Letters in Applied NanoBioScience
Open-Access Journal (ISSN: 2284-6808)

Article
Volume 15, Issue 1, 2026, 34
https://doi.org/10.33263/LIANBS151.034

rPLS-LDA: A Robust Partial Least Squares Linear Discriminant Analysis for Transcriptomics and Metabolomics Data Analysis

Md. Rabiul Auwul [1,*], Md. Shahjaman [2,*]

[1] Department of Agricultural and Applied Statistics, Faculty of Agricultural Economics and Rural Development, Gazipur Agricultural University, Gazipur 1706, Bangladesh; rabiulauwul@gau.edu.bd;
[2] Department of Statistics, Begum Rokeya University, Rangpur-5400, Bangladesh; shahjaman.stat@brur.ac.bd;
* Correspondence: rabiulauwul@gau.edu.bd (M.R.A.); shahjaman.stat@brur.ac.bd (M.S.);

Received: 10.08.2024; Accepted: 22.02.2026; Published: 30.03.2026

Abstract: Classification is a supervised learning approach for allocating unlabeled dichotomous patient samples into labeled dichotomous patient samples. Conventional classification methods are clinically based and therefore have limited applications, with lower prediction accuracy. Partial Least Squares Linear Discriminant Analysis (PLS-LDA) is one of the most popular classifiers for predicting unlabeled dichotomous patient samples. Nevertheless, this method yields misleading results when an outlying data vector is present under classical maximum likelihood estimators. In this paper, we developed a robust PLS-LDA approach using the minimum beta divergence method. The efficiency of the proposed approach was evaluated through comparisons with existing classification methods using both simulated and real datasets (transcriptomics and metabolomics datasets). Results from 10-fold cross-validation showed that the proposed method significantly outperforms other methods across various performance indices at three outlier rates (10%, 20%, and 30%). Whereas, in the absence of outliers, it keeps almost equal performance with the traditional PLS-LDA method. The protein-protein interaction (PPI), Gene Ontology (GO), and KEGG analyses of the 10 genes identified in prostate cancer data by the proposed procedure demonstrated the significance of the proposed methods. The computational tool has been implemented in an R package, which is publicly available from https://github.com/MdRabiulAuwul/rPLS-LDA.

Keywords: classification; PLS-LDA; minimum beta-divergence estimator; robustness; outlier; transcriptomics; metabolomics.

1. Introduction

The key issue in today's world is big data for its miscellaneous, shapeless, and fast-changing nature. OMICS datasets are the big biological datasets. Both transcriptomics (gene expression datasets, GEDs) and metabolomics datasets are fast-growing and high-dimensional, representing information at two different levels of biological systems in the fields of transcriptomics and metabolomics[1–3]. A single gene expression dataset comprises thousands of features relative to a smaller number of samples, where some features are heterogeneous [2,4,5. Therefore, the complexity of analyzing these datasets is increasing day by day. These characteristics of the OMICS datasets are also known as the curse of dimensionality. Hence,

downstream analyses using these types of datasets often suffer from the curse of dimensionality, a phenomenon that refers to the problem that occurs when adding more features or variables makes the data space so large and sparse that it becomes harder to find meaningful patterns, often leading to increased noise and mistakes in analysis [6]. To solve these problems, supervised machine learning methods have been developed for mining useful information. Classification is also referred to as supervised machine learning (SML) approaches for separating multivariate data into various populations (e.g., normal or cancer) based on training datasets whose class is known in advance [7–9]. The main objective of classification is to estimate the parameters and construct the model using the training dataset in the training phase, and to predict class labels for the test dataset usinghe model in the classification phase [10–12].

Numerous SML approaches have been developed and employed in both GED and metabolomics data [13–19]. The earliest and commonly used method is Fisher's linear discriminant analysis (LDA) [20]. One of the major drawbacks of LDA is that it suffers from the curse of dimensionality and collinearity problems. K-Nearest Neighbor (KNN) [21] searches for the nearest neighbors using the distance metric for allocating the unlabeled samples into the labeled categories. A non-linear predictor has been formulated using a regression-based model, namely logistic regression with dichotomous dependent variables, to fulfill the SML task. Naïve Bayes Classifier [22] is one of the most popular classifiers in the Bayesian platform. Support vector machine (SVM) [23] has been comprehensively employed for analyzing the different OMICS data. There are several dimension reduction methods available in the literature in bioinformatics research [24–27]. Among them, the classical principal component analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and linear discriminant analysis (LDA) are more popular [28]. The partial least squares (PLS) regression has its popularity especially where there are a large number of features than samples, as in OMICS data. It has also overcome the problem of multicollinearity by transforming the input variables into latent variables. The most popular variant of PLS-regression is Partial Least Squares Discriminant Analysis (PLS-LDA) [29], and it is used when the response variable is dichotomous. PLS-LDA comprises PLS-regression and LDA. In PLS-LDA, firstly, the PLS algorithm is used to solve the dimensionality and the collinearity problems [30,31. At the same time, it exhibits excellent performance for solving the curse of dimensionality problem [32]. Secondly, the LDA algorithm is used to classify the test dataset based on the training dataset. Nevertheless, most of the SML algorithms discussed earlier, including PLS-LDA, suffer from the problem of outliers and produce misleading results in their presence. There are various reasons for contamination of GE and metabolomics datasets by outliers as they go through different data-generating steps.

Therefore, despite the powerful capacity of PLS-LDA to overcome the problems of high dimensionality and multicollinearity in OMICs data, it fails to achieve the goal in the presence of outlying data vectors [33,34]. To address this, the present study proposed a robust PLS-LDA classifier (rPLS-LDA) based on the minimum β-divergence approach [35]. The proposed method produces a weight function, the beta-weight function, for detecting outlying data vectors in the robust PLS-LDA approach. To investigate and compare the performance of the proposed method we consider six popular classifiers namely, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) [20], K-Nearest Neighbor (KNN) [21], Naïve

Bayes (NB) [22], Logistic Regression (LR) [36] and Partial Least Square Linear Discriminant Analysis (PLS-LDA) [29]. We employed these methods in both simulated and real datasets.

The downstream analyses for the identified bioimarkers in transcriptomics and metabolomics datasets may play a crucial role in discovering therapeutic targets [37–42]. In the next section, we describe the robustification method of the PLS-LDA classifier. In Section 3, the performance of the proposed method is demonstrated through simulation studies and real-data analysis, compared with some existing methods. Finally, we end this paper with a conclusion.

2. Materials and Methods

Let πc be the cth p-variate normal population with density function $f_c(x_c) = N(x_c|\mu_c, \Lambda_c)$, where μ_c is the mean vector and Λ_c is the covariance matrix for the cth population (c=1,2,…, K). Assume that we have a training sample of vectors $\{x_{kc} = (x_{1kc}, x_{2kc}, \dots\dots, x_{Pkc})^T ; k = 1, 2, \dots\dots, N_c\}$ of size N_c which is obtained from $N(x_c|\mu_c, \Lambda_c)$ for c = 1,2,…,K where, x_{gkc} denotes the kth observation of the gth variable in the cth population.

2.1. Classical maximum likelihood estimators.

Suppose that the maximum likelihood estimators (MLEs) $\hat{\mu}_c$ and $\hat{\Lambda}_c$ of μ_c and Λ_c are obtained based on the training datasets are given below:

$$\hat{\mu}_c = \frac{1}{N_c}\sum_{k=1}^{N_c} x_{kc} \quad (1)$$

$$\hat{\Lambda} = \frac{1}{N}\sum_{c=1}^{K} N_c\hat{\Lambda}_c \quad (2)$$

$$\hat{\Lambda}_c = \text{diag}(\hat{\sigma}^2{}_{1c}, \hat{\sigma}^2{}_{2c}, \dots\dots, \hat{\sigma}^2{}_{Pc}) \quad (3)$$

Where, $\hat{\sigma}^2{}_{gc} = \frac{1}{N_c}\sum_{k=1}^{N_c}(x_{gkc} - \hat{\mu}_{gc})^2$, $\hat{\mu}_{gc} = \frac{1}{N_c}\sum_{k=1}^{N_c} x_{gkc}$ and $N = \sum_{c=1}^{K} N_c$; g = 1,2,……..,P.

2.2. Partial least squares linear discriminant analysis (PLS-LDA).

PLS-LDA assumes that the response vector Y contains categorical values. The multiple linear regression (MLR) model approach, based on matrix notation, is used:

$$Y=XB+F \quad (4)$$

Where X is a matrix of dimension (N × P), B is a (P×1) regression coefficient matrix, F is a vector of (N × 1), and Y is a response vector of dimension (N × 1). The least square estimates of this approach are, $B = (X^TX)^{-1}X^TY$. But when the number of genes (variables) is larger than the number of samples (observations), which usually occurs in microarray gene expression or metabolomics data, the matrix of X^TX becomes non-singular. As a result, we suffer from computational complexity or the curse of dimensionality problem. PLS-LDA overcomes this problem by decomposing the data matrices X and Y. The fundamental PLS-LDA equations are as follows:

$$X=TRT+E \quad (5)$$

$$Y=TQT+F \quad (6)$$

Where T is an orthogonal matrix of $(N \times p)$, R and Q are loading matrices of dimension $(N \times p)$ and $(1 \times p)$. Note that here $p < P$. E and F are the error matrices corresponding to the data matrix X and response vector Y with dimensions$(N \times p)$ and $(N \times 1)$, respectively.

Now, let the weights matrix W be $(P \times p)$, and the scores matrix can be written as:

$$T=XW(RTW)^{-1} \quad (7)$$

Now, substitute equation (4) in the PLS-DA model of equation (3), we obtain:

$$Y=XW(RTW)^{-1}QT+F \quad (8)$$

Hence, the regression coefficient of the PLS-DA approach is:

$$\widehat{B}=W(RTW)^{-1}QT \quad (9)$$

Therefore, using the PLS-DA algorithm, an unknown sample of Y can be predicted by:

$$\widehat{Y}=XW(RTW)^{-1}QT \quad (10)$$

However, by the assumption of the PLS-LDA procedure, the data matrix X must be mean-centered [29]. For mean centering, the PLS-LDA algorithm uses the classical maximum likelihood estimators (MLEs) of the mean ($\hat{\mu}_c$) which is calculated from the training datasets. Therefore, the PLS-LDA approach based on the classical mean centering procedure produces misleading results in the presence of outliers [43,44]. To overcome this problem, the PLS-LDA algorithm has been robustified using the minimum β-divergence method, as outlined below:

2.3. Minimum β-divergence estimators.

Let for cth condition (c=1,2,…, K), the true density and the model density are denoted by $p(x_c)$ and $q(x_c|\theta_c)$, respectively, then the β-divergence of two densities can be defined by:

$$D_\beta\big(p(x_c), q(x_c|\theta_c)\big) \int \left[\frac{1}{\beta}\{p^\beta(x_c) - q^\beta(x_c|\theta_c)\}p(x_c) - \frac{1}{\beta+1}\{p^{\beta+1}(x_c) - q^{\beta+1}(x_c|\theta_c)\}\right] dx_c$$

for $\beta > 0$ and $D_\beta\big(p(x_c), f(x_c|\theta_c)\big) \geq 0$. Equality holds if and only if $p(x_c) = f(x_c|\theta_c)$ for all x_c.

The β-divergence reduces to the Kullback-Leibler (K-L) divergence if $\beta \to 0$, that is:

$$\lim_{\beta \downarrow 0} D_\beta\big(p(x_c), q(x_c|\theta_c)\big) = \int p(x_c)\log\frac{p(x_c)}{q(x_k|\theta_c)}dx_c = D_{KL}\big(p(x_c), q(x_c|\theta_c)\big)$$

The minimum β-divergence estimator is obtained by minimizing:

$$\hat{\theta}_c = \underset{c}{\operatorname{agrmin}}\, \mathcal{D}_\beta(p(x_c), q(x_c|\theta'_c))$$

The minimum β-divergence estimators $\hat{\mu}_{c,\beta}$ and $\widehat{\Lambda}_{c,\beta}$ for the mean vector μ_c and the diagonal covariance matrix Λ_c respectively, are obtained iteratively as follows:

$$\hat{\mu}_{c,\beta}^{(r+1)} = \frac{\sum_{k=1}^{N_c} \varphi_\beta(x_{kc}|\hat{\mu}_c^{(r)},\hat{\Lambda}_c^{(r)})x_{kc}}{\sum_{k=1}^{N_c} \varphi_\beta(x_{kc}|\hat{\mu}_c^{(r)},\hat{\Lambda}_c^{(r)})} \quad (11)$$

$$\hat{\Lambda}_{c,\beta}^{(r+1)} = \frac{1}{N}\sum_{c=1}^{2} N_c \, diag(\hat{\sigma}_{1c,\beta}^2, \hat{\sigma}_{2c,\beta}^2, \dots\dots\dots, \hat{\sigma}_{Pc,\beta}^2) \quad (12)$$

$$\text{Where,} \hat{\sigma}_{gc,\beta}^2 = (\beta+1)\frac{\sum_{k=1}^{N_c} \varphi_\beta\left(x_{kc}\middle|\hat{\mu}_c^{(r)},\hat{\Lambda}_c^{(r)}\right)(x_{gkc}-\hat{\mu}_{gc}^{(r)})^2}{\sum_{k=1}^{N_c} \varphi_\beta(x_{kc}|\hat{\mu}_c^{(r)},\hat{\Lambda}_c^{(r)})} \quad (13)$$

And:

$$\varphi_\beta(x_{kc}|\hat{\mu}_c^{(r)},\hat{\Lambda}_c^{(r)}) = \exp\left\{-\frac{\beta}{2}\left(x_{kc}-\hat{\mu}_c^{(r)}\right)^T \hat{\Lambda}_c^{(r)^{-1}}\left(x_{kc}-\hat{\mu}_c^{(r)}\right)\right\} \quad (14)$$

The function given in (14) is called the β-weight function, and it plays an important role in robust parameter estimation. The optimum β value is tuned via grid search combined with k-fold cross-validation to optimize model performance and prevent overfitting. The final selected β value is reported along with the tuning procedure to ensure transparency and reproducibility.

2.4. Outlier detection for PLS-LDA using β-weight function.

The β-weight function ranges from 0 to 1 and assigns higher weights to normal data points while downweighting outliers and unusual observations. Hence, we partitioned the data matrix based on the β-weight function as follows:

$$\varphi_{c,\beta}(x_{kc}|\hat{\mu}_{c,\beta}^{(r)},\hat{\Lambda}_{c,\beta}^{(r)}) = \exp\left\{-\frac{\beta}{2}\left(x_{kc,\beta}-\hat{\mu}_{c,\beta}^{(r)}\right)^T \hat{\Lambda}_{c,\beta}^{(r)^{-1}}\left(x_{kc.\beta}-\hat{\mu}_{c,\beta}^{(r)}\right)\right\};\ \beta > 0$$

The choice of the tuning parameter β plays an important role in the performance of the proposed method; we select this parameter β by cross-validation as discussed in the robust Naïve Bayes classifier [44].

Let $D^0 = \{x \in D_c\}$ and $D^* = \{x \notin D_c\}$ be the sets of usual/goof and unusual/outlying data points, respectively. Then we partition the entire data space D into D^0 and D^* using the partition rule given by:

$$\varphi_{c,\beta}(x_{kc}|\hat{\mu}_{c,\beta}^{(r)},\hat{\Lambda}_{c,\beta}^{(r)}) = \begin{cases} > \delta_c, \text{if } x \in D_c \\ \leq \delta_c, \text{if } x \notin D_c, \text{ or } x \text{ is outlying} \end{cases} \quad (15)$$

One way to compute the cut-off value δ_c using the empirical distribution of β-weight function and by the values of quantile [44] for k = 1,2,…, N_c with probability:

$$\Pr\left\{\varphi_{c,\beta}(x_{kc}|\hat{\mu}_{c,\beta}^{(r)},\hat{\Lambda}_{c,\beta}^{(r)}) \leq \delta_c\right\} \leq \vartheta = 0.01 \quad (16)$$

Then, whether the data vector x is contaminated or not can be defined as follows:

$$\varphi_\beta(x) = \sum_{c=1}^{K} \varphi_{c,\beta}(x_{kc}|\hat{\mu}_{c,\beta}^{(r)},\hat{\Lambda}_{c,\beta}^{(r)}) = \begin{cases} \geq \delta, \text{if } x \text{ is not outlying} \\ < \delta, \text{if } x \text{ is outlying} \end{cases} \quad (17)$$

Where, $\delta = \sum_{c=1}^{K} \delta_c$

However, in this study, the threshold value of δ is chosen directly as follows:

$$\delta = (1-\eta)\min_{x\in\mathfrak{D}} \varphi_\beta(x) + \eta \max_{x\in\mathfrak{D}} \varphi_\beta(x) \quad (18)$$

With heuristically η = 0.10. For the detection of outliers, equation (18) was used to choose the threshold value in the previous works [43,44].

3. Results and Discussion

In this section, the performance of our proposed algorithm (rPLS-LDA) has been checked and compared with the six popular existing algorithms (SVM, LDA, KNN, Naïve Bayes, Logistic Regression, and PLS-LDA). To evaluate the performance of each algorithm, we implemented it on both simulation and real OMICS datasets, and the measurements were performed using the e1071, class, knn, rpart, and plsgenomics packages in R. The R package MASS was used to evaluate the performance of these algorithms. The Comprehensive R Archive Network (CRAN) or Bioconductor are the main sources of these packages. The summary description of the datasets used in this study is presented in Table 1.

Table 1. Description of the datasets used in this study.

Dataset	Numbers of sample	Numbers of features
Small simulation	20	1000
Large simulation	60	1000
Colon cancer transcriptomic	62	6500
Prostate cancer transcriptomic	472	6144
Lung cancer metabolic	82	158

3.1. Simulated data analysis n_1=n_2.

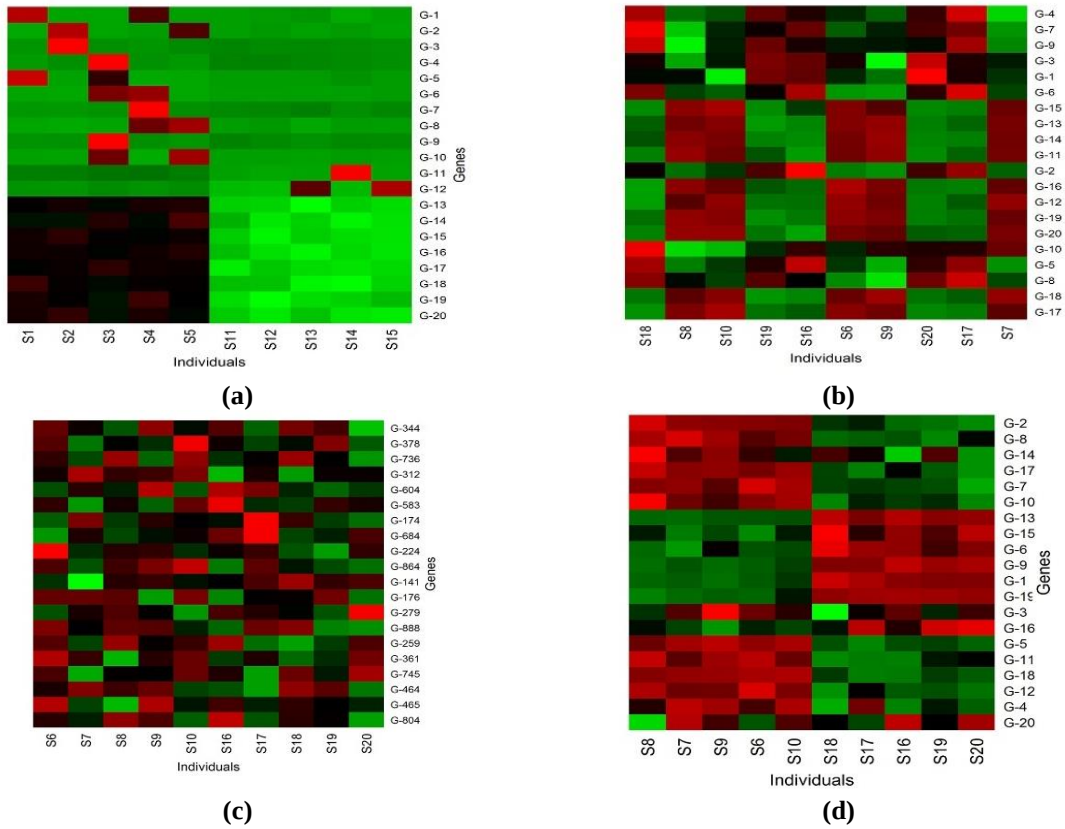
The simulation data that replicate the characteristics of real gene expression data have been generated for two (k=2) groups, in the case of those with and without outliers. Table 2 describes the model for generating simulation data. The dataset was contaminated with Gaussian noise for the randomization. The columns and rows of these simulated data represent the sample groups (normal and cancer) and genes, respectively. We generated two categories of datasets: the first one contains G=1,000 genes with a small sample size, 10 in each group (N_1=N_2=10), and the second one contains G=1,000 genes with a large sample size, 30 in each group (N_1=N_2=30). The gene expression profiles of 1,000 genes with N= (N_1+N_2) samples are represented by both datasets. Among the 1000 genes, we generated 20 (2%) differentially expressed (DE) genes and 980 equally expressed (EE) genes (pattern 3) that were represented in both datasets. These 20 DE genes were divided into P1 =10, up-regulated (pattern 1), and P2=10, down-regulated (pattern 2) DE genes of two groups. We fixed the value of the Gaussian noise parameter, σ2= 0.1, and the parameter μ as 0 to generate both of these datasets.

Table 2. Simulated data-generating model.

Gene group	Sample group	
	Normal (N_1)	Cancer (N_2)
p_1	$N(\mu, \sigma^2)$	$N(+\mu, \sigma^2)$
p_2	$N(-\mu, \sigma^2)$	$N(+\mu, \sigma^2)$
p_3	$N(0, \sigma^2)$	$N(0, \sigma^2)$

To create the classification environment, the training and test datasets were constructed by randomly dividing the samples of these datasets. Firstly, we evaluate the efficiency of the proposed rPLS-LDA algorithm by comparing it with existing classical PLS-LDA approach in

the presence of 10% outliers in the training dataset. Figure 1(a-b) shows the contaminated training and test datasets after selecting the DE gene by t-test. Figure 1(c-d) illustrates the classified test dataset by traditional PLS-LDA and rPLS-LDA, respectively. From this figure, we can conclude that the proposed rPLS-LDA outperformed PLS-LDA, correctly classifying the test dataset samples.



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

Figure 1. Performance comparison between PLS-LDA and rPLS-LDA for simulated dataset (n_1=n_2=10): **(a)** Contaminated training dataset; **(b)** test dataset; **(c)** classified test data by PLS-LDA; **(d)** classified test data by rPLS-LDA.

Next, the proposed rPLS-LDA algorithm compared with the popular classifiers namely, SVM, LDA, K-NN, Naïve Bayes, logistic regression, random forest and PLS-LDA for sample classification as normal or cancer groups have been employed with 100 simulated datasets that are generated with Table 2 for each of small (n=20, n1=10, n2=10) and large (n=60, n1=30, n2=30) sample cases, respectively. To generate outlying datasets, we multiply a constant, c, by the maximum value of the gene expressions within the groups $g^*_{ijk} = u + c^*(x_{ijk}; k = 1,2; i = 1,2,\dots,G; j = 1,2,\dots n_k)$. Here, g_{ijk} symbolizes the i^{th} gene expression of j^{th} samples in k^{th}group, $u \in$ (5, 10) and $c \in$ (2, 4). We considered different outlying percentages of genes (10%, 20%, and 30%) with one or two randomly selected samples.

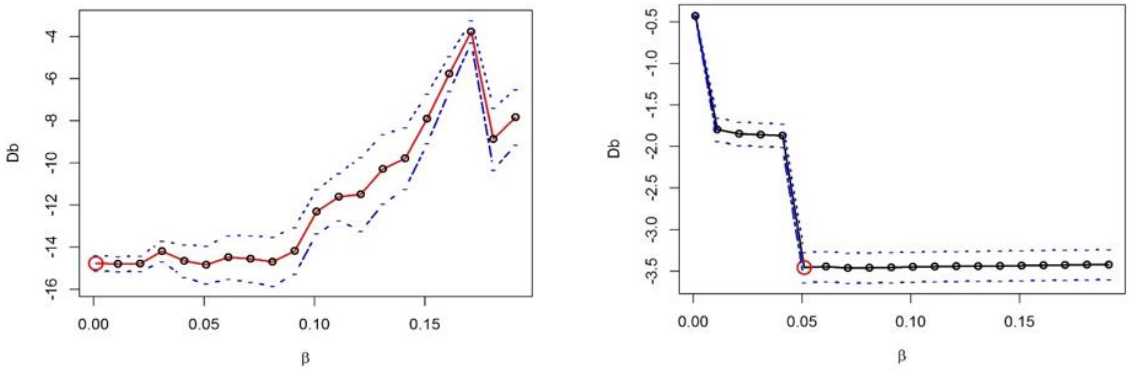
We computed average values of different performance measurements of these seven classifiers, such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and detection rate (DR), based on the estimated top 20 DE genes for each of the 100 datasets. We summarized these performance measures by averaging the 100 values for the small- and large-sample cases in Tables 3 and 4, respectively. We noticed in Table 3 and Table 4 that our proposed rPLS-LDA algorithm gives better results than the SVM, LDA, KNN, Naïve Bayes, Logistic Regression, and PLS-LDA classifiers in all cases (dataset with the absence of outlier, in the presence of 10%, 20%, and 30% outliers). We noticed in

Table 3 that, the average accuracies of SVM, LDA, KNN, Naïve Bayes, LR, PLS-LDA and rPLS-LDA classifiers are 0.636, 0.700, 0.820, 0.602, 0.718, 0.904 and 0.914 respectively in case of absence of outliers; 0.624, 0.622, 0.796, 0.572, 0.684, 0.434 and 0.866 respectively in case of presence of 10% outliers; 0.652, 0.650, 0.756, 0.600, 0.684, 0.444 and 0.878 respectively in case of presence of 20% outliers and 0.594, 0.566, 0.698, 0.544, 0.612, 0.400 and 0.754 respectively in case of presence of 30% outliers. Similarly, from Table 4, we observed that the accuracy of the proposed rPLS-LDA outperformed those of the SVM, LDA, KNN, Naïve Bayes, LR, and PLS-LDA algorithms across all cases involving large outlying datasets.

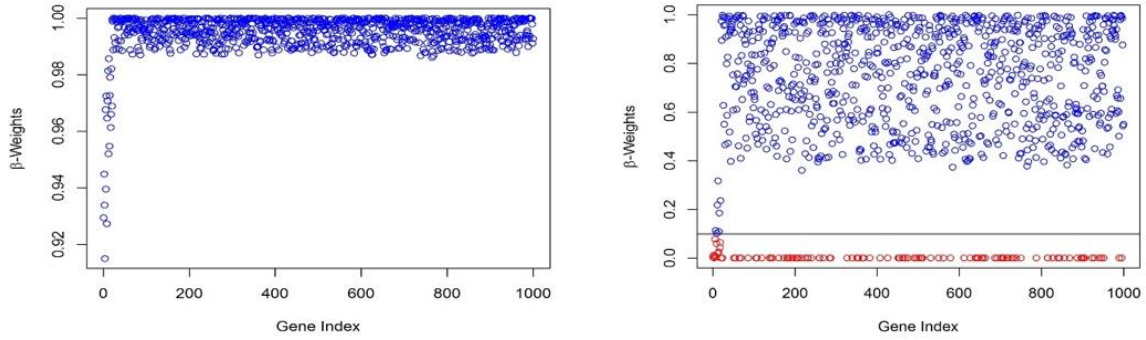
Table 3. Performance results of the seven classifiers with a small sample size (n_1=n_2=10) simulation data.

Performane metrics	In the absence of outliers							In the presence of 10% outliers						
	SVM	**LDA**	**KNN**	**NB**	**LR**	**PLS-LDA**	**rPLS-LDA**	**SVM**	**LDA**	**KNN**	**NB**	**LR**	**PLS-LDA**	**rPLS-LDA**
Accuracy	0.636	0.700	0.820	0.602	0.718	0.904	0.914	0.624	0.622	0.796	0.572	0.684	0.434	0.866
Sensitivity	0.615	0.703	0.836	0.601	0.742	0.887	0.908	0.760	0.681	0.773	0.687	0.705	0.380	0.869
Specificity	0.797	0.765	0.863	0.742	0.694	0.967	0.963	0.625	0.650	0.880	0.605	0.669	0.620	0.927
PPV	0.811	0.781	0.877	0.777	0.739	0.960	0.957	0.706	0.690	0.893	0.675	0.656	0.395	0.932
NPV	0.698	0.717	0.841	0.674	0.713	0.895	0.913	0.803	0.690	0.818	0.709	0.725	0.458	0.892
DR	0.615	0.703	0.836	0.601	0.742	0.887	0.908	0.760	0.681	0.773	0.687	0.705	0.380	0.869
	In the presence of 20% outliers							**In the presence of 30% outliers**						
Accuracy	0.652	0.650	0.756	0.600	0.684	0.444	0.878	0.594	0.566	0.698	0.544	0.612	0.400	0.754
Sensitivity	0.655	0.630	0.702	0.589	0.676	0.120	0.880	0.783	0.681	0.671	0.671	0.636	0.400	0.889
Specificity	0.778	0.758	0.869	0.751	0.701	0.880	0.943	0.582	0.590	0.801	0.611	0.606	0.600	0.756
PPV	0.810	0.757	0.884	0.785	0.692	0.383	0.947	0.660	0.612	0.776	0.653	0.605	0.340	0.799
NPV	0.737	0.671	0.745	0.662	0.689	0.452	0.900	0.750	0.663	0.747	0.686	0.670	0.440	0.896
DR	0.655	0.630	0.702	0.589	0.676	0.120	0.880	0.783	0.681	0.671	0.671	0.636	0.400	0.889

Figure 2 shows the beta selection process via 5-fold cross-validation for our proposed algorithm. Figure 2(a) and Figure 2(b) showed the beta selection for the small sample size (n_1=n_2=10) simulation data in scenarios with the absence of outliers and with the 10% outliers, respectively. We identified that the optimal beta values are close to zero (0.001) in the absence of outliers, and in the presence of 10% outliers, they become 0.051. The analysis revealed that, in the absence of outliers, the proposed classifier tends to the classical MLE classifier as the beta value approaches zero. With these appropriately estimated beta values, the beta weights were calculated for each observation in the training set. Figure 2(c) shows the beta weight graph with β =0.001 in the case of the absence of an outlier. We observed that there were no outliers in the absence of outliers, as all weighted scores were much larger than zero. Figure 2(d) shows the beta weight graph with β =0.051 in the presence of 10% outliers. We observed that it clearly identified outlying samples in the data, as the weighted scores of each outlying sample were close to zero.



(a) **(b)**



(c) **(d)**

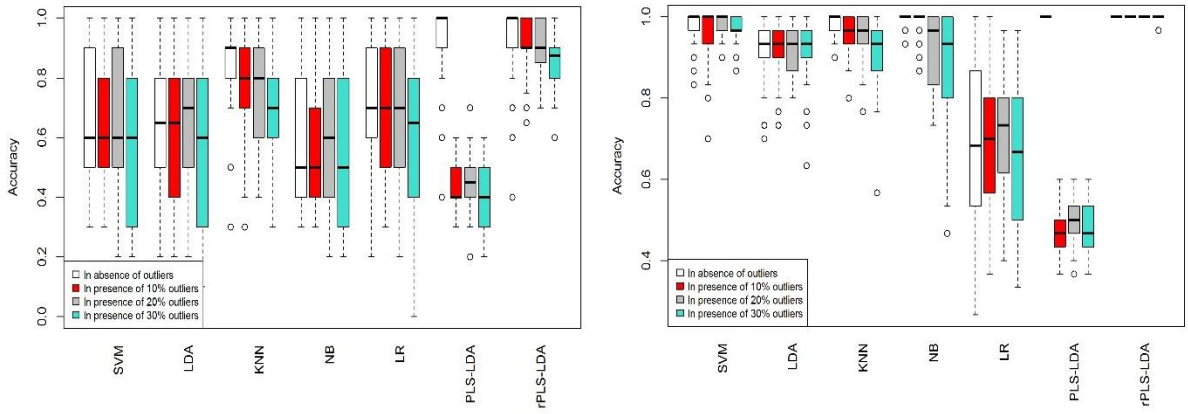
Figure 2. Beta selection for the simulation (n_1=n_2=10) data: **(a)** in the absence of outliers; **(b)** in the presence of 10% outliers. Based on these selected beta values, the weight plots for finding outliers; **(c)** in the absence of outliers;
(d) in presence of 10% outliers.

The boxplot of the test accuracies for the small sample size and large sample size simulation datasets is represented in Figure 3(a) and Figure 3(b), respectively. In Figure 3, the white, red, gray, and turquoise boxes represent the datasets with the absence of outliers, presence of 10%, 20%, and 30% outliers, respectively. These boxplots also describe the same results as drawn from Tables 3 and 4. Hence, we may conclude from this simulation experiment that the proposed rPLS-LDA algorithm outperformed in all datasets.

3.2. Real data analysis.

3.2.1. Colon cancer data analysis.

The colon cancer gene expression dataset consists of 6,500 transcripts from 22 healthy normal and 40 cancer tissue samples, generated using Affymetrix technology. Among 6,500, the gene expression profiles of 2000 genes were filtered out by selecting the maximum minimal intensity across the samples. This dataset can be downloaded from the plsgenomics [45] R package and also from https://microarray.princeton.edu/oncology.



(a) **(b)**

Figure 3. Performance evaluation of the classifiers using boxplot of accuracies of simulation data: **(a)** for small sample size (n_1=n_2=10); **(b)** for large sample size (n_1=n_2=30).

Table 4. Performance results of the seven classifiers with a large sample size (n_1=n_2=30) simulation data.

Performane metrics	In the absence of outliers							In the presence of 10% outliers						
	SVM	**LDA**	**KNN**	**NB**	**LR**	**PLS-LDA**	**rPLS-LDA**	**SVM**	**LDA**	**KNN**	**NB**	**LR**	**PLS-LDA**	**rPLS-LDA**
Accuracy	0.974	0.927	0.983	0.996	0.695	0.999	0.999	0.963	0.926	0.963	0.987	0.689	0.460	0.999
Sensitivity	0.975	0.924	0.985	0.995	0.711	0.999	0.999	0.970	0.925	0.950	0.976	0.667	0.160	0.999

Performane metrics	In the absence of outliers							In the presence of 10% outliers						
	SVM	**LDA**	**KNN**	**NB**	**LR**	**PLS-LDA**	**rPLS-LDA**	**SVM**	**LDA**	**KNN**	**NB**	**LR**	**PLS-LDA**	**rPLS-LDA**
Specificity	0.982	0.932	0.983	0.999	0.684	0.999	0.999	0.967	0.932	0.980	0.999	0.713	0.840	0.999
PPV	0.978	0.936	0.983	0.999	0.699	0.999	0.999	0.966	0.934	0.978	0.999	0.706	0.429	0.999
NPV	0.969	0.924	0.984	0.992	0.700	0.999	0.999	0.964	0.926	0.950	0.975	0.677	0.466	0.999
DR	0.975	0.924	0.985	0.995	0.711	0.999	0.999	0.970	0.925	0.950	0.976	0.667	0.160	0.999
	In the presence of 20% outliers							**In the presence of 30% outliers**						
Accuracy	0.980	0.920	0.952	0.914	0.694	0.502	0.999	0.972	0.919	0.917	0.878	0.661	0.483	0.999
Sensitivity	0.988	0.909	0.923	0.827	0.670	0.222	0.999	0.970	0.909	0.868	0.765	0.664	0.180	0.999
Specificity	0.976	0.932	0.980	0.998	0.713	0.778	0.999	0.979	0.936	0.971	0.984	0.653	0.820	0.999
PPV	0.971	0.928	0.975	0.997	0.689	0.461	0.999	0.978	0.933	0.971	0.990	0.672	0.470	0.999
NPV	0.987	0.918	0.936	0.876	0.708	0.514	0.999	0.967	0.913	0.889	0.852	0.651	0.485	0.999
DR	0.988	0.909	0.923	0.827	0.670	0.222	0.999	0.970	0.909	0.868	0.765	0.664	0.180	0.999

To demonstrate the performance of the SVM, LDA, KNN, Naïve Bayes, LR, PLS-LDA, and rPLS-LDA algorithms for the classification of normal and colon samples, the entire dataset was randomly split into training (70%) and test (30%) sets. Firstly, we selected the top 20 features from the training dataset using the t-test, ranking the adjusted p-values. The adjusted p-values were computed with the Benjamini-Hochberg method [46]. After that, we examined the classifiers' performances based on the top 20 features. The selected 20 features were visualized via a heatmap in Figure S1. We implemented this process 20 times with 5-fold cross-validation, and the performance measures were evaluated both in the absence and in the presence of 10% outliers. We corrupted 10% genes by the outliers as mentioned before. Table 5 summarizes the average values of these measures. From this Table, we noticed that the proposed rPLS-LDA (proposed) algorithm gives better accuracy than the SVM, LDA, KNN, Naïve Bayes, LR, and PLS-LDA algorithms. For example, SVM, LDA, KNN, Naïve Bayes, LR, and PLS-LDA produced accuracies 0.855 (0.814), 0.815 (0.722), 0.871 (0.809), 0.832 (0.792), 0.733 (0.549), 0.874 (0.828), respectively, and our proposed rPLS-LDA produced an accuracy of 0.876 (0.832), which is larger than the other classifiers. The bracketed values in this Table represent the estimated average accuracies by the seven methods in the presence of outliers. Figure 4(a) displays the boxplot of test accuracies using 100 times implementation; it also describes the same results as Table 5. For the determination of our proposed algorithm's performance, the beta weight plots have been identified in Figure S2. Figures S2a and S2b represent that the beta value is 0.01 and 0.06, respectively, in the case of original and corrupted data. Their corresponding weight plots showed the identified outliers in Figure S2c and Figure S2d, respectively. It identified that the beta weight also identified a few outliers in the case of the original data (as we know, the colon data contains outliers).

Table 5. Performance evaluation of the seven classifiers based on colon cancer microarray and prostate cancer datasets.

Performane metrics	Colon cancer microarray data							Prostate cancer data						
	SVM	**LDA**	**KNN**	**NB**	**LR**	**PLS-LDA**	**rPLS-LDA**	**SVM**	**LDA**	**KNN**	**NB**	**LR**	**PLS-LDA**	**rPLS-LDA**
Accuracy	0.855 (0.814)	0.815 (0.722)	0.871 (0.809)	0.832 (0.792)	0.549 (0.733)	0.874 (0.828)	0.876 (0.832)	0.667 (0.658)	0.656 (0.650)	0.656 (0.652)	0.610 (0.620)	0.659 (0.650)	0.671 (0.656)	0.677 (0.671)
Sensitivity	0.787 (0.715)	0.726 (0.680)	0.847 (0.743)	0.789 (0.719)	0.529 (0.589)	0.837 (0.789)	0.837 (0.732)	0.963 (0.942)	0.932 (0.916)	0.911 (0.912)	0.762 (0.786)	0.932 (0.914)	0.961 (0.942)	0.963 (0.996)
Specificity	0.898 (0.880)	0.866 (0.752)	0.887 (0.848)	0.863 (0.829)	0.557 (0.814)	0.896 (0.857)	0.896 (0.891)	0.042 (0.057)	0.072 (0.085)	0.115 (0.100)	0.287 (0.268)	0.081 (0.088)	0.055 (0.051)	0.052 (0.053)
PPV	0.792 (0.786)	0.786 (0.581)	0.786 (0.742)	0.750 (0.733)	0.367 (0.693)	0.797 (0.776)	0.797 (0.808)	0.680 (0.679)	0.680 (0.680)	0.686 (0.683)	0.694 (0.695)	0.682 (0.680)	0.683 (0.678)	0.683 (0.679)
NPV	0.893 (0.837)	0.845 (0.826)	0.922 (0.851)	0.894 (0.844)	0.720 (0.767)	0.919 (0.878)	0.919 (0.846)	0.394 (0.353)	0.380 (0.350)	0.375 (0.351)	0.367 (0.368)	0.363 (0.332)	0.423 (0.354)	0.427 (0.358)
DR	0.787 (0.715)	0.726 (0.680)	0.847 (0.743)	0.789 (0.719)	0.529 (0.589)	0.837 (0.789)	0.837 (0.732)	0.963 (0.942)	0.932 (0.916)	0.911 (0.912)	0.762 (0.786)	0.932 (0.914)	0.961 (0.942)	0.963 (0.996)

3.2.2. Prostate cancer dataset.

This dataset was used in the study [47] and can be downloaded from Gene Expression Omnibus of the National Center for Biotechnology Information website (GEO, http://www.ncbi.nlm.nih.gov/geo/) with GEO series accession number GSE8402. This dataset contains the expression profiles of 6144 genes obtained from 455 prostate cancer tumors. Among 455 tumor samples, there were 103 fusion status-positive samples, and 352 were fusion status-negative samples. We randomly split the entire dataset into two independent sets (training and test) so that the number of tumor samples in each set was equal. For convenience, we first selected the top 1000 features using a t-test on the training dataset.

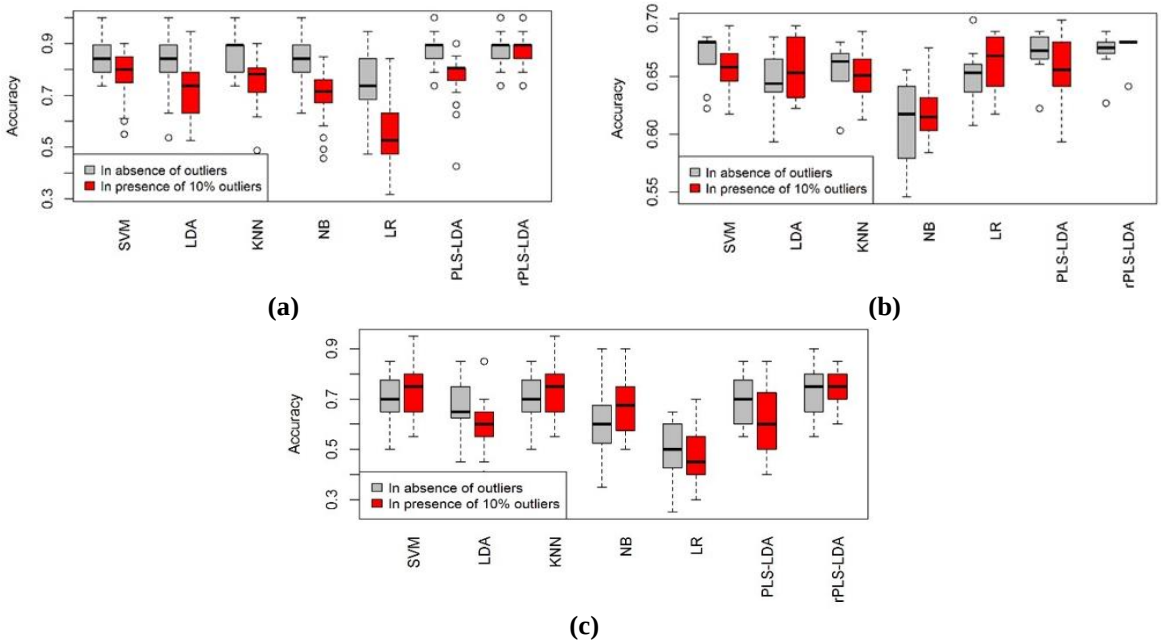


Figure 4. Performance evaluations of the classifiers using boxplots of the accuracies for: **(a)** colon cancer microarray data; **(b)** prostate cancer microarray data; **(c)** lung cancer metabolic data.

We computed average values of different performance indices, such as accuracy, sensitivity, specificity, PPV, NPV, and DR, based on training datasets by the seven methods (SVM, LDA, KNN, Naïve Bayes, LR, PLS-LDA, and rPLS-LDA), both in the absence and in the presence of 10% outliers. We corrupted 10% genes by the outliers as mentioned before. These measures were summarized in Table 4. From this Table, we observe that, in the absence of outliers, the proposed method produced results similar to those of other existing methods used in this study. In contrast, in the presence of outliers, the proposed method achieves higher test accuracy than the other six methods. For example, the proposed method produced accuracies of 0.677 (0.671), which is larger than 0.671 (0.656), 0.667 (0.658), 0.656 (0.652), and 0.656 (0.650) for the four nearest competitors, PLS-LDA, SVM, and KNN. The bracketed values in this Table represent the estimated average accuracies by the seven methods in the presence of outliers. Figure 4(b) shows the boxplot of estimated test accuracies by eight methods using 10-fold cross-validation. From this boxplot, it is clear that the proposed method has less variability than the other methods. This figure also depicts a similar interpretation as in Table 5.

The feature selection procedure of the proposed method identified 10 significant genes (RAB30, DEK, CD55, STK39, CHN2, CDK2AP1, NDRG1, CASP2, CHRM3, and KLF6) in this dataset, compared to the classical procedure. The Venn diagram of the identified DE

(Differentially Expressed) genes by the classical and proposed procedures has been visualized in Figure 5(a). To assess the significance of these ten genes, a PPI network was computed with NetworkAnalyst [48] and visualized in Figure 5(b). It showed that there is a significant interaction among these genes. Among these genes, the Chimerin 2 (CHN2) and Kruppel Like Factor 6 (KLF6) were identified as having a significant association with cancers, including prostate cancer [49–51]. The RAB30 (Member RAS Oncogene Family) gene is associated with Lymphoblastic Leukemia or Lymphoma with Etv6-Runx1 [52]. The DEK (DEK Proto-Oncogene) and Serine/Threonine Kinase 39 (STK39) have been identified as a potential therapeutic target for neuroendocrine prostate cancer [53,54]. The N-Myc Downstream Regulated 1 (NDRG1) gene has been detected as a suppressor of prostate cancer metastasis [55,56]. The Cholinergic Receptor Muscarinic 3 (CHRM3) genes play a significant role in forming cancer, including bladder cancer, identified in a survey among a Chinese Han Population in Kaohsiung City [57,58].

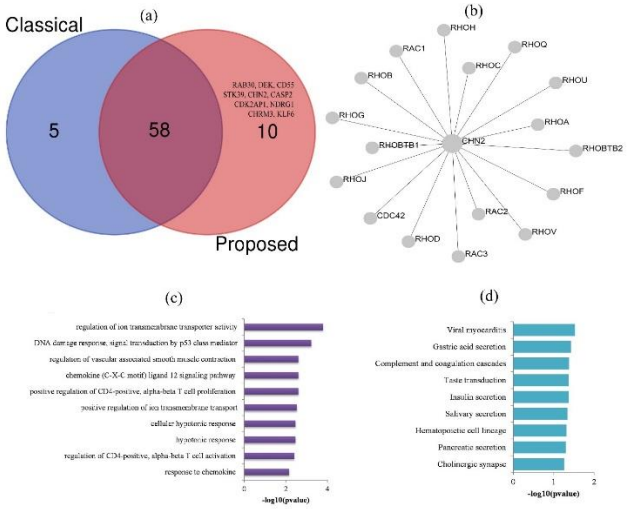


Figure 5. Enrichment analyses of the ten different genes identified by the proposed method: **(a)** the Venn diagram showed the ten different genes; **(b)** the PPI network showed the significant interaction among these genes; **(c)** the MF of GO analyses; **(d)** KEGG pathways interrelated with these genes.

The Molecular Function (MF) of the Gene Ontology (GO) and the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways were significantly analyzed for these ten genes by Enrichr web-tools [59]. The most significant MF found in 'regulation of ion transmembrane transporter activity', 'chemokine', 'regulation of vascular associated smooth muscle contraction', 'positive regulation of ion transmembrane transport', and so on, and these were visualized in Figure 5(c). The KEGG pathways related to these ten genes were significantly associated with 'Viral myocarditis', 'Gastric acid secretion', 'Insulin secretion', 'Taste transduction', 'Pancreatic secretion', 'Cholinergic synapse', and 'Apoptosis' (Figure 5(d)). These MF and pathways were significantly associated with several cancer diseases, including prostate cancer.

3.2.3. Lung cancer metabolic data.

The lung cancer metabolic dataset consists of 158 metabolites and 82 blood samples (41 control and 41 cancer) produced by Oliver Fiehn [60] using the Gas chromatography time of flight mass spectrometry (GC-TOF-MS) approach. These blood samples were collected

through EDTA tubes (stored at −80°C), and approved protocols were used to prepare the samples (serum and plasma).

To demonstrate the performance of the SVM, LDA, KNN, Naïve Bayes, LR, PLS-LDA, and the proposed rPLS-LDA algorithms for the classification of cancer and control samples, we split the dataset into 82 serum samples (training set) and 82 plasma samples (test set). Both the training and test datasets consist of 158 metabolites with 82 samples. Firstly, we randomly selected 20 samples from 82 training samples through bootstrapping. Secondly, we selected the top 50 features from the training dataset using the t-test, ranking the Benjamini-Hochberg [19] adjusted p-values. After that, we examined the classifiers' performances based on the top 20 features. We implemented this process 20 times, a 5-fold cross-validation process. Table 5 summarizes the average values of these performance measures both for the absence and presence of 10% outliers. From this Table, we observed that the proposed rPLS-LDA algorithm achieves higher accuracies than SVM, LDA, KNN, Naïve Bayes, LR, and PLS-LDA. For example, SVM, LDA, KNN, Naïve Bayes, LR, and PLS-LDA produce accuracies 0.738 (0.700), 0.668 (0.603), 0.748 (0.705), 0.683 (0.613), 0.498 (0.475), 0.700 (0.615), respectively, and our proposed rPLS-LDA produces an accuracy of 0.735 (0.725) which is larger than the other classifiers. The bracketed values in this Table represent the estimated average accuracies by the seven methods in the presence of outliers. Figure 4(c) displays the boxplot of test accuracies using 100 times implementation; it also describes the same results as Table 6.

Table 6. Performance evaluation using the estimated values difference measure by seven methods based on the lung cancer metabolic dataset.

Performane metrics	**SVM**	**LDA**	**KNN**	**NB**	**LR**	**PLS-LDA**	**rPLS-LDA**
Accuracy	0.738 (0.700)	0.668 (0.603)	0.748 (0.705)	0.683 (0.613)	0.498 (0.475)	0.700 (0.615)	0.735 (0.725)
Sensitivity	0.904 (0. 763)	0.710 (0.686)	0.904 (0.763)	0. 612 (0.279)	0.575 (0.456)	0.776 (0.563)	0. 897 (0.828)
Specificity	0.627 (0.563)	0. 579 (0.487)	0.627 (0.563)	0.858 (0.704)	0.447 (0.561)	0.572 (0.594)	0.621 (0.528)
PPV	0. 706 (0.692)	0. 689 (0.561)	0.692 (0.706)	0.611 (0.739)	0.504 (0.550)	0. 657 (0.632)	0. 698 (0.687)
NPV	0. 865 (0.754)	0.705 (0.615)	0.754 (0.865)	0.601 (0.656)	0.511 (0.466)	0.786 (0.675)	0. 835 (0.819)
DR	0. 904 (0.763)	0.710 (0.686)	0.763 (0.904)	0.279 (0.612)	0.575 (0.456)	0.776 (0.563)	0. 897 (0.828)

4. Conclusions

Correct classification of microarray gene expression and metabolomics data is a significant issue. These datasets often contain outliers due to multiple steps in the data-generating processes, which may affect the performance of downstream analysis. Among the existing methods of classification, the PLS-LDA is the most popular one. However, it gives misleading results in the presence of outliers. In this study, we robustify the PLS-LDA classifier with a minimum beta-divergence estimator. The performance of the proposed method depends on the value of the betas, and it converges to the classical PLS-LDA as the beta values approach zero. The performance of the proposed methods was compared with six commonly used classifiers, namely, SVM, LDA, KNN, NB, LR, and PLS-LDA, for both simulated and real data analyses, and it improved the performance over other methods. This proposed method will enable handling outliers at large scales in gene expression and metabolomic data. The DE

genes identified by the proposed methods have been found to have a significant association with cancer. The PPI, KEGG, and GO analyses showed the significance of these genes and the proposed method. One limitation of our approach is that we did not compare it with the latest deep learning models or other advanced classification algorithms, and our experiments were conducted on a limited number of real datasets. Moving forward, we plan to extend our work by evaluating the proposed method against these state-of-the-art techniques across a larger and more diverse set of datasets. This will help us better understand its strengths and areas for improvement.

Author Contributions

Conceptualization, M.R.A. and M.S.; methodology, M.R.A.; software, M.R.A.; validation, M.R.A.; formal analysis, M.R.A.; investigation, M.R.A.; resources, M.R.A. and M.S.; data curation, M.R.A.; writing—original draft preparation, M.R.A.; writing—review and editing, M.R.A. and M.S.; visualization, M.R.A.; supervision, M.S. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data presented in this study are openly available at https://github.com/MdRabiulAuwul/rPLS-LDA. Additional data are available upon reasonable request to the corresponding author.

Funding

This research received no external funding.

Acknowledgments

We are very thankful to all faculty members of the Department of Statistics, Begum Rokeya University, and Guangzhou University for providing all facilities for completing this manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Poinsignon, T.; Poulain, P.; Gallopin, M.; Lelandais, G. Working with Omics Data: An Interdisciplinary Challenge at the Crossroads of Biology and Computer Science. In Machine Learning for Brain Disorders, Colliot, O., Ed.; Springer US: New York, NY, **2023**; pp. 313-330, https://doi.org/10.1007/978-1-0716-3195-9_10.

2. He, K.; Wang, Q.; Gao, X.; Tang, T.; Ding, H.; Long, S. Transcriptomic and metabolomic analyses reveal the essential nature of Rab1B in Toxoplasma gondii. *Parasites Vectors* **2023**, *16*, 409, https://doi.org/10.1186/s13071-023-06030-6.
3. Deng, L.; Li, W.; Liu, W.; Liu, Y.; Xie, B.; Groenen, M.A.M.; Madsen, O.; Yang, X.; Tang, Z. Integrative metabolomic and transcriptomic analysis reveals difference in glucose and lipid metabolism in the longissimus muscle of Luchuan and Duroc pigs. *Front. Genet.* **2023**, *14*, 1128033, https://doi.org/10.3389/fgene.2023.1128033.
4. Omae, K.; Komori, O.; Eguchi, S. Quasi-linear score for capturing heterogeneous structure in biomarkers. *BMC Bioinformatics* **2017**, *18*, 308, https://doi.org/10.1186/s12859-017-1721-x.
5. Piccolo, S.R.; Mecham, A.; Golightly, N.P.; Johnson, J.L.; Miller, D.B. The ability to classify patients based on gene-expression data varies by algorithm and performance metric. *PLoS Comput. Biol.* **2022**, *18*, e1009926, https://doi.org/10.1371/journal.pcbi.1009926.
6. Osmak, G.J.; Pisklova, M.V. Transcriptomics and the "Curse of Dimensionality": Monte Carlo Simulations of ML-Models as a Tool for Analyzing Multidimensional Data in Tasks of Searching Markers of Biological Processes. *Mol. Biol.* **2025**, *59*, 135-141, https://doi.org/10.1134/S0026893324700778.
7. Kigo, S.N.; Omondi, E.O.; Omolo, B.O. Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model. *Sci. Rep.* **2023**, *13*, 17315, https://doi.org/10.1038/s41598-023-44326-w.
8. Auwul, M.R.; Zhang, C.; Shahjaman, M. A Robust Procedure for Machine Learning Algorithms Using Gene Expression Data. *Biointerface Res. Appl. Chem.* **2022**, *12*, 2422–2439, https://doi.org/10.33263/BRIAC122.24222439.
9. Auwul, M.R.; Hakim, M.A.; Dhonno, F.T.; Shilpa, N.A.; Sohag, A.; Abedin, M.Z. Using Outlier Modification Rule for Improvement of the Performance of Classification Algorithms in the Case of Financial Data. In Novel Financial Applications of Machine Learning and Deep Learning: Algorithms, Product Modeling, and Applications, Abedin, M.Z., Hajek, P., Eds.; Springer International Publishing: Cham, **2023**; Volume 336, pp. 75-92, https://doi.org/10.1007/978-3-031-18552-6_5.
10. Singh, R.K.; Sivabalakrishnan, M. Feature Selection of Gene Expression Data for Cancer Classification: A Review. *Procedia Comput. Sci.* **2015**, *50*, 52-57, https://doi.org/10.1016/j.procs.2015.04.060.
11. Mangat, P.K.; Saini, D.K.S. Chapter 26 - Relevance of data mining techniques in real life. In System Assurances, Johri, P., Anand, A., Vain, J., Singh, J., Quasim, M., Eds.; Academic Press: **2022**; pp. 477-502, https://doi.org/10.1016/B978-0-323-90240-3.00026-6.
12. Auwul, M.R.; Hakim, M.A.; Dhonno, F.T.; Shilpa, N.A.; Abedin, M.Z. Knowledge Mining from Health Data: Application of Feature Selection Approaches. In Novel Financial Applications of Machine Learning and Deep Learning: Algorithms, Product Modeling, and Applications, Abedin, M.Z., Hajek, P., Eds.; Springer International Publishing: Cham, **2023**; Volume 336, pp. 217-231, https://doi.org/10.1007/978-3-031-18552-6_13.
13. Lyons-Weiler, J.; Patel, S.; Bhattacharya, S. A Classification-Based Machine Learning Approach for the Analysis of Genome-Wide Expression Data. *Genome Res.* **2003**, *13*, 503-512, https:///doi.org/10.1101/gr.104003.
14. Vanitha, C.D.A.; Devaraj, D.; Venkatesulu, M. Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection. *Procedia Comput. Sci.* **2015**, *47*, 13-21, https://doi.org/10.1016/j.procs.2015.03.178.
15. Tarek, S.; Abd Elwahab, R.; Shoman, M. Gene expression based cancer classification. *Egypt. Inform. J.* **2017**, *18*, 151-159, https://doi.org/10.1016/j.eij.2016.12.001.
16. Ghosh, T.; Zhang, W.; Ghosh, D.; Kechris, K. Predictive Modeling for Metabolomics Data. In Computational Methods and Data Analysis for Metabolomics, Li, S., Ed.; Springer US: New York, NY, **2020**; Volume 2104, pp. 313-336, https://doi.org/10.1007/978-1-0716-0239-3_16.
17. Assawamakin, A.; Prueksaaroon, S.; Kulawonganunchai, S.; Shaw, P.J.; Varavithya, V.; Ruangrajitpakorn, T.; Tongsima, S. Biomarker Selection and Classification of "-*Omics*" Data Using a Two-Step Bayes Classification Framework. *BioMed Res. Int.* **2013**, *2013*, 148014, https://doi.org/10.1155/2013/148014.
18. Auwul, M.R.; Rahman, M.R.; Gov, E.; Shahjaman, M.; Moni, M.A. Bioinformatics and machine learning approach identifies potential drug targets and pathways in COVID-19. *Brief. Bioinform.* **2021**, *22*, bbab120, https://doi.org/10.1093/bib/bbab120.
19. Zakaria, L.; Ebeid, H.M.; Dahshan, S.; Tolba, M.F. Analysis of Classification Methods for Gene Expression Data. In Proceedings of The International Conference on Advanced Machine Learning Technologies and

Applications (AMLTA2019), Cairo, Egypt, 28–30 March 2019; Springer: Cham, **2020**; Volume 921, pp. 190-199, https://doi.org/10.1007/978-3-030-14118-9_19.
20. Fisher, R.A. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Ann. Eugen.* **1936**, *7*, 179–188, https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.
21. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185, https://doi.org/10.1080/00031305.1992.10475879.
22. John, G.H.; Langley, P. Estimating continuous distributions in Bayesian classifiers, Proceedings of the Eleventh Conference on Uncertainty in artificial intelligence, Montréal, Qué, Canada, 18 - 20 August 1995; Besnard, P., Hanks, S., Eds.; Morgan Kaufmann Publishers Inc: San Francisco, CA, United States, **1995**; pp. 338–345.
23. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers, Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, USA, 27 - 29 July 1992; Association for Computing Machinery: New York, NY, United States, **1992**; pp. 144–152, https://doi.org/10.1145/130385.130401.
24. Downing, T.; Angelopoulos, N. A primer on correlation-based dimension reduction methods for multi-omics analysis. *J. R. Soc. Interface* **2023**, *20*, 20230344, https://doi.org/10.1098/rsif.2023.0344.
25. Li, J.; Wang, Y. nPCA: a linear dimensionality reduction method using a multilayer perceptron. *Front. Genet.* **2023**, *14*, 1290447, https://doi.org/10.3389/fgene.2023.1290447.
26. Zhou, H.; Kang, Y.; Liu, G.; You, G. An improved LDA dimension reduction algorithm for multivariate time series classification. *J. Appl. Stat.* **2025**, 1-14, https://doi.org/10.1080/02664763.2025.2530580.
27. Sudibyo, U.; Rustad, S.; Andono, P.N.; Fanani, A.Z.; Supriyanto, C. iLDA: A new dimensional reduction method for non-Gaussian and small sample size datasets. *Egypt. Inform. J.* **2024**, *27*, 100533, https://doi.org/10.1016/j.eij.2024.100533.
28. Saiyed, N.; Kantipudi, M.V.V.P. Advancing Breast Cancer Detection: A Comparison of PCA and LDA Methods in Analyzing Ultrasound Imagery. *J. Curr. Sci. Technol.* **2025**, *15*, 125, https://doi.org/10.59796/jcst.V15N3.2025.125.
29. Marigheto, N.A.; Kemsley, E.K.; Defernez, M.; Wilson, R.H. A comparison of mid-infrared and raman spectroscopies for the authentication of edible oils. *J. Am. Oil Chem. Soc.* **1998**, *75*, 987-992, https://doi.org/10.1007/s11746-998-0276-4.
30. Boulesteix, A.L. PLS Dimension Reduction for Classification with Microarray Data. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, 392, https://doi.org/10.2202/1544-6115.1075.
31. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109-130, https://doi.org/10.1016/S0169-7439(01)00155-1.
32. Goodhue, D.L.; Lewis, W.; Thompson, R. Does PLS Have Advantages for Small Sample Size or Non-Normal Data?. *Manag. Inf. Syst. Q.* **2012**, *36*, 981-1001, https://doi.org/10.2307/41703490.
33. Pokhrel, D.R.; Sirisomboon, P.; Khurnpoon, L.; Posom, J.; Saechua, W. Comparing Machine Learning and PLSDA Algorithms for Durian Pulp Classification Using Inline NIR Spectra. *Sensors* **2023**, *23*, 5327, https://doi.org/10.3390/s23115327.
34. Mooijman, P.; Catal, C.; Tekinerdogan, B.; Lommen, A.; Blokland, M. The effects of data balancing approaches: A case study. *Appl. Soft Comput.* **2023**, *132*, 109853, https://doi.org/10.1016/j.asoc.2022.109853.
35. Mollah, M.N.H.; Eguchi, S.; Minami, M. Robust Prewhitening for ICA by Minimizing β-Divergence and Its Application to FastICA. *Neural Process. Lett.* **2007**, *25*, 91-110, https://doi.org/10.1007/s11063-006-9023-8.
36. Kleinbaum, D.G.; Klein, M. *Logistic Regression : A Self-Learning Text*; 2nd ed.; Springer : New York, NY, USA, 2005.
37. Aerqin, Q.; Wang, Z.-T.; Wu, K.-M.; He, X.-Y.; Dong, Q.; Yu, J.-T. Omics-based biomarkers discovery for Alzheimer's disease. *Cell. Mol. Life Sci.* **2022**, *79*, 585, https://doi.org/10.1007/s00018-022-04614-6.
38. Qiu, S.; Cai, Y.; Yao, H.; Lin, C.; Xie, Y.; Tang, S.; Zhang, A. Small molecule metabolites: discovery of biomarkers and therapeutic targets. *Signal Transduct. Target. Ther.* **2023**, *8*, 132, https://doi.org/10.1038/s41392-023-01399-3.
39. Chen, W.; Guo, W.; Li, Y.; Chen, M. Integrative analysis of metabolomics and transcriptomics to uncover biomarkers in sepsis. *Sci. Rep.* **2024**, *14*, 9676, https://doi.org/10.1038/s41598-024-59400-0.
40. Ali Hossain, M.; Asa, T.A.; Rabiul Auwul, M.; Aktaruzzaman, M.; Mahfizur Rahman, M.; Rahman, M.Z.; Moni, M.A. The pathogenetic influence of smoking on SARS-CoV-2 infection: Integrative transcriptome

and regulomics analysis of lung epithelial cells. *Comput. Biol. Med.* **2023**, *159*, 106885, https://doi.org/10.1016/j.compbiomed.2023.106885.
41. Shahjaman, M.; Rezanur Rahman, M.; Rabiul Auwul, M. A network-based systems biology approach for identification of shared Gene signatures between male and female in COVID-19 datasets. *Inform. Med. Unlocked* **2021**, *25*, 100702, https://doi.org/10.1016/j.imu.2021.100702.
42. Auwul, M.R.; Zhang, C.; Rahman, M.R.; Shahjaman, M.; Alyami, S.A.; Moni, M.A. Network-based transcriptomic analysis identifies the genetic effect of COVID-19 to chronic kidney disease patients: A bioinformatics approach. *Saudi J. Biol. Sci.* **2021**, *28*, 5647-5656, https://doi.org/10.1016/j.sjbs.2021.06.015.
43. Shahjaman, M.; Rahman, M.R.; Islam, S.M.S.; Mollah, M.N.H. A Robust Approach for Identification of Cancer Biomarkers and Candidate Drugs. *Medicina* **2019**, *55*, 269, https://doi.org/10.3390/medicina55060269.
44. Ahmed, M.S.; Shahjaman, M.; Rana, M.M.; Mollah, M.N.H. Robustification of Naïve Bayes Classifier and Its Application for Microarray Gene Expression Data Analysis. *BioMed Res. Int.* **2017**, *2017*, 3020627, https://doi.org/10.1155/2017/3020627.
45. Durif, G. plsgenomics: PLS Analyses for Genomics. R package version 1.5-3, **2005**, https://cran.r-project.org/package=plsgenomics.
46. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289-300, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
47. Setlur, S.R.; Mertz, K.D.; Hoshida, Y.; Demichelis, F.; Lupien, M.; Perner, S.; Sboner, A.; Pawitan, Y.; Andrén, O.; Johnson, L.A.; Tang, J.; Adami, H.-O.; Calza, S.; Chinnaiyan, A.M.; Rhodes, D.; Tomlins, S.; Fall, K.; Mucci, L.A.; Kantoff, P.W.; Stampfer, M.J.; Andersson, S.-O.; Varenhorst, E.; Johansson, J.-E.; Brown, M.; Golub, T.R.; Rubin, M.A. Estrogen-Dependent Signaling in a Molecularly Distinct Subclass of Aggressive Prostate Cancer. *J. Natl. Cancer Inst.* **2008**, *100*, 815-825, https://doi.org/10.1093/jnci/djn150.
48. Xia, J.; Gill, E.E.; Hancock, R.E.W. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat. Protoc.* **2015**, *10*, 823-844, https://doi.org/10.1038/nprot.2015.052.
49. Treeck, O.; Buechler, C.; Ortmann, O. Chemerin and Cancer. *Int. J. Mol. Sci.* **2019**, *20*, 3750, https://doi.org/10.3390/ijms20153750.
50. Narla, G.; Heath, K.E.; Reeves, H.L.; Li, D.; Giono, L.E.; Kimmelman, A.C.; Glucksman, M.J.; Narla, J.; Eng, F.J.; Chan, A.M.; Ferrari, A.C.; Martignetti, J.A.; Friedman, S.L. *KLF6*, a Candidate Tumor Suppressor Gene Mutated in Prostate Cancer. *Science* **2001**, *294*, 2563-2566, https://doi.org/10.1126/science.1066326.
51. Hu, K.; Ma, C.; Ma, R.; Zheng, Q.; Wang, Y.; Zhang, N.; Sun, Z. Roles of Krüppel-like factor 6 splice variant 1 in the development, diagnosis, and possible treatment strategies for non-small cell lung cancer. *Am. J. Cancer Res.* **2022**, *12*, 4468-4482.
52. Chen, D.; Guo, J.; Miki, T.; Tachibana, M.; Gahl, W.A. Molecular cloning of two novel *rab* genes from human melanocytes. *Gene* **1996**, *174*, 129-134, https://doi.org/10.1016/0378-1119(96)00509-4.
53. Lin, D.; Dong, X.; Wang, K.; Wyatt, A.W.; Crea, F.; Xue, H.; Wang, Y.; Wu, R.; Bell, R.H.; Haegert, A.; Brahmbhatt, S.; Hurtado-Coll, A.; Gout, P.W.; Fazli, L.; Gleave, M.E.; Collins, C.C.; Wang, Y. Identification of DEK as a potential therapeutic target for neuroendocrine prostate cancer. *Oncotarget* **2014**, *6*, 1806, https://doi.org/10.18632/oncotarget.2809.
54. Hendriksen, P.J.M.; Dits, N.F.J.; Kokame, K.; Veldhoven, A.; van Weerden, W.M.; Bangma, C.H.; Trapman, J.; Jenster, G. Evolution of the Androgen Receptor Pathway during Progression of Prostate Cancer. *Cancer Res.* **2006**, *66*, 5012-5020, https://doi.org/10.1158/0008-5472.CAN-05-3082.
55. Sharma, A.; Mendonca, J.; Ying, J.; Kim, H.-S.; Verdone, J.E.; Zarif, J.C.; Carducci, M.; Hammers, H.; Pienta, K.J.; Kachhap, S. The prostate metastasis suppressor gene NDRG1 differentially regulates cell motility and invasion. *Mol. Oncol.* **2017**, *11*, 655-669, https://doi.org/10.1002/1878-0261.12059.
56. Joshi, V.; Lakhani, S.R.; McCart Reed, A.E. NDRG1 in Cancer: A Suppressor, Promoter, or Both? *Cancers* **2022**, *14*, 5739, https://doi.org/10.3390/cancers14235739.
57. Wang, C.-T.; Chen, T.-M.; Mei, C.-T.; Chang, C.-F.; Liu, L.-L.; Chiu, K.-H.; Wu, T.-M.; Lan, Y.-C.; Liu, W.-S.; Chen, Y.-H.; Lin, Y.-M.J. The Functional Haplotypes of *CHRM3* Modulate mRNA Expression and Associate with Bladder Cancer among a Chinese Han Population in Kaohsiung City. *BioMed Res. Int.* **2016**, *2016*, 4052846, https://doi.org/10.1155/2016/4052846.
58. Calaf, G.M.; Crispin, L.A.; Muñoz, J.P.; Aguayo, F.; Bleak, T.C. Muscarinic Receptors Associated with Cancer. *Cancers* **2022**, *14*, 2322, https://doi.org/10.3390/cancers14092322.

59. Kuleshov, M.V.; Jones, M.R.; Rouillard, A.D.; Fernandez, N.F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S.L.; Jagodnik, K.M.; Lachmann, A.; McDermott, M.G.; Monteiro, C.D.; Gundersen, G.W.; Ma'ayan, A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **2016**, *44*, W90-W97, https://doi.org/10.1093/nar/gkw377.
60. Miyamoto, S.; Taylor, S.L.; Barupal, D.K.; Taguchi, A.; Wohlgemuth, G.; Wikoff, W.R.; Yoneda, K.Y.; Gandara, D.R.; Hanash, S.M.; Kim, K.; Fiehn, O. Systemic Metabolomic Changes in Blood Samples of Lung Cancer Patients Identified by Gas Chromatography Time-of-Flight Mass Spectrometry. *Metabolites* **2015**, *5*, 192-210, https://doi.org/10.3390/metabo5020192.

Publisher's Note & Disclaimer

Supplementary Materials

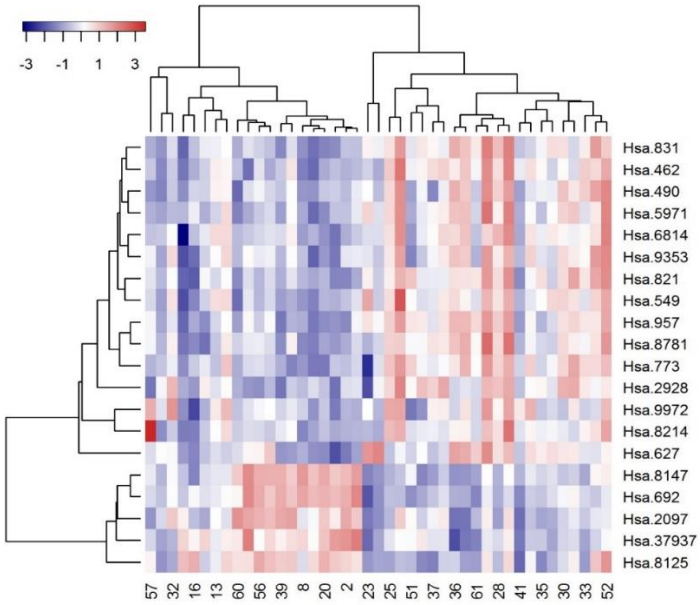


Figure S1. Heatmap for the Colon cancer data with selected twenty features.

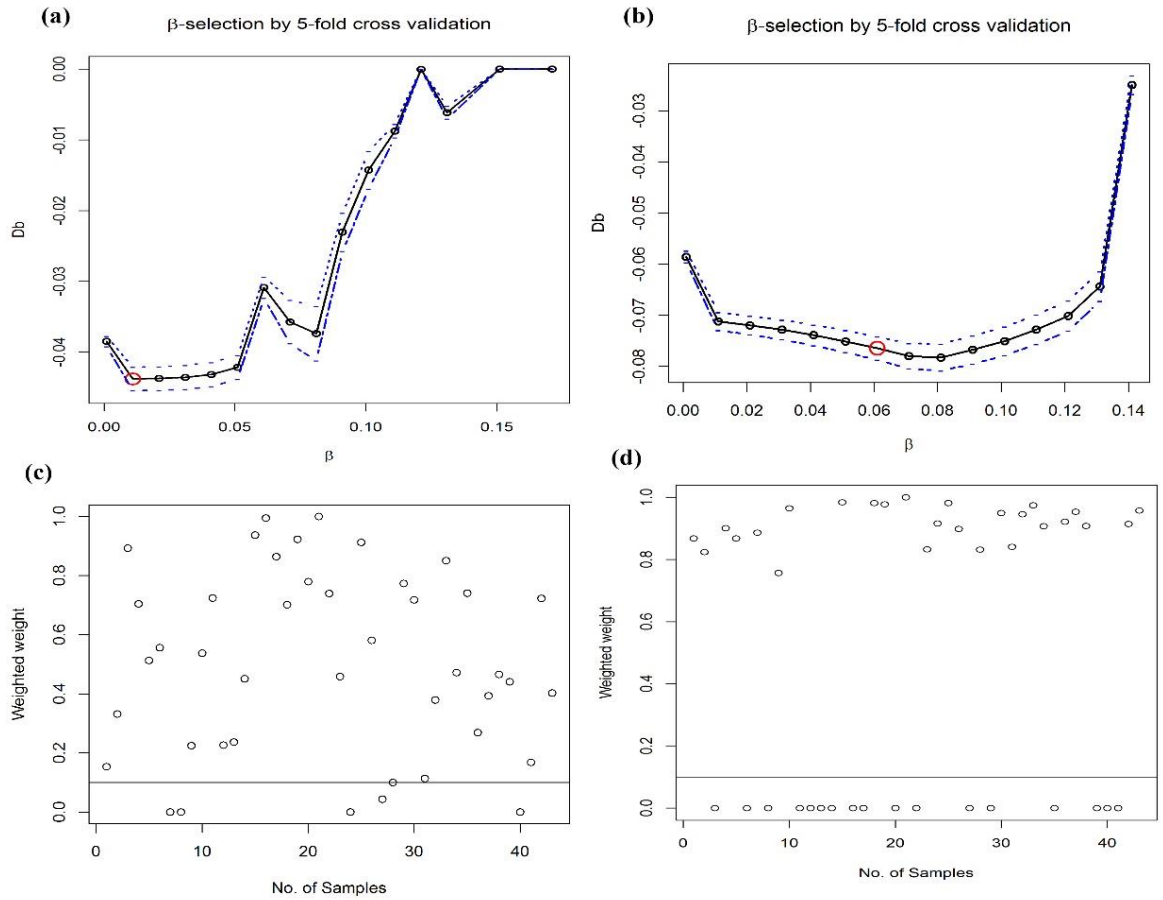


Figure S2. Beta selection for the Colon data: (a) with original data; (b) with corrupted data via 10% outliers. Based on this selected beta values, the weight plots for finding outliers; (c) with original data; (d) with corrupted data via 10% outliers.