IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received March 24, 2022, accepted April 17, 2022, date of publication May 3, 2022, date of current version May 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3172281

Gene Selection in Binary Classification Problems Within Functional Genomics Experiments via Robust Fisher Score

MUHAMMAD HAMRAZ[1], ZARDAD KHAN[1], DOST MUHAMMAD KHAN[1], (Senior Member, IEEE), NAZ GUL[1], AMJAD ALI[1], AND SAEED ALDAHMANI[2]

[1]Department of Statistics, Abdul Wali Khan University Mardan, Khyber Pakhtunkhwa 23200, Pakistan
[2]Department of Analytics in the Digital Era, United Arab Emirates University, United Arab Emirates

Corresponding authors: Zardad Khan (zardadkhan@awkum.edu.pk) and Amjad Ali (amjad.ali@awkum.edu.pk)

ABSTRACT This study proposes a supervised feature selection technique for classification in high dimensional binary class problems by adding robustness in the conventional Fisher Score. The proposed method utilizes the more robust measure of location i.e. the Median and measure of dispersion known as Rousseeuw and Croux statistic (Q_n). Initially minimum subset of genes is identified by the Greedy search approach, which is then combined with the top ranked genes obtained via the proposed Robust Fisher Score (RFish). Finally to remove redundancy in the selected genes, Least Absolute Shrinkage and Selection Operator (LASSO) has been applied. The proposed method is validated on five publicly available datasets. The results of the proposed method are compared with six well known feature selection methods based on prediction performance via Random Forest (RF), Support Vector Machine (SVM) and k Nearest Neighbour (k-NN) classifiers. Box-plots and Bar-plots of the results of the proposed method and all the other methods considered in the manuscript are also given. The Results show that the proposed method (RFish) performs better than all the other methods in majority of the cases. The paper gives a detailed simulation study to further assess the proposed method.

INDEX TERMS Classification, feature selection, high dimensional gene expression datasets, Fisher Score, Rousseeuw and Croux statistic.

I. INTRODUCTION

Recent revolution in the functional genomic technologies is producing a huge amount of data. Microarray and other high-throughput technologies are capable of studying thousands of genes generated by these technologies, simultaniously. The understanding of such microarray datasets has paved a way in developing new statistical learning tools. The main problem in such type of datasets is the curse of dimensionality, where tens of thousands of genes are measured on a very small number, tens to few hundreds of samples or observations. These problems are also referred to as ($n < p$) problems. Such high-dimensional gene expression datasets are usually not good for classification purpose because of the curse of dimensionality [1]. Moreover most of the genes in such type of datasets are noisy and do not contribute in the classification of observations to their ture classes (phenotypes). This problem can be solved by using a technique called feature selection, [2], [3]. Feature selection reduces dimension of the data by selecting a subset of important genes while removing the redundant and irrelevant ones from the entire features/genes space, thereby reducing complexity and cost of the computation. Moreover, a small number of features are also useful in mitigating the training time, increasing the generalizability of models by minimizing their variances, as well as in mitigating the curse of dimensionality in ($n < p$) problems.

Generally speaking, feature selection methods can be divided into three categories i.e. Wrapper, Embedded, and Filter methods [3]. The filter methods rank the genes according to their importance before the learning algorithm is employed. Filter methods selects those genes for classification that have higher ranking scores. Examples of such methods could be found in [1]–[13]. Wrapper methods assign ranks to genes according to their importance with the help of learning algorithms that will ultimately be employed.

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

The method given in [14] comes under the category of Wrapper methods. In Embedded methods the learning algorithm and feature selection methods are combined to select the most informative genes. The specific learning algorithm and embedded methods are tightly coupled, thereby reducing the application of these feature selection methods to other learning algorithms. The commonly used embedded procedures are decision tree algorithm, regression with LASSO and Ridge regression. The LASSO and ridge methods shrink the coefficients of non-informative features to zero and almost zero, respectively, by applying some penalty on the features coefficients. Method given in [15] comes under the category of Embedded methods. This paper is based on the idea that falls under the category of filtering methods. Motivated by [16] and [17], the proposed method selects discriminative genes or features by using the Robust Fisher Score approach which are then combined with the minimum subset of genes obtained via greedy search approach given in [12], for binary class problems. To remove the redundancy in the selected genes, this paper uses the (LASSO). The proposed method is applied on five publicly available benchmark gene expression datasets. To evaluate the performance of the selected genes via the proposed method, the results are compared with six well known feature selection methods that are: proportion overlapping score (POS) [12], maximum relevance and minimum redundancy (mRmR) [18], Wilcoxon rank sum test [19], sigF [20], GClust [13] and [16], based on classification error rate. For this purpose, three popular classifiers namely Random Forest (RF) [21], k-Nearest Neighbors (k-NN) [22] and Support Vector Machine (SVM) [23] are used to check the performance of genes selected through the proposed method in comparison with the other feature/gene methods. Motivated by the Fisher score given in [16], a new statistic called robust fisher score (RFish) has been proposed which is based on the robust measures of location and dispersion i.e., median and Rousseeuw and Croux statistic. The advantage of using these robust measures is that the value of (RFish) will remain stable in the presence of outliers. It shows the actual discriminative ability of the genes avoiding the effect of outliers. Moreover, to remove the redundancy in the selected genes, (LASSO) has been used, which shrinks the coefficients of uninformative and redundant genes to zero and retains the informative ones.

The remainder of this paper is organized as follows. Section 2 provides a summary of the related work done in the literature. In Section 3, the proposed method is discussed thoroughly. Section 4, gives experiments on the benchmark datasets and results of the proposed method. This section also gives a detailed simulation study to support the argument given in this paper. Finally, Section 5 gives the conclusion based on the work done in the paper.

II. RELATED WORK

Recent years have witnessed a lot of work on feature selection in high dimensional gene expression datasets. The main objective of feature selection is to identify the most discriminative and important features [3]. Feature selection reduces computational complexity as well the training time of the learning models by enhancing or retaining their accuracy [24]–[26]. According to [2], the accuracy and prediction power of classifier could be enhanced if it is provided with the important and discriminative genes. According to [27]–[29], it is NP-hard problem to select an appropriate subset of features and has attracted a lot of researchers to use stochastic and heuristic algorithms. Authors in [30], [31] have introduced methods that allocate the important feature set to the local behaviour of data by using different regions of feature space. Similarly studies in [32]–[34] and [35], [36] have selected the important features by ranking them while using the aggregate sample data. The (LASSO) is an another feature selection method, which selects the discriminative features by setting the coefficients of non informative features/genes equal to zero [35]. A method called least angle regression (LAR) given in [33] is based on the LASSO method which computes all the LASSO estimates, and then chooses those genes for model construction which are highly correlated with the genes already selected. A semi-supervised feature selection method introduced in [36], called Rescaled Linear Square Regression (RLSR) have incorporated rescaling factors in order to rank the features and exploit the least square regression model. Authors in [34] introduced feature selection method known as Hilbert–Schmidt independence criterion least absolute shrinkage and selection operator (HSIC-Lasso). This method selects the most discriminative and non-redundant features by using set of kernel functions. A technique known as ‘relative importance’ was proposed in [37], in which important genes were identified by growing a large number of trees. A method known as minimal redundancy maximal relevance (mRMR) given in [18] declares those genes as important which have maximum relevance with the response class and minimum redundancy. The ensemble version of [18] can be found in [38]. Authors in [39] have used the Principal component analysis technique for the selection of informative features/genes. Genes that corresponds to the component with less variation were declared as important. Similarly, another study given in [40] has used the factor analysis technique in order to select a set of discriminative genes. A comparison of various feature selection techniques is given in [41], [42]. Several studies have used the p-value of statistical tests for the selection of important genes. Examples are the Wilcoxon rank-sum test and t-test for feature selection as can be found in [19], [20]. Selection of important genes with the help of impurity measures like Gini index, max minority and information gain can be found in [43]. Informative genes can also be selected by investigating the overlapping degree of genes across the different classes. Authors in [44] proposed a method by investigating the overlapping degree of genes across the different classes. Genes that have smaller overlapping degree between the classes are declared informative. Authors in [45] extended

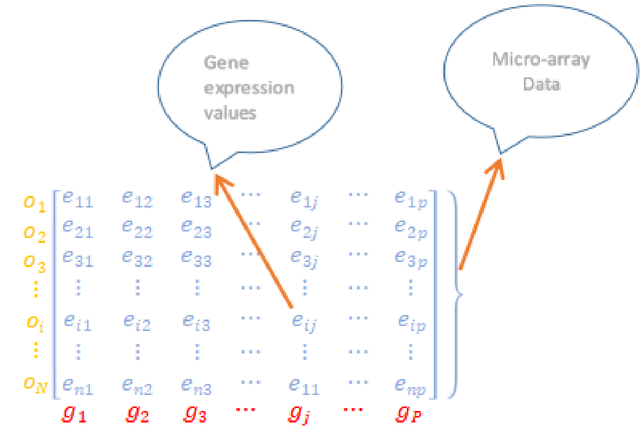
Gene expression values
Micro-array Data
o_1 e_{11} e_{12} e_{13} ⋯ e_{1j} ⋯ e_{1p}
o_2 e_{21} e_{22} e_{23} ⋯ e_{2j} ⋯ e_{2p}
o_3 e_{31} e_{32} e_{33} ⋯ e_{3j} ⋯ e_{3p}
o_i e_{i1} e_{i2} e_{i3} ⋯ e_{ij} ⋯ e_{ip}
o_N e_{n1} e_{n2} e_{n3} ⋯ e_{11} ⋯ e_{np}
g_1 g_2 g_3 ⋯ g_j ⋯ g_P

FIGURE 1. A general layout of the gene expression data.

the idea of [44] by introducing one additional factor i.e. the number of overlapped observations in the overlapping area for each gene. The authors in [45] combined the minimum subset of genes obtained via Greedy search approach with the top ranked genes according to their proposed method. The idea of [45] was further extended by [12] named as 'POS' score, by adding a factor i.e. proportion of overlapping samples in each class for each gene. Genes with the smallest 'POS' scores are considered as informative. Authors in [46] introduced a criteria for feature selection by performing extensive experiments to investigate the efficacy of their proposed method on the publicly available datasets related to computational biology. Feature selection for multi class problems in high dimensional datasets can be found in [47], which is an extended version of decomposition-based multi-objective optimization approach. Further methods relating to feature/gene selection in high dimensional gene expression datasets could be found in [46], [48]–[57] and the references cited therein. Moreover, methods used in different applications associated to features selection can be found in [58]–[67].

III. ROBUST FEATURE SELECTION METHOD

Microarray gene expression datasets are generally represented in the form of a matrix and it is given by $E = [e_{ij}]$, where $E \in \Re^{N\times P}$ and e_{ij} is the observed value of gene expression for i^{th} gene and for j^{th} tissue sample or observation, for $j = 1, 2, 3, \dots, P$ and $i = 1, 2, 3, \dots, N$. In binary class gene expression datasets, each tissue sample is categorized into one of the two categories, 0 or 1. Suppose $Y \in \Re^N$ is class labels vector, then its i^{th} element will have a unique value c, which is either 0 or 1. The matrix representation of a gene expression datasets is given in Figure 1. Samples are given in the rows while genes are listed in the colums of the give figure. Each cell in the matrix represents gene expression values of various genes for the corresponding samples.

The necessary definitions used in this paper are as follows:

Median:

The median of any jth gene, for class ($c = 0, 1$) is given by $Med_{(j,c)}$ where $j = 1, 2, 3, \dots, P$.

Rousseeuw and Croux Statistic(Q_n):

The Rousseeuw and Croux Statistic (Q_n) given in [68], for any jth gene and class $c = 0, 1$ is given as:

$$Q_{n(j,c)} = c_{n(j,c)} * Q_{1(j,c)}(|x_{l(j,c)} - x_{m(j,c)}| : l < m), \tag{1}$$

where $c_{n(j,c)}$ is a constant which depends on the length of j^{th} gene and class c. The quantity $|x_{l(j,c)} - x_{m(j,c)}|$ represents the absolute pairwise differences of gene expression values of j^{th} gene for class $c = 0, 1$.

Gene Masks:

The matrix of gene masks $M \in \Re^{N\times P}$ is computed by the method given in [12]. The gene masks show the ability of genes to correctly classify the tissue samples to their true classes. In other words it shows the classification power of gene. The elements of genes mask matrix M is computed as follows.

$$m_{ij} = \begin{cases} 1, & \text{if } e_{ij} \in \text{(non-overlapping region as defined,} \\ & \text{in [12])} \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $j = 1, 2, 3, \dots, P$ and $i = 1, 2, 3, \dots, N$.

Minimum subset of genes:

The minimum subset of genes that correctly classify the tissue samples to their true classes are identified with the help of the method given in [12]. These are the genes that correctly classify the maximum number of observations to their true response classes avoiding the effect of outliers in the training phase.

Robust Fisher Score(RFish):

Motivated by [16] and [17], a Robust Fisher Score (RFish) is proposed in this paper for binary class problems and is given by;

$$RFish_{(j)} = \frac{n_1|Med_{(j,1)} - Med_{(j)}| + n_0|Med_{(j,0)} - Med_{(j)}|}{Q_{n(j)}}, \tag{3}$$

where $RFish_{(j)}$ represents the robust fisher score for j^{th} gene, $Med_{(j,1)}$ and $Med_{(j,0)}$ are the medians of j^{th} gene for class 1 and 0, respectively. n_0 and n_1 are the number of tissue samples, for a particular gene in class 0 and class 1 respectively. $Med_{(j)}$ is the overall median of j^{th} gene for both the classes. Furthermore, $Q_{n(j)} = n_1 Q_{n(j,1)} + n_0 Q_{n(j,0)}$, where $Q_{n(j,1)}$ and $Q_{n(j,0)}$ are the Rousseeuw and Croux Statistics for measuring the variability in j^{th} gene for class 1 and 0, respectively.

Least Absolute Shrinkage and Selection Operator(LASSO):

One of the main problems in ordinary Fisher Score [16] is that, it does not handle the redundancy problem [17]. To overcome this problem the (LASSO) [35] has been used.

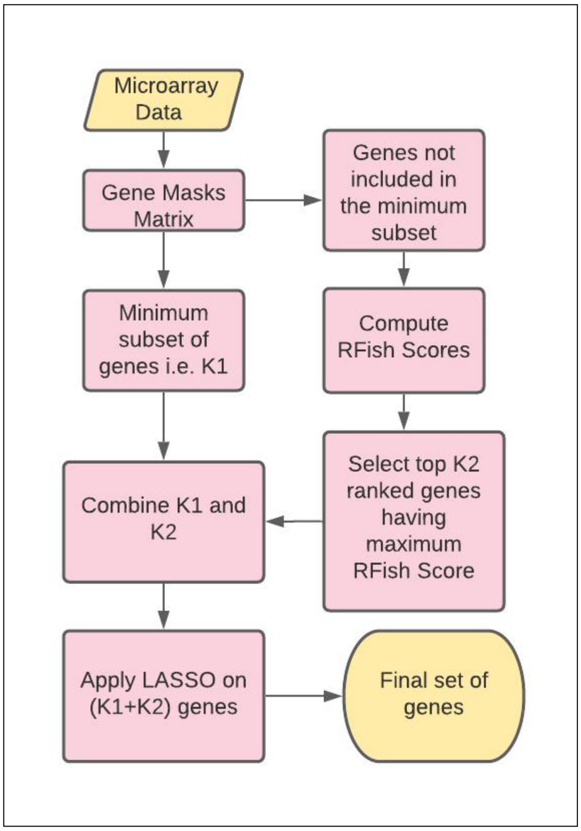
Microarray Data
Gene Masks Matrix
Genes not included in the minimum subset
Minimum subset of genes i.e. K1
Compute RFish Scores
Select top K2 ranked genes having maximum RFish Score
Combine K1 and K2
Apply LASSO on (K1+K2) genes
Final set of genes

FIGURE 2. **Flowchart of the proposed method.**

LASSO shrinks the coefficients of the uninformative genes to zero and retains those genes in the model that are informative for the classification purpose. This technique takes into account the following expression for the estimation of the coefficients of genes.

$$\max_{\beta_1,\beta_0} \sum_{i=1}^{N} [y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0 + \beta^T x_i})] - \lambda \sum_{j=1}^{P} |\beta_j|, \tag{4}$$

where β_j is the regression coefficient of j^{th} gene, λ is the amount of penalty imposed on the genes coefficient and y_i represent the target class in the binary class problems. For further details on the LASSO technique, see [35].

The proposed method (RFish) takes into account the following steps in selecting the discriminative set of genes in high dimensional gene expression datasets.

1) In the first phase the proposed method identifies the minimum subset of genes K_1 in the training phase with the help of the greedy search approach given in [12] by avoiding the effect of outliers. These are the genes that classify the tissue samples to their true classes without any ambiguity. The details of greedy approach are given in [12].
2) Those genes which are not selected in the minimum subset of genes in the first phase are then subjected to

TABLE 1. **Datasets description showing number of samples, number of genes, class wise distribution of samples in the data.**

Dataset	Samples	Genes	Class sizes	Source
Leukeamia	68	7029	47/25	https://cran.r-project.org/web/packages/propOverlap/index.html
Colon	62	2000	40/22	https://www.openml.org/d/1432
Srbct	54	2308	29/25	https://file.biolab.si/biolab/supp/bi-cancer/projections/info/SRBCT.html
GSE4045	37	22215	29/8	https://www.ncbi.nlm.nih.gov/gds/?term=GSE4045
Breastcancer	78	4948	33/43	http://www.bioconductor.org/

the calculation of the proposed robust fisher (RFish) score. Genes with the maximum (RFish) scores are the ones that have maximum discriminative ability. By this way the genes are arranged in decreasing order according to the (RFish) scores and the top K_2 genes are selected.

3) In the third phase of the proposed strategy the K_1 and K_2 genes in Steps (1) and (2) are combined.
4) To remove the redundancy in the genes selected in Step (3), LASSO is applied, which shrinks the coefficients of non-informative genes to zero while retaining those which are informative. Genes are then arranged in decreasing order of their coefficients.
5) Finally for the model construction, the required top number of genes are selected from the set of genes in Step (4).

The general work flow and the pseudo-code of the proposed method (RFish), are given in Figure 2 and Algorithm 1, respectively.

Algorithm 1 Algorithm of the Proposed Method (RFish) for Gene Selection

```
1:  Inputs: Let input feature space is $X^{N \times P}$ and response
    vector is $Y$, and let the number genes to be selected is
    $(r)$.
2:  Output: The ordered discriminative genes set $T$.
3:  for $j \rightarrow P$ do
4:      Using [12], to compute the gene mask for each gene
        i.e. $m_{ij}$.
5:  end for
6:  let $M \in \Re^{N \times P}$ be the gene mask matrix $M = [m_{ij}]$, where
    its $i^{th}$ value for $j^{th}$ gene is either 0 or 1.
7:  Let $M_{..}(\mathbb{H})$ be the aggregate or total mask of genes.
8:  Using the greedy search approach given [12] for the
    selection of the minimum subset of genes from $M$ and
    $M_{..}(\mathbb{H})$ and denote it by $K_1$.
9:  Exclude the minimum subset of genes $K_1$ from the whole
    set of genes $H$ to get $H'$ i.e. $H' = H - K_1$.
10: for $j \rightarrow H'$ do
11:     Apply the proposed method (RFish).
12: end for
13: Select the top ranked $K_2$ genes that have maximum RFish
    score.
14: Combine the genes obtained in step (8) and (13) i.e. $K_1 +$
    $K_2$.
15: Apply LASSO on selected genes in step (14).
16: Order the genes according to the model's coefficients in
    step (15).
17: Select the $|T| = r$ top ranked genes in step (16) for model
    construction.
```

The intuition behind the efficient performance of the proposed method (RFish) is that, the Fisher score given in [16] is based on the ordinary measures of location and dispersion i.e. mean and standard deviation. Since the mean and standard deviation are sensitive to extreme values or outliers, therefore large value of Fisher Score in [16] does not necessarily imply that the gene has higher discriminative ability or the lower score of any gene indicates the poor discriminative ability. Furthermore, the Fisher method is not capable of handling the similarity or redundancy among the genes. Keeping in view these weaknesses, a robust version of Fisher Score is purposed. The proposed method is based on the robust measures of location and dispersion i.e. median and Rousseeuw and Croux statistic. The advantage of using these robust measures is that the value of RFish will remain stable in the presence of outliers. It shows the actual discriminative ability of the genes avoiding the effect of outliers. Moreover, to remove the redundancy in the selected genes, (LASSO) has been used, which shrinks the coefficients of uninformative and redundant genes to zero and retains the informative ones.

IV. EXPERIMENTS AND RESULTS

This section provides a complete details of the experiments conducted in order to derive and compare the results of the proposed method with the other different gene selection methods on various benchmark datasets. A method with smallest classification error rate is considered as the best method. Furthermore, research in [69] has used various gene selection methods, which are given in [55], and it has been shown that a classifier's accuracy is significantly effected by different gene selection methods. Also, this approach has been widely used by different other studies as given in [12], [16], [17], [45]. A brief description of the datasets used in this study is given below.

A. MICROARRAY GENE EXPRESSION DATASETS

Five publicly available gene expression datasets have been considered in this study. All the datasets are taken from various open sources. The detailed description of these datasets are given in Table 1. The small, round blue cell tumors (SRBCTs) of childhood, which include NeuroBlastoma(NB), RhabdoMyoSarcoma (RMS), Non-Hodgkin Lymphoma (NHL) and the EWing family of tumors (EWS) are so named because of their similar appearance on routine histology.The dataset consists of 83 observations, 29, 25, 11 and 18 observations of NB, RMS, NHL and EWS respectively described by 2308 genes. Since this thesis considers only binary classification problems, the two classes with the topmost number of observations, i.e. NB and RMS, are only considered for the analysis. Also, breastcancer data includes 4948 genes measured in 78 patients: 34 with Distant Metastases (DM); 44 without distant Metas-tases (NODM). It is available in the [Bioconductor] repository, `http://{\penalty-\@M}www.bioconductor.org/` from the R package 'cancerdata'.

B. EXPERIMENTAL SETUP

The experimental setup of the manuscript is given in this section. Five gene expression datasets have been considered

TABLE 2. Classification error rates produced by different methods on various classifiers for Colon dataset.

Genes	RFish	Fish	POS	GClust	Wilc	mRmR	sigF
				RF			
5	**0.146**	0.179	0.386	0.284	0.334	0.196	0.197
10	0.191	**0.176**	0.392	0.281	0.293	0.211	0.205
15	**0.173**	0.177	0.389	0.247	0.303	0.211	0.186
20	**0.139**	0.178	0.393	0.264	0.293	0.203	0.185
25	**0.152**	0.177	0.378	0.247	0.290	0.199	0.191
30	0.155	**0.144**	0.356	0.246	0.286	0.166	0.197
				kNN			
5	0.186	**0.159**	0.414	0.300	0.209	0.211	0.218
10	0.205	0.174	0.425	0.290	0.253	0.316	**0.161**
15	0.196	**0.184**	0.438	0.292	0.287	0.316	0.200
20	**0.122**	0.184	0.427	0.342	0.366	0.316	0.199
25	**0.169**	0.179	0.413	0.382	0.469	0.316	0.214
30	0.166	**0.151**	0.405	0.324	0.494	0.221	0.210
				SVM			
5	**0.148**	0.167	0.416	0.288	0.364	0.207	0.193
10	0.159	0.163	0.439	0.295	0.346	0.232	**0.119**
15	**0.124**	0.186	0.467	0.277	0.338	0.245	0.154
20	0.126	0.191	0.439	0.248	0.329	0.158	**0.118**
25	0.140	0.151	0.442	0.263	0.336	0.211	**0.100**
30	0.144	0.124	0.408	0.224	0.331	0.160	**0.109**

TABLE 3. Classification error rates produced by different methods on various classifiers for Srbct dataset.

Genes	RFish	Fish	POS	GClust	Wilc	mRmR	sigF
				RF			
5	0.039	0.044	0.048	0.096	**0.021**	0.390	0.040
10	0.032	0.026	0.018	0.027	**0.013**	0.086	0.035
15	0.025	0.026	0.004	0.016	0.013	0.165	**0.001**
20	0.027	0.032	**0.009**	0.016	**0.009**	0.081	0.010
25	0.028	0.029	**0.009**	0.017	**0.009**	0.067	0.011
30	0.022	0.027	0.006	0.023	**0.005**	0.075	0.007
				kNN			
5	0.031	0.035	0.078	0.100	0.074	0.078	**0.000**
10	0.027	0.020	0.039	0.055	0.071	0.069	**0.000**
15	0.027	0.030	0.039	0.075	0.074	0.071	**0.000**
20	0.014	0.034	0.036	0.053	0.066	0.071	**0.000**
25	0.031	0.028	0.038	0.060	0.066	0.074	**0.000**
30	0.032	0.032	0.034	0.047	0.064	0.065	**0.000**
				SVM			
5	0.025	0.042	0.086	0.035	0.328	0.412	**0.007**
10	0.028	0.034	0.016	0.029	0.204	0.143	**0.006**
15	0.011	0.020	0.004	0.015	0.188	0.182	**0.002**
20	0.004	0.034	0.011	0.020	0.144	0.130	**0.002**
25	0.011	0.029	0.011	0.030	0.134	0.098	**0.000**
30	0.002	0.014	0.009	0.018	0.131	0.129	**0.000**

for the analysis. Each dataset is divided into (70%) training and (30%) testing parts. Seventy percent (70%) of the observations are taken randomly without replacement form each dataset as training parts, while the remaining (30%) of each dataset are taken as testing parts. A split sample analysis of 500 runs is used for each combination of gene selection methods and the corresponding classifier. The classifier which are considered for the analysis are k-Nearest Neighbours (k-NN), Random Forest (RF) and Support Vector Machine (SVM). Furthermore, R library `randomForest` [70] is used for the implementation of random forest with default parameters, `ntree` $= 500$, `mtry` $= \sqrt{p}$ and `nodesize` $= 1$. R package `kernlab` [71] is used for the implementation of Support Vector Machine (SVM) classifier with default parameters. Similarly, for k-Nearest neighbour classifier, R package `caret` [72] is used with default parameter value of $k = 5$.

The training parts of each dataset are used for the selection of various sets of informative genes that are 5, 10, 15 20, 25 and 30 by various feature selection methods to train the classifiers. Performance metric i.e. classification error rate is used to check the classifiers' performance based on the selected set of genes.

C. RESULTS AND DISCUSSION

This section provides the results of the proposed method (RFish) and six other well known gene selection methods included in this study. All the methods are validated on five publicly available datasets, and these are "Leukemia", "Colon", "Srbct", "GSE4045" and "Breastcancer" by using three different classifiers namely, Random Forest (RF), k-Nearest Neighbours (k-NN) and Support Vector Machine (SVM). The results are given in Tables 2, 3, 4, 5 and 6. The details of all the other gene selection methods used in this paper can be found in [12], [13], [18]–[20] and [16]. Table 2 shows the results of the proposed method and the other methods for "Colon" dataset on the classifiers. It is evident from Table 2 that the proposed method (RFish) performs better than all the other methods on random forest classifier except for the set of genes 5, 15,20 and 25, where the method "Fish" performs better for the subset of genes 10 and 30. On k-NN classifier the method "Fish" wins over all the other methods for subset of genes 5 and 15. For the subset of genes 20 and 25 the proposed method is performing better. Similarly, on SVM classifier the proposed method performs better than all the other methods, except for the number of genes 10, 20, 25 and 30 where the method "sigF" is producing minimum error rate among all the other methods. Similarly, Table 3 shows the results of the dataset "Srbct" for the different gene

TABLE 4. Classification error rates produced by different methods on various classifiers for GSE4045 dataset.

Genes	RFish	Fish	POS	GClust	Wilc	mRmR	sigF
			RF				
5	**0.067**	0.080	0.108	0.291	0.170	0.311	0.077
10	0.077	0.081	0.066	0.240	0.165	0.216	0.092
15	0.065	0.088	**0.062**	0.236	0.190	0.211	0.098
20	**0.058**	0.082	0.069	0.218	0.127	0.201	0.062
25	0.072	0.078	**0.040**	0.231	0.172	0.175	0.088
30	0.075	0.076	**0.053**	0.227	0.162	0.144	0.053
			kNN				
5	**0.006**	0.053	0.091	0.265	0.122	0.346	0.083
10	**0.013**	0.074	0.121	0.275	0.220	0.315	0.064
15	**0.005**	0.081	0.099	0.236	0.255	0.308	0.031
20	**0.000**	0.086	0.154	0.271	0.210	0.215	0.036
25	**0.015**	0.091	0.122	0.281	0.235	0.189	0.072
30	**0.000**	0.085	0.143	0.265	0.223	0.128	0.061
			SVM				
5	**0.038**	0.093	0.162	0.429	**0.155**	0.247	0.088
10	**0.033**	0.092	0.053	0.251	0.135	0.213	0.086
15	0.031	0.089	**0.026**	0.245	0.153	0.206	0.073
20	**0.018**	0.089	0.040	0.256	0.125	0.205	0.054
25	**0.018**	0.088	**0.018**	0.264	0.133	0.196	0.043
30	**0.005**	0.088	0.012	0.260	0.142	0.152	0.032

TABLE 5. Classification error rates produced by different methods on various classifiers for Breastcancer dataset.

Genes	RFish	Fish	POS	GClust	Wilc	mRmR	sigF
			RF				
5	**0.213**	0.259	0.296	0.261	0.390	0.455	0.490
10	**0.169**	0.252	0.308	0.261	0.360	0.462	0.514
15	**0.157**	0.288	0.323	0.202	0.337	0.414	0.519
20	**0.167**	0.266	0.290	0.199	0.377	0.468	0.481
25	**0.183**	0.287	0.300	0.223	0.366	0.473	0.495
30	**0.215**	0.270	0.268	0.242	0.350	0.454	0.411
			kNN				
5	**0.000**	**0.000**	0.314	0.313	0.405	0.402	0.448
10	**0.000**	**0.000**	0.276	0.297	0.390	0.396	0.501
15	**0.003**	**0.003**	0.297	0.241	0.391	0.401	0.514
20	**0.001**	**0.001**	0.279	0.257	0.408	0.395	0.473
25	**0.015**	0.017	0.256	0.271	0.404	0.393	0.462
30	**0.026**	0.033	0.261	0.258	0.387	0.394	0.436
			SVM				
5	0.241	**0.234**	0.310	0.512	0.384	0.367	0.522
10	**0.156**	0.218	0.272	0.522	0.351	0.456	0.484
15	**0.167**	0.255	0.262	0.515	0.350	0.427	0.462
20	**0.185**	0.246	0.225	0.542	0.386	0.474	0.409
25	**0.186**	0.279	0.250	0.523	0.377	0.427	0.406
30	**0.230**	0.260	0.249	0.455	0.351	0.457	0.418

TABLE 6. Classification error rates produced by different methods on various classifiers for Leukemia dataset.

Genes	RFish	Fish	POS	GClust	Wilc	mRmR	sigF
			RF				
5	0.011	**0.002**	0.003	0.040	0.050	0.241	0.171
10	**0.000**	0.006	0.002	0.029	0.021	0.215	0.173
15	**0.000**	0.012	0.006	0.025	0.040	0.223	0.156
20	**0.000**	0.010	0.006	0.025	0.149	0.216	0.166
25	**0.000**	0.012	0.006	0.031	0.007	0.213	0.113
30	**0.000**	0.008	0.002	0.030	0.006	0.192	0.110
			kNN				
5	**0.000**	0.001	0.079	0.089	0.130	0.286	0.224
10	**0.000**	0.006	0.093	0.077	0.136	0.219	0.237
15	**0.000**	0.011	0.093	0.069	0.277	0.201	0.219
20	**0.000**	0.013	0.107	0.222	0.130	0.243	0.232
25	**0.004**	0.021	0.089	0.113	0.302	0.313	0.172
30	0.022	**0.015**	0.070	0.109	0.124	0.193	0.134
			SVM				
5	**0.031**	0.008	0.067	0.076	0.070	0.257	0.166
10	0.019	**0.018**	0.075	0.133	0.046	0.241	0.159
15	**0.014**	0.022	0.070	0.137	0.061	0.219	0.166
20	**0.017**	0.022	0.110	0.138	0.150	0.211	0.187
25	**0.005**	0.017	0.114	0.126	0.028	0.201	0.136
30	**0.012**	0.017	0.073	0.086	0.028	0.191	0.129

selection methods. It is clear from the Table 3 that the method "sigF" outperforms all the other methods on k-NN and SVM classifiers. In case of random forest classifier the method "Wilc" is winning in majority of the cases. The results of the dataset "GSE4045" are displayed in Table 4. The proposed method "RFish" is outperforming all the other methods on k-NN and SVM classifiers, while on random forest classifier, the method "POS" is producing minimum classification error rates than all the other methods except for the set of genes 5 and 20 where the proposed method is winning over all the other gene selection methods. The results on datasets "Breastcancer" on various gene selection methods are given in Table 5. The proposed method has outperformed all the other methods on all the classifiers and producing comparable results with the method "Fish" on k-NN classifier. Finally

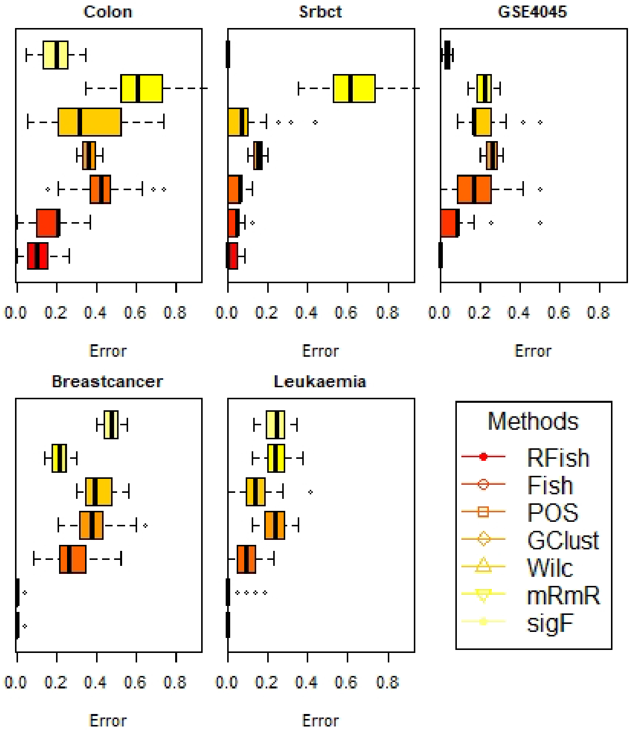

FIGURE 3. Boxplots of classification error rates for 20 number of genes.

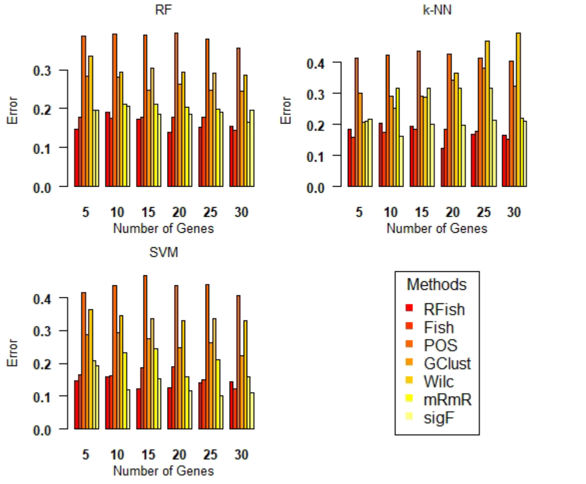

FIGURE 4. Classification error rates of the methods for different number of genes for Colon dataset.

Table 6 shows the results of the proposed method "RFish" and all the other feature/gene selection methods on the dataset "Leukemia". It can be seen in Table 6 that the proposed "RFish" is outperforming all the other methods on random forest and k-NN classifiers. On SVM classifier the method "Fish" is producing minimum error rate for the subset of genes 10, while for the remaining subset of informative genes the proposed method is producing minimum classification error rates.

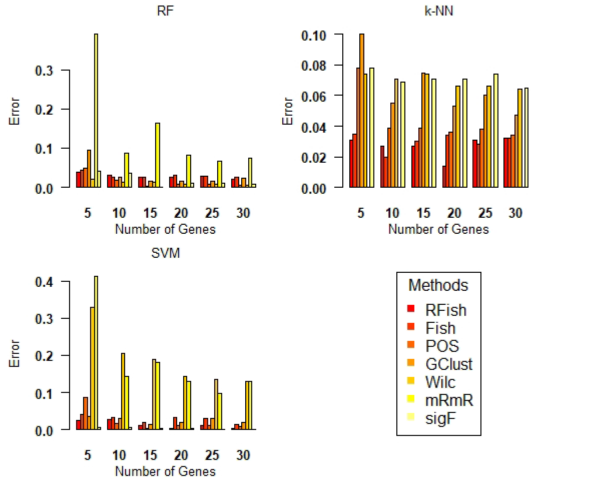

FIGURE 5. Classification error rates of the methods for different number of genes for Srbct dataset.

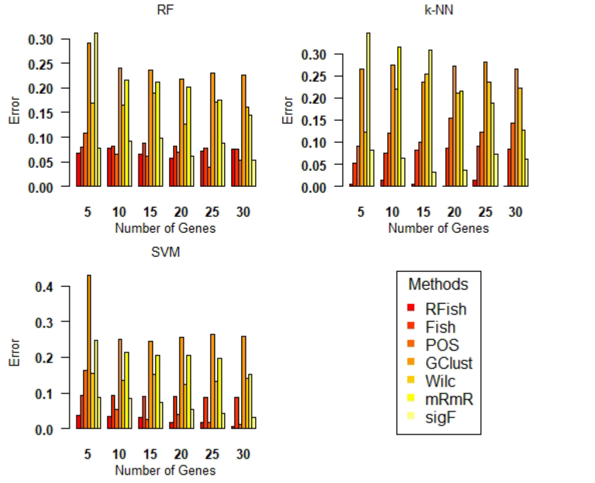

FIGURE 6. Classification error rates of the methods for different number of genes for GSE4045 dataset.

To further assess the efficiency of various gene selection methods, boxplots of the results are constructed. The boxplots of the results of the proposed method (RFish) and all the other gene selection methods considered in this paper for twenty number of genes on k-NN classifier are constructed and they are given in Figure 3. It is evident from the boxplots that the method (Fish) is outperforming all the other methods in the case of "Colon" dataset. For the remaining datasets, the proposed method (RFish) is producing minimum classification errors than all the other gene selection methods except for the dataset "Breastcancer" where it is almost similar to the method (Fish). Moreover for the dataset "Srbct" the method "sigF" is producing minimum error rate. Overall the proposed method (RFish) is winning on 3 out of 5 datasets and producing similar results on one dataset.

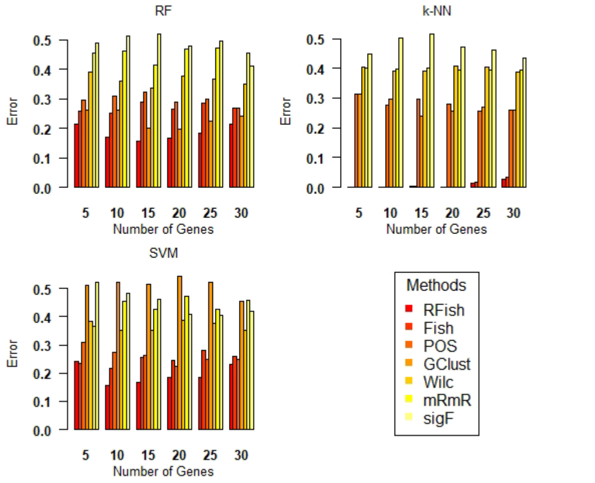

FIGURE 7. Classification error rates of the methods for different number of genes for Breastcancer dataset.

TABLE 7. Classification error rates by random forest and support vector machine classifiers on simulated data having outliers in the important variables.

Genes	POS	RFish	Fish	GClust	Wilc	mRmR	sigF
				RF			
5	0.395	0.381	0.404	0.398	**0.367**	0.379	0.398
10	0.356	**0.335**	0.390	0.401	0.339	0.371	0.387
15	0.338	**0.331**	0.355	0.400	0.334	0.360	0.377
20	0.348	**0.318**	0.388	0.418	0.319	0.352	0.372
25	0.328	**0.313**	0.391	0.414	0.319	0.354	0.360
30	**0.324**	0.326	0.382	0.416	0.328	0.361	0.361
				SVM			
5	0.453	**0.355**	0.381	0.444	0.339	0.428	0.400
10	0.399	**0.325**	0.368	0.451	0.333	0.409	0.392
15	0.369	**0.322**	0.343	0.466	0.327	0.392	0.384
20	0.388	**0.322**	0.372	0.439	0.323	0.390	0.390
25	0.340	0.325	0.379	0.452	**0.293**	0.340	0.381
30	0.371	0.313	0.362	0.431‘	**0.307**	0.356	0.371

Similarly for quick insights into the results of the proposed method (RFish) and all the other methods included in this study, the bar plots are also constructed. These bar plots are given in Figures 4, 5, 6, 7 and 8 respectively. From Figure 4 it is clear that the proposed method "RFish" is producing minimum error rates as compared to all the other methods, in majority of the cases, on random forest classifier. On k-NN classifier the proposed method "RFish" has minimum error rates for the subset of genes 20 and 25. Similarly from Figure 5 it is evident that the method "sigF" is winning over all the other methods in majority of the cases in the case of "Srbct" dataset. Bars in Figure 6 corresponding to the method "POS" indicates that it is performing better than all the other methods in majority of the cases on random forest classifier, while on k-NN and SVM classifiers the proposed method is producing minimum classification error rates. Finally from Figures 7 and 8, it can be observed that the proposed method is outperforming all the other methods on various classifiers for the datasets "Breastcancer" and "Leukemia" respectively.

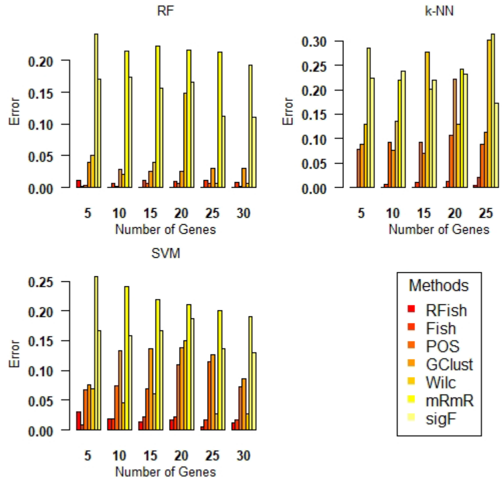

FIGURE 8. Classification error rates of the methods for different number of genes for Leukaemia dataset.

TABLE 8. Classification error rates by random forest and support vector machine classifiers on simulated data having no outliers in the important variables.

Genes	POS	RFish	Fish	GClust	Wilc	mRmR	sigF
				RF			
5	0.359	0.319	0.394	0.371	**0.264**	0.387	0.333
10	0.316	0.327	0.388	0.391	**0.263**	0.365	0.321
15	0.320	0.350	0.384	0.390	**0.223**	0.352	0.326
20	0.302	0.332	0.379	0.319	**0.273**	0.364	0.319
25	0.329	0.296	0.393	0.315	**0.277**	0.325	0.300
30	0.326	**0.279**	0.365	0.316	0.299	0.312	0.291
				SVM			
5	0.400	0.313	0.381	0.343	**0.262**	0.348	0.363
10	0.353	0.314	0.374	0.352	**0.239**	0.351	0.358
15	0.338	0.342	0.369	0.366	**0.238**	0.348	0.369
20	0.323	0.342	0.358	0.339	**0.256**	0.354	0.343
25	0.331	0.291	0.381	0.357	**0.255**	0.324	0.335
30	0.363	**0.252**	0.342	0.334‘	0.286	0.329	0.327

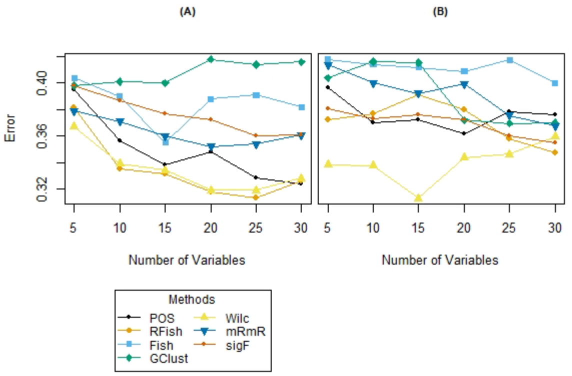

FIGURE 9. Plots of error rates of of the methods for different subsets of genes for the datasets; (A): when there are outliers in the variables, (B): when there are no outliers in the variables.

V. SIMULATION

This section describes two simulation scenarios for the proposed method. The first scenario is designed to mimic a situation where the proposed method is useful, whereas the second scenario shows a data generating environment that might not favour the proposed method. For this purpose two different models are used, one for each scenario. The class probabilities of the Bernoulli response $Y = Bernoulli(p)$ given $N \times P$ dimensional matrix X of identically and independently distributed (iid) N observations from $Normal(0, 1)$ and $Uniform(0, 1)$ distributions, in each scenario are generated by using the following equation.

$$p(y|X) = \frac{exp(b \times X - a)}{1 + exp(b \times X - a)}. \tag{5}$$

The values of a and b are both fixed at 1.5. A vector of coefficients, β is generated from $Uinform(5, -5)$ distribution to fit the linear predictor given as

$$Y = X\beta + \epsilon. \tag{6}$$

Top five i.e. $K = 5$, important variables are identified from the above model based on their coefficients β^s. In order to contaminate the data, outliers are added to these top five variables from $Normal(20, 60)$ distribution. in addition some noisy variables are also added to the data from $Normal(5, 10)$ distribution. By this way a simulated data having $N = 100$ observations and $P = 120$ variables is generated. The second model is also constructed in a similar fashion. The difference between the two models is that, the former contains 20% of the total observations as outliers in the important variables, while the later does not contain outliers. A total of 500 realizations are made in this paper for data simulation. For running the algorithms, the same experimental setup is used as given for benchmark datasets.

The results for all the methods included in this paper are computed on the basis of two classifiers, random forest (RF) and support vector machine (SVM) only. The results of both the scenarios i.e. data having outliers and data that do not contain outliers are given in Tables 7 and 8, respectively. It is clear from the tables that when the data contain outliers or extreme values in the features, the proposed method (RFish) outperforms all the other methods, whereas the performance of the proposed method is poor when there are no outliers in the data. Thus the simulated data analysis supports our argument that the proposed method is performing better in the presence of extreme values in the data. Furthermore, plots of error rates for different subsets of gene are also constructed as given in Figure 9.

VI. CONCLUSION

This paper has discussed the idea of identifying the set of most discriminative genes in high dimensional gene expression datasets by combining the minimum subset of genes obtained via greedy search approach [12] and the top ranked genes obtained by the proposed method (RFish). Moreover, to remove the redundancy in the selected set of genes a well known technique (LASSO) has been used. The proposed method (RFish) takes into account the more robust measures of location and dispersion i.e. the median and Rousseeuw and Croux statistic (Q_n), thereby reducing the effect of extreme values or outliers in gene expression values. Larger value of the proposed method (RFish) for any gene indicates that the gene has more discriminative ability in classifying the tissue samples to their true response classes.

The analysis of the paper, based on two classifiers, revealed that feature selection methods might be adversely affected by the presence of outliers in expression values. Such outliers in the expression values might lead to misleading ranks to genes by gene selection method. The proposed RFish method attempted to solve this problem by using robust statistics while ranking genes.

Moreover one can use other various types of robust measures of location and dispersion available in the literature rather than the median and Rousseeuw and Croux statistic (Q_n). This work can also be extended to the situation where the response variable is of continuous nature. The proposed method can also be extended to the unsupervised machine learning scenario, where one can divide the entire set of genes into various clusters and then apply the proposed method in each cluster to get a set of discriminative genes/features [13], [53]. The final set of discriminative genes in this case will be the combination of top ranked genes in each cluster. Also, the proposed gene selection method can be validated by extending the performance assessment of the selected genes to other classification methods [73], [74].

The proposed method, however, is only designed for binary class problems which limits its applicability. Furthermore, the method might become time consuming in case of ultra high dimensional settings. This problem can be solved by using parallel computing.

REFERENCES

[1] M. Hamraz, N. Gul, M. Raza, D. M. Khan, U. Khalil, S. Zubair, and Z. Khan, "Robust proportional overlapping analysis for feature selection in binary classification within functional genomic experiments," *PeerJ Comput. Sci.*, vol. 7, p. e562, Jun. 2021.

[2] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. ICML*, vol. 97, Nashville, TN, USA, 1997, p. 35.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[4] A. Bommert, T. Welchowski, M. Schmid, and J. Rahnenführer, "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data," *Briefings Bioinf.*, vol. 23, no. 1, Jan. 2022, Art. no. bbab354.

[5] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, no. 1, pp. 23–69, 2003.

[6] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 18, 2005, pp. 1–8.

[7] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection," in *Proc. AAAI*, vol. 2, 2008, pp. 671–676.

[8] L. Song, A. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo, "Supervised feature selection via dependence estimation," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 823–830.

[9] K. K. Ghosh, S. Ahmed, P. K. Singh, Z. W. Geem, and R. Sarkar, "Improved binary sailfish optimizer based on adaptive β-hill climbing for feature selection," *IEEE Access*, vol. 8, pp. 83548–83560, 2020.

[10] I. M. El-Hasnony, S. I. Barakat, M. Elhoseny, and R. R. Mostafa, "Improved feature selection model for big data analytics," *IEEE Access*, vol. 8, pp. 66989–67004, 2020.

[11] H. Seo and D.-H. Cho, "Cancer-related gene signature selection based on boosted regression for multilayer perceptron," *IEEE Access*, vol. 8, pp. 64992–65004, 2020.

[12] O. Mahmoud, A. Harrison, A. Perperoglou, A. Gul, Z. Khan, M. Metodiev, and B. Lausen, "A feature selection method for classification within functional genomics experiments based on the proportional overlapping score," *BMC Bioinform.*, vol. 15, no. 1, p. 274, 2014.

[13] Z. Khan, M. Naeem, U. Khalil, D. M. Khan, S. Aldahmani, and M. Hamraz, "Feature selection for binary classification within functional genomics experiments via interquartile range and clustering," *IEEE Access*, vol. 7, pp. 78159–78169, 2019.

[14] J. Chen, S. Yuan, D. Lv, and Y. Xiang, "A novel self-learning feature selection approach based on feature attributions," *Expert Syst. Appl.*, vol. 183, Nov. 2021, Art. no. 115219.

[15] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[16] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification* Hoboken, NJ, USA: Wiley, 2001.

[17] Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," 2012, *arXiv:1202.3725*.

[18] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, Apr. 2005.

[19] B. Lausen, T. Hothorn, F. Bretz, and M. Schumacher, "Assessment of optimal selected prognostic factors," *Biometrical J., J. Math. Methods Biosci.*, vol. 46, no. 3, pp. 364–374, Jul. 2004.

[20] P. Das, A. Roychowdhury, S. Das, S. Roychoudhury, and S. Tripathy, "SigFeature: Novel significant feature selection method for classification of gene expression data using support vector machine and t statistic," *Frontiers Genet.*, vol. 11, p. 247, Apr. 2020.

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

[22] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[23] C. Liao, S. Li, and Z. Luo, "Gene selection for cancer classification using Wilcoxon rank sum test and support vector machine," in *Proc. Int. Conf. Comput. Intell. Secur.*, Nov. 2006, pp. 368–373.

[24] M. Gaudioso, E. Gorgone, M. Labbé, and A. M. Rodríguez-Chía, "Lagrangian relaxation for SVM feature selection," *Comput. Oper. Res.*, vol. 87, pp. 137–145, Nov. 2017.

[25] Y. Xue, L. Zhang, B. Wang, Z. Zhang, and F. Li, "Nonlinear feature selection using Gaussian kernel SVM-RFE for fault diagnosis," *Int. J. Speech Technol.*, vol. 48, no. 10, pp. 3306–3331, Oct. 2018.

[26] P. Chaudhari and H. Agarwal, "Improving feature selection using elite breeding QPSO on gene data set for cancer classification," in *Intelligent Engineering Informatics*. New York, NY, USA: Springer, 2018, pp. 209–219.

[27] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.

[28] Z. Wang, M. Li, and J. Li, "A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure," *Inf. Sci.*, vol. 307, pp. 73–88, Jun. 2015.

[29] S. Paul and S. Das, "Simultaneous feature selection and weighting—An evolutionary multi-objective optimization approach," *Pattern Recognit. Lett.*, vol. 65, pp. 51–59, Nov. 2015.

[30] Y. Sun, S. Todorovic, and S. Goodison, "Local-learning-based feature selection for high-dimensional data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1610–1626, Sep. 2010.

[31] N. Armanfard, J. P. Reilly, and M. Komeili, "Local feature selection for data classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1217–1227, Jun. 2016.

[32] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[33] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[34] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, "High-dimensional feature selection by feature-wise kernelized lasso," *Neural Comput.*, vol. 26, no. 1, pp. 185–207, 2014.

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

[36] X. Chen, G. Yuan, F. Nie, and J. Z. Huang, "Semi-supervised feature selection via rescaled linear regression," in *Proc. IJCAI*, Aug. 2017, pp. 1525–1531.

[37] M. Dramiński, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, ''Monte Carlo feature selection for supervised classification,'' *Bioinformatics*, vol. 24, no. 1, pp. 110–117, Jan. 2008.

[38] J. De, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains, ''MRMRe: An R package for parallelized mRMR ensemble feature selection,'' *Bioinformatics*, vol. 29, no. 18, pp. 2365–2368, Sep. 2013.

[39] J. Lu, R. T. Kerns, S. D. Peddada, and P. R. Bushel, ''Principal component analysis-based filtering improves detection for affymetrix gene expression arrays,'' *Nucleic Acids Res.*, vol. 39, no. 13, p. e86, Jul. 2011.

[40] W. Talloen, D.-A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass, and H. W. H. Göhlmann, ''I/NI-calls for the exclusion of non-informative genes: A highly effective filtering tool for microarray data,'' *Bioinformatics*, vol. 23, no. 21, pp. 2897–2902, 2007.

[41] A. Ultsch, C. Pallasch, E. Bergmann, and H. Christiansen, ''A comparison of algorithms to find differentially expressed genes in microarray data,'' in *Advances in Data Analysis, Data Handling and Business Intelligence*. New York, NY, USA: Springer, 2009, pp. 685–697.

[42] H.-C. Liu, P.-C. Peng, T.-C. Hsieh, T.-C. Yeh, C.-J. Lin, C.-Y. Chen, J.-Y. Hou, L.-Y. Shih, and D.-C. Liang, ''Comparison of feature selection methods for cross-laboratory microarray analysis,'' *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 3, pp. 593–604, May 2013.

[43] Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif, ''RankGene: Identification of diagnostic genes based on expression data,'' *Bioinformatics*, vol. 19, no. 12, pp. 1578–1579, 2003.

[44] D. Apiletti, E. Baralis, G. Bruno, and A. Fiori, ''The painter's feature selection for gene expression data,'' in *Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2007, pp. 4227–4230.

[45] D. Apiletti, E. Baralis, G. Bruno, and A. Fiori, ''MaskedPainter: Feature selection for microarray data analysis,'' *Intell. Data Anal.*, vol. 16, no. 4, pp. 717–737, Jul. 2012.

[46] D. Nardone, A. Ciaramella, and A. Staiano, ''A sparse-modeling based approach for class specific feature selection,'' *PeerJ Comput. Sci.*, vol. 5, p. e237, Nov. 2019.

[47] A. A. Bidgoli, H. Ebrahimpour-Komleh, and S. Rahnamayan, ''An evolutionary decomposition-based multi-objective feature selection for multi-label classification,'' *PeerJ Comput. Sci.*, vol. 6, p. e261, Mar. 2020.

[48] H. MotieGhader, Y. Masoudi-Sobhanzadeh, S. H. Ashtiani, and A. Masoudi-Nejad, ''MRNA and microRNA selection for breast cancer molecular subtype stratification using meta-heuristic based algorithms,'' *Genomics*, vol. 112, no. 5, pp. 3207–3217, Sep. 2020.

[49] M. Dashtban and M. Balafar, ''Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts,'' *Genomics*, vol. 109, no. 2, pp. 91–107, Mar. 2017.

[50] H. Nematzadeh, R. Enayatifar, M. Mahmud, and E. Akbari, ''Frequency based feature selection method using whale algorithm,'' *Genomics*, vol. 111, no. 6, pp. 1946–1955, 2019.

[51] M. Maghsoudloo, S. A. Jamalkandi, A. Najafi, and A. Masoudi-Nejad, ''An efficient hybrid feature selection method to identify potential biomarkers in common chronic lung inflammatory diseases,'' *Genomics*, vol. 112, no. 5, pp. 3284–3293, Sep. 2020.

[52] M. Rostami, S. Forouzandeh, K. Berahmand, and M. Soltani, ''Integration of multi-objective PSO based feature selection and node centrality for medical datasets,'' *Genomics*, vol. 112, no. 6, pp. 4370–4384, Nov. 2020.

[53] E. Shamsara and J. Shamsara, ''Bioinformatics analysis of the genes involved in the extension of prostate cancer to adjacent lymph nodes by supervised and unsupervised machine learning methods: The role of SPAG1 and PLEKHF2,'' *Genomics*, vol. 112, no. 6, pp. 3871–3882, Nov. 2020.

[54] C. Ao, W. Zhou, L. Gao, B. Dong, and L. Yu, ''Prediction of antioxidant proteins using hybrid feature representation method and random forest,'' *Genomics*, vol. 112, no. 6, pp. 4666–4674, Nov. 2020.

[55] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, ''A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis,'' *Bioinformatics*, vol. 21, no. 5, pp. 631–643, Mar. 2005.

[56] H. K. Rana, M. S. Azam, M. R. Akhtar, J. M. W. Quinn, and M. A. Moni, ''A fast iris recognition system through optimum feature extraction,'' *PeerJ Comput. Sci.*, vol. 5, p. e184, Apr. 2019.

[57] M. A. P. Chamikara, A. Galappaththi, R. D. Yapa, R. D. Nawarathna, S. R. Kodituwakku, J. Gunatilake, A. A. C. A. Jayathilake, and L. Liyanage, ''Fuzzy based binary feature profiling for modus operandi analysis,'' *PeerJ Comput. Sci.*, vol. 2, p. e65, Jun. 2016.

[58] P. S. A. Babu, C. S. R. Annavarapu, and S. Dara, ''Clustering-based hybrid feature selection approach for high dimensional microarray data,'' *Chemometric Intell. Lab. Syst.*, vol. 213, Jun. 2021, Art. no. 104305.

[59] Z. Y. Algamal, M. K. Qasim, M. H. Lee, and H. T. M. Ali, ''Improving grasshopper optimization algorithm for hyperparameters estimation and feature selection in support vector regression,'' *Chemometric Intell. Lab. Syst.*, vol. 208, Jan. 2021, Art. no. 104196.

[60] V. H. A. Ribeiro and G. Reynoso-Meza, ''Feature selection and regularization of interpretable soft sensors using evolutionary multi-objective optimization design procedures,'' *Chemometric Intell. Lab. Syst.*, vol. 212, May 2021, Art. no. 104278.

[61] N. A. Al-Thanoon, Z. Y. Algamal, and O. S. Qasim, ''Feature selection based on a crow search algorithm for big data classification,'' *Chemometric Intell. Lab. Syst.*, vol. 212, May 2021, Art. no. 104288.

[62] J. Wang and S. Zhang, ''PA-PseU: An incremental passive-aggressive based method for identifying RNA pseudouridine sites via Chou's 5-steps rule,'' *Chemometric Intell. Lab. Syst.*, vol. 210, Mar. 2021, Art. no. 104250.

[63] M. Tahir, H. Tayara, M. Hayat, and K. T. Chong, ''KDeepBind: Prediction of RNA-proteins binding sites using convolution neural network and *k*-gram features,'' *Chemometric Intell. Lab. Syst.*, vol. 208, Jan. 2021, Art. no. 104217.

[64] A. Ahmad, S. Akbar, S. Khan, M. Hayat, F. Ali, A. Ahmed, and M. Tahir, ''Deep-AntiFP: Prediction of antifungal peptides using distant multi-informative features incorporating with deep neural networks,'' *Chemometric Intell. Lab. Syst.*, vol. 208, Jan. 2021, Art. no. 104214.

[65] S. Rose, S. Nickolas, and S. Sangeetha, ''A recursive ensemble-based feature selection for multi-output models to discover patterns among the soil nutrients,'' *Chemometric Intell. Lab. Syst.*, vol. 208, Jan. 2021, Art. no. 104221.

[66] S. Dilmi and M. Ladjal, ''A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques,'' *Chemometric Intell. Lab. Syst.*, vol. 214, Jul. 2021, Art. no. 104329.

[67] M. Pérez-Cova, C. Bedia, D. R. Stoll, R. Tauler, and J. Jaumot, ''MSroi: A pre-processing tool for mass spectrometry-based studies,'' *Chemometric Intell. Lab. Syst.*, vol. 215, Aug. 2021, Art. no. 104333.

[68] P. J. Rousseeuw and C. Croux, ''Alternatives to the median absolute deviation,'' *J. Amer. Statist. Assoc.*, vol. 88, no. 424, pp. 1273–1283, Dec. 1993.

[69] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, ''Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,'' *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[70] A. Liaw and M. Wiener, ''Classification and regression by randomforest,'' *R News*, vol. 2, no. 3, pp. 18–22, Nov. 2002.

[71] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, ''Kernlab—An S4 package for kernel methods in R,'' *J. Stat. Softw.*, vol. 11, no. 9, pp. 1–20, 2004.

[72] M. Kuhn, ''Building predictive models in R using the caret package,'' *J. Stat. Softw., Articles*, vol. 28, no. 5, pp. 1–26, 2008. [Online]. Available: https://www.jstatsoft.org/v028/i05, doi: 10.18637/jss.v028.i05.

[73] Z. Khan, A. Gul, A. Perperoglou, M. Miftahuddin, O. Mahmoud, W. Adler, and B. Lausen, ''Ensemble of optimal trees, random forest and random projection ensemble classification,'' *Adv. Data Anal. Classification*, vol. 14, no. 1, pp. 97–116, Mar. 2020.

[74] A. Gul, A. Perperoglou, Z. Khan, O. Mahmoud, M. Miftahuddin, W. Adler, and B. Lausen, ''Ensemble of a subset of kNN classifiers,'' *Adv. Data Anal. Classificat.*, vol. 12, no. 4, pp. 827–840, 2018.