STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **12**, March 2024, pp 325–342.
Published online in International Academic Press (www.IAPress.org)

Proposed Two-Steps Procedure of Classification High Dimensional Data with Regularized Logistic Regression

Omar Q. Alshebly [1,*], Suhail N. Abdullah [2]

[1] *Department of Statistics and Informatics, University of Mosul, Iraq*
[2] *Department of Statistics, University of Baghdad, Iraq*

Abstract The field of Bioinformatics has developed in response to the rapid increase in biological data, particularly high-dimensional gene expression data. Bioinformatics utilizes optimization, computational science, and statistical methods to effectively address challenges in the field of molecular biology. Numerous genes (variables) in gene expression are irrelevant to their study. Gene selection has been demonstrated to be an effective means of enhancing the performance of numerous methods of classification. The job of acquiring significant variables via the use of ranking variable selection (RVS) techniques and then picking the most effective classifier is an enormous challenge in the context of high-dimensional data. In this study, we proposed a new ranking filter method using smooth clipped absolute deviation depending on the resampling technique (RSVS) to obtain a proficient subset of genes with strong classification abilities. This is achieved by merging a screening technique employed as a filtering method in conjunction with regularized logistic regression (RLR), such as LASSO, ALASSO, ENET, and MCP. The study involved the utilization of both simulated and real datasets to conduct an empirical evaluation of the proposed approach. The findings indicated that the proposed method outperformed other established methods. It was tested using three publicly available data sets about cancer. The results demonstrate that the suggested approach is highly effective and viable, thus showing a strong level of performance with regard to accuracy, geometric mean, and area under the curve. Furthermore, the findings suggest that the genes most often chosen are physiologically associated with a specific form of cancer. Therefore, the method which has been suggested has potential advantages for the classification of cancer via the use of DNA gene expression data within a clinical setting.

Keywords Classification, Regularized Logistic Regression, Ranking Variable Selection, Gene Expression Data, Accuracy, Smooth Clipped Absolute Deviation, RSVS.

AMS 2010 subject classifications 62J07 , 62H30 , 92C40

DOI: 10.19139/soic-2310-5070-1846

1. Introduction

Recent development of DNA microarray technology has emerged as a pivotal advancement in the domains of biology, medicine, bioinformatics, and genetics [1, 2, 3]. The advent of high-throughput techniques has resulted in the generation of high-dimensional data, which has significant relevance in the domain of machine learning and data mining research [4]. The microarray datasets are extremely sparse and possess a high dimension. Redundancy, noise, and dimensionality are a few of the difficulties associated with these datasets [5]. The ”curse of dimensionality” is a prominent challenge seen in several biological data types, in situations when the number of variables exceeds the number of samples ($p > n$). One common issue encountered when dealing with high-dimensional data is the occurrence of redundant and superfluous variables.

[*] Correspondence to: Omar Q. Alshebly (Email: omarqusay@uomosul.edu.iq). Department of Statistics and Informatics, College of Computer Sciences and Mathematics, University of Mosul, Nineveh, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

There are many statistical issues that are encountered when modeling high-dimensional datasets. These challenges include overfitting, computational complexity, and estimator instability. The constraints outlined above make the deployment of statistical microarray classification [6]. One of the study topics in statistical applications is the reduction of dimensionality since this is often the most effective method for working with large datasets of high complexity [7, 8]. Over the last decade, the disciplines of machine learning and statistical methods have seen the introduction of many two-stage procedures, such as ranking variable selection (RVS) and classification algorithms, aimed at effectively tackling these challenges [5, 9, 10]. The application of RVS techniques in classification methodologies consistently improves performance by reducing data dimensionality through the exclusion of irrelevant genes (variables).

The first phase of the study involves the implementation of the variable selection strategy based on ranking, which aims to exclude a significant portion of the genes that have low expression levels. The genes that exhibit considerable expression are then used by the classifiers in order to enhance their predictive capabilities. Several strategies have been widely utilized in the literature to decrease the dimensionality of data by eliminating noisy and unnecessary variables. These methods include Information Gain (IG) [11], Fisher Score (FS) [12], and Chi-square (CS) [13]. The SIS (Sure Independent Screening) strategy, as proposed by [14], was designed to guarantee that all significant variables are retained after variable screening, with a probability that approaches one. Parametric models have become more prominent in high-dimensional experiments [15]. Several parametric models of regularized logistic regression (RLR), such as the least-absolute selection shrinkage operator (LASSO) [16] and ridge [17], use ℓ_1-norm and ℓ_2-norm penalties. The Adaptive Lasso (ALASSO) algorithm incorporates ridge technique weights. The Elastic Net (ENET) approach, as proposed by [18], is a regularization method that combines both ℓ_1-norm and ℓ_2-norm penalties. The two remaining parametric models are the non-concave and concave penalty-based smooth clipped absolute deviation (SCAD) model, as described by [19], and the minimum concave penalty (MCP) model, which was introduced by [20].

A study aimed to improve classification and prediction accuracy in gene-dependent cancers by identifying significant variables using a new proposed filter via the resampling-based SCAD variable selection (RSVS) technique. The RSVS uses selection probability with the SCAD penalty and the sure independence screening (SIS) approach to determine the threshold for selecting variables with the highest ranks. These variables are then used in regularized logistic regression models.

The subsequent sections of this work are structured in the following manner: Section 2 provides an explanation of the filtering methods and the SIS approach. Section 3 provides a description of the Regularized Logistic Regression, including five models. In Section 4, the suggested method is explained, along with the algorithmic process. Section 5 focuses on defining the metrics used and evaluating the effectiveness of the research. The experimental findings derived from the simulation and real data from gene expression datasets are reported in Sections 6 and 7. Lastly, Section 8 presents overarching findings.

2. Filtering Methods

Variable ranking (VR) approaches may be categorized into two main types: supervised [21] and unsupervised [22]. The classification is determined by the availability of class labels. Filtering techniques or methods do not depend on classification algorithms, which leads to quicker processing performance in comparison to the wrapper and embedding methods. The identification of relevant attributes is established by assessing measures such as distance, entropy, and uncertainty. A multitude of algorithms have been devised. An exemplary instance is the Relief method, as given by [23], which utilizes a metric function grounded on distance. This study aims to examine many commonly used filtering methods in the domain of gene expression research, including the Fisher score, information gain, and chi-square.

2.1. Fisher score

The Fisher Score (FS) [24] is a widely used strategy in the realm of variable selection for identifying significant variables from a larger set with the intention of using them in subsequent classification tasks. In order to achieve this objective, discriminative and statistical models are used. The fundamental criterion for variable selection is to minimize the intra-class distance, i.e., the distance between data points within the same class, while maximizing the inter-class distance, i.e., the distance between data points belonging to different classes. The Fisher score for each variable F$_j$ is computed using this idea:

$$F\left(j\right)=\frac{\sum_{k=1}^{c}n_k\left(m_k^j-m^j\right)^2}{\sum_{k=1}^{c}n_k\left(\sigma_k^j\right)^2} \tag{1}$$

The value F(j) represents the computed (FS) for each variable j,class $c = 0or1$, m_k^j is the mean of k-th class, m^j is the mean and σ_k^j is the variance of all dataset.

2.2. *The Chi-square test*

The Chi-square test, sometimes referred to as (CS), is categorized as a non-parametric test and is generally used to determine statistical significance when investigating the association between two qualitative variables. During the pre-processing phase, the ”equal interval width” method converts numerical variables into their respective categorical equivalents. The method referred to as ”equal interval width” first divides the data into q intervals, with each interval having the same width. The formula $w = (max - min)/q$, is used for the purpose of calculating the interval width. Additionally, the phrases $min + w, min + 2w, ...,$ and $min + (q - 1)w$ are utilized to ascertain the limits of each individual interval [25, 26].
A training set comprises of x$_j$ =(x$_{j1}$, ...,x$_{jp}$), where g represents each variable in a set of p variables. The CS score for a specific variable g with r distinct variable values [26] can be computed.

$$\widetilde{x}^2\left(g\right)=\sum_{j=1}^{r}\sum_{s=1}^{p}\frac{\left(O_{js}-E_{js}\right)^2}{E_{js}} \tag{2}$$

O_{js} refers to the count of instances that show the jth variable value, provided the value of variable g.
E_{js} framework denotes the expected presence of data instances that encompass the specific value assigned to the variable denoted as g.

2.3. *Information Gain*

The use of Information Gain (IG) is employed to quantitatively measure the amount of information that a variable gives about the class [11]. The Information Gain approach is often used in Ranking Variable Selection (RVS) due to its notable computing efficiency and straightforward interpretability. The presence of unconnected or inaccurate variables does not provide any meaningful information, whereas the inclusion of variables that are properly divisible provides the maximum amount of information. The metric denoting the degree of purity is called information, whereas the metric representing the degree of impurity is known as entropy. One limitation of the information gain measure is its suboptimal performance in the presence of duplicate variables [27].
The following equation can be used to calculate (IG) between the gth variable in x$_i$ and the response variable y$_i$:

$$IG(x_i;y_i)=H(x_i)-H(x_i|y_i) \tag{3}$$

$$H(x_i)=\sum_{i}p\left(x_i\right)\log_2\left(p\left(x_i\right)\right)$$

$$H(x_i|y_i)=\sum_{y_i\in Y}p\left(y_i\right)\sum_{x_i\in X}p\left(x_i|y_i\right)\log_2 p\left(x_i|y_i\right)$$

The symbol H(x$_i$) represents the entropy of variable x$_i$, while H(x$_i$|y$_i$) represents the conditional entropy of x$_i$ given y$_i$.

2.4. Sure Independence Screen (SIS)

The approach used for variable selection or screening is Sure Independence Screening. The SIS approach was presented by the author in reference [14]. A specific methodology is used to decrease the number of variables, represented as d,from a significant magnitude to a smaller subset, designated as m.
Ranking variable selection approaches assign scores to individual variables based on their respective states. The variables are arranged in ascending order of significance based on the given scores. The procedure for determining the variables with the highest rank is accomplished by using the SIS condition, which is specifically delineated as follows: [28].

$$\frac{s}{\log s} \tag{4}$$

The variable s is the total number of samples.

3. Regularized Logistic Regression(RLR)

Logistic regression is a well-accepted and essential statistical method used in the fields of medicine and other sciences for solving binary classification issues. These problems include a response variable that is dichotomous, meaning it can only take on two distinct values, either zero or one [29]. Researchers encounter significant hurdles when using logistic regression on datasets with large dimensions in many disciplines such as biomedical imaging, DNA microarrays, and genomics, especially when the number of variables p surpasses the sample size n.
Let $y_i \in \{0, 1\}$,Create a vector of dimensions $n \times 1$ consisting of tissues., and let x_j be a $p \times 1$ vector of variables. The vector of probability estimates undergoes a logistic transformation $\pi_i = p(y_i = 1|x_j)$ and the relationship can be represented by a linear function that has undergone a logit transformation: [6]

$$\ln[\frac{\pi_i}{1-\pi_i}] = \beta_0 + \sum_{j=1}^{p} {x_j}^T \beta_j, \quad i = 1, 2, ..., n, \tag{5}$$

where β_0 is the intercept as well as β_j is a $p \times 1$ the unexplored coefficients are represented as a vector.
The log-likelihood function is defined as:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \{y_i \ln \pi(x_{ij}) + (1 - y_i) \ln(1 - \pi(x_{ij})\}\,, \; j = 1, 2, ...p. \tag{6}$$

One advantageous characteristic of logistic regression is its capacity to simultaneously estimate probabilities. $\pi(x_{ij})$ and $1 - \pi(x_{ij})$ for each class .
The probability of classifying the i^{th} sample in class is estimated by $\hat{\pi}_i = \exp(\beta_0 + \sum_{j=1}^{p} {x_j}^T \beta_j)/1 + \exp(\beta_0 + \sum_{j=1}^{p} {x_j}^T \beta_j)$.
The approach of regularized logistic regression (RLR) involves the inclusion of a non-negative regularization term in equation (5), which allows for the control of gene coefficients in scenarios with a large number of dimensions. The idea made by Tibshirani [16] about the use of ℓ_1-norm regularization is widely acknowledged and often referenced in academic literature. The simultaneous processes of gene selection and estimation are performed using the ℓ_1-norm regularization approach. This technique includes applying restrictions to the log-likelihood function of gene (variable) coefficients. The RLR [6] is formally defined in the following manner:

$$RLR = \sum_{i=1}^{n} \{y_i \ln \pi(x_{ij}) + (1 - y_i) \ln(1 - \pi(x_{ij})\} + \lambda P(\beta). \tag{7}$$

The vector β was estimated through the process of minimizing.

$$\hat{\beta}_{RLR} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \{y_i \ln \pi(x_{ij}) + (1 - y_i) \ln(1 - \pi(x_{ij})\} + \lambda\, P(\beta) \right] \tag{8}$$

The variable $\lambda P(\beta)$ represents the regularization term, used to regularize the estimates. The penalty term is determined by the value of the positive tuning parameter λ, which controls the balance between accurately fitting the data to the model and the effect of the regularization.

The Lasso method, first proposed by [16], is a modified version of conventional least squares that has been extended to include a broader scope of applications. The primary objective of using the Lasso method is to conduct variable selection and regularization techniques, hence enhancing the predictive accuracy of the model.

The Lasso method provides a notable advantage in terms of its computing feasibility when used in the analysis of classification data characterized by a large number of dimensions. On the other hand, in cases where there are significant relationships and a noticeable clustering pattern among relevant genes or variables, the Lasso technique has a propensity to arbitrarily choose one gene while discarding the others. Moreover, the Lasso technique demonstrates inconsistent gene selection as a result of uniformly applying shrinkage to every gene coefficient.

The estimation of the vector β through the utilization of the lasso method is formally defined as follows:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta}\left[-\sum_{i=1}^{n}\{y_i \ln \pi(x_{ij}) + (1-y_i)\ln(1-\pi(x_{ij})\} + \lambda\sum_{j=1}^{p}|\beta_j|\right] \tag{9}$$

The variable λ serves as a tuning parameter in this context. The maximum likelihood estimator (MLE) converges to a specific value when the parameter λ takes on the value of zero.

The Adaptive Lasso technique or method was devised as a strategy to alleviate the inherent bias present in the Lasso method. The amount of bias in the model estimate is contingent upon the numerical value assigned to the variable λ. The Alasso technique utilizes adaptive weights to impose penalties on individual coefficients within the $\ell 1$ penalty [14]. The incorporation of a weight penalty is a technique used to mitigate the bias included in the Lasso method. This is accomplished by allocating comparatively lower weights to variables that have bigger coefficients and greater weights to variables that have smaller coefficients. On the other hand, variables that possess substantial coefficients are assigned reduced weights. Preserving the sparsity characteristic of the Lasso method is most important when mitigating its selection bias. The mathematical formulation of the regularized logistic regression using adaptive LASSO is given by the following equation:

$$\hat{\beta}_{ALASSO} = \arg\min_{\beta}\left[-\sum_{i=1}^{n}\{y_i \ln \pi(x_{ij}) + (1-y_i)\ln(1-\pi(x_{ij})\} + \lambda\sum_{j=1}^{p}w_j\,|\beta_j|\right] \tag{10}$$

Where $w_j = (w_1, ..., w_p)^T$ is $p \times 1$weight vector. It is dependent on the consistent initial values of $\hat{\beta}$ and $w_j = (|\hat{\beta}_j|)^{-\gamma}$, where γ is a positive constant.

The Elastic Net regularization approach was proposed by [30] as a means to address the initial limitations of the LASSO method in gene selection. The elastic net regularization method aims to combine the ℓ_2-norm and ℓ_1-norm regularization techniques. It does this by using the ridge regression penalty to address issues related to high correlation while also leveraging the variable selection aspect of LASSO regularization. The RLR model, including the elastic net penalty, may be mathematically expressed as:

$$\hat{\beta}_{Elastic} = \arg\min_{\beta}\left[-\sum_{i=1}^{n}\{y_i \ln \pi(x_{ij}) + (1-y_i)\ln(1-\pi(x_{ij})\} + \lambda_1\sum_{j=1}^{p}w_j\,|\beta_j| + \lambda_2\sum_{j=1}^{p}{\beta_j}^2\right] \tag{11}$$

Let y$_i$ and x$_i$ represent the ith and jth vectors of the augmented matrix, respectively.

The regression approach referred to as smoothly Clipped Absolute Deviation (SCAD) is well recognized. The suggested approach was introduced as a means to fully address the constraint of variable selection in situations where there are a substantial number of variables. It exhibits desirable oracle qualities in comparison to Lasso and Elastic Net [8]. The use of the SCAD method, which incorporates a concave penalty function, effectively addresses the inherent drawbacks of the Lasso technique. The penalty functions used in Lasso regression do not satisfy the mathematical requirements of continuity, sparsity, and unbiasedness simultaneously.

The regression coefficients are obtained from the regularized linear regression using SCAD as follows:

$$\hat{\beta}_{SCAD} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \{ y_i \ln \pi(x_{ij}) + (1 - y_i) \ln(1 - \pi(x_{ij}) \} + \lambda \sum_{i=1}^{p} p_\lambda \left(|\beta_i| \right) \right] \tag{12}$$

Where $p_\lambda \left(. \right)$ is the limit of the SCAD penalty and defined:

$$|\beta_i| I_{|\beta_j| \leq \lambda} + (\frac{\left\{ (c^2 - 1)\lambda^2 - (c\lambda - |\beta_j)^2 + \right\} I(\lambda \leq |\beta_j|)}{2(c-1)} , C > 2 and \lambda \geq 0. \tag{13}$$

The abbreviation MCP is often used to refer to the Minimax Concave Penalty [20]. The motivation for the use of the MCP method is to exclude insignificant predictors from the model while allowing significant predictors to remain unpenalized. The MCP demonstrates efficiency in its use of Oracle technology and has the desirable attribute of being Oracle-compatible. Therefore, the penalty function used in the MCP method has a concave or non-convex structure. The behaviour of MCP is characterized by a somewhat lower level of group cohesion and a tendency to prefer smaller group sizes.

The regression coefficients are obtained from the regularized linear regression using MCP as follows:

$$\hat{\beta}_{MCP} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \{ y_i \ln \pi(x_{ij}) + (1 - y_i) \ln(1 - \pi(x_{ij}) \} + \lambda \sum_{i=1}^{p} p_\lambda \left(|\beta_i| \right) \right] \tag{14}$$

Where $p_\lambda \left(. \right)$ is the limit of the MCP penalty and defined by:

$$\frac{2c\lambda \left| \beta_j \right| - \beta_j^2}{2c} I \left(\left| \beta_j \right| \leq c\lambda \right) + \frac{c\lambda^2}{2} I \left(\left| \beta_j \right| > c\lambda \right), c > 1 \; and \; \lambda \geq 0. \tag{15}$$

4. The Proposed Method based on Resampling

The primary aim of gene or variable selection is to improve the classification performance, expedite and optimize the gene selection procedure, and get a more complete comprehension of the underlying classification issue.

In this study, we provide a novel filter that incorporates an approach to classification based on resampling techniques. The filtering approach incorporates the introduction of the resampling-based SCAD variable selection method (RSVS).

The RSVS technique is derived from the SCAD penalized regression method and incorporates a resampling strategy to identify important variables based on their frequency ranking. SCAD formula based on the equation number (12) at this study.

The below equation represents the probability of selection in each gene (variable) according to the SCAD method.

$$S \left(v_n \right) = \frac{1}{R} \sum_{i=1}^{R} \frac{1}{L} \sum_{j=1}^{L} I \left(\beta_i \neq 0 \right), for \; i = 1, 2, \ldots, p \tag{16}$$

The symbol R represents the total number of resampling iterations, L represents the total number of λ values, v$_n$ represents the variable indexed as i, p represents the total number of variables, n represents the total number of samples, and $I()$ denotes an indicator variable. The variable selection model is constructed by considering a certain number of resamples, denoted as R, and a certain number of values of λ, denoted as L. The model construction process incorporates the use of 10-fold cross-validation.

After using the RSVS method to rank the variables, we use the SIS method to select the most important variables based on Equation (4). In each iteration, the computation involves determining the count of true variables picked from the top-ranked variables according to the SIS method. This count is then averaged across a total of 100 iterations. The previously mentioned genes are then utilized in the framework of LASSO, ALASSO, ENET, and

MCP. The regularized logistic regression technique evaluates the outcomes of the proposed method in conjunction with a new filtering approach.

Algorithm 1:The proposed method (RSVS)

Step 1: The dataset is divided into a training set (70%) and a testing set (30%), with the training set used for model construction and the testing set for performance evaluation.

Step 2: Calculating the frequency of each gene from 100 different models of λ values.

Step 3: Perform Step 1 and Step 2 iteratively for a total of 100 repetitions.

Step 4: Equation (16)

$$S\left(v_n\right) = \frac{1}{R}\sum_{i=1}^{R}\frac{1}{L}\sum_{j=1}^{L} I\left(\beta_i \neq 0\right), for\ i = 1, 2, \ldots, p,$$

used to calculate the probability of selecting each variable and then rank them accordingly.

Step 5: Selecting the genes with the highest frequency depending on

$$\frac{s}{\log s}$$

Step 6: Apply them to regularized logistic regression methodologies to create prediction models.

Figure 1 presents the workflow of the proposed method(RSVS) in our study.

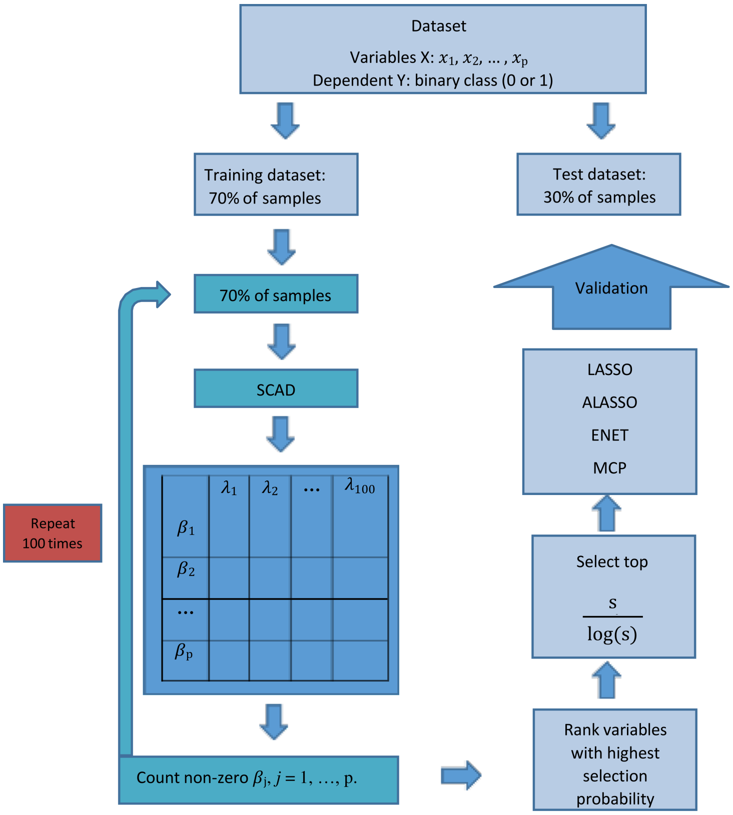
Dataset
Variables X: $x_1, x_2, \ldots, x_p$
Dependent Y: binary class (0 or 1)
Training dataset: 70% of samples
Test dataset: 30% of samples
Validation
70% of samples
SCAD
λ_1
λ_2
…
λ_{100}
β_1
β_2
…
β_p
Repeat 100 times
Count non-zero β_j, $j = 1, \ldots, p$.
LASSO
ALASSO
ENET
MCP
Select top $\frac{s}{\log(s)}$
Rank variables with highest selection probability

Figure 1. The workflow of the proposed method(RSVS).

5. Performance Metrics

Accuracy is a commonly used criterion for evaluating the efficacy of classification models when applied to balanced datasets. The computation entails the determination of the proportion of correctly classified samples relative to the overall number of samples used in the classification model [31].
Sensitivity is used as a diagnostic measure to assess the accuracy of accurately identifying individuals as ”diseased”. The concept of specificity is used to evaluate the accuracy of accurately classifying a person as ”normal”. The metrics of Accuracy, Sensitivity, and Specificity are often defined in relation to the quantities of true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP) [32].

$$\begin{aligned} Accuracy &= \tfrac{TP+TN}{N} \\ Sensitivity &= \tfrac{\mathrm{TP}}{\mathrm{TP+FN}} \\ Specificity &= \tfrac{\mathrm{TN}}{\mathrm{TN+FP}} \end{aligned} \tag{17}$$

Also, we used Geometric Mean is the joint performance evaluated by utilizing the Geometric Mean of sensitivity and specificity [6].

$$G - Mean = \sqrt{\text{Sensitivity } \times \text{ Specificity}} \tag{18}$$

The Area under Curve (AUC) is a metric that quantifies the aggregate performance of classifier scores throughout the whole range of feasible threshold values.
The area under the receiver operating characteristic (ROC) curve is often used as a metric to assess the efficacy of probability categorization. The calculation of the area under a curve involves the use of the following formula:

$$A_{ROC} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} = \frac{1}{PN}\int_0^N TP * \, dFP \tag{19}$$

6. Results of Simulation

The utilization of the multivariate normal distribution within an auto-regressive correlation structure enables the generation of modelling data. The correlation matrix is a mathematics concept that is described formally as follows:

$$\sum = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots\cdots\cdots \rho_{1g} \\ \rho_{21} & \rho_{22} & \cdots\cdots\cdots \rho_{2g} \\ & \cdots\;\cdots & \\ \rho_{s1} & \rho_{s2} & \cdots\cdots\cdots \rho_{sg} \end{pmatrix} \tag{20}$$

The equation that is used to produce the binary response variables is as follows:

$$\pi_i \left(y_i = 1|x_i\right) = \frac{\exp\left(x_i\beta\right)}{1+\exp\left(x_i\beta\right)} \tag{21}$$

The variable of interest is denoted as a set of significant variables($\mathrm{x_i}$). In contrast, the response variable y_i is produced through Bernoulli experiments, as demonstrated in Equation (21). The experiment was repeated 100 times, with a sample size of n=200 and a variable dimension of p=1000. The regression coefficients were derived from a uniform distribution with a lower bound of 2 and an upper bound of 4 [34]. The equation that represents the correlation between the i^{th} and j^{th} variables is $\rho_{ig} \;\; = \rho^{|\mathrm{i}-\mathrm{j}|}$ where i and j are integers ranging from 1 to g. We obtained data exhibiting correlations of 0.1, 0.4, and 0.7 by employing this high-dimensional configuration. The observed values in the datasets exhibit correlation structures that range from low to high, resembling patterns often seen in other forms of data within the area of bioinformatics, such as gene expression data.
Microarray data often demonstrates a substantial level of correlation, and the use of simulated data allows for the evaluation of suggested approaches on correlated data of this kind. The simulation data is first divided into two subsets: a training dataset, which accounts for 70% of the total data, and a test dataset, which accounts for the

remaining 30% of the entire data. The Ranking Variable Selection (RVS) methods and the new filter proposed (RSVS) are utilized on the training dataset, where the variables are ranked according to their degree of importance. The SIS is then utilized to select the most significant variables, which are ranked at the top. The selected variables are subsequently inputted into the regularized regression models. To comprehensively evaluate the performance of the models and prevent the impact of data partitioning, a 10-fold cross-validation approach was employed for all RLR models. The computations were conducted by averaging the results from 100 random partitioning iterations. All the implementations of the study on simulation and real data applications are carried out using R and Origin2023b.

Table 1. The true variable selection was determined by taking the average of the filtering methods used across all of the simulation data with its various correlation sets

Correlation	RSVS(SD)	FS(SD)	CS(SD)	IG(SD)
0.1	**5.89(0.345)**	4.25(0.912)	5.15(0.833)	4.12(0.977)
0.4	**4.56(0.925)**	3.80(0.930)	4.01(0.944)	3.27(0.973)
0.7	**2.89(1.385)**	0.32(2.961)	1.12(2.010)	0.25(3.112)

According to the result presented in Table 1. It was observed that the RSVS method, which employs the SIS criteria to pick the top-ranked variables, includes a larger set of significant variables in contrast to other existing RVS methods such as FS, CS, and IG, which were used for comparative purposes. The RSVS approach reliably identified an average of 5.89, 4.56, and 2.89 true variables from a set of six significant variables that possess true significance. Furthermore, it demonstrated the lowest standard deviation(SD) values compared to the other methods. The mentioned result was demonstrated in three unique correlation structures offering strong proof in favour of the method presented in our research.

Figure 2 presented an illustration of boxplots showing the utilization of the RVS method in conjunction with the SIS technique for analyzing correlation data values of 0.1, 0.4, and 0.7 with 100 iterations. The boxplots demonstrate that the RSVS method exhibits a superior average in true variable selection compared to (FS), (CS) and (IG) methods. Through the proposed method, it is observed that the data order within the Box-plot Graph is more consistent and the variables show good correlation with each other, with only a few extreme values. CS method in the second order in terms of variable ranks.

Table 2 presented the computed values of accuracy, and geometric mean (GM-mean) for all the methods, in order to demonstrate the efficacy of the proposed RSVS method in combination with regularized logistic regression models, we conducted an analysis using simulation data comprising three distinct correlation values.

We demonstrated the superior performance of RVS as a proposed filter (RSVS) across all metrics. This demonstrates that the variable selection with (RSVS) filter exhibits superior accuracy, and G-mean compared to other methods. The variables are created with a low correlation structure, characterized by a correlation coefficient of r = 0.1. The performance of the proposed RSVS filter surpasses that of current filters across all classifier methods, including LASSO, ALASSO, ENET, and MCP. Furthermore, the suggested integration of the RSVS method and ENET classifier exhibits superior performance when compared to various combinations involving the RVS method and classifier, such as RSVS-LASSO, RSVS-ALASSO, RSVS-MCP, and other combinations. This superiority is evident in the accuracy and G-mean metrics, which attain values of 0.888 and 0.886, respectively. Additionally, the standard deviations (SD) associated with these metrics are 0.058 and 0.059, respectively. Which are relatively small values. Alternatively, the predictor variables can be obtained using a moderate correlation structure characterized by a correlation coefficient of r = 0.4. The combination of the RSVS technique and ENET classifier, which achieved an accuracy of 0.89 and a G-mean of 0.88, demonstrated superior performance compared to other combinations of the RVS method and classifiers. The standard deviations (SD) for accuracy and G-mean were 0.038 and 0.03, respectively. However, the data is created using a high correlation data structure with a correlation coefficient (r) of 0.7. The comparative analysis reveals that the performance of the proposed RSVS filter with ENET model surpasses that of other combinations using RVS methods and classifiers.

Figures 3 and 4 presented the boxplot illustrating the area under the curve (AUC) for the four filtering methods employed in the study. These box plots represent the performance of all classifiers at correlation coefficient

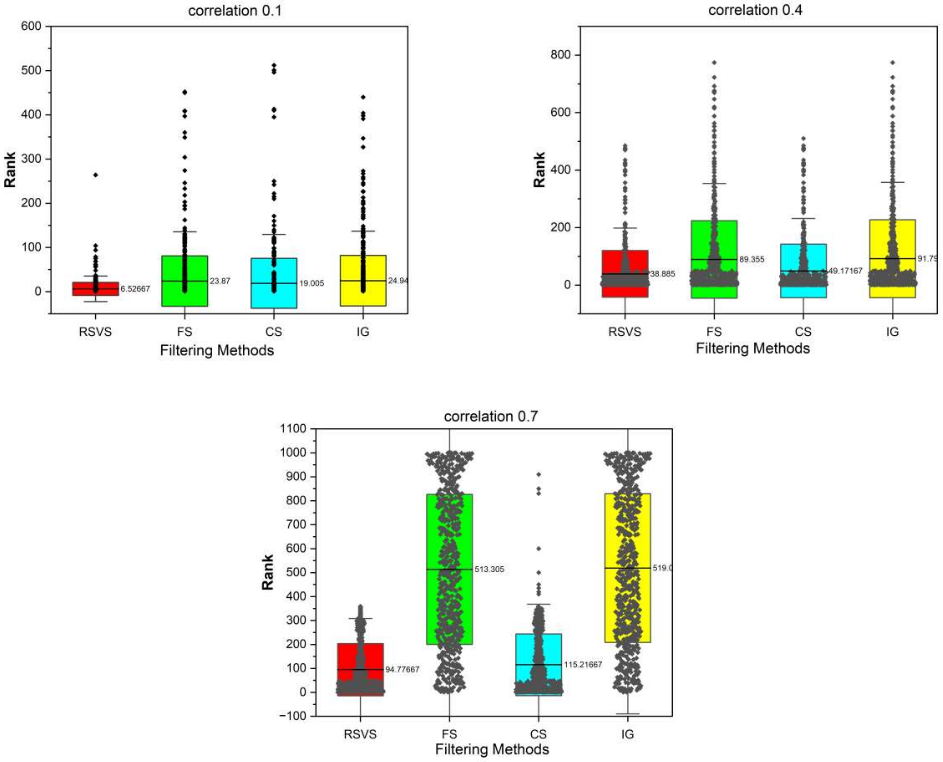

Figure 2. The boxplot of ranking true variables for filtering methods with correlation coefficients of 0.1, 0.4, and 0.7 with 100 iterations.

values of 0.1 and 0.4, respectively. Notably, the proposed method utilizing the ENET classifier exhibited a distinct superiority over the other methods when applied to the testing dataset.

7. Results of Real Dataset

The results of the simulation showed significant and obvious superiority of the proposed method in our study, so this method will be applied only with real data. The datasets included in the present study consisted of microarray gene expression data. The datasets provided in this study are associated with three publicly available high-dimensional gene expression datasets that have been previously used by other researchers. These datasets include colon cancer [36]. https://github.com/ramhiser/datamicroarray/blob/master/data/alon.RData, leukemia data [35]. https://github.com/ramhiser/datamicroarray/blob/master/data/golub.RData, and prostate cancer [37]. https://github.com/ramhiser/datamicroarray/blob/master/data/singh.RData.The datasets under consideration include a response variable categorized into two distinct classes. The use of the previous dataset in this investigation was motivated by the absence of relevant data of comparable kinds within the geographical context of Iraq.

The dataset is divided into two subsets, with 0.70 of the samples allocated for training and the remaining 0.30 reserved for testing. RVS methods utilize the training data to eliminate genes that are not relevant and prefer genes of significance. The most significant genes(variables) are chosen using the SIS technique. These genes(variables) are then utilized in conjunction with different classifiers. The model-building process uses a 10-fold cross-validation (CV) approach with 100 iterations. Finally, the testing data set is utilized to estimate different performance metrics.

Table 2. The assessment of various filtering and classification methods in analyzing simulation data characterized by three distinct correlation values

Correlation	RVS+Classifiers	Accuracy(SD)	G-mean(SD)
0.1	RSVS-LASSO	**0.841(0.059)**	**0.838(0.061)**
	RSVS-ALASSO	**0.849(0.077)**	**0.843(0.067)**
	RSVS-ENET	**0.888(0.058)**	**0.886(0.059)**
	RSVS-MCP	**0.82(0.062)**	**0.817(0.063)**
	FS-LASSO	0.811(0.053)	0.826(0.082)
	FS-ALASSO	0.819(0.050)	0.821(0.067)
	FS-ENET	0.858(0.066)	0.847(0.060)
	FS-MCP	0.801(0.077)	0.821(0.081)
	CS-LASSO	0.826(0.065)	0.832(0.067)
	CS-ALASSO	0.825(0.050)	0.833(0.066)
	CS-ENET	0.861(0.072)	0.869(0.063)
	CS-MCP	0.816(0.065)	0.821(0.067)
	IG-LASSO	0.818(0.052)	0.829(0.075)
	IG-ALASSO	0.811(0.059)	0.822(0.082)
	IG-ENET	0.839(0.077)	0.846(0.082)
	IG-MCP	0.812(0.066)	0.823(0.078)
0.4	RSVS-LASSO	**0.877(0.047)**	**0.875(0.049)**
	RSVS-ALASSO	**0.881(0.040)**	**0.87(0.05)**
	RSVS-ENET	**0.891(0.038)**	**0.88(0.03)**
	RSVS-MCP	**0.873(0.045)**	**0.872(0.047)**
	FS-LASSO	0.841(0.075)	0.855(0.077)
	FS-ALASSO	0.862(0.067)	0.86(0.067)
	FS-ENET	0.863(0.052)	0.868(0.066)
	FS-MCP	0.851(0.070)	0.859(0.072)
	CS-LASSO	0.861(0.051)	0.86(0.053)
	CS-ALASSO	0.872(0.052)	0.868(0.059)
	CS-ENET	0.871(0.047)	0.872(0.048)
	CS-MCP	0.863(0.048)	0.862(0.049)
	IG-LASSO	0.842(0.072)	0.859(0.078)
	IG-ALASSO	0.861(0.053)	0.863(0.060)
	IG-ENET	0.866(0.037)	0.868(0.050)
	IG-MCP	0.860(0.048)	0.866(0.066)
0.7	RSVS-LASSO	**0.912(0.039)**	**0.911(0.039)**
	RSVS-ALASSO	**0.914(0.038)**	**0.916(0.040)**
	RSVS-ENET	**0.928(0.030)**	**0.92(0.03)**
	RSVS-MCP	**0.906(0.041)**	**0.895(0.042)**
	FS-LASSO	0.881(0.066)	0.875(0.066)
	FS-ALASSO	0.882(0.065)	0.872(0.067)
	FS-ENET	0.901(0.050)	0.89(0.06)
	FS-MCP	0.881(0.070)	0.881(0.07)
	CS-LASSO	0.892(0.038)	0.891(0.039)
	CS-ALASSO	0.897(0.037)	0.895(0.037)
	CS-ENET	0.921(0.039)	0.917(0.040)
	CS-MCP	0.88(0.041)	0.879(0.041)
	IG-LASSO	0.88(0.067)	0.875(0.066)
	IG-ALASSO	0.887(0.064)	0.872(0.067)
	IG-ENET	0.90(0.051)	0.891(0.06)
	IG-MCP	0.879(0.072)	0.881(0.07)

Table 3. Presents a comprehensive summary of the three gene expression datasets

Dataset	No. of samples	No. of genes	Class
Colon Cancer	62	2000	40 tumor / 22 normal
Leukemia	73	7129	48 ALL / 25 AML
Prostate Cancer	102	12600	52 tumor / 50 normal

The histogram in Figure 5 provides a summary of the pairwise correlation. The average correlation coefficient among genes of colon cancer is 0.44. The evidence suggests a strong link between genes, as shown by the values observed in the simulation experiments. The mean correlation between genes (variables) in cases of leukemia and

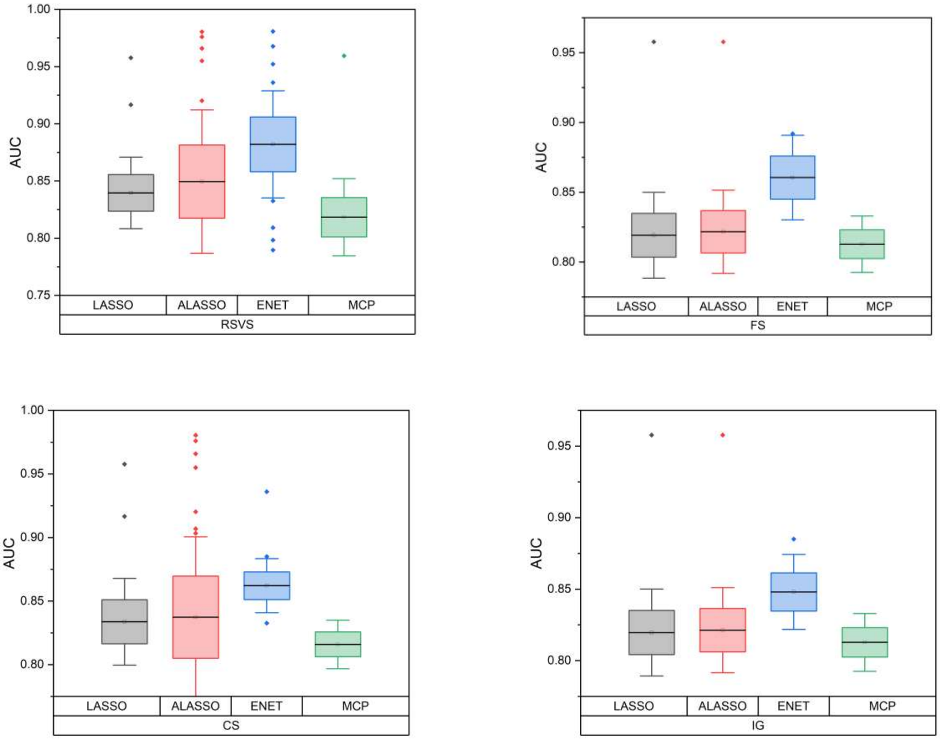

Figure 3. The boxplot of the area under the curve for filtering methods with a correlation coefficient of 0.1 with 100 iterations of the testing dataset.

prostate cancer is 0.03 and 0.04, respectively. This is attributed to the use of complete gene expression data, as opposed to the colon gene-expression dataset.

Table 4. Classifier Evaluation Using the proposed RSVS Method for a real testing dataset with an average of 100 iterations

Cancer data	RVS+Classifiers	Accuracy(SD)	AUROC(SD)	G-Mean(SD)
Colon	RSVS-LASSO	0.797(0.092)	0.88(0.086)	0.725(0.161)
	RSVS-ALASSO	0.807(0.084)	0.885(0.084)	0.742(0.140)
	RSVS-ENET	0.809(0.081)	0.889(0.081)	0.751(0.138)
	RSVS-MCP	0.784(0.084)	0.854(0.101)	0.712(0.157)
Leukemia	RSVS-LASSO	0.975(0.043)	0.996(0.021)	0.964(0.066)
	RSVS-ALASSO	0.979(0.040)	0.999(0.017)	0.975(0.055)
	RSVS-ENET	0.983(0.037)	0.994(0.025)	0.975(0.058)
	RSVS-MCP	0.953(0.069)	0.988(0.03)	0.925(0.121)
Prostate	RSVS-LASSO	0.919(0.051)	0.964(0.039)	0.918(0.054)
	RSVS-ALASSO	0.921(0.050)	0.967(0.037)	0.921(0.053)
	RSVS-ENET	0.924(0.049)	0.968(0.036)	0.923(0.052)
	RSVS-MCP	0.897(0.061)	0.956(0.043)	0.898(0.061)

Based on the results presented in Table 4, it is clear in all three real datasets, the accuracy, AUC, and G-mean of the individual classifiers are significantly better when they are used with the RSVS filter depend-on resampling than when they are used with the FS, CS, and IG methods.The RSVS with ENET classifier had the greatest average accuracy of 0.809, an AUC of 0.889, and a G-mean of 0.751 when applied to colon cancer. The RSVS technique has high performance in classifying leukemia and prostate cancer across all individual classifiers. However, when comparing the CS technique to other RVS approaches, it was found that applying the CS method to all individual

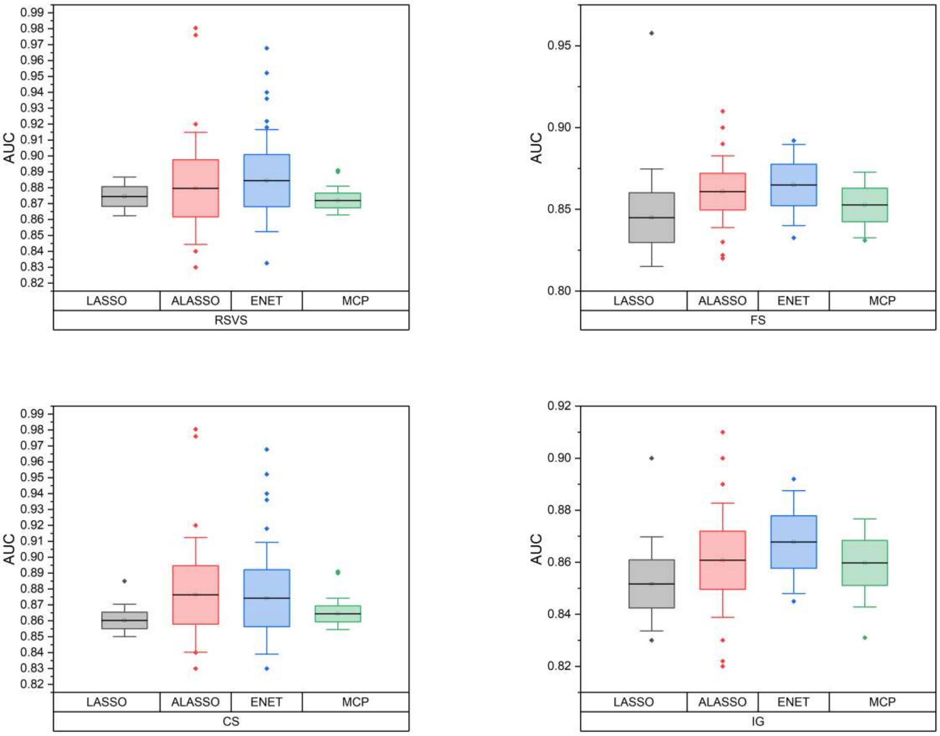

Figure 4. The boxplot of the area under the curve for filtering methods with a correlation coefficient of 0.4 with 100 iterations of the testing dataset.

classifiers resulted in much superior performance compared to applying the FS and IG methods to the individual classifiers. Conversely, it is seen that the standard deviation values obtained using the suggested method were much lower compared to those obtained using other methods. In conclusion, the combined approach of the RSVS-ENET exhibited superior performance compared to the other combinations, including the RVS method and classifier. This clarifies that the suggested framework exhibits superior performance in cases where there is a substantial correlation among the predictor variables.

Figure 6 shows a clear advantage of the ENET method when the suggested filter is used to look at three real datasets that depend on the genetic matrix, which is a major threat to human life. This is what leads us to develop this method for greater accuracy in future studies.

7.1. Results of Statistical Significance Test

In order to make sure that the suggested method really does work at finding highly relevant genes and classifying them well, a paired two-tailed t-test was used to compare the suggested method to each competing method pairwise. The test was conducted based on the calculation of the area under the curve (AUC) of the training dataset. In our investigation, the elastic net classifier was used to evaluate the performance of all four filters. It was observed that the ENET filter yielded the most favorable outcomes. Figure 7 displays the boxplot representing the Area Under the Curve (AUC) for each method used across all three datasets in the context of the elastic net (ENET) classifier. It is clearly seen that the AUC of the proposed method is comparable to the results obtained from FS-ENET, CS-ENET, and IG-ENET.

The paired two-tailed t-test findings at a specified significance level $alpha = 0.05$ are shown in Table 5. Based on the information in Table 5, the suggested method performs statistically better in the area under the curve

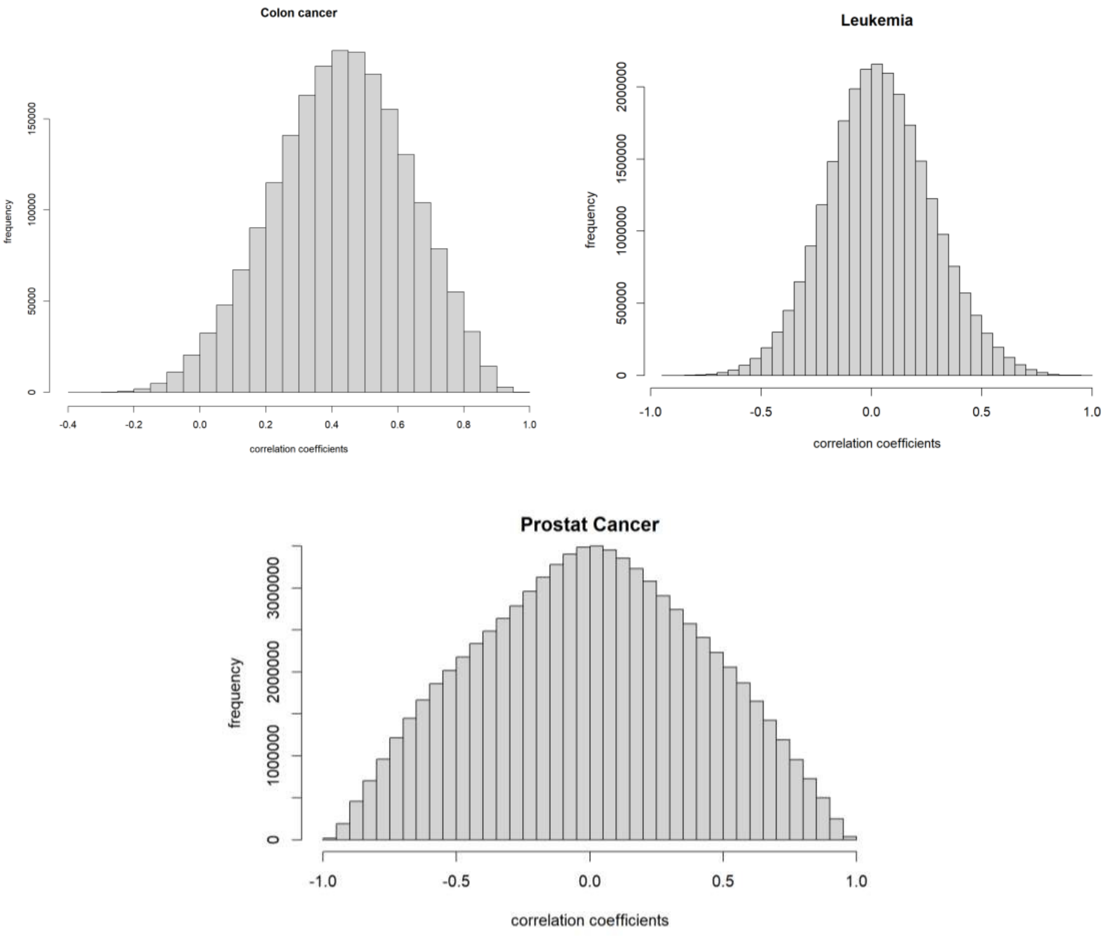

Figure 5. The histogram of pairwise correlation coefficients of a real dataset

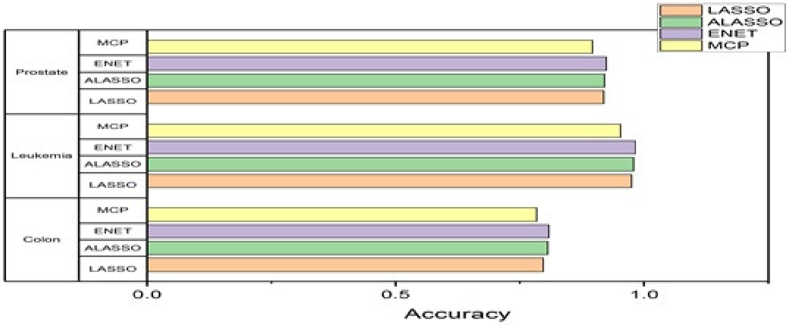

Figure 6. Average Accuracy of classification methods with RSVS method on the real data.

(AUC) compared to FS-ENET, CS-ENET, and IG-ENET in all datasets. The proposed method shows a statistically significant rise in the area under the curve (AUC) when compared to FS-ENET, CS-ENET, and IG-ENET for the training dataset.

In table 5 (*) means that the two methods have significant differences.

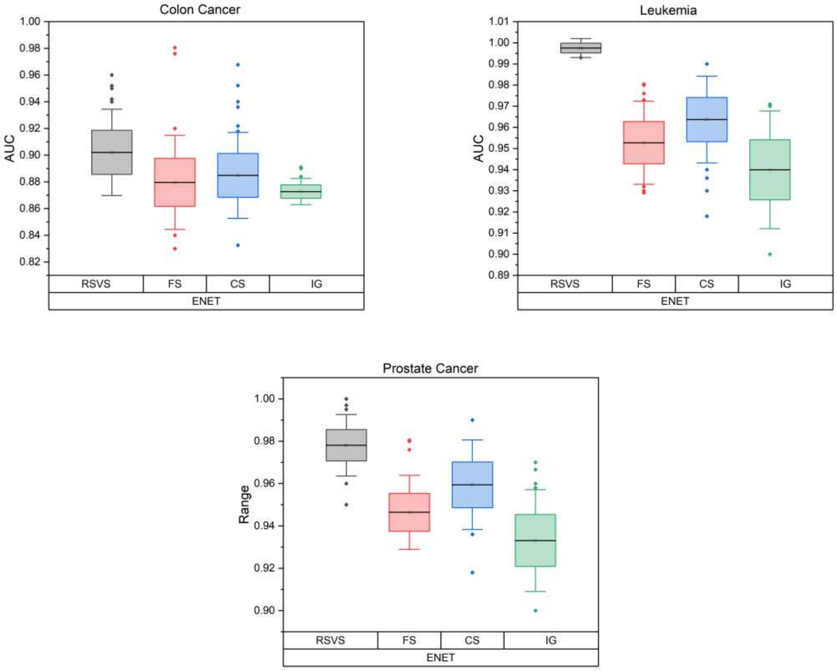

Figure 7. Boxplot of the AUC for the three datasets achieved by the four used methods of ENET classifier.

Table 5. P-values for the paired t-test of our suggested method against three other methods on three sets of real data.

Dataset	RSVS-ENET vs FS-ENET	RSVS-ENET vs CS-ENET	RSVS-ENET vs IG-ENET
Colon Cancer	0.0069(*)	0.0042(*)	0.0075(*)
Leukemia	0.0010(*)	0.0008(*)	0.0012(*)
Prostate Cancer	0.0033 (*)	0.0029(*)	0.0035(*)

7.2. The Results of Biological Research on The Most Important Genes

After the discovery of the human genome and the identification of many genes causing various types of diseases, the term molecular diagnosis appeared, which is more accurate than a clinical diagnosis in determining the disease because molecular diagnosis does not determine the disease by symptoms as much as it targets the root cause of this disease by examining and testing the genes in which the mutations causing the disease occur even before the onset of symptoms[45].

The completeness of the human genome sequence and the knowledge of all genes would lead to a qualitative breakthrough in bioinformatics. It is expected that by the next few years, in many developed countries, molecular diagnostics will become a standard practice through which blood samples are taken to extract the DNA sequence, and then this material is examined to determine the likelihood of certain types of diseases. Also, molecular diagnostics is carried out in many cases through two stages: the first is the extraction of genes and the level of their expression; the second is the use of computer algorithms that represent intelligent statistical and software procedures to deal with the big data generated by these genes. Based on this principle, we have introduced a number of intelligent techniques and statistical treatments represented by these techniques in order to try to improve the performance of the molecular diagnostic process and give good results. By identifying the most important genes responsible for the disease, which will provide a roadmap for doctors in diagnosing these diseases, early screening for these cancers may give a preliminary diagnosis in the event of genetic mutations in patients likely to be affected, and by paying attention to the important moral genes that appeared in the study as an assistant

to specialists and those interested in this field. These qualitative diagnostic tests enable the early detection of a wide range of pathological conditions, and the forthcoming advancements in medicine will significantly enhance the accuracy, confidence, efficiency, and preventive measures associated with diagnosing genetic diseases.
It is of interest to conduct further validation to ascertain the biological significance of the genes that have been identified as the most probable. The approach described in this study, namely RSVS using classifier methods, was implemented for each route. Tables 6, 7, and 8 provide a comprehensive overview of the gene symbols and their corresponding frequencies as chosen for analysis. The identified genes for colon cancer, leukemia, and prostate cancer exhibit consensus in their selection and demonstrate biological relevance to their respective cancer types, as shown by several studies [38, 39, 40, 41, 42, 43, 44].

Table 6. 8-Top significant genes selected for the colon dataset by the proposed method

variable selection with no. of appeared of all classifiers	Gene symbol and Pathway name( )
x765(4)	AJUBA (Hippo signaling)
x377(4)	TEAD4 (Hippo signaling)
x249(2)	TCF7 (Hippo signaling)
x1772(4)	IL8 (cytokine–cytokine receptor)
x1870(3)	CXCL2 (chemokine signaling)
x66(4)	CXCL1 (TNF signaling)
x493(2)	CD44 (microRNAs in Cancer)
x1423(3)	MMP3 (TNF signaling)

*( ) The shared genes (variables) present in all four gene lists that have been graded.

Table 7. 8-Top significant genes selected for the leukemia dataset by the proposed method

variable selection with no. of appeared of all classifiers	Gene symbol and Pathway name( )
x5039(4)	KIT (Hematopoietic cell lineage)
x461 (3)	RFXAP (Primary immunodeficiency)
x3320(4)	CD19 (Hematopoietic cell lineage)
x6539(3)	CD3E (Primary immunodeficiency)
x3847(3)	ITGA2B (Hematopoietic cell lineage)
x4847(4)	LCK (Primary immunodeficiency)
x2020(3)	CD8(B cell receptor signaling pathway)
x1779(2)	DNTT(Hematopoietic cell lineage)

Table 8. 8-Top significant genes selected for the prostate dataset by the proposed method

variable selection with no. of appeared of all classifiers	Gene symbol and Pathway name( )
x6185(4)	ALDH3A2 Pentose and glucuronate interconversions)
x10234(4)	GUSB (Pentose and glucuronate interconversions)
x8965(3)	GSTP1 (Glutathione metabolism)
x5890(4)	UGP2(Pentose and glucuronate interconversions)
x7623(4)	CFD (Complement and coagulation cascades)
x9172(3)	XYLB(Pentose and glucuronate interconversions)
x11858(3)	ALDH2(Pentose and glucuronate interconversions)
x9850(2)	ODC1 (Glutathione metabolism)

8. Conclusion

In the domain of genomics, a considerable quantity of genes often surpasses the size of the sample under investigation. From a biological perspective, it is noteworthy to mention that only a restricted number of these genes demonstrate a direct correlation with the condition under investigation. This study introduces the suggested approach, referred to as the RSVS method, which has been created with the purpose of filtering high-dimensional data that is relevant to the field of bioinformatics. The RSVS approach has superior effectiveness in identifying relevant factors when compared to other competing RVS methods, as shown by comprehensive simulation tests. In order to demonstrate the efficacy of classification, several classifiers, including LASSO, ALASSO, ENET, and MCP, were used on the highest-ranked features produced from the proposed RSVS approach, in addition to other widely-used RVS methods such as FS, CS, and IG. The classifiers used in the proposed RSVS technique exhibited superior performance in terms of accuracy, AUC, and G-Mean when compared to other combinations of RVS with classifiers, as shown in both simulated and real datasets. Furthermore, the genes identified by the suggested RSV technique on the real dataset have been shown to be statistically significant in terms of their frequency in their association with cancer. It provides support to doctors and specialists in the initial diagnosis of these potentially fatal malignancies. In our future research, we will prioritize the development of a technique for the adaptation of the ENET approach through a sparse support vector machine as well as studying the possibility of building expert systems for medical diagnosis and prediction of cancers via fuzzy logic systems.

REFERENCES

1. Honrado, E., Osorio, A., Palacios, J., and Benítez, J. *Pathology and gene expression of hereditary breast tumors associated with BRCA1, BRCA2 and CHEK2 gene mutations*, Oncogene,vol. 25,no. 43, pp. 5837-5845,2006.
2. Hussein, N.A.K., and Al-Sarray, B, *Deep Learning and Machine Learning via a Genetic Algorithm to Classify Breast Cancer DNA Data*, Iraqi Journal of Science, pp. 3153-3168,2022.
3. Bhola, A., and Singh, S., *Gene selection using high dimensional gene expression data: an appraisal*, Current Bioinformatics,vol. 13,no. 3, pp. 225-233,2018.
4. Bourgon, R., Gentleman, R., and Huber, W. *Independent filtering increases detection power for high-throughput experiments*, Proceedings of the National Academy of Sciences,vol. 107, no. 21, pp. 9546-9551,2010.
5. Kim, S., and Kim, J.-M. *Two-stage classification with sis using a new filter ranking method in high throughput data*, Mathematics,vol. 7,no. 6, pp. 493,2019.
6. Algamal, Z.Y., and Lee, M.H. *Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification*, Expert Systems with Applications,vol. 42,no. 23, pp. 9326-9332,2015.
7. Piao, Y., Piao, M., Park, K., and Ryu, K.H. *An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data*, Bioinformatics,vol. 28, no. 24, pp. 3306-3315,2012.
8. Chen, W., and Angelia, S. *Classification consistency analysis for bootstrapping gene selection*, Journal of Statistics,vol. 39, pp. 7270-7280,2013.
9. Sun, J., Wu, Q., Shen, D., Wen, Y., Liu, F., Gao, Y., Ding, J., and Zhang, J. *TSLRF: two-stage algorithm based on least angle regression and random forest in genome-wide association studies*, Scientific reports,vol. 9, no. 1, pp. 18034,2019.
10. Algamal, Z.Y., and Lee, M.H. *A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification*, Advances in data analysis and classification,vol. 13,no. 3, pp. 753-771,2019.
11. Ross, Q.J. *C4. 5: programs for machine learning*, San Mateo, CA, 1993.
12. Guyon, I., and Elisseeff, A. *An introduction to variable and feature selection*, Journal of machine learning research,vol. 3, (Mar), pp. 1157-1182,2003.
13. UM, O. *Estimating the Fisher's Scoring Matrix Formula from Logistic Model*, American Journal of Theoretical and Applied Statistics,2013.
14. Fan, J., and Lv, J. *Sure independence screening for ultrahigh dimensional feature space*, Journal of the Royal Statistical Society Series B: Statistical Methodology,vol. 70, no. 5, pp. 849-911,2008.
15. Hastie, T., Tibshirani, R., Friedman, J.H., and Friedman, J.H. *The elements of statistical learning: data mining, inference, and prediction*, Springer, 2009.
16. Tibshirani, R. *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society Series B: Statistical Methodology,vol. 58,no. 1, pp. 267-288,1996.
17. Marquardt, D.W., and Snee, R.D. *Ridge regression in practice*, he American Statistician,vol. 29, no. 1, pp. 3-20,1975.
18. Wang, Y., Yang, X.-G., and Lu, Y. *Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information*, Applied Mathematical Modelling,vol. 71, pp. 286-297,2019.
19. Fan, J., and Li, R. *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American statistical Association,vol. 96, no. 456, pp. 1348-1360,2001.
20. Zhang, C.-H. *Nearly unbiased variable selection under minimax concave penalty*, Annals of Statistics,2010.
21. Song, L., Smola, A., Gretton, A., Borgwardt, K.M., and Bedo, J. *Supervised feature selection via dependence estimation*, CML, pp. 823-830,2007.
22. Pabitra Mitra, C. A. Murthy, and Sankar K. Pal. *Unsupervised feature selection using feature similarity*, IEEE Trans. Pattern Anal. Mach. Intell., 24:301–312, 2002.
23. Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S., and Moore, J.H. *Relief-based feature selection: Introduction and review*, Journal of biomedical informatics,vol. 85, pp. 189-203,2018.
24. Duda, R.O., Hart, P.E., and Stork, D.G. *Solution Manual to accompany: Pattern Classification*, second edition, 2000.

25. Mahdi, G.J., and Salih, O.M. *Variable Selection Using aModified Gibbs Sampler Algorithm with Application on Rock Strength Dataset*, Baghdad Science Journal,vol. 19, no. 3, pp. 0551-0559,2022.
26. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., and Liu, H. *Feature selection: A data perspective*, ACM computing surveys (CSUR),vol. 50, no. 6, pp. 1-45,2017.
27. Abd Algafore, H.A., and Hashem, S.H. *Spam filtering based on naïve Bayesian with information gain and ant colony system*, Iraqi Journal of Science, pp. 719-727,2016.
28. Dash, R. *A two stage grading approach for feature selection and classification of microarray data using Pareto based feature ranking techniques: A case study*, Journal of King Saud University-Computer and Information Sciences,vol. 32, no. 2, pp. 232-247,2020.
29. Al-Tai, A.A., and Al-Kazaz, Q.N.N. *Semi parametric Estimators for Quantile Model via LASSO and SCAD with Missing Data*, Journal of Economics and Administrative Sciences,vol. 28, no. 133, pp. 82-96,2022.
30. Zou, H. and T. Hastie. *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society Series Bvol. 67,pp. 301-320,2005.
31. Patil, A.R., Chang, J., Leung, M.-Y., and Kim, S. *Analyzing high dimensional correlated data using feature ranking and classifiers*, Computational and Mathematical Biophysics,vol. 7,no. 1, pp. 98-120,2019.
32. Sokolova, M., Japkowicz, N., and Szpakowicz, S. *Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation*, In Australian Conference on Artificial Intelligence,2006.
33. Hosmer, D., and Lemeshow, S. *Applied Logistic Regression*, 2nd edition, Johnson Wiley and Sons, New York,2000.
34. Pi, L., and Halabi, S. *Combined performance of screening and variable selection methods in ultra-high dimensional data in predicting time-to-event outcomes*, Diagnostic and prognostic research,vol 2,no. 1, pp. 1-12,2018.
35. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., and Caligiuri, M.A. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, science,vol. 286, no. 5439, pp. 531-537,1999.
36. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Proceedings of the National Academy of Sciences,vol. 96, no. 12, pp. 6745-6750,1999.
37. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., and Richie, J.P. *Gene expression correlates of clinical prostate cancer behavior*, Cancer cell,vol. 1, no. 2, pp. 203-209,2002.
38. Yang, W., Ma, J., Zhou, W., Li, Z., Zhou, X., Cao, B., Zhang, Y., Liu, J., Yang, Z., and Zhang, H *Identification of hub genes and outcome in colon cancer based on bioinformatics analysis*, Cancer Management and Research, pp. 323-338,2018.
39. Shukir, F.S *Class Prediction Methods Applied to Microarray Data for Classification*, Iraqi Journal of Science,vol. 53, no.4, pp. 1193-1206,2012.
40. Chen, Y., Wang, L., Li, L., Zhang, H., and Yuan, Z. *'Informative gene selection and the direct classification of tumors based on relative simplicity*, BMC bioinformatics,vol. 17, no. 1, pp. 1-16,2016.
41. Mao, Z., Cai, W., and Shao, X. *Selecting significant genes by randomization test for cancer classification using gene expression data*, Journal of biomedical informatics,vol. 46, no. 4, pp. 594-601,2013.
42. Han, B., Li, L., Chen, Y., Zhu, L., and Dai, Q. *A two step method to identify clinical outcome relevant genes with microarray data*, Journal of Biomedical Informatics,vol. 44, no. 2, pp. 229-238,2011.
43. Liang, Y., Liu, C., Luan, X.-Z., Leung, K.-S., Chan, T.-M., Xu, Z.-B., and Zhang, H. *Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification*, BMC bioinformatics,vol 14, no. 1, pp. 1-12,2013.
44. Wang, S.-L., Li, X., Zhang, S., Gui, J., and Huang, D.-S. *Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction*, Computers in Biology and Medicine,vol. 40, no. 2, pp. 179-189,2010.
45. Cheung, K., Ma, H., Tse, F., Yeung, F., Tsang,F., Chu, M., Kan, M., Cho, S., Ng, W. and Chan, C. *The applications of metabolomics in the molecular diagnostics of cancer*, Expert review of molecular diagnostics, vol. 19, pp. 785-793,2019.