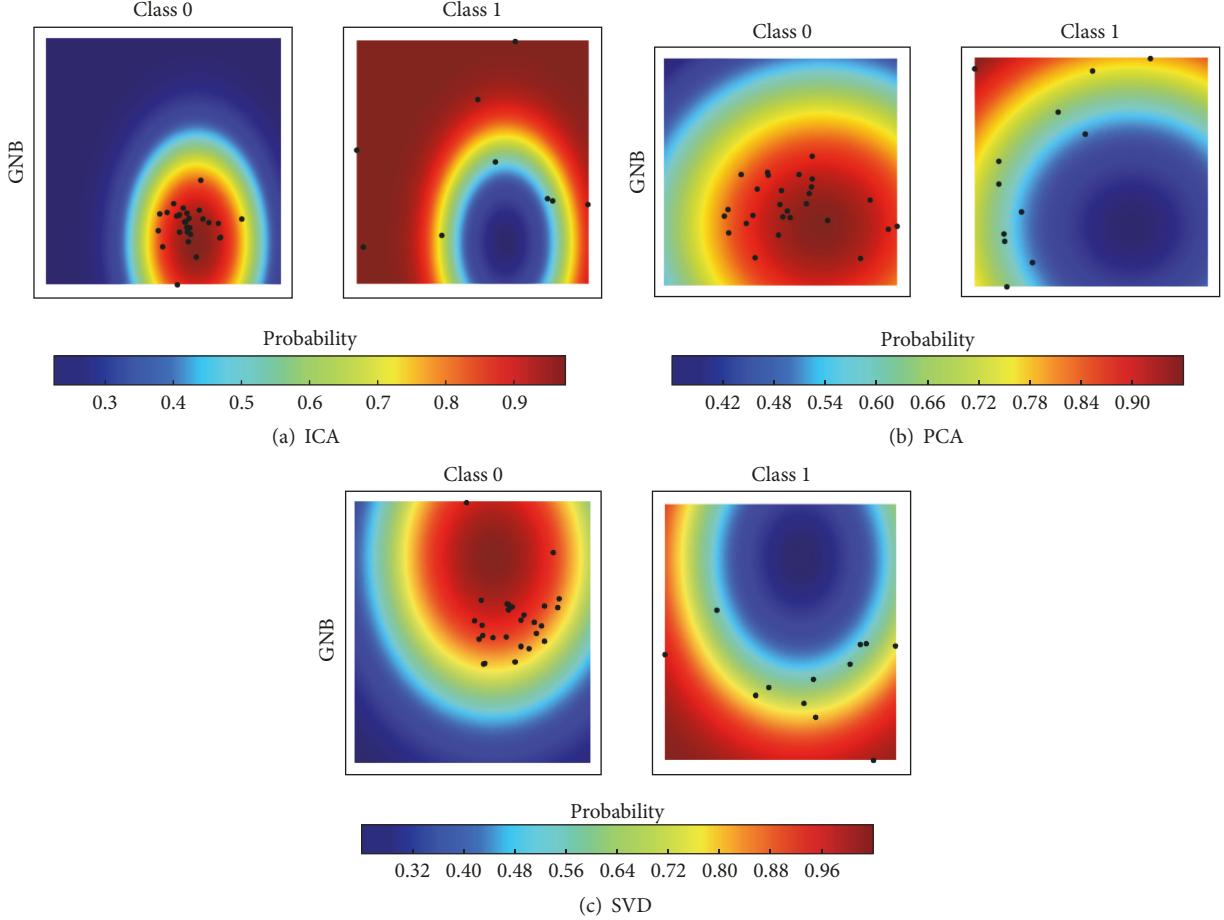
Figure 9: Contour plots of probability surface for classification using the Gaussian Naive-Bayes for ICA, PCA, and SVD data reduction techniques.

As shown in Figure 4(a), the cross-validation error rate started increasing right after the first terminal node. The dotted black line is the minimum error plus one standard deviation of the training error was defined as a metric. In this regard, the desired estimated error rate using cross-validation has to be below this line. As shown, the PCA has the largest margin with the defined metric line and the lowest error rate with 0.256 with respect to ICA and SVD data reduction algorithms.

This result matches the confusion matrices presented in Figure 14. As previously discussed, it is desired to predict the relapse of the subjects in the class 1. However, based on the confusion matrices, the accuracy of 74.4% was achieved, but the CART classifier did not predict the subjects in the nonrelapse class very well. Only 3 out of 13 subjects in the nonrelapse class were predicted by the CART along with the PCA feature extraction scheme. This is also shown in Figure 15(c), where the subjects in the class 1 (nonrelapse) were technically predicted on the edge of the probability surfaces and closer to 0 (blue color). This might be due to the nature of the decision trees which are usually suitable for balanced binary-class problems [53].

Alternatively, the Naive-Bayes classifier based on its probabilistic nature solves the prediction classification problems implicitly. This would be an alternative to predict the subjects in the nonrelapse class better. Table 4 illustrates the statistical parameters of the estimated true error rate using 10-fold cross-validation after 51 runs for optimized Naive-Bayes classifier.

As presented, the best error was achieved along with the SVD data reduction technique with a cross-validation error of 0.410 with a prediction accuracy of 82.1% as depicted in Figure 16(c). Employing a probabilistic classifier while dealing with a nonbalanced binary-class problem outperformed the CART with an accuracy of 69.2% and 85.2% for prediction of the nonrelapse and relapse class, respectively. This result is clearly presented in Figure 17(b), where the GNB predicted 8 subjects out of 9 subjects in the class 1 as shown in the counter plot with probability value of 1 (red color).

By optimizing the Naive-Bayes algorithm after 51 runs, the classifier was able to predict the subjects in the nonrelapse even better. As shown in Figure 18, the solid line is the observed error rate and the dashed line is the cross-validation estimate of the true error rate. After about 25 function

TABLE 4: Statistics of 51 runs of the optimized and Gaussian Naive-Bayes classifier with 10-fold cross-validation for different data reduction algorithms.

Algorithm	GNB	ONB				
	Total	Min	Mean	Median	Max	STD
ICA	0.461	0.282	0.329	0.333	0.384	0.020
PCA	0.589	0.307	0.344	0.333	0.435	0.021
SVD	0.410	0.282	0.319	0.307	0.410	0.024

(a) ICA

(b) PCA

(c) SVD
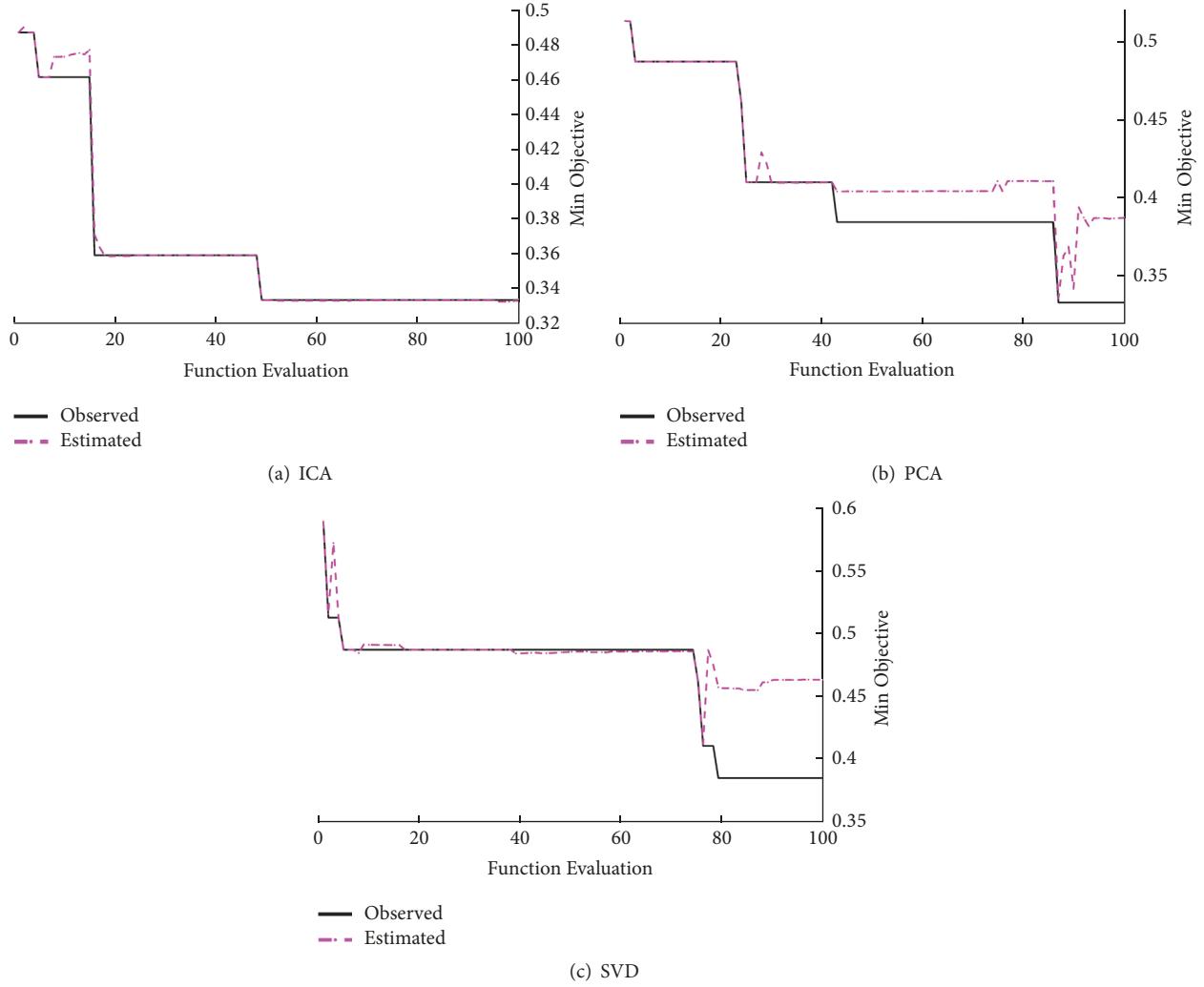
FIGURE 10: Convergence rates of the optimized Naive-Bayes classifier for ICA, PCA, and SVD data reduction techniques.

evaluations for ICA, the estimated error rate did not match the observed error rate. However, it reached the minimum estimated error of 0.282 as presented in Table 4. On the other hand, the estimate of the true rate exactly matched the observed error rate employing SVD data reduction scheme and reached the minimum of 0.282.

Both ICA and SVD algorithms reached the minimum estimated error, but the overall prediction accuracy of SVD was 13% better than ICA. It is true that the ICA reached an estimated error better than PCA, but the ONB classifier along with the PCA algorithm even predicted the subjects

with better accuracy. This could be due to the same nature of the PCA and SVD as previously discussed in Section 4. The ONB along with the ICA algorithm predicted all the subjects in the relapse class correctly. However, it failed at the prediction of 61.5% of the subjects in the nonrelapse class as shown in Figure 19(a). In conclusion, ICA could not extract enough structural information to be employed in prediction of nonrelapse subjects for this study.

Figure 20 visualizes the performance of the classifiers and the trade-off between sensitivity and specificity in which any increase in sensitivity will be accompanied by a decrease

(a) ICA

(b) PCA

(c) SVD
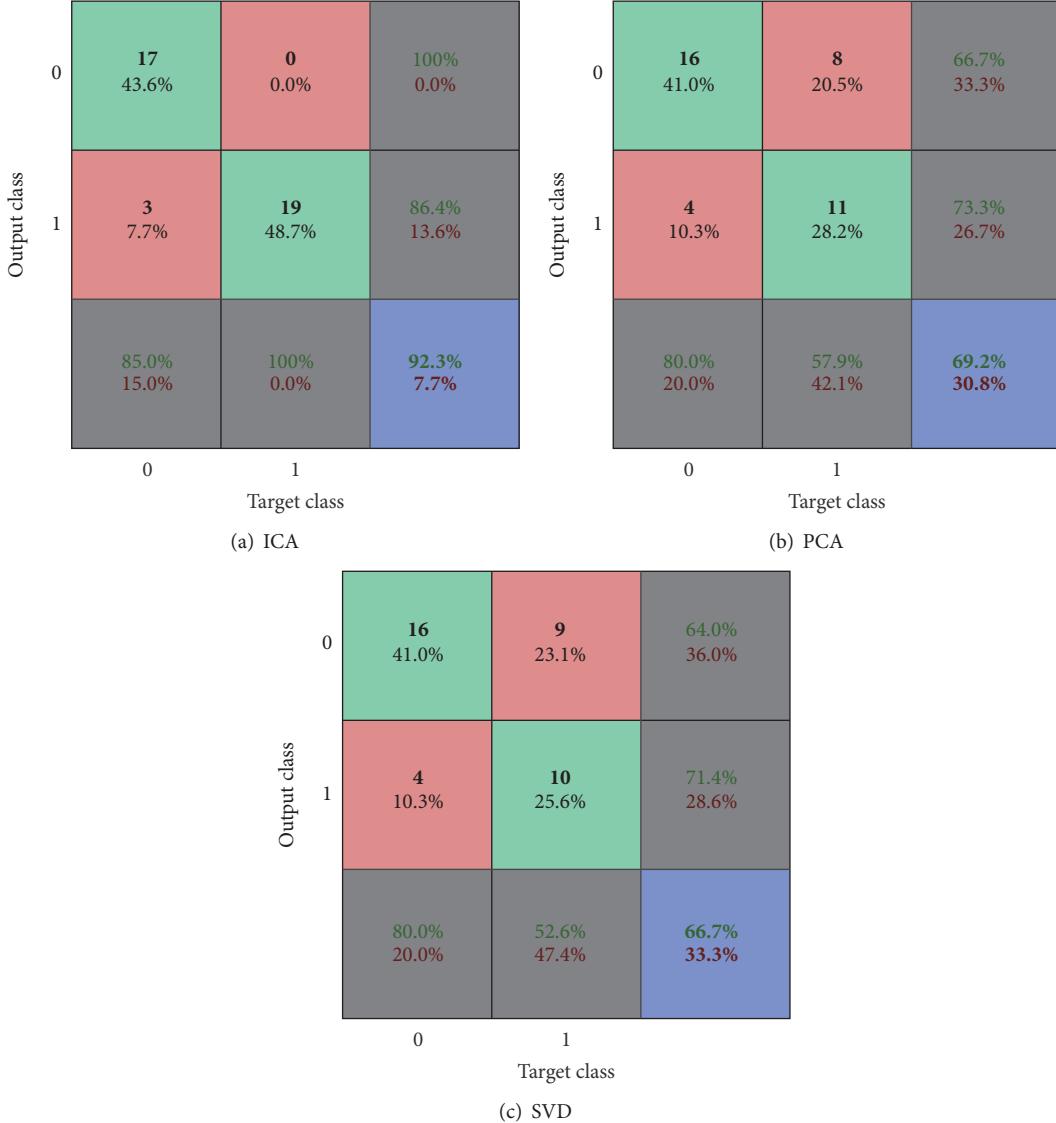
FIGURE 11: Confusion matrices of classification using the optimized Naive-Bayes for ICA, PCA, and SVD data reduction techniques.

CART, (AUC = 0.95)
GNB, (AUC = 0.73)
ONB, (AUC = 0.99)

(a) ICA

CART, (AUC = 0.86)
GNB, (AUC = 0.64)
ONB, (AUC = 0.79)

(b) PCA
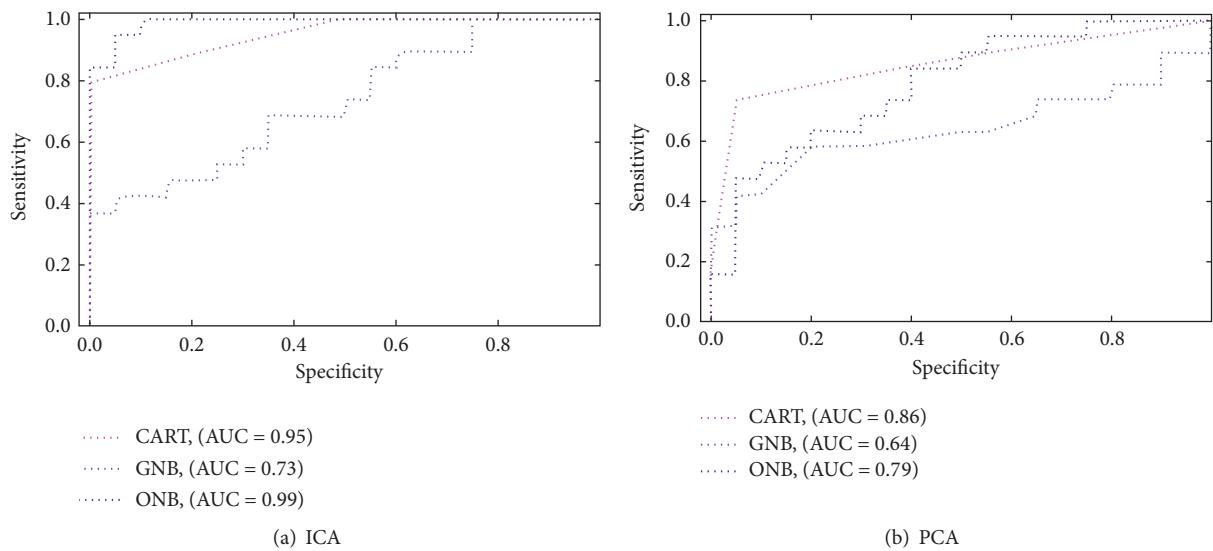
FIGURE 12: Continued.

(c) SVD
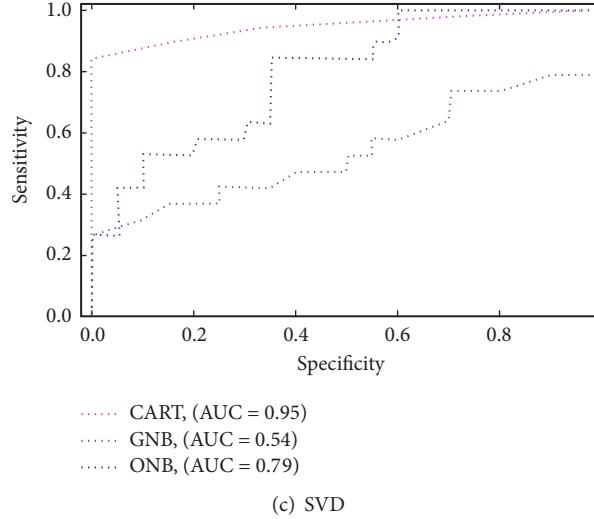
Figure 12: ROC curves for classification using the CART, Gaussian, and optimized Naive-Bayes with ICA, PCA, and SVD data reduction techniques.


(a) ICA

(b) PCA

(c) SVD
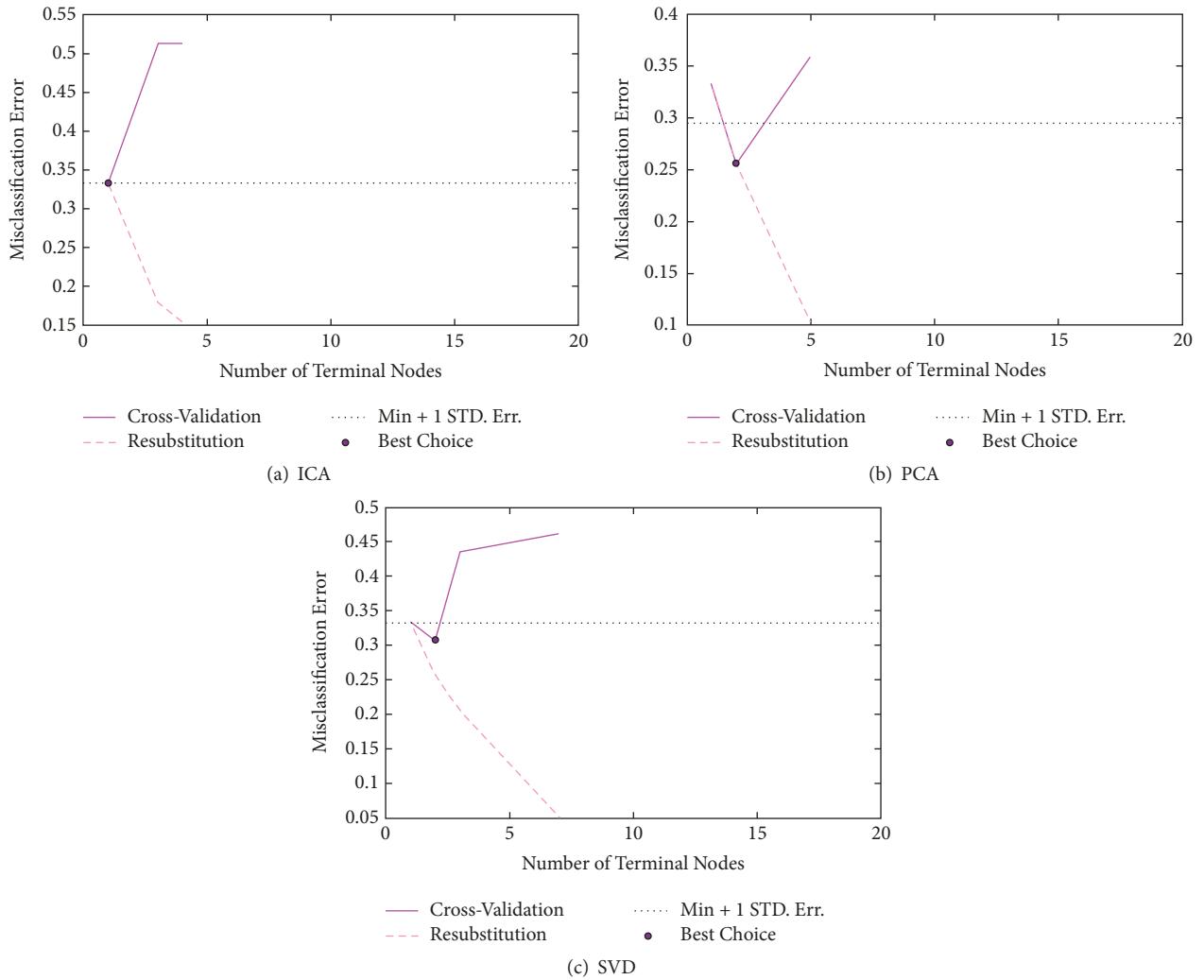
Figure 13: Misclassification error for the CART for different numbers of terminal nodes with ICA, PCA, and SVD data reduction techniques.

(a) ICA

(b) PCA

(c) SVD
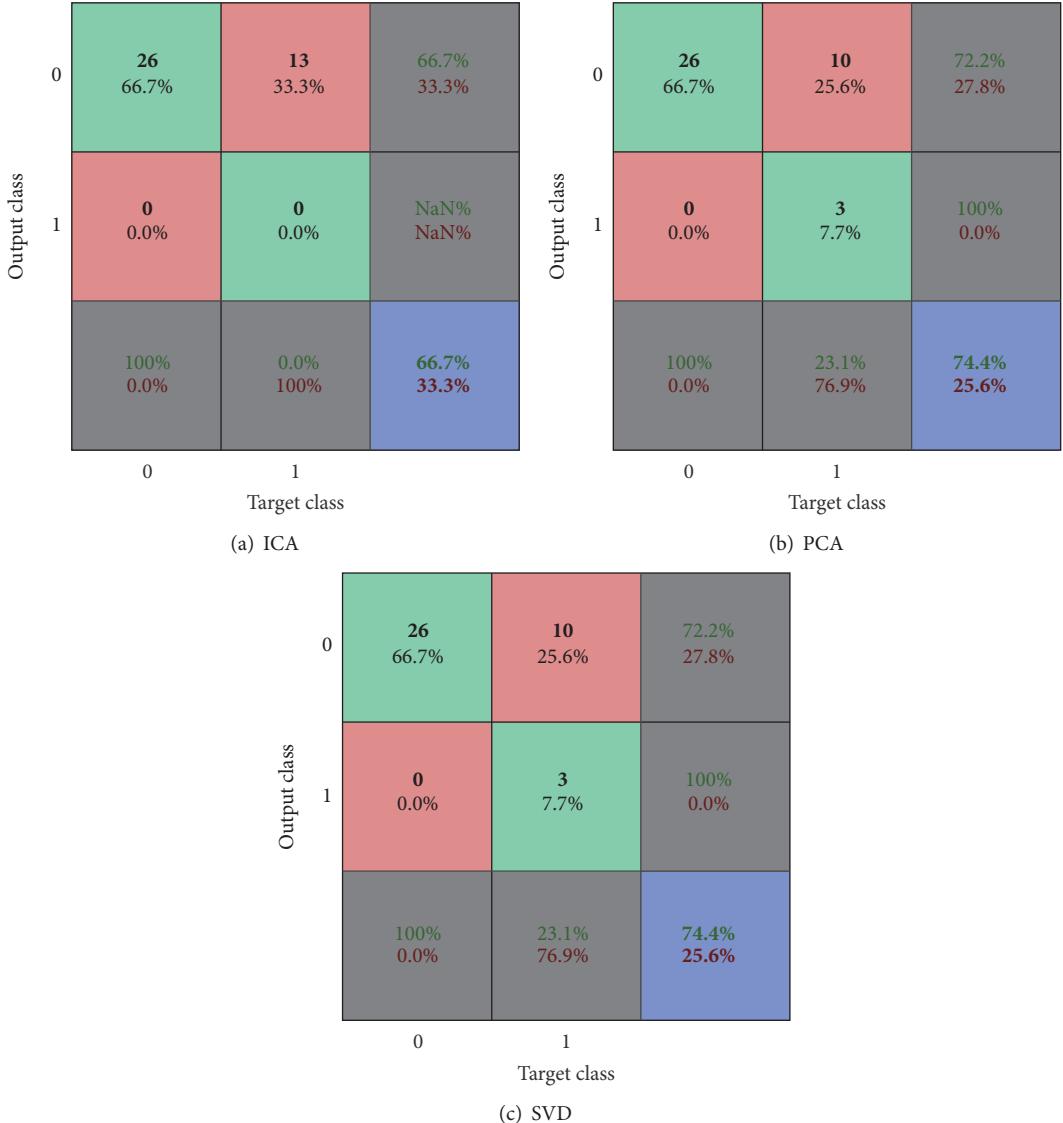
FIGURE 14: Confusion matrices of classification using CART for ICA, PCA, and SVD data reduction techniques.


(a) ICA

(b) PCA
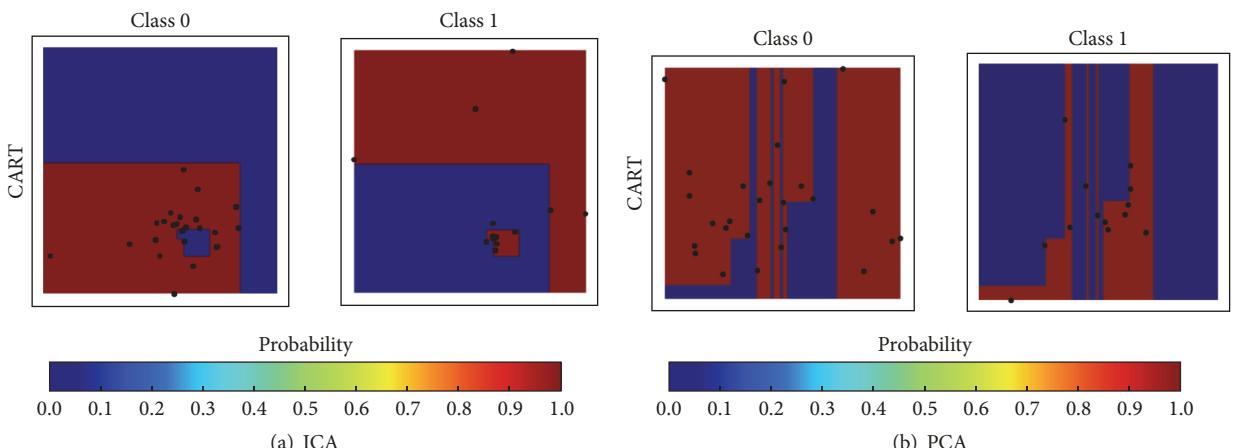
FIGURE 15: Continued.

(c) SVD
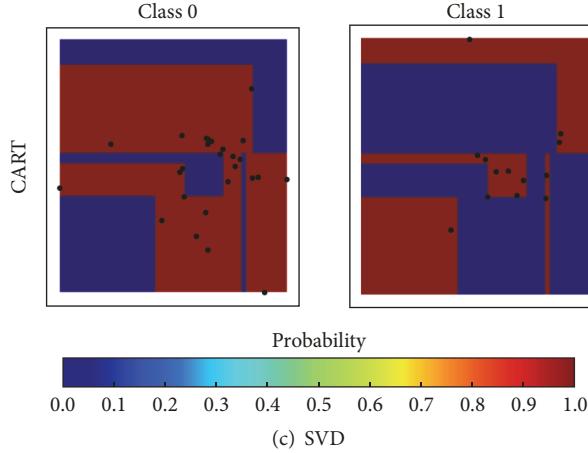
FIGURE 15: Contour plots of probability surface for classification using the CART for ICA, PCA, and SVD data reduction techniques.

(a) ICA

(b) PCA

(c) SVD

FIGURE 16: Confusion matrices of classification using the Gaussian Naive-Bayes for ICA, PCA, and SVD data reduction techniques.
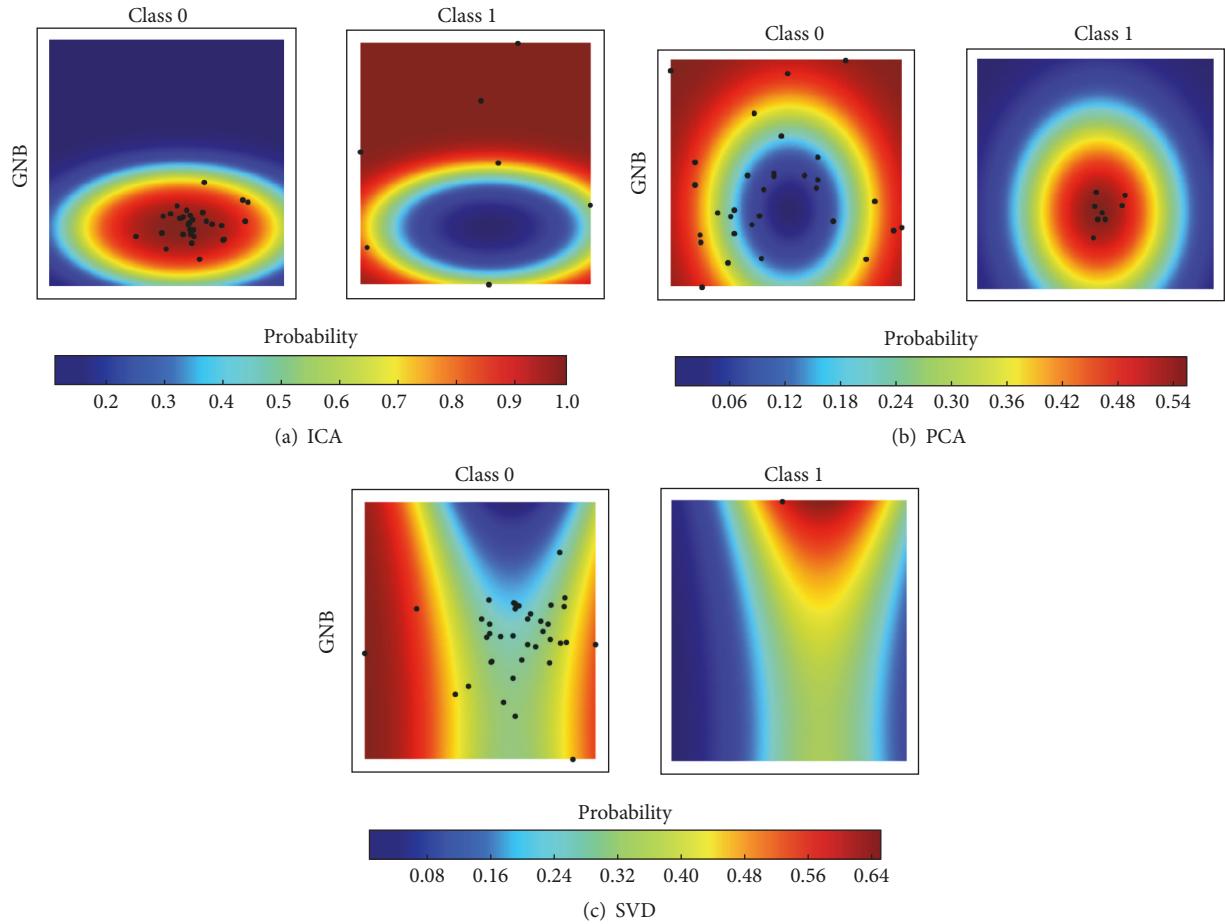
FIGURE 17: Contour plots of probability surface for classification using the Gaussian Naive-Bayes for ICA, PCA, and SVD data reduction techniques.


(a) ICA

(b) PCA
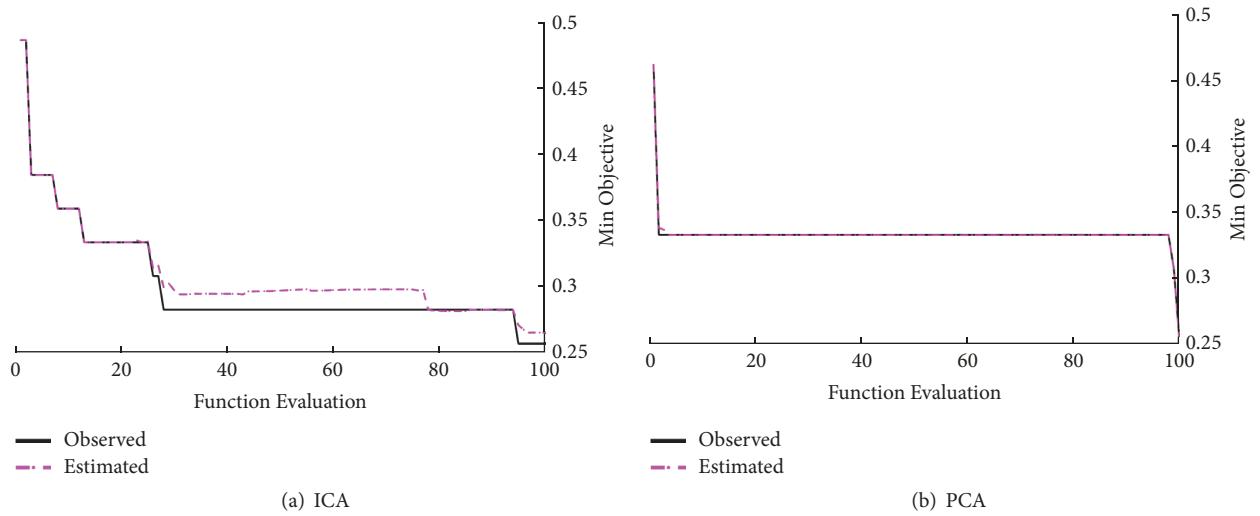
FIGURE 18: Continued.

(c) SVD
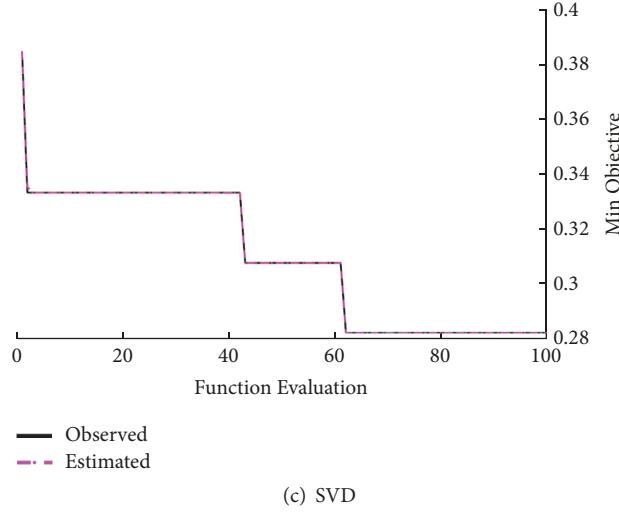
FIGURE 18: Convergence rates of the optimized Naive-Bayes classifier for ICA, PCA, and SVD data reduction techniques.

(a) ICA

(b) PCA

(c) SVD
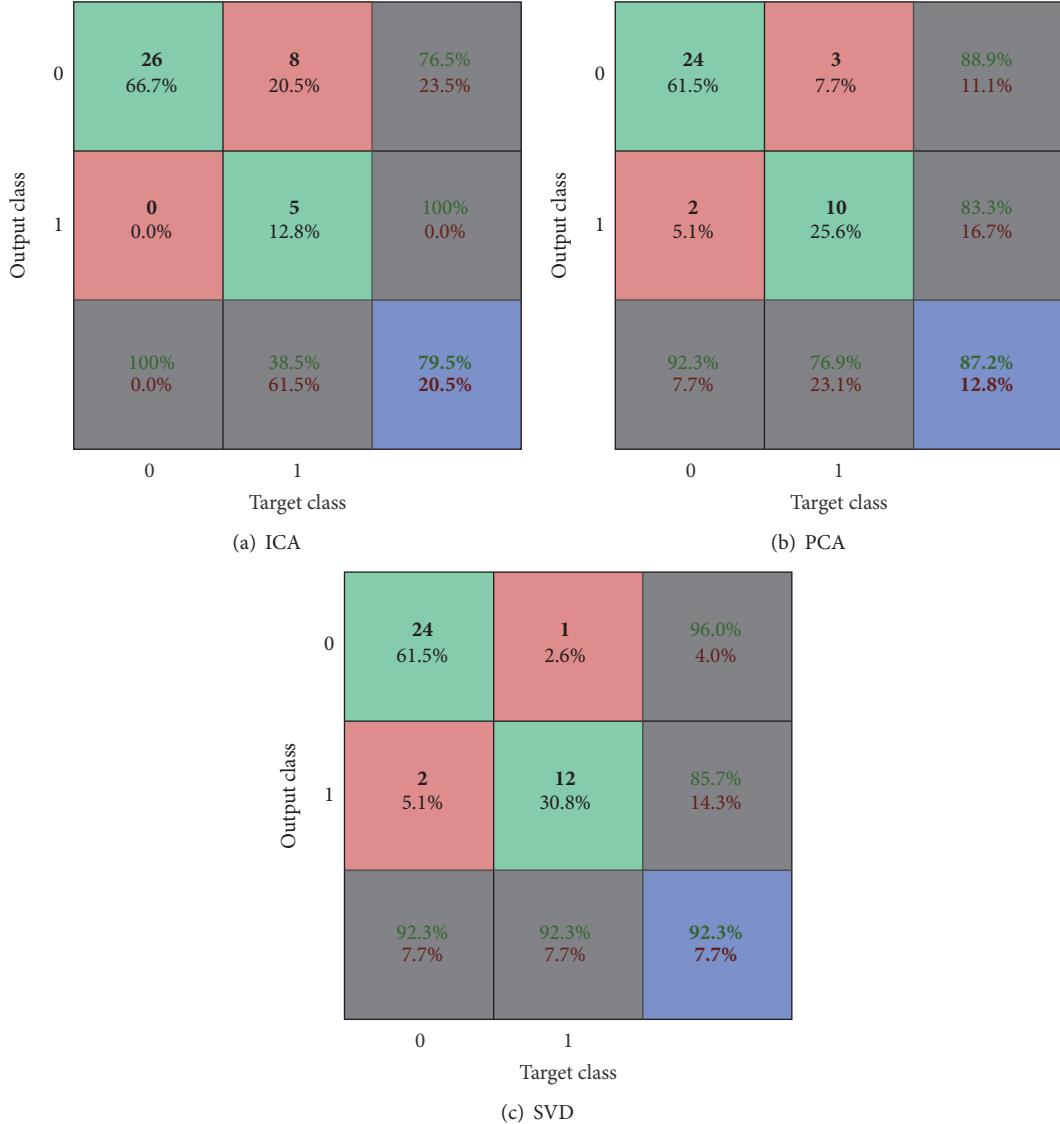
FIGURE 19: Confusion matrices of classification using the optimized Naive-Bayes for ICA, PCA, and SVD data reduction techniques.

(a) ICA

(b) PCA

(c) SVD
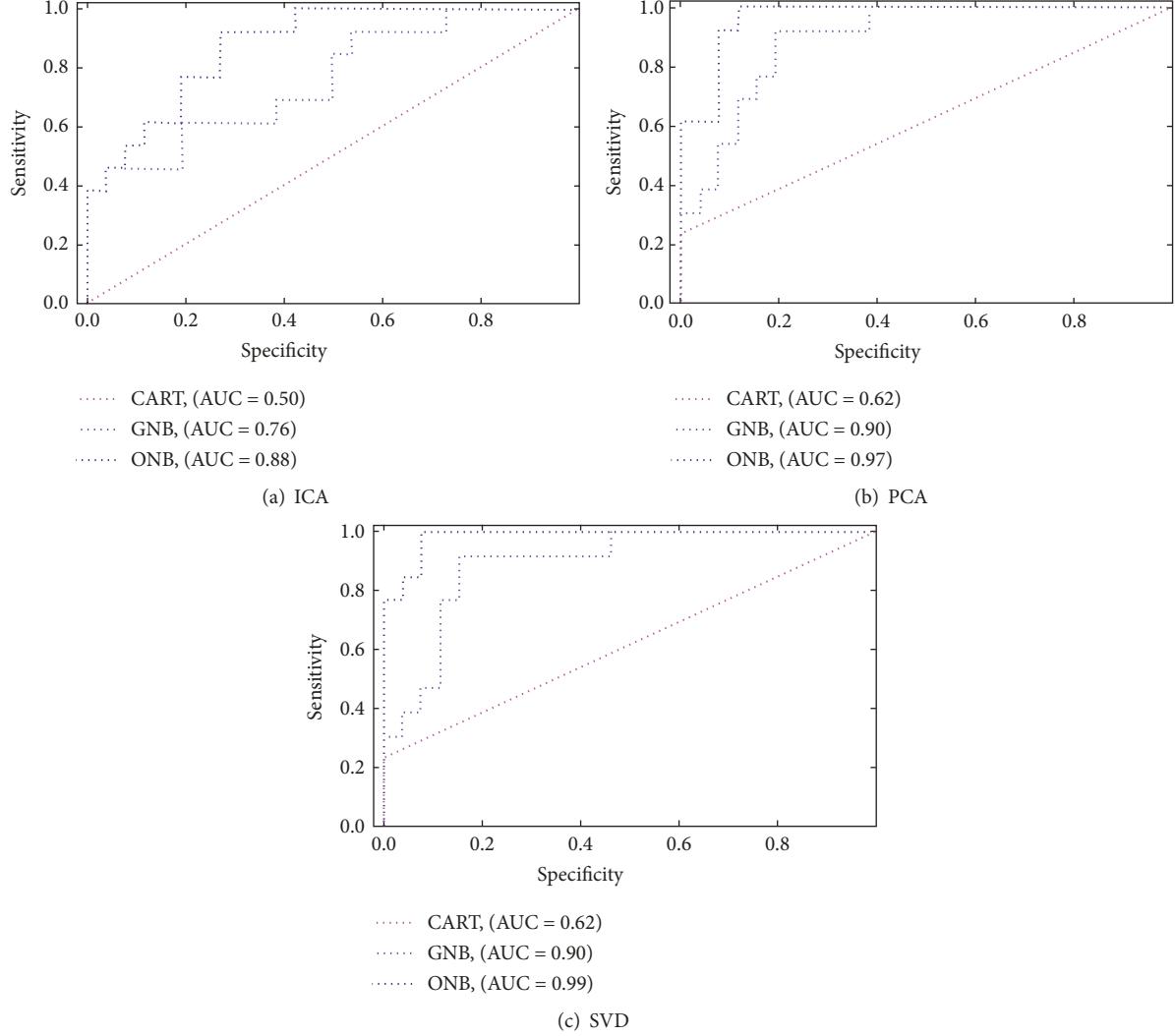
FIGURE 20: ROC curves for classification using the CART, Gaussian, and optimized Naive-Bayes with ICA, PCA, and SVD data reduction techniques.

in specificity. As seen, the ONB classifier along with SVD predicted the subjects in relapse and nonrelapse classes with an AUC of 0.99, which is an extraordinary result. In addition to this, PCA along with the ONB classifier also showed a reasonable accuracy with an AUC of 0.97. This might be due to the similar nature of the principal components and singular values, which showed similar performance for this database. As seen, the ONB classifier that did a phenomenal job separating the relapse and nonrelapse classes has an ROC curve that comes close to the upper left corner of the plot. It should be noted that even ICA, along with the ONB classifier, showed reasonable results in predicting subjects in the relapse class. However, it seemed that ICA could not extract structural features to detect the subjects in the nonrelapse class. This might be the reason that the FN values were high. This is so obvious by considering ICA along with the CART classifier, where the line represents a classifier that did not do better than random prediction.

7. Conclusions and Future Work

The scientific goal of this study was to develop new theory-driven biomarkers by implementing and evaluating novel techniques from resting-state brain scans that can be used in relapse prediction for nicotine-dependent patients and future treatment efficacy. Two classes of patients were studied, one took the drug N-acetylcysteine and the other took a placebo. The patients underwent a double-blind smoking cessation treatment and the resting-state fMRI of their brains before and after treatment was recorded. The high dimensionality of the fMRI data taken from patients made it difficult for the classification tasks to employ the original preprocessed data as input. In this regard, data reduction algorithms including ICA, PCA, and SVD were employed. The CART decision tree and the Naive-Bayes classifier with two different implementations were chosen for the classification tasks. Based on the results, the following conclusions can be drawn:

(1) The proposed model, including the features extracted from the resting-state fMRI brain scans, was validated by classifying the subjects into NAC and placebo classes. The optimized Naive-Bayes along with independent component analysis gave an accuracy of 93% with sensitivity-specificity of 99% for the validation set. The validated feature model based on singular value decomposition along with optimized Naive-Bayes classifier gave an accuracy of 93% with sensitivity-specificity of 99% for the relapse set.

(2) The validation results indicated that independent component analysis can be employed to extract structural information to be used in a balanced-unbiased classification problem using both explicit and implicit classification algorithms.

(3) The relapse results showed that singular value decomposition would extract critical features to be used in an unbalanced-biased classification problem employing implicit classification algorithms.

(4) The relapse results showed that interpreting fMRI connectivity maps based on machine learning algorithms might result in developing novel theory-driven biomarkers with clinical applications in future.

(5) All the analysis was based on the difference between baseline and follow up scans, which were acquired in the resting-state. Thus, it can be assumed that any difference between the groups is the result of NAC, since this was the only variable that differed between the groups. In addition to this, there were no differences between the groups at baseline, except the motivation to change their behavior.

In future work, deep learning and convolutional neural networks (CNN) can be employed to maximize the information for the training process. As discussed, the fMRI data is given in 4-dimensional NIFTI format (three spatial dimensions, one temporal dimension). In fact, a 3D movie for each subject for 200 snapshots before and 200 snapshots after the treatment is given. In this approach, the CNN model can be trained employing different kernels, subsampling, max-pooling, and padding, along with fully connected networks to maximize the amount of meaningful information in the training process. The trained model can be used in relapse prediction based on fMRI scans without using any voxel selection schemes.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

References

[1] U. S. General, *The health benefits of smoking cessation*, Department of Health and Human Services, Wash, USA, 1990.

[2] N. E. Paterson and A. Markou, "Prolonged nicotine dependence associated with extended access to nicotine self-administration in rats," *Psychopharmacology*, vol. 173, no. 1, pp. 64–72, 2004.

[3] U. S. General, U. S. General, The health consequences of smoking: nicotine addiction, A report of the Surgeon General, 1988.

[4] W. A. Corrigall, K. M. Coen, and K. L. Adamson, "Self-administered nicotine activates the mesolimbic dopamine system through the ventral tegmental area," *Brain research*, vol. 65, p. 3, 1994.

[5] M. C. Fiore, T. E. Novotny, J. P. Pierce et al., "Methods Used to Quit Smoking in the United States: Do Cessation Programs Help?" *Journal of the American Medical Association*, vol. 263, no. 20, pp. 2760–2765, 1990.

[6] L. Schmaal, D. J. Veltman, A. Nederveen, W. Van Den Brink, and A. E. Goudriaan, "N-acetylcysteine normalizes glutamate levels in cocaine-dependent patients: A randomized crossover magnetic resonance spectroscopy study," *Neuropsychopharmacology*, vol. 37, no. 9, pp. 2143–2152, 2012.

[7] S. D. LaRowe, P. Mardikian, R. Malcolm et al., "Safety and tolerability of N-acetylcysteine in cocaine-dependent individuals," *American Journal on Addictions*, vol. 15, no. 1, pp. 105–110, 2006.

[8] L. Schmaal, L. Berk, K. P. Hulstijn, J. Cousijn, R. W. Wiers, and W. Van Den Brink, "Efficacy of N-acetylcysteine in the treatment of nicotine dependence: A double-blind placebo-controlled pilot study," *European Addiction Research*, vol. 17, no. 4, pp. 211–216, 2011.

[9] P. N. Mardikian, S. D. LaRowe, S. Hedden, P. W. Kalivas, and R. J. Malcolm, "An open-label trial of N-acetylcysteine for the treatment of cocaine dependence: A pilot study," *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, vol. 31, no. 2, pp. 389–394, 2007.

[10] S. D. LaRowe, P. W. Kalivas, J. S. Nicholas, P. K. Randall, P. N. Mardikian, and R. J. Malcolm, "A double-blind placebo-controlled trial of N-acetylcysteine in the treatment of cocaine dependence," *American Journal on Addictions*, vol. 22, no. 5, pp. 443–452, 2013.

[11] G. Deshpande, P. Wang, D. Rangaprakash, and B. Wilamowski, "Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data," *IEEE Transactions on Cybernetics*, vol. 45, no. 12, pp. 2668–2679, 2015.

[12] P. Yan, Y. Cao, Y. Yuan, B. Turkbey, and P. L. Choyke, "Label image constrained multiatlas selection," *IEEE Transactions on Cybernetics*, vol. 45, no. 6, pp. 1158–1168, 2015.

[13] T. M. Mitchell, S. V. Shinkareva, A. Carlson et al., "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[14] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fMRI) 'brain reading': detecting and classifying distributed patterns of fMRI activity in human visual cortex," *NeuroImage*, vol. 19, no. 2, pp. 261–270, 2003.

[15] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: multi-voxel pattern analysis of fMRI data," *Trends in Cognitive Sciences*, vol. 10, no. 9, pp. 424–430, 2006.

[16] D. M. Schnyer, P. C. Clasen, C. Gonzalez, and C. G. Beevers, "Evaluating the diagnostic utility of applying a machine

learning algorithm to diffusion tensor MRI measures in individuals with major depressive disorder," *Psychiatry Research: Neuroimaging*, vol. 264, pp. 1–9, 2017.

[17] T. A. Carlson, P. Schrater, and S. He, "Patterns of activity in the categorical representations of objects," *Cognitive Neuroscience*, vol. 15, no. 5, pp. 704–717, 2003.

[18] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: a tutorial overview," *NeuroImage*, vol. 45, no. 1, pp. S199–S209, 2009.

[19] T. M. Mitchell, R. Hutchinson, R. S. Niculescu et al., "Learning to decode cognitive states from brain images," *Machine Learning*, vol. 57, no. 1-2, pp. 145–175, 2004.

[20] J. Mourão-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data," *NeuroImage*, vol. 28, no. 4, pp. 980–995, 2005.

[21] A. Smitha, A. Ehtemami, D. Frattea et al., "Functional connectivity analysis of resting-state fmri networks in nicotine dependent patients in nicotine dependent patients," in *PIE Medical Imaging, International Society for Optics and Photonics*, 2016.

[22] A. Ehtemami, *Statistical data analysis of resting state fMRI: A study of nicotine addiction treatment [Ph.D. thesis]*, The Florida State University, A study of nicotine addiction treatment, 2016.

[23] L. I. Kuncheva, J. J. Rodriguez, C. O. Plumpton, D. E. J. Linden, and S. J. Johnston, "Random subspace ensembles for fMRI classification," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 531–542, 2010.

[24] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[25] L. I. Kuncheva and J. J. Rodríguez, "Classifier ensembles for fMRI data analysis: an experiment," *Magnetic Resonance Imaging*, vol. 28, no. 4, pp. 583–593, 2010.

[26] P. K. Douglas, S. Harris, A. Yuille, and M. S. Cohen, "Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief," *NeuroImage*, vol. 56, no. 2, pp. 544–553, 2011.

[27] J. Kuzilek, V. Kremen, F. Soucek, and L. Lhotska, "Independent component analysis and decision trees for ECG holter recording de-noising," *PLoS ONE*, vol. 9, no. 6, Article ID e98450, 2014.

[28] A. Heinz, J. Wrase, T. Kahnt et al., "Brain activation elicited by affectively positive stimuli is associated with a lower risk of relapse in detoxified alcoholic subjects," *Alcoholism: Clinical and Experimental Research*, vol. 31, no. 7, pp. 1138–1147, 2007.

[29] A. R. Schneeberger, C. G. Huber, A. Seixas et al., "Alcohol consumption and use of health care services in people with severe mental illness and stressful childhood experiences," *Journal of Addictive Diseases*, vol. 36, no. 2, pp. 97–104, 2017.

[30] A. Tahmassebi, A. H. Gandomi, I. McCann et al., "An evolutionary approach for fMRI big data classification," in *Proceedings of the 2017 IEEE Congress on Evolutionary Computation, CEC 2017*, pp. 1029–1036, esp, June 2017.

[31] B. Froeliger, P. A. McConnell, N. Stankeviciute, E. A. McClure, P. W. Kalivas, and K. M. Gray, "The effects of N-Acetylcysteine on frontostriatal resting-state functional connectivity, withdrawal symptoms and smoking abstinence: A Double-blind, Placebo-Controlled fMRI Pilot Study," *Drug and Alcohol Dependence*, vol. 156, article 5746, pp. 234–242, 2015.

[32] M. H. J. Schulte, A. E. Goudriaan, A. M. Kaag et al., "The effect of N-acetylcysteine on brain glutamate and gamma-aminobutyric acid concentrations and on smoking cessation: A randomized, double-blind, placebo-controlled trial," *Journal of Psychopharmacology*, vol. 31, no. 10, pp. 1377–1379, 2017.

[33] M. A. Lindquist, "Functional causal mediation analysis with an application to brain connectivity," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1297–1309, 2012.

[34] H. Xie, V. D. Calhoun, J. Gonzalez-Castillo et al., "Whole-brain connectivity dynamics reflect both task-specific and individual-specific modulation: A multitask study," *NeuroImage*, 2017.

[35] A. K. Seth, P. Chorley, and L. C. Barnett, "Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling," *NeuroImage*, vol. 65, pp. 540–555, 2013.

[36] T. Zhang, F. Li, L. Beckes, C. Brown, and J. A. Coan, "Nonparametric inference of the hemodynamic response using multi-subject fMRI data," *NeuroImage*, vol. 63, no. 3, pp. 1754–1765, 2012.

[37] Y.-J. Lin and A. P. Koretsky, "Manganese ion enhances T1-weighted MRI during brain activation: an approach to direct imaging of brain function," *Magnetic Resonance in Medicine*, vol. 38, no. 3, pp. 378–388, 1997.

[38] N. W. Churchill, A. Oder, H. Abdi et al., "Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods," *Human Brain Mapping*, vol. 33, no. 3, pp. 609–627, 2012.

[39] E. E. Tripoliti, D. I. Fotiadis, M. Argyropoulou, and G. Manis, "A six stage approach for the diagnosis of the Alzheimer's disease based on fMRI data," *Journal of Biomedical Informatics*, vol. 43, no. 2, pp. 307–320, 2010.

[40] K. J. Friston, A. P. Holmes, J.-B. Poline et al., "Analysis of fMRI time-series revisited," *NeuroImage*, vol. 2, no. 1, pp. 45–53, 1995.

[41] K. J. Worsley and K. J. Friston, "Analysis of fMRI time-series revisited—again," *NeuroImage*, vol. 2, no. 3, pp. 173–181, 1995.

[42] J. L. Lancaster, M. G. Woldorff, L. M. Parsons et al., "Automated Talairach Atlas labels for functional brain mapping," *Human Brain Mapping*, vol. 10, no. 3, pp. 120–131, 2000.

[43] J. A. Maldjian, P. J. Laurienti, R. A. Kraft, and J. H. Burdette, "An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets," *NeuroImage*, vol. 19, no. 3, pp. 1233–1239, 2003.

[44] A. Tahmassebi, A. H. Gandomi, and A. Meyer-Bäse, "High Performance GP-Based Approach for fMRI Big Data Classification," in *Proceedings of the the Practice and Experience in Advanced Research Computing 2017*, pp. 1–4, New Orleans, LA, USA, July 2017.

[45] R. A. Poldrack, "Region of interest analysis for fMRI," *Social Cognitive and Affective Neuroscience*, vol. 2, no. 1, pp. 67–70, 2007.

[46] P. Comon, "Independent component analysis, A new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[47] V. D. Calhoun, J. Liu, and T. Adali, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *NeuroImage*, vol. 45, no. 1, supplement 1, pp. S163–S172, 2009.

[48] A. Meyer-Baese, A. Wismueller, and O. Lange, "Comparison of two exploratory data analysis methods for fMRI: Unsupervised clustering versus independent component analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 8, no. 3, pp. 387–398, 2004.

[49] M. J. McKeown, S. Makeig, G. G. Brown et al., "Analysis of fMRI data by blind separation into independent spatial components," *Human Brain Mapping*, vol. 6, no. 3, pp. 160–188, 1998.

[50] A. Meyer-Baese, *Pattern recognition for medical imaging*, Academic Press, 2004.

[51] A. Tahmassebi, A. H. Gandomi, I. McCann et al., "fMRI smoking cessation classification using genetic programming," *In Workshop on Data Science Meets Optimization*, 2017, http://dso.org/images/Workshop_papers/Gandomi.pdf.

[52] J. A. Etzel, V. Gazzola, and C. Keysers, "An introduction to anatomical ROI-based fMRI classification analysis," *Brain Research*, vol. 1282, pp. 114–125, 2009.

[53] P.-N. Tan, *Introduction to data mining*, Pearson Education India, 2006.

[54] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, Mass, USA, 1984.

[55] M. Gori, *C4. 5: Programming for machine learning*, vol. 38, Morgan Kaufmann Publishers, Cambridge, UK, 1993.

[56] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[57] J. R. Quinlan, "Simplifying decision trees," *International Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221–234, 1987.

[58] M. Mehta, R. Agrawal, and J. Rissanen, "Sliq: A fast scalable classifier for data mining," *Advances in Database Technology*, pp. 18–32, 1996.

[59] J. Shafer, R. Agrawal, and M. Mehta, "Sprint: A scalable parallel classi er for data mining," in *Int. Conf. Very Large Data Bases*, pp. 544–555, Citeseer, 1996.

[60] G. Langs, B. H. Menze, D. Lashkari, and P. Golland, "Detecting stable distributed patterns of brain activation using Gini contrast," *NeuroImage*, vol. 56, no. 2, pp. 497–507, 2011.

[61] E. B. Hunt, H. A. Simon, J. Marin, and P. Stone, "Experiments in Induction," *The American Journal of Psychology*, vol. 80, no. 4, p. 651, 1967.

[62] F. Pedregosa, G. Varoquaux, and A. Gramfort, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[63] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning, series in statistics*, vol. 1, Springer, New York, NY, USA, 2001.

[64] MATLAB, version 8.5.0.197613 R2015a, The MathWorks Inc., Natick, Massachusetts, 2015.

[65] L. Buitinck, G. Louppe, M. Blondel et al., "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[66] L. Wasserman, *All of statistics: a concise course in statistical inference*, USA, Springer Science & Business Media, 2013.