IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 November 2022, accepted 20 November 2022, date of publication 30 November 2022,
date of current version 5 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3225685

RESEARCH ARTICLE

A Group Feature Ranking and Selection Method Based on Dimension Reduction Technique in High-Dimensional Data

IQBAL MUHAMMAD ZUBAIR AND BYUNGHOON KIM
Department of Industrial and Management Engineering, Hanyang University, Ansan 15588, South Korea

Corresponding author: Byunghoon Kim (byungkim@hanyang.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MIST) under Grant NRF-2022R1F1A1063273.

ABSTRACT Group feature selection methods select the important group features by removing the irrelevant group features for reducing the complexity of the model. To the best of our knowledge, there are few group feature selection methods that provide the relative importance of each feature group. For this purpose, we developed a sparse group feature ranking method based on the dimension reduction technique for high dimensional data. Firstly, we applied relief to each group to remove irrelevant individual features. Secondly, we extract the new feature that represents each feature group. To this end, we reduce the multiple dimension of the group feature into a single dimension by applying Fisher linear discriminant analysis (FDA) for each feature group. At last, we estimate the relative importance of the extracted feature by applying random forest and selecting important features that have larger importance scores compared with other ones. In the end, machine-learning algorithms can be used to train and test the models. For the experiment, we compared the proposed with the supervised group lasso (SGL) method by using real-life high-dimensional datasets. Results show that the proposed method selects a few important group features just like the existing group feature selection method and provides the ranking and relative importance of all group features. SGL slightly performs better on logistic regression whereas the proposed method performs better on support vector machine, random forest, and gradient boosting in terms of classification performance metrics.

INDEX TERMS Dimension reduction, feature extraction, group feature ranking, group feature selection, high dimensional data.

I. INTRODUCTION

Feature selection is an essential task in high-dimensional data analysis. In these types of datasets, the number of features is greater than the number of observations. For instance, the gene expression dataset used in bioinformatics includes thousands of features, whereas the number of samples is much smaller than that. In the problem, the existing researchers strived to identify the most important features because there are many irrelevant or redundant features in the dataset [1]. Feature selection techniques are categorized into three methods: filter, wrapper, and embedded-based methods [2], [3], [4]. Filter methods are independent of the learning process, which identifies the relevant subset of features to a target variable by using statistical analysis [5]. The wrapper method finds optimal features that maximize a classification performance [6]. The embedded methods are between the filter and wrapper methods to take advantage of both types of methods. It also combines the learning process and feature selection to determine the subset of features [7].

In many applications, features are correlated with each other because they stem from the same source. For instance, many features in microarray gene data have a relation with each other and possess a group structure [6]. Therefore, these features have similar effects on the target variable. In this case, it is more appropriate to select the correlated features in the same group than to select individual features. The correlated features form a group feature; in this case, feature selection corresponds to group feature selection instead of individual feature selection because the individual feature selection does not consider the structure information.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

Group feature selection techniques aim to remove irrelevant and redundant group features. By selecting only relevant group features, we can improve the computational efficiency and the classification performance of machine learning models [6]. Many studies such as [8], [9], [10], [11], and [12] developed group feature selection methods. Bakin [8] proposed a method for group feature selection that was further expanded by Yuan and Lin [9], known as the group lasso. Group lasso is the natural extension of the lasso and uses the L2 norm of coefficient as a penalty function related to group features [13]. It selects a group of features that are highly correlated to each other instead of an individual feature. Meier et al. [10] further extended this method for logistic regression by applying it to DNA sequence data. Group lasso and its extension select the sparse set of group features, but these methods do not select relevant individual features for each feature group. Later, Simon et al. [11] filled the gap and developed the sparse group lasso method, which can perform sparsity on both a group and individual level. Fang et al. [14] extended the sparse group lasso method and developed the adaptive sparse group lasso. Moreover, Vincent and Hansen [15] extended the sparse group lasso to the multinomial sparse group lasso. Group lasso, sparse group lasso, and their extensions are effective methods for gene selection and classification. Supervised group lasso [12] is also another sparse group feature selection method. It has two steps. In the first step, it applies a lasso within each group to find important features. In the second step, it uses group lasso for selected important groups. Therefore, it merges the lasso and group lasso method to perform sparsity on the group level and individual level.

However, there is a limitation to the existing studies. The existing approaches do not provide the relative importance of the selected group features. Because they simply select a few group features, it is unknown how much a group feature is more important than others. Therefore, the existing studies do not identify the relative importance of each group feature. Moreover, many existing methods such as [10] and [12] developed their group feature selection methods that can be only applicable to logistic regression which is a linear model.

To overcome this drawback, we propose a new, sparse group feature ranking technique that is developed based on dimension reduction techniques. In the first step, we remove irrelevant individual features in each feature group by using the relief algorithm. In the second step, we apply an FDA to reduce the multiple dimension of group features into a single dimension for each feature group. Lastly, we compute the relative importance of the new feature that represents a feature group by using random forest algorithm. The key finding of this paper is summarized as follows:

- This method estimates the importance of each group feature and selects only a few highly important group features.
- The new method is proposed to extract a single-dimensional feature that captures the characteristics of multidimensional group features adopting a dimension reduction technique such as FDA.
- This method is not model specific. Any classifier, such as logistic regression or support vector machine, can be used for this method.

The rest of the material presented in this study is organized as follows. Section 2 presents a literature review on dimension reduction (feature extraction) and feature ranking. In section 3, the methodology of this paper is described. In section 4, results obtained from experiments are discussed. Section 5 contains the conclusion and discussion.

II. RELATED WORK

A. DIMENSION REDUCTION

Dimension reduction, also known as feature extraction, is an important step in pre-processing for high-dimensional data analysis in the field of data analysis and machine learning. Feature extraction is a widely used method for dimension reduction in which high-dimension space is transformed into a low-dimensional space containing relevant information [16].

However, machine learning models trained based on high dimensional data can have high variance, which causes overfitting of the machine learning models to the training dataset. We can avoid the issue by reducing the dimension of the high-dimensional data. Consequently, we need to reduce the dimension of the data to decrease the variance of the fitted model without losing much information. This dimension reduction reduces the risk of overfitting and training costs of the machine learning algorithms. Machine learning algorithms perform well on low-dimensional data and yield better results. Because of these benefits, dimension reduction is necessary [17]. Dimension reduction increases the bias of the trained machine learning models, but it is appropriate to use a less complex model with a higher bias for high dimensional low sample sized (hereafter HDLSS) data [18].

Dimension reduction is essential in many areas, especially in microarray gene expression datasets. Usually, in gene datasets, the number of features is much larger than the number of samples. High dimensional data requires much memory to train a machine learning algorithm and deteriorates the performance of the trained model due to the overfitting problem [19]. Therefore, dimension reduction is needed to address the curse of dimensionality problem by transforming the original high dimensional space into the lower dimension of a new subspace [20]. Principle component analysis (PCA) and Fisher linear discriminant analysis (FDA) are popular methods that reduce the dimension of data. The fundamental difference between these two techniques is that PCA is an unsupervised learning technique that maximizes the variance of the extracted features, whereas FDA is a supervised learning technique that maximizes the class separability of the new features [17]. This study employed Fisher's linear discriminant analysis (FDA) to reduce dimensions.

FDA is one of the tools for supervised learning [21]. It also gives the low-dimensional projection to a discriminative direction of original data, which can be valuable for interpretation [22]. FDA calculates variance between classes in the first step to measure class separability. After calculating variance between classes, FDA calculates variance within the class. In the last stage, FDA finds a new subspace that maximizes the ratio between variance between classes and variance within classes [20].

B. FEATURE RANKING

By ranking individual features, we can select relevant features and figure out the relative importance of all the features. The selected features based on their ranking have enough information to improve the performance of the classification model [23]. Also, by removing the irrelevant features that have low rankings, we can train machine learning algorithms with less computational cost.

Filter methods are usually employed to rank features before selecting features. The filter methods usually include two steps. Firstly, features are ranked by some measures, such as relevancy with the target class or class separability. Secondly, features that have low rankings are removed from the data. The filter methods based on ranking techniques do not depend on machine learning algorithms [24].

Features can be ranked by different techniques such as information theory [25] and statistical distance [26]. In distance-based feature ranking methods, relief [27] is one of the important methods that ranks individual features by estimating their importance scores [23]. Other methods based on information theory used Entropy, information gain (IG), gain ratio (GR), and symmetrical uncertainty (SU) for feature ranking.

The random forest can be applied to rank features. It uses Gini or permutation importance as a measure to calculate the importance of features [24]. There are some other ranking methods like fisher ratio [28], T-test [29], Z-score [30], fold-change ratio [31], and rank product [32]. However, all the existing ranking techniques were developed to rank individual features.

III. METHODOLOGY

In this section, we present a new method of group feature ranking technique and how to select group features using ranking. As mentioned in the introduction, in the first step, the relief algorithm is used for removing irrelevant individual features in each group. After removing irrelevant features in each group, FDA is applied for dimension reduction of each group feature. Through this process, each group feature transforms into a single dimension. A new extracted feature represents corresponding group features. In the third step, we apply a random forest to determine the ranking of the features and select those features which pass the threshold. The selected features can be applied to different machine learning algorithms. Figure 1 describes the procedure of this study.

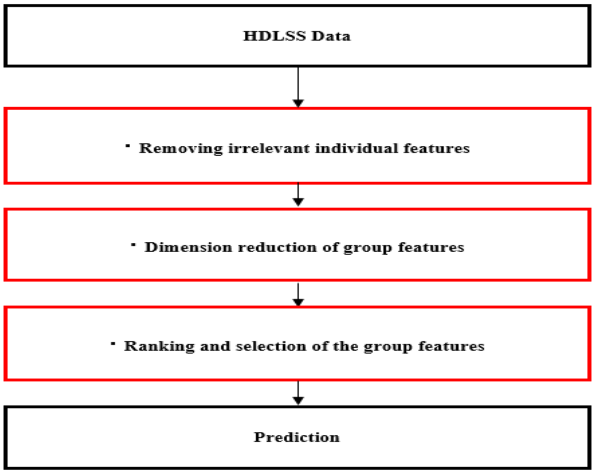

HDLSS Data
· Removing irrelevant individual features
· Dimension reduction of group features
· Ranking and selection of the group features
Prediction

FIGURE 1. The framework of the study.

A. RELIEF FOR REMOVING IRRELEVANT INDIVIDUAL FEATURES

High-dimensional data includes many irrelevant features to the target variable, but these features may not be useful to compute the importance of the feature group. Therefore, we simply remove that individual feature before computing the importance of the group features. There are many existing individual feature selection approaches. In this study, we employ the relief algorithm for selecting individual features.

Kira and Rendell [27] described the original relief algorithm as an effective technique for individual feature weighting. As a filter method in the pre-processing stage to remove irrelevant features, relief calculates the weights of all individual features by the proxy statistic that indicates the quality of that feature. This feature weight, referred to as feature score, shows the 'relevance' of that feature to the target variable. The range of feature scores can be -1 to $+1$, -1 shows the lowest quality of the feature to predict the target variable, and $+1$ shows the highest quality of the feature. Remarkably, the original relief algorithm was designed only for binary classification problems.

We suppose that there is a high dimensional dataset $\boldsymbol{D}$ including n instances that can be denoted by $\boldsymbol{D} = \{\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n\}$ where $\boldsymbol{O}_i$ represents the i-th instance. Each instance $\boldsymbol{O}_i$ has K dimensional feature vector $\boldsymbol{F}$. We note that all the components of the feature vector $\boldsymbol{F}$ are grouped into G feature groups. In other words, feature vector $\boldsymbol{F}$ is a set of G feature groups $\boldsymbol{F} = \left\{\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_g, \ldots, \boldsymbol{F}_G\right\}$ Firstly, we compute the importance score of a feature f that is denoted by $W[f]$ where $f \in \boldsymbol{F}$. Before calculating the score of each feature f, Relief search for the nearest hit and miss neighbors after computing the distance of d_{ij} between the instance $\boldsymbol{O}_i$ and another instance $\boldsymbol{O}_j$ based on Manhattan ($q = 1$) or Euclidian ($q = 2$) metric as shown in equation (1) [33].

$$d_{ij} = \left(\sum_{f \in F} \left| diff\left(f, \left(O_i, O_j\right)\right) \right|^q \right)^{1/q} \tag{1}$$

where *diff* $(f, (O_i, O_j))$ denotes that the normalized distance between O_i and O_j only for the given feature f. We can express it by using the following equation:

$$diff\ (f, (O_i, O_j)) = \frac{|value\ (f, O_i) - value\ (f, O_j)|}{\max(f) - \min(f)} \tag{2}$$

By using the distance d_{ij} of equations (1) and (2), we are searching for nearest hits and miss neighbors. For a given instance O_i, we search for its neighbor hit whose class is the same with the O_i (hereafter H) and search for its neighbor miss whose class is different from the O_i (hereafter M). After searching for M and H, the relief-based score for feature f can be updated as follows:

$$W[f] := W[f] - \frac{diff\ (f, O_i, H)}{n} + \frac{diff\ (f, O_i, M)}{n} \tag{3}$$

We have summarized the overall procedure of the relief in the appendix. After computing the importance score of each feature, we select the top m features for each feature group based on the scores to select the relevant ones to the target variable.

B. DIMENSION REDUCTION OF GROUP FEATURES BASED ON FDA

After selecting the relevant features in the previous section, we obtain a new dataset $\boldsymbol{X}$ whose dimension is much less than K. The new dataset $\boldsymbol{X}$ still has a high dimensional feature. We also note that the new feature vector also has G feature groups (i.e. $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_g, \ldots, \boldsymbol{X}_G\}$). n_1 samples of $\boldsymbol{X}$ belong to class 1 whereas n_2 samples of them belong to class 2 (i.e. $n_1 + n_2 = n$). We reduce the dimension of $\boldsymbol{X}_g$ into a single dimension by employing FDA. FDA is a supervised technique for dimension reduction that is used in various fields of data mining and machine learning. We adopt the dimension reduction technique to maximize the separation between different classes.

The goal of the FDA is to obtain a scalar y_g by projecting the g-th feature group of $\boldsymbol{X}$ (i.e. $\boldsymbol{X}_g$) onto a line:

$$y_g = \theta_g^T X_g \tag{4}$$

We estimate the optimal $\boldsymbol{\theta}_g$ that maximize the ratio of 'between-class variance' and 'within-class variance' as follows:

$$\theta_g^* = argmax_\theta \frac{(\theta^T S_{B(g)} \theta)}{(\theta^T S_{W(g)} \theta)} \tag{5}$$

where $\mathbf{S}_{B(g)}$ denotes the between-class scatter matrix of $\boldsymbol{X}_g$ and $\boldsymbol{S}_{W(g)}$ denotes the within-class scatter matrix of $\boldsymbol{X}_g$. The between-class scatter matrix shows the distinction between different classes. We can estimate the scatter matrix by computing distances between the means of classes as follows:

$$S_{B(g)} = (\mu_{2(g)} - \mu_{1(g)})(\mu_{2(g)} - \mu_{1(g)})^T \tag{6}$$

where $\boldsymbol{\mu}_{1(g)}$ and $\boldsymbol{\mu}_{2(g)}$ denote the mean vectors of $\boldsymbol{X}_g$ whose class labels are 1 and 2 respectively. On the other hand, we can estimate the within-class scatter matrix $\boldsymbol{S}_{W(g)}$ of $\boldsymbol{X}_g$.

$$\begin{aligned} S_{W(g)} = & \sum_{(x_i \in c_1)} (x_{i(g)} - \mu_{1(g)})(x_{i(g)} - \mu_{1(g)})^T \\ & + \sum_{(x_i \in c_2)} (x_{i(g)} - \mu_{2(g)})(x_{i(g)} - \mu_{2(g)})^T \end{aligned} \tag{7}$$

where $\boldsymbol{x}_{i(g)}$ denotes the g-th feature group vector of the i–th sample. The sample can belong to either the class $\boldsymbol{c}_1$ or the class $\boldsymbol{c}_2$. By using this process, we can transform a high-dimensional feature of $\boldsymbol{x}_{i(g)}$ into a single-dimensional feature of y_g.

C. RANKING AND SELECTION OF THE NEW GROUP FEATURES

After reducing the dimension of $\boldsymbol{X}_g$ into a single dimension by employing FDA, we obtain the new dataset $\boldsymbol{Y}$ that has G single-dimension features including n instances. The new dataset (i.e. $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_g, \ldots, \boldsymbol{y}_G\}$) is the reduced dataset of $\boldsymbol{F}$ including G feature groups (i.e. $\boldsymbol{F} = \{\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_g, \ldots, \boldsymbol{F}_G\}$). Each feature y_g in the new dataset $\boldsymbol{y}$ represents the feature group $\boldsymbol{F}_g$. Now, we can compute the relative importance of the new features to estimate the importance of the group feature by using a random forest.

Random forest is a machine learning technique that can be used to compute the relative importance of individual features. It combines many decision trees and can be used for classification and regression problems in machine learning [34]. For making each tree, it selects a random number of instances and features from the dataset.

We can estimate the relative importance of the new features by employing Random forest (hereafter RF). In this study, we estimate the importance of the new features by using the Gini method, also called "mean decrease in impurity (MDI)." In RF, features are selected multiple times to split a node in all trees. MDI is calculated by taking the sum of the decrease in node impurity and averaging it over all trees. The g-th feature y_g of the new dataset $\boldsymbol{y}$ will be considered important if it creates a large decrease in node impurity in all splits during the splitting process.

After calculating the importance of features, one can rank them based on their importance, and select the new features that have a higher ranking. As the threshold of the selection, we can compute the average score of the feature importance

IV. RESULTS

A. DATA DESCRIPTION

This study uses four datasets. Three datasets GSE16446, GSE25066, and GSE2034 taken from the Affymetrix human genome U133A Array (HG-U133A). The remaining one is the simulated data taken from the [35]. These are breast cancer datasets whose predictors are genes with a binary response variable that shows the survival of breast cancer patients after chemotherapy. GSE16446 has 46478 features, GSE25066 has 21362 features, and GSE2034 has 12634 features, whereas the simulated dataset has 10000 total features.

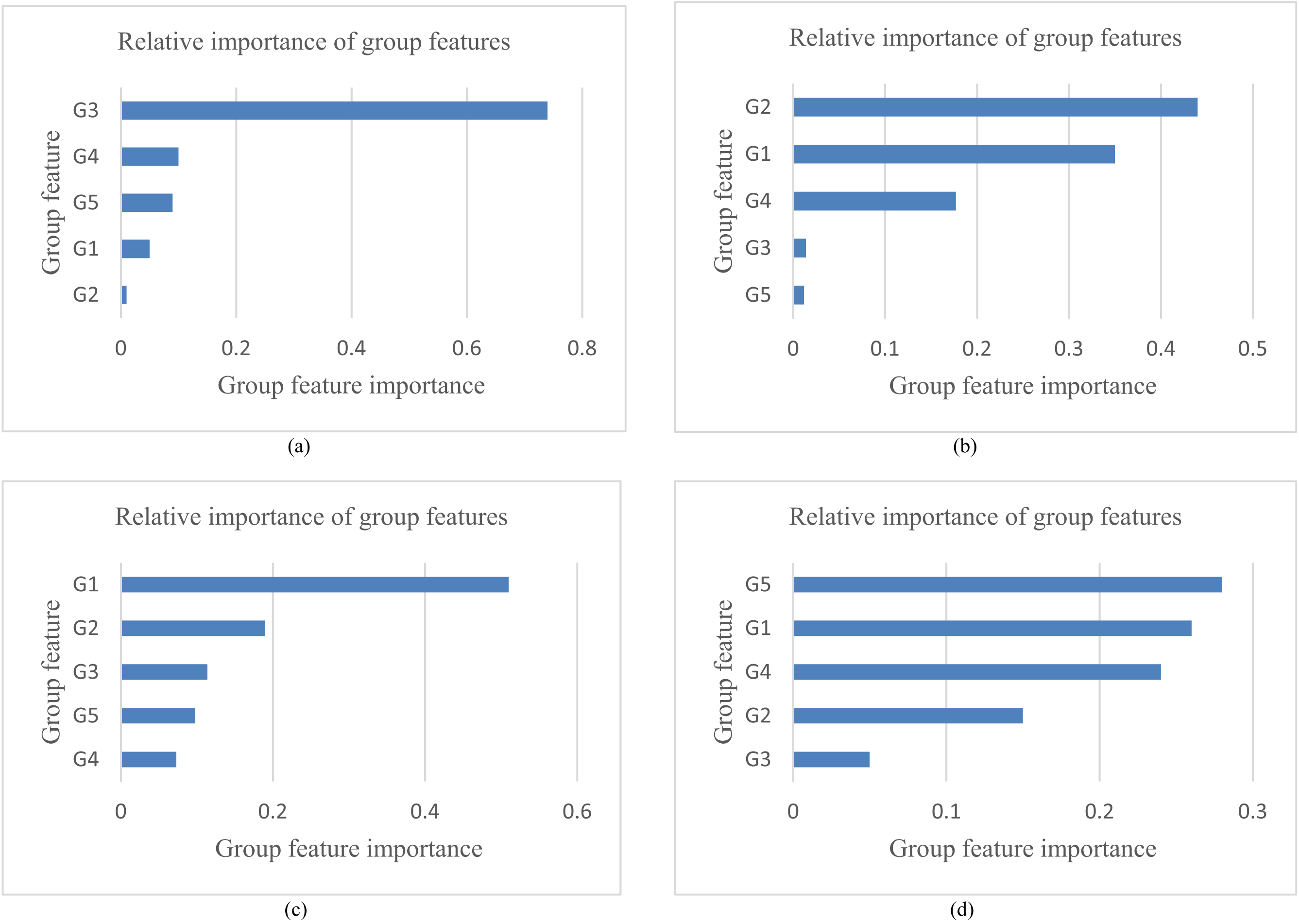


FIGURE 2. Relative importance among group features of GSE16446 (a), GSE25066 (b), GSE2034 (c), and simulated dataset (d).

Many features in GSE16446 and GSE25066 datasets have duplicates with the same name. Therefore, an average of these duplicates is taken and made into a single feature. After taking the average of duplicate features, GSE16446 has 23265 features, whereas GSE25066 has 13470 features as shown in TABLE 1.

TABLE 1. Description of datasets.

Microarray	Series	No. of samples	No. of features	No. of classes
HG-U133	GSE16446	80	23265	2
HG-U133	GSE25066	170	13470	2
HG-U133	GSE2034	286	12634	2
Simulated	-	200	10000	2

In the dataset of GSE16446, 80 patients were involved in chemotherapy. Among the patients, eight were diagnosed as normal whereas 72 patients got breast cancer again. Important genes of these patients were stored in Affymetrix Gene Chip 2.0 array. In the dataset of GSE25066, 170 patients were involved; 57 were diagnosed as normal, whereas 113 patients had cancer again. In the third dataset, GSE2034, 286 patients were involved; of which 179 were found normal and 107 has cancer again. In the remaining simulated dataset, there are 200 samples with the same number of positive and negative samples. In these datasets, the number of samples is too small compared to the number of features. So, these are high dimensional low sample sized (HDLSS) datasets.

B. RELATIVE IMPORTANCE AND SELECTION OF GROUP FEATURES

As mentioned above, the datasets used in this study are high dimensional. Therefore, many features are relevant to each other. Relevant features form a group structure. So, agglomerative hierarchical clustering is applied to datasets to find groups of features. Each dataset is divided into five groups. After splitting data into groups or clusters, relief is applied to each group to remove irrelevant or less important features. The top 500 features were selected after relief. Then, FDA

TABLE 2. Selected group features on GSE16446 data.

Group	G1	G2	G3	G4	G5
No. of features	8605	5156	2774	4941	1768
Selected groups by the proposed method			✓	✓	
Selected groups by SGL			✓	✓	

TABLE 3. Selected group features on GSE25066 data.

Group	G1	G2	G3	G4	G5
No. of features	3766	3032	2057	2351	2252
Selected groups by the proposed method	✓	✓			
Selected groups by SGL	✓	✓			

TABLE 4. Selected group features on GSE2034 data.

Group	G1	G2	G3	G4	G5
No. of features	2528	3463	2451	3433	746
Selected groups by the proposed method	✓	✓			
Selected groups by SGL		✓	✓		

TABLE 5. Selected group features on simulated data.

Group	G1	G2	G3	G4	G5
No. of features	1722	2513	1307	2006	2452
Selected groups by the proposed method	✓			✓	✓
Selected groups by SGL	✓			✓	

was applied to each group to reduce dimension. Afterward, a random forest was used to determine the relative importance and ranking of these group features. Figures 2 show the relative importance of group features. The proposed method also gave the ranking of group features. Existing methods are lacking in this area; those methods can't provide the ranking and relative importance of group features. The proposed method cannot only select the highest significant group features but also give the ranking and relative importance of all group features.

Top-ranked group features were selected after calculating relative importance. Tables 2, 3, 4, and 5 represent the total number of features in each group feature, as well as selected group features after applying the proposed method.

C. PERFORMANCE METRICS

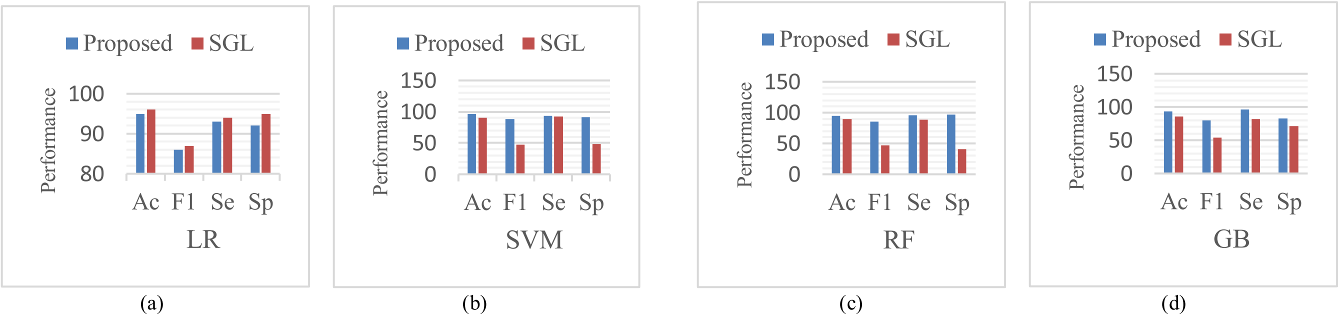
The proposed and supervised group lasso methods' performance is compared using different machining learning algorithms. For comparison of the proposed method with the existing method, support vector machine (SVM), logistic regression (LR), random forest (RF), and gradient boosting (GB) classifiers by using 5-fold cross-validation were used. To determine the performance of these methods the following performance metrics are used: accuracy, F1 score, sensitivity, and specificity. The mathematical equation of accuracy can be shown as:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{8}$$

In equation 8, *tp* denotes true positive, *tn* indicates true negative, *fp* represents false positive, and *fn* shows false negative.

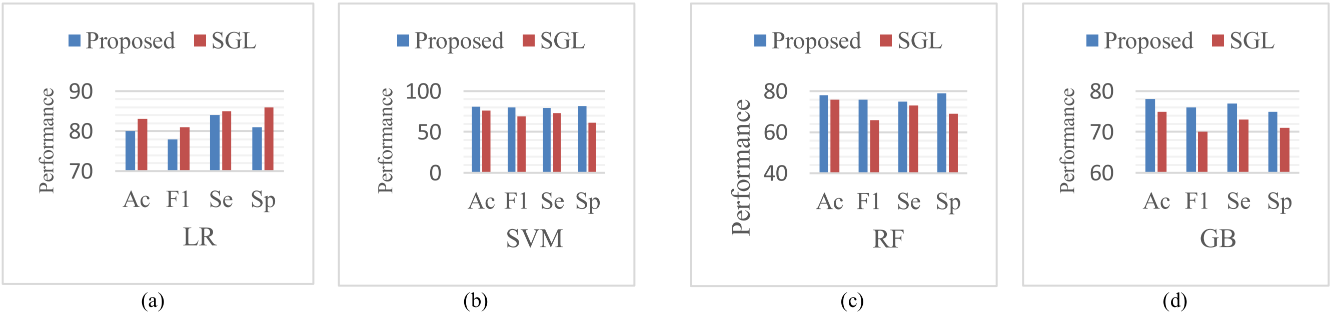
The F1 score is another important parameter for evaluating the performance of machine learning approaches, and it's the weighted average of precision and recall. F1 score is calculated by:

$$F1\,score = 2 * \frac{precision * recall}{precision + recall} \tag{9}$$



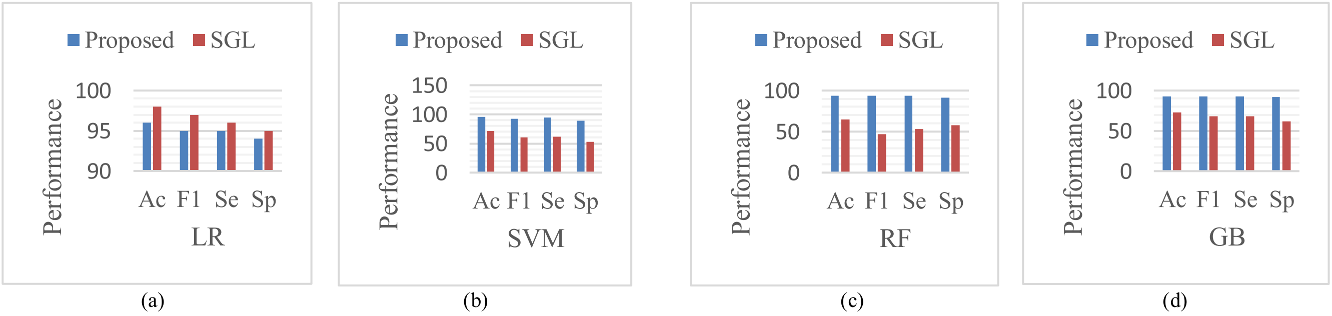
(a) (b) (c) (d)

FIGURE 3. Performance of the proposed and SGL method on LR (a), SVM (b), RF (c), and GB (d) on the GSE16446 dataset.



(a) (b) (c) (d)

FIGURE 4. Performance of the proposed and SGL method on LR (a), SVM (b), RF (c), and GB (d) on the GSE25066 dataset.



(a) (b) (c) (d)

FIGURE 5. Performance of the proposed and SGL method on LR (a), SVM (b), RF (c), and GB (d) on the GSE2034 dataset.

In this equation (9), precision is the positive predicted value, and it can be estimated by:

$$precision = \frac{tp}{tp + fp} \tag{10}$$

whereas recall in equation (9) can be calculated as:

$$recall = \frac{tp}{tp + fn} \tag{11}$$

The three datasets which are used in this study are imbalanced. So, only accuracy and F1 score cannot completely evaluate the proposed method. Therefore, sensitivity and specificity measures are also used. These metrics can be determined by:

$$Sensitivity = \frac{tp}{tp + fn} \tag{12}$$

$$Specificity = \frac{tn}{tn + fp} \tag{13}$$

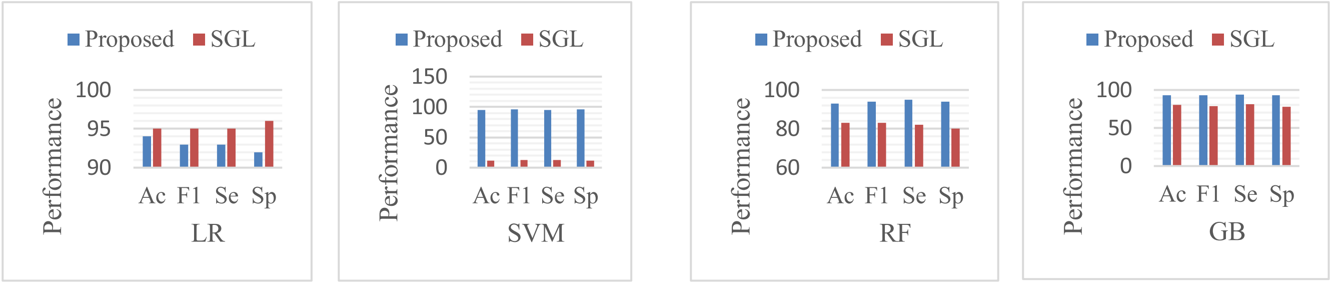
These performance metrics are used for the evaluation of the proposed and existing method.

D. COMPARISON BASED ON PERFORMANCE METRICS

After selecting highly ranked group features, machine learning algorithms were applied to these selected group features to check the proposed method's performance.

Similarly, a supervised group lasso was applied to each group feature. After finding the important groups and features, and combining them, machine learning approaches were applied to that data to check the performance of the supervised group lasso method.

The proposed method was applied to each group with 5-fold cross-validation and then took the average of these



(a) (b) (c) (d)

FIGURE 6. Performance of the proposed and SGL method on LR (a), SVM (b), RF (c), and GB (d) on the simulated dataset.

five values. In GSE16446 data, the accuracy of the proposed method on LR, SVM, RF, and GB was 95, 96, 95, and 93 percent respectively. Moreover, the F1 score of the proposed method was 86, 88, 85, and 80 on these classifiers. However, SGL had 96, 90, 89.9, and 86 percent accuracy on these classifiers respectively. Furthermore, the F1 score of SGL on these classifiers were 87, 47, 48, and 54 respectively. Figures 3-6 show the performance of the proposed method with SGL for accuracy, F1 score, sensitivity, and specificity measures. In these results, it can be seen that on the logistic regression classifier SGL performs slightly better on the accuracy, F1 score, sensitivity, and specificity than the proposed method. It is because SGL is specially designed for LR. The performance of the proposed method is not bad on LR. However, on all other three classifiers proposed method outperforms the SGL.

E. TIME COMPLEXITY ANALYSIS

Figure 7 shows the time complexity analysis between the proposed and SGL method. On GSE16446, the total features were 23265. SGL took 12.3 minutes to implement on that data whereas the proposed method took 11.6 minutes. On the other three datasets, SGL running time was 7.8, 7.6, and 6.6 minutes respectively. However, the proposed method took 7.2, 7.1, and 6.3 minutes respectively. Therefore, the proposed method takes less time compared with the existing method such as SGL.

In this figure, it can be seen that the proposed method is more computationally efficient than SGL. All these experiments were done on a desktop with 8GB RAM and Intel Core 8th generation i5 CPU.

The proposed method performs better because after removing irrelevant features by relief, the dimension was reduced to 1 by applying FDA. In each group feature, FDA reduced its dimension to 1. As the total number of clusters is 5 in each dataset. Therefore, after applying FDA, only five features remained. Each feature represents the group feature. However, SGL did not reduce the dimension of each group to 1. After applying SGL on GSE16446, GSE25066, GSE2034, and simulated data, the total number of features was 32, 63, 164, and 55 respectively.

So, by applying the proposed method, the dimension was reduced to 5 but after applying the SGL dimension was high as compared to the proposed method. If the dimension will high, there will be more overfitting, and the model will become more complex. As a result, machine learning algorithms did not perform well on overfitting and complex models. On the other hand, if the dimension is reduced, the chance of overfitting ends, and the model becomes simple. Therefore, machine learning approaches perform better on low-dimensional data.

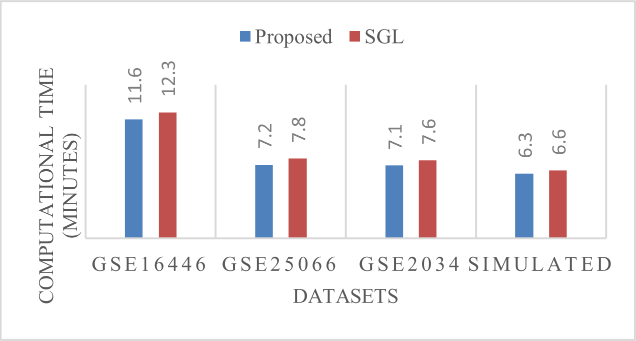


FIGURE 7. Time complexity analysis among the proposed and SGL method on GSE16446, GSE25055, GSE2034, and simulated dataset.

If the data dimension is high, the machine learning model will have high variance but low bias. Conversely, low-dimensional data has low variance with high bias. In machine learning, using a model with a high bias but a low variance is more appropriate than a high variance with a low bias. Machine learning algorithms do not give better performance if the variance is high. During the dimension reduction process, it should be noted that a decrease in variance should be larger than the increase in bias. The proposed method performs better than SGL because it has low variance and high bias. Biasness is negligible as compared to the reduction in variance.

F. COMPARISON BASED ON THE ROC CURVE AND AUC SCORE

The receiver operating characteristic (ROC) curve and area under the curve (AUC) are other important metrics for measuring methods' performance. The ROC curves of the proposed method and SGL on different machine learning techniques on all datasets are shown in Figures 8-11. Machine

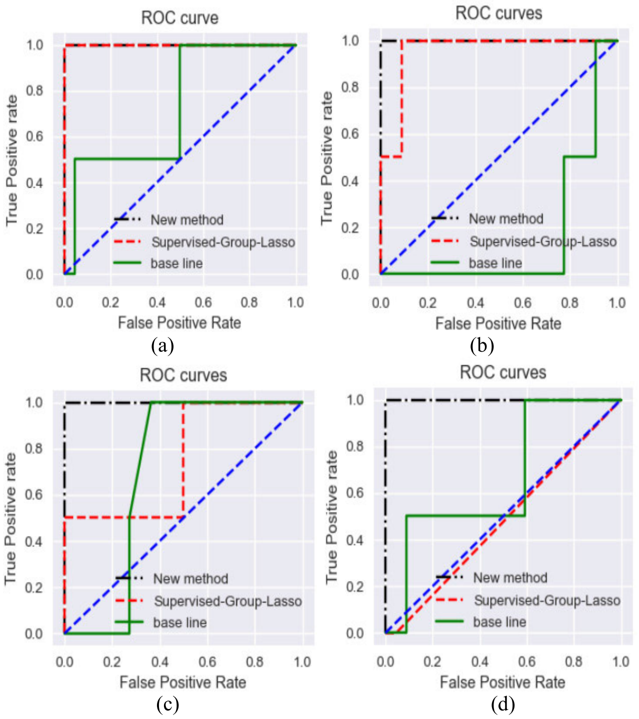


FIGURE 8. Performance comparison between proposed and SGL methods on logistic regression LR (a), SVM (b), RF (c), and GB (d) on the GSE16446 dataset. In (a) AUC score of the proposed and existing method is 1. In (b) AUC score of the proposed and existing methods is 1 and 0.95, respectively. In (c) AUC score of the proposed and existing methods is 1 and 0.75, respectively. In (d) AUC score of the proposed and existing methods is 1 and 0.47, respectively.

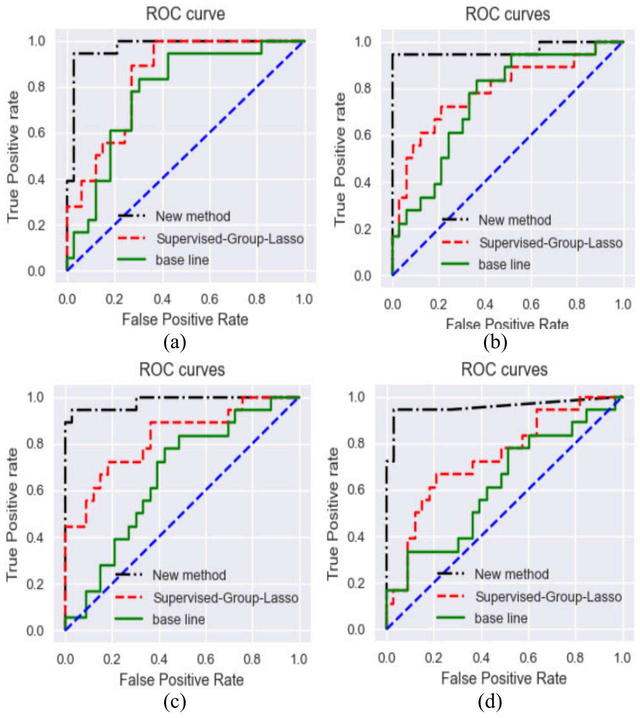


FIGURE 9. Performance comparison between proposed, and SGL methods on logistic regression (a), SVM (b), RF (b), and GB (c) on the GSE25066 dataset. In (a) AUC score of the proposed and existing method is 0.96 and 0.99, respectively. In (b) AUC score of the proposed and existing methods is 0.93 and 0.89, respectively. In (c), the AUC scores of the proposed and existing methods are 0.96 and 0.86, respectively. Finally, in (d), the AUC score of the proposed and existing methods is 0.95 and 0.84, respectively.

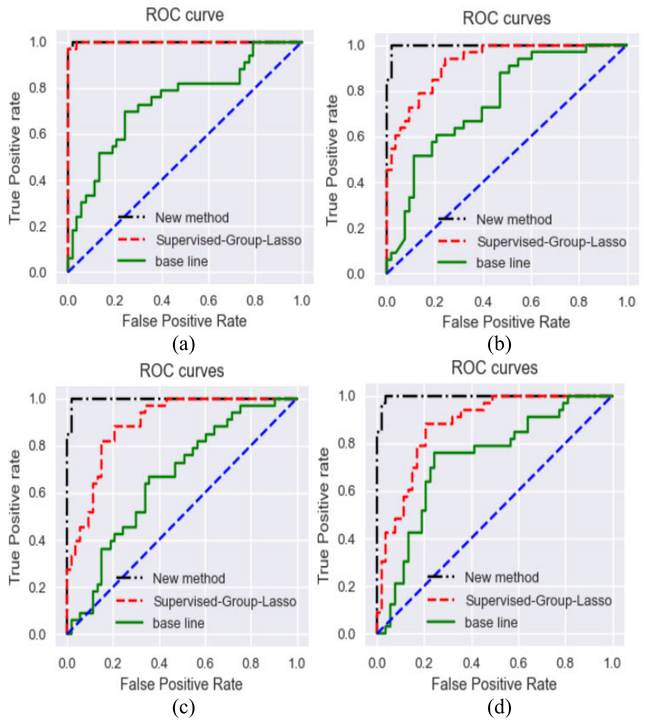
learning algorithms were also applied to the original data. The baseline represents the results of machine learning algorithms on original data. In these figures, the graphical performance of these two methods can be seen on different thresholds. In these figures, it can be seen that SGL performs equal to or better than the proposed method on logistic regression, but on all other three classifiers (SVM, RF, GB) proposed method outperforms the SGL. SGL is specially designed for logistic regression and linear classifiers, but the proposed method is not specific to linear classifiers. Therefore, the proposed method can not only perform on linear methods but also gives almost the same performance on nonlinear classifiers like SVM. On logistic regression, the proposed method performs the same in Figures 8, but in Figures 9-11, its performance is slightly lower than SGL. So, one can say that on LR, both methods perform equally. On SVM, the proposed method performs better than SGL in all cases. However, on the other two classifiers (RF and GB) performance of the proposed method is also better than SGL.

G. VALIDATION OF RESULTS BY STATISTICAL NON-PARAMETRIC TESTS

To validate the proposed method, we compared the proposed one with the counterparts such as NO-FS (i.e. no feature selection) and SGL. Classification accuracies are obtained by using four different machine learning algorithms LR, SVM, RF, and GB on data obtained by the proposed and its counterparts. Each method has four accuracies on each dataset. For instance, the proposed method was applied to the GSE16446 dataset and selects the important group features. On those important group features, four machine learning algorithms were applied and got classification accuracies. Similarly, the proposed method achieved accuracies with these machine learning algorithms on the other three datasets. Four datasets were used for this study. Therefore, in the end, each method has sixteen accuracies.

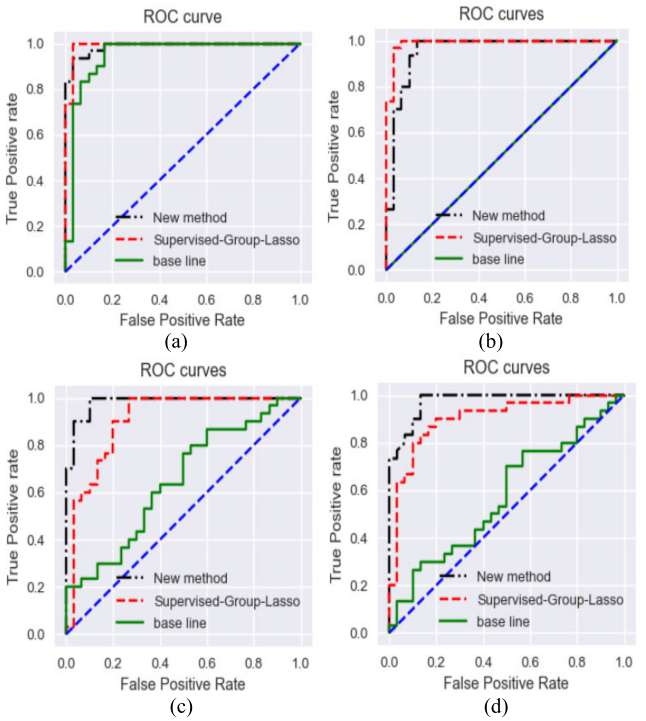
On these accuracies, the non-parametric test is applied. Non-parametric tests are very useful for determining the significance of the proposed method against the existing method [36]. For this purpose, we applied a pairwise Wilcoxon signed rank test on the accuracies for confirming and validating the results. This test is used for the paired data and calculates the difference score for each pair. It also analyzes the sign and magnitude of the difference score. After calculating the absolute value of the difference score, it ranks the absolute value of the difference score. The last step is to assign the sign of difference score to each rank. Therefore, some ranks have positive sigh and some have negative signs.

Table 6 computes the ranks of the Wilcoxon signed rank test between the proposed and SGL method based on



(a) (b) (c) (d)

FIGURE 10. Performance comparison between proposed, and SGL methods on logistic regression (a), SVM (b), RF (b), and GB (c) on the GSE2034 dataset. In (a) AUC score of the proposed and existing method is 0.99. In (b) AUC score of the proposed and existing methods is 0.99 and 0.92, respectively. In (c) AUC score of the proposed and existing methods is 0.99 and 0.89, respectively. In (d) AUC score of the proposed and existing methods is 0.99 and 0.87, respectively.



(a) (b) (c) (d)

FIGURE 11. Performance comparison between proposed, and SGL methods on logistic regression (a), SVM (b), RF (b), and GB (c) on the simulated dataset. In (a) AUC score of the proposed and existing method is 0.98 and 0.99, respectively. In (b) AUC score of the proposed and existing methods is 0.96 and 0.99, respectively. In (c) AUC score of the proposed and existing methods is 0.98 and 0.90, respectively. In (d) AUC score of the proposed and existing methods is 0.98 and 0.90, respectively.

accuracies. In this table, each row represents the positive rank against the given method in the column. The last row of the table shows the positive rank of the proposed method against the other methods.

TABLE 6. Calculated ranks by Wilcoxon signed-rank test.

	[a]No-FS	SGL	Proposed
No-FS	-	8.5	0
SGL	69.5	-	18.5
Proposed	137.4	117.5	-

[a] No-FS shows the results without applying any group feature selection method.

TABLE 7. The rejection map of Wilcoxon signed rank test among the methods used in this study.

	No-FS	SGL	Proposed
No-FS	-	blue	blue
SGL	red	-	blue
Proposed	red	red	-

Table 7 shows the rejection map. In this table, the red indicates that the method given in the row performs better than the method given in the column. Moreover, blue represents the method given in the column performs better than the method given in the row. For example, in the last row red indicates that the proposed method performs better than the No-FS and SGL. Consequently, the proposed method has a significant difference from the other methods. In other words, the proposed method outperforms the other methods.

Data used in this study have binary target variables; future work can be done on a continuous variable. The first goal of this study is to develop a group feature selection and ranking method that can perform on any type of classifier. Moreover, check the performance of the proposed method with the existing group feature selection method. The proposed method performs equal to or better than SGL on almost all performance metrics.

V. CONCLUSION AND DISCUSSION

To the best of our knowledge, few studies proposed a feature extraction method for group feature selection. Existing group feature selection methods do not extract new features from existing features for group feature selection. The proposed method can also perform a ranking among group features according to their significance and importance. This proposed method was applied to real-world data to check its power in group selecting and ranking.

The datasets used in this study have a low sample size and large feature size. Consequently, it is necessary and difficult to reduce the dimension to a low subspace. This study suggests that if the data dimension goes down, the data performance will increase. In contrast, performance will go down as the dimension will be high. One can adjust the dimension reduction rate according to the data, and, if the data has a large sample size, then the researcher is not bound to reduce the dimension of each group to 1.

The proposed method improves the adaptability of the existing method with the classifier because it can be used with any classifier. It also enhances the existing method in terms of ranking of groups because existing methods select a few groups. Still, they do not provide information on which group is most important and which one is less important. The proposed method is also better for machine learning algorithms because it transforms all features of the group into one dimension, so if the dimension of data will low, the learning algorithm becomes easier and faster. On the other hand, the existing methods are complex compared to the proposed method.

In this study, a new feature extraction method for group feature selection and ranking has been proposed for small samples, but high dimensional data. After splitting data into training and testing data, the proposed method removed the irrelevant features from each group and then reduced each group's dimension by applying FDA. First, machine learning models were trained on training data. Then, after the training of models, test data was given to the models to check the performance of the proposed data. Results showed that the performance of the proposed method on linear classifiers is almost equal to the existing method and has better performance on nonlinear classifiers.

The proposed method has many required properties. The proposed method extracted only one feature in each group to attain the same classification accuracy, sensitivity, specificity, and F1 score as in the original high dimensional data. The learning process of machine learning algorithms is also accessible after applying the proposed method because it converts the original data into a low dimension compared to the existing method.

Although this method was applied to HDLSS data, one can use this method for large sample-size data as well. This study used data that has binary target variable with 0 and 1 class. For future work, the proposed method can be extended to multiple classes other than the bioinformatics field.

APPENDIX

Algorithm 1 is the pseudocode of the relief method. The distance matrix is found by calculating distance among instances (Algorithm, line 6). After that hits and misses of each instance are computed (Algorithm, line 11). In line 16, the score of each feature is calculated.

Algorithm 1: Pseudo Code of Relief Algorithm

```
1      $n \leftarrow$ number of training instances
2      $f \leftarrow$ features
3      $K \leftarrow$ Total number of features
4      $P \leftarrow$ number of nearest hits and misses
5      Pre-process dataset $D$
6      Compute the distance for each instance and make a
         distance matrix
7      Initialize all feature score $W[f] = 0$
8      for $i = 1$ to $n$ do
9          for $j = 1$ to $n$ do
10             Randomly select an instance $O_i$
11             Find the $P$ nearest hits and $P$ nearest misses
                 for each instance
12         end
13         for all misses and hits do
14             # features weight update
15             for $f := 1$ to $K$ do
16             $W[f] := W[f] - \frac{diff(f, O_i, H)}{n} + \frac{diff(f, O_i, M)}{n}$
17         end
18     end
19     return vector $W$ of features score
20 end
```

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MIST) (NRF-2022R1F1A1063273). The authors would like to thank the editor and reviewers for their valuable and constructive comments.

REFERENCES

[1] W. Gao, L. Hu, P. Zhang, and F. Wang, "Feature selection by integrating two groups of feature evaluation criteria," *Exp. Syst. Appl.*, vol. 110, pp. 11–19, Nov. 2018.

[2] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[3] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, 1997.

[4] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*. Berlin, Germany: Springer, 2008.

[5] A. Dehnavi, M. Sehhati, H. Rabbani, and S. Javanmard, "Using protein interaction database and support vector machines to improve gene signatures for prediction of breast cancer recurrence," *J. Med. Signas Sensors*, vol. 3, no. 2, p. 87, 2013.

[6] F. Tang, L. Adam, and B. Si, "Group feature selection with multiclass support vector machine," *Neurocomputing*, vol. 317, pp. 42–49, Nov. 2018.

[7] T. Abeel, T. Helleputte, Y. Van De Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2010.

[8] S. Bakin, "Adaptive regression and model selection in data mining problems," Doctor Philosophy, Austral. Nat. Univ., Canberra, Australia, May 1999.

[9] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc., Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.

[10] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 70, no. 1, pp. 53–71, Feb. 2008.

[11] S. Noah, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J. Comput. Graph. Stat.*, vol. 22, no. 2, pp. 231–245, May 2013.

[12] S. Ma, X. Song, and J. Huang, "Supervised group lasso with applications to microarray data analysis," *BMC Bioinf.*, vol. 8, no. 1, pp. 1–17, Dec. 2007.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B*, vol. 58, no. 1, pp. 267–288, 1996.

[14] K. Fang, X. Wang, S. Zhang, J. Zhu, and S. Ma, "Bi-level variable selection via adaptive sparse group lasso," *J. Stat. Comput. Simul.*, vol. 85, no. 13, pp. 2750–2760, Sep. 2015.

[15] M. Vincent and N. R. Hansen, "Sparse group lasso and high dimensional multinomial classification," *Comput. Statist. Data Anal.*, vol. 71, pp. 771–786, Mar. 2014.

[16] N. Chumerin and M. Van Hulle, "Comparison of two feature extraction methods based on maximization of mutual information," in *Proc. 16th IEEE Signal Process. Soc. Workshop Mach. Learn. Signal Process.*, Sep. 2006, pp. 343–348.

[17] Y. Xie and T. Zhang, "A fault diagnosis approach using SVM with data dimension reduction by PCA and LDA method," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2015, pp. 869–874.

[18] P. Mehta, M. Bukov, C. H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, "A high-bias, low-variance introduction to machine learning for physicists," *Phys. Rep.*, vol. 810, pp. 1–124, May 2019.

[19] H. Zhang, J. Wang, Z. Sun, J. M. Zurada, and N. R. Pal, "Feature selection for neural networks using group lasso regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 4, pp. 659–673, Apr. 2020.

[20] D. Ö. Sahin, O. E. Kural, S. Akleylek, and E. Kilic, "Permission-based Android malware analysis by using dimension reduction with PCA and LDA," *J. Inf. Secur. Appl.*, vol. 63, Dec. 2021, Art. no. 102995.

[21] D. J. Hand, "Classifier technology and the illusion of progress," *Stat. Sci.*, vol. 21, no. 1, pp. 1–14, Feb. 2006.

[22] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[23] M. Yassi and M. H. Moattar, "Robust and stable feature selection by integrating ranking methods and wrapper technique in genetic data classification," *Biochem. Biophys. Res. Commun.*, vol. 446, no. 4, pp. 850–856, Apr. 2014.

[24] A. U. Haq, D. Zhang, H. Peng, and S. U. Rahman, "Combining multiple feature-ranking techniques and clustering of variables for feature selection," *IEEE Access*, vol. 7, pp. 151482–151492, 2019.

[25] B. Bonev, F. Escolano, D. Giorgi, and S. Biasotti, "Information-theoretic selection of high-dimensional spectral features for structural recognition," *Comput. Vis. Image Understand.*, vol. 117, no. 3, pp. 214–228, 2013.

[26] S. W. Card, "Information distance based fitness and diversity metrics," in *Proc. 12th Annu. Conf. Companion Genetic Evol. Comput. (GECCO)*, 2010, pp. 1851–1854.

[27] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. AAAI*, 1992, vol. 2, no. 1992, pp. 129–134.

[28] J. Ye, T. Xiong, and D. Madigan, "Computational and theoretical analysis of null space and orthogonal linear discriminant analysis," *J. Mach. Learn. Res.*, vol. 7, no. 7, pp. 1–22, 2006.

[29] J. G. Thomas, J. M. Olson, S. J. Tapscott, and L. P. Zhao, "An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles," *Genome Res.*, vol. 11, no. 7, pp. 1227–1236, Jul. 2001.

[30] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, and M. A. Caligiuri, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[31] H. Tao, C. Bausch, C. Richmond, F. R. Blattner, and T. Conway, "Functional genomics: Expression analysis of Escherichia coli growing on minimal and rich media," *J. Bacteriology*, vol. 181, no. 20, pp. 6425–6440, Oct. 1999.

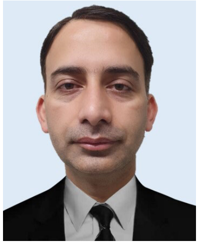
[32] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, "Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments," *FEBS Lett.*, vol. 573, nos. 1–3, pp. 83–92, Aug. 2004.

[33] T. T. Le, R. J. Urbanowicz, J. H. Moore, and B. A. McKinney, "STatistical inference relief (STIR) feature selection," *Bioinformatics*, vol. 35, no. 8, pp. 1358–1365, Apr. 2019.

[34] L. Breiman, "Random forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

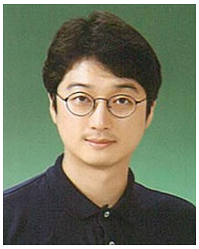
[35] H. Xie, J. Li, Q. Zhang, and Y. Wang, "Comparison among dimensionality reduction techniques based on random projection for cancer classification," *Comput. Biol. Chem.*, vol. 65, pp. 165–172, Dec. 2016.

[36] M. K. Ebrahimpour and M. Eftekhari, "Ensemble of feature selection methods: A hesitant fuzzy sets approach," *Appl. Soft Comput.*, vol. 50, pp. 300–312, Jan. 2017.



IQBAL MUHAMMAD ZUBAIR received the M.S. degree in engineering management from the University of Engineering and Technology, Taxila, Pakistan, in 2015. He is currently pursuing the Ph.D. degree in industrial and management engineering with Hanyang University, ERICA, Ansan, South Korea.

His research interests include statistical learning, data mining, and machine learning application, especially in bioinformatics.



BYUNGHOON KIM received the Ph.D. degree in industrial and systems engineering from Rutgers University, New Brunswick, NJ, USA, in 2015.

He is currently an Assistant Professor with the Department of Industrial and Management Engineering, Hanyang University, Ansan, South Korea. His research interests include data mining in the semiconductor manufacturing processes, network data analysis, and high-dimensional data analysis.