ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 14 (2024) No. 6, pp. 359-379
DOI:10.33168/JSMS.2024.0623

Gene Selection and Cancer Type Prediction via Filtering and Weighted Regularized Logistic Regression

Omar Q. Alshebly, Suhail N. Abdullah

Department of Statistics, College of Administration and Economics, University of Baghdad, Baghdad, Iraq

omarqusay@uomosul.edu.iq (Corresponding Author)

Abstract. The technological advancements in recent years have led to high-dimensional data becoming a prominent focus of research in genetics, bioinformatics, and biostatistics. This study develops a two-step approach using regularized logistic regression for cancer classification and gene selection with high-dimensional gene expression data. The method combines sure independence screening (SIS) using filtering methods for initial gene selection, followed by an adaptive LASSO (ALASSO) technique with a proposed weighting scheme. The model was applied to three cancer gene expression datasets - colon, leukemia and prostate. The ALASSO with the Fisher score filter demonstrated the best performance, achieving classification accuracy up to 98.2% and AUC of 96.7% for leukemia data. The top genes selected were biologically relevant to their cancer types. This study demonstrates the promise of integrating SIS filtering and weighted ALASSO for informative gene selection and accurate cancer prediction from high-dimensional gene expression data.

Keywords: Classification, Regularized logistic regression, Filtering methods, Adaptive LASSO, Gene expression data, Accuracy.

1. Introduction

The recent development of DNA microarray technology has emerged as a pivotal advancement in the domains of biology, medicine, bioinformatics, and genetics (Honrado et al., 2006; Khalid and Basad, 2022; Bhola and Singh, 2018). Microarray gene expression data is extensively utilized in the classification and detection of cancer (Algamal and Lee, 2015). The classification of normal and abnormal cellular patterns plays a crucial role in cancer diagnosis and treatment development, making it a very critical process within cancer classification (Korkmaz et al., 2014; Zhang et al., 2015). The use of cancer prediction has significant potential in enhancing patient healthcare, hence contributing to the improvement of treatment and therapy outcomes (Chen et al., 2016; Mao et al., 2013). 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 (Kim and Kim, 2019).

Statistical challenges that arise when modeling high-dimensional datasets include overfitting, computational complexities, and estimation instability. The aforementioned challenges render the implementation of statistical microarray classification techniques highly arduous (Piao et al., 2012). When trying to apply classification difficulties in a high-dimensional situation, the most typical issues include overfitting and multi-collinearity. 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 (Chen and Angelia, 2013). The selection of variables is a critical component in statistical modeling involving high-dimensional data. The primary objective is to discern a subset of significant variables from an extensive set of explanatory variables. By reducing the number of explanatory variables to only those that contain pertinent information, the objective of optimal subset selection is to enhance statistical models. It is imperative to consider not only the predictive performance but also the interpretability of this finding (Fan and Lv, 2010). In recent years, a multitude of rank-based variable selection (RVS) techniques have been developed as a means of addressing these issues. The application of RVS techniques in classification methodologies consistently improves performance by reducing data dimensionality through the exclusion of irrelevant genes (variables).

Information gain (IG), Fisher score (FS), and chi-square (CS) are only a few of the well-known RVS approaches (Ross, 1993; Okeh and Oyeka, 2013; Guyon and Elisseeff, 2003). The SIS (Sure Independent Screening) strategy, as proposed by Ross (1993), was designed to guarantee that all significant variables are retained after variable screening, with a probability that approaches one. The process of variable (gene) selection involves the identification and selection of a limited set of genes from a gene dataset characterized by high dimensionality. Within the framework of these datasets, it is frequently observed that a considerable proportion of genes are classified as irrelevant or redundant, which has the potential to weaken the effectiveness of the classification model. Hence, it is recommended to reduce the dimensionality of these datasets. Parametric models have become more prominent in high-dimensional experiments (Hastie et al., 2009). Several parametric models of regularized logistic regression (RLR), such as the least-absolute selection shrinkage operator (LASSO) (Tibshirani, 1996) and ridge (Marquardt and Snee, 1975), useℓ_1 -norm and ℓ_2-norm penalties. Also the Adaptive Lasso (ALASSO) algorithm incorporates ridge technique weights, as proposed by Zou (2006). Lasso has achieved tremendous success, but it is biased and inconsistent in its selection of significant genes. The alasso method has garnered significant attention as a potential solution to address this particular issue (Hastie et al., 2009).

The current work introduces a two-stage regularized logistic regression technique, which aims to identify a subset of genes with strong classification skills. This is achieved by combining a screening approach as a filter method and an adaptive lasso with a novel weight as an embedding method. During the first stage, a screening methodology is used to choose genes that have a strong individual association with the cancer class level. This step effectively filters out several nonessential genes and significantly decreases the computing time required for subsequent stages. During the second stage, the adaptive lasso is used concurrently as both the gene selection and model estimation technique for the most

optimal subset identified. This is achieved by including a new weight. Through this phase, the genes with the highest discriminative power are chosen. A series of comprehensive comparisons were conducted between our proposed gene selection approach and other competing methods, using various benchmark gene expression datasets. These comparisons were carried out by experimental means. The experimental findings provide evidence that the suggested methodology is very successful in the identification of genes that are most relevant, exhibiting a notable degree of accuracy in categorization.

The subsequent sections of this work are structured in the following manner: Section 2 provides the most important previous studies related to the subject of the manuscript. Section 3 provides an explanation of the filtering methods and the SIS approach. It also provides a description of regularized logistic regression, including two models. As well, the suggested method is explained, along with the algorithmic process. Furthermore, it focuses on defining the metrics used and evaluating the effectiveness of the research. The experimental findings derived from the real data from gene expression datasets are reported in Section 4. Lastly, Section 5 presents a conclusion study.

2. Literature Review

Researchers have been working hard over the past few years to figure out how to choose which variables to use in datasets with high dimensions, which have a small number of observations and a lot of variables. The prevalence of current applications in medical studies, genetic research, bioinformatics, and several other fields is primarily responsible for this occurrence. The empirical validity of many conventional approaches to variable selection has been demonstrated via extensive research conducted over the course of several years.

Generalized linear model (GLM) is known for its poor performance in high-dimensional data prediction and classification. Penalized likelihood methods, which control coefficient variances in exchange for increased bias, are attractive for this purpose. These methods shrink regression coefficients toward zero by imposing a size constraint on parameters and adding a penalty to the loss function. A growing literature is dedicated to variable selection using penalized methods in various situations.

In 1996, Tibshirani introduced LASSO, a penalizing method that uses the ℓ_1-norm instead of the $\ell_2 - \text{norm}$. This method reduces regression factors to almost zero, making models understandable. LASSO is used in various GLM models, such as logistic regression. (Park and Hastie, 2008)

Zou (2006) introduced the adaptive LASSO, a refined version of the LASSO. This method penalizes coefficients in the ℓ_1-norm using adaptive weights. The study suggests using ordinary least squares (OLS) estimates as an initial weight in the -norm penalty, but ridge regression estimates are recommended for correlated variables. However, in high-dimensional data, ordinary least squares estimates are not accessible, making the adaptive LASSO method inapplicable. (Zou, 2006)

Huang et al. (2008) investigated the asymptotic properties of adaptive LASSO estimators in sparse linear regression models with high-dimensional data, proving their oracle property by using marginal regression estimators as initial weight. (Huang et al., 2008)

Bunea (2009) and Wang et al. (2010) studied variable selection consistency in limited samples using ℓ_1-norm and ℓ_2-norm approaches in linear and logistic regression models. They found varied selection consistency under acceptable settings, while the bridge penalty technique addressed variable proliferation relative to data amount. (Bunea, 2009; Wang et al., 2010)

Zeng and Xie (2012) suggested using the SCAD penalty and ℓ_1-norm to tackle strongly correlated variables. Dong et al. (2013) proposed a composite penalization approach that combines SCAD and ridge penalties in logistic regression to address multicollinearity and benefit from desirable oracle qualities, effectively addressing multicollinearity in logistic regression. (Zeng and Xie, 2012; Dong et al., 2013)

Sequential LASSO is a further development of the LASSO method, as introduced by Luo and Chen

(2014). The underlying concept of their suggested methodology is the resolution of a series of partly penalized problems. Under a set of plausible assumptions, it has been shown that the resultant estimator has the oracle qualities. (Luo and Chen, 2014).

Wang et al. (2015) conducted an investigation of the convergence rate of the LASSO and its sparse characteristic in high-dimensional generalized linear models (GLM), given appropriate assumptions. In addition, Song and Liang (2015) introduced the concept of reciprocal LASSO as a novel approach to penalized likelihood estimation. They also demonstrated the efficacy of this technique by establishing its variable selection consistency. (Wang et al.; Song and Liang, 2015)

Jiang et al. (2015) introduced an extension of the Least Absolute Shrinkage and Selection Operator (LASSO) for Generalized Linear Models (GLM), offering superior enhancements in asymptotic theory. Mao (2015) introduced a penalized approach with high efficiency in linear regression models, using a weighted variant of the residual sum of squares, demonstrating consistency and asymptotic efficiency properties. (Jiang et al.; Mao, 2015)

Kock (2016) examined adaptive LASSO oracle features in both stationary and non-stationary auto-regressions. Bayesian information criteria yielded consistent variable selection. Chatterjee et al. (2015) demonstrated that the residual empirical process, contingent on adaptive LASSO, provides correct inference in high-dimensional regression scenarios where the LASSO method is unsuccessful. (Kock; Chatterjee et al., 2015)

High-dimensional data handling often faces issues with redundant and irrelevant components. To address this, two-stage machine learning frameworks, variable selection and classification, have been introduced. Variable selection techniques eliminate redundant and undesired noisy variables, reducing data dimensionality. Classification algorithms are then enhanced with notable attributes to enhance their predictive capabilities.

Elyasigomari et al. (2017) introduced the MRMR-COA-HS technique, a two-stage approach for gene selection. The technique uses MRMR feature selection to identify relevant genes, which are then integrated with a support vector machine as a classifier. The method demonstrated superior performance in microarray datasets. (Elyasigomari et al., 2017)

Kim and Kim 2019 study introduces a filter ranking method called "PF" that combines elastic net penalty with sure independence screening (SIS) based on resampling techniques. Compared to MMLR, SIS-LASSO, SIS-MCP, and SIS-SCAD with PF showed superior accuracy, AUROC, geometric mean, and true positive detection. This method was used to analyze gene expression data for colon and lung cancer. (Kim and Kim, 2019)

Sun et al. (2019) introduced a two-stage algorithm called TSLRF, which uses least angle regression and random forest to identify quantitative trait nucleotides (QTNs) associated with target traits. The method outperforms existing methods in simulation experiments and real data analyses, identifying 60 genes and multiple gene clusters. (Sun et al.,2019)

Salesi et al. (2021) introduced TAGA, a hybrid feature selection algorithm that enhances generalization power. TAGA combines a long-term memory tabu search with an asexual genetic algorithm, employing integer-encoded solution representation, a unique mutation operator, a tabu list encoding scheme, and an information theory-based fitness function. Experiments on high-dimensional datasets show TAGA outperforms conventional and state-of-the-art algorithms. (Salesi et al.,2021)

Last but not least, Qin et al. (2022) developed a two-stage feature selection framework that uses Wrapper, embedding, and filtering methods to tackle high-dimensional data challenges. The first stage uses WGCNA, random forest, and mRMR, while the second stage introduces a novel gene selection method combining Salp Swarm Algorithm and machine learning techniques. The framework consistently achieves over 97.6% accuracy on ten datasets. (Qin et al.,2022)

In summary, many researchers are focusing on supervised machine learning algorithms to identify

and predict biomarkers in cancer and disease datasets. This study aimed to improve cancer prediction accuracy by identifying significant genes responsible for a disease and improving classification accuracy. The study involved two steps: applying Sure Independence Screening (SIS) using the Fisher score, information gain, and chi-square. In addition to creating a regularized logistic regression model with new weights of ALASOO.

3. Material and Methods

3.1. Filtering Methods

There are two primary kinds of variable ranking (VR) approaches: supervised (Song et al., 2007) and unsupervised (Mitra et al., 2002). The categorization is established based on the presence or absence of class labels. Filtering techniques or procedures work well no matter what classification algorithms are used. This makes them more efficient than wrapper and embedding methods for processing. The evaluation of metrics like distance, entropy, and uncertainty allows for the determination of relevant characteristics. Numerous algorithms have been developed. One notable example is the relief approach, as described by (Urbanowicz et al., 2018), which employs a metric function based on distance. The objective of this work is to investigate several widely used filtering techniques in the field of gene expression research, including the Fisher score, information gain, and chi-square.

3.1.1. Fisher score

The Fisher Score (FS) (Hart et al., 2000) is a commonly used technique in the field of variable selection that aims to discover relevant variables from a larger collection for use in future classification tasks. Discriminative and statistical models are used to attain this purpose. The primary criteria for selecting variables is to reduce the distance between data points within the same class, known as intra-class distance, while simultaneously maximizing the distance between data points belonging to different classes, referred to as inter-class distance. The computation of the Fisher score for each variable Fj is based on the following concept:

$$\text{G-Mean}=\sqrt{\text{Sensitivity} \times \text{Specificity}} \qquad (1)$$

The value F (j) represents the computed (FS) for each variable j, class C= 0 or 1, 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.

3.1.2. The Chi-square test

The Chi-square test, often known as CS, is classified as a non-parametric test and is commonly used to establish statistical significance when examining the relationship between two categorical variables. In the pre-processing step, the "equal interval width" approach is used to transform numerical variables into their corresponding category counterparts. The methodology, often known as "equal interval width," involves dividing the dataset into q intervals, ensuring that each interval has the same width. The formula *w = (max - min)/q* is used to compute the width of an interval. Furthermore, the expressions *min + w, min + 2w, and min + (q - 1)* w are used to determine the boundaries of each respective interval (Mahdi and Salih, 2022; Li et al., 2017).

A training set comprises $x_j = (x_{j1}, ...x_{jp})$, where g represents each variable in a set of p variables. The CS score for a specific variable *g* with *r* distinct variable values can be computed: (Li et al., 2017)

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

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*.

3.1.3. Information gain

Information gain (IG) is utilized to quantify the quantity of information that a variable provides regarding a class (Ross, 1993). Information gain is frequently used in ranking variable selection (RVS) due to its high computing efficiency and simple interpretability. The presence of unrelated or inaccurate variables provides no meaningful information, whereas the inclusion of properly divisible variables provides the most information. The metric that represents the degree of purity is known as information, while the metric that represents the degree of impureness is known as entropy. The suboptimal performance of the information gain measure in the presence of duplicate variables is one of its limitations (Algafore and Hashem, 2016).

The following equation can be used to calculate the IG between the g[th] variable in x_i and the response variable y_i:

$$IG(x_i; y_i) = H(x_i) - H(x_i|y_i) \qquad (3)$$

$$H(x_i) = \sum_{i} \quad p(x_i)\, log \quad _2 \left(p(x_i)\right)$$

$$H(x_i|y_i) = \sum_{y_{i\epsilon Y}} \quad p(y_i) \sum_{x_{i\epsilon X}} \quad p(x_i|y_i)\, log \quad _2\, p(x_i|y_i)$$

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.

3.1.4. Sure Independence Screen (SIS)

The method used for the purpose of variable selection or screening is known as Sure Independence Screening. The author introduced the SIS technique as a reference (Fan and Lv, 2008). A certain approach is used to reduce the number of variables, denoted as *d*, from a substantial magnitude to a smaller subset, denoted as *m*.

Using RVS techniques, scores are assigned to individual variables based on their respective conditions. Based on the provided scores, the variables are arranged in ascending order of significance. Using the condition, which is described in detail below, the procedure for determining the variables with the highest rank is carried out: (Dash, 2020)

$$\frac{s}{log\, s} \qquad (4)$$

The variable *s* denotes the total number of samples.

3.2. Regularized Logistic Regression (RLR)

Logistic regression is widely recognized as a crucial technique in medicine and science for addressing binary classification problems, wherein the response variable is dichotomous and takes on values of either zero (0) or one (1). In various domains, including biomedical imaging, DNA microarrays, and genomics, researchers face notable challenges when employing logistic regression on datasets with high dimensions if the number of variables (p) is greater than the sample size (n).

Let $y_i \in \{0,1\}$ Create a vector of dimensions $n * 1$ consisting of tissue., and let xi be a $p * 1$ vector of variables. The vector of probability estimates undergoes a logistic transformation $\pi_i = p(y_i = 1|x_i)$ and the relationship can be represented by a linear function that has undergone a logit transformation (Algamal and Lee,2015).

$$ln[\tfrac{\pi_i}{1-\pi_i}] = \beta_0 + \sum_{j=1}^{p} \boldsymbol{x}_j{}^T \beta_j\,, \quad i = 1,2,\ldots,n \qquad (5)$$

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 defined as:

$$\boldsymbol{\ell}(\beta_0, \beta) = \sum_{i=1}^{n} \{\boldsymbol{y}_i\, ln\, \pi\, (\boldsymbol{x}_{ij}) + (1 - \boldsymbol{y}_i)\, ln(1 - \pi(\boldsymbol{x}_{ij})\}, j = 1,2,\ldots,p \qquad (6)$$

One advantageous characteristic of logistic regression is its capacity to simultaneously estimate probabilities. $\pi(\boldsymbol{x}_{ij})$ and $1-\pi(\boldsymbol{x}_{ij})$ for each class. The probability of classifying the i^{th} sample in class 1 is estimated by$\hat{\pi}_i = exp(\beta_0 + \sum_{j=1}^{p} \boldsymbol{x}_j{}^T \beta_j)/\, 1 + exp(\beta_0 + \sum_{j=1}^{p} \boldsymbol{x}_j{}^T \beta_j)$.

The regularized logistic regression (RLR) method adds a non-negative regularization part to equation (5), which makes it possible to handle gene coefficients in situations with a lot of dimensions. Tibshirani's (1996) proposal of ℓ_1-norm regularization is a commonly used regularization term in academic literature. The process of gene selection and estimation is carried out concurrently by the ℓ_1-norm regularization technique, which involves the imposition of constraints on the log-likelihood function of gene (variable) coefficients. The RLR is defined as follows:

$$RLR = \sum_{i=1}^{n}\{\boldsymbol{y}_i\, ln\, \pi\,(\boldsymbol{x}_{ij}) + (1-\boldsymbol{y}_i)\, ln(\,1-\pi(\boldsymbol{x}_{ij})\} + \lambda P(\beta). \qquad (7)$$

The vector β was estimated through the process of minimizing:

$$\hat{\beta}_{RLR} = argmin \quad _{\beta}\left[\sum_{i=1}^{n}\{\boldsymbol{y}_i\, ln\, \pi\,(\boldsymbol{x}_{ij}) + (1-\boldsymbol{y}_i)\, ln(\,1-\pi(\boldsymbol{x}_{ij})\} + \lambda P(\beta)\right] \qquad (8)$$

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, which Tibshirani (1996) introduced, is a variation of ordinary least squares that has expanded to include a wider range of applications. The primary objective of utilizing the Lasso method is to engage in variable selection and regularization techniques, thereby enhancing the predictive accuracy of the model.

When working with classification data that has a lot of dimensions, the Lasso method is very helpful because it is easy to compute. Conversely, in instances where there are strong correlations and a discernible grouping pattern among pertinent genes or variables, the Lasso method tends to exhibit a tendency to select one gene randomly while disregarding the remaining genes. Furthermore, the Lasso method exhibits inconsistent gene selection due to the uniform application of shrinkage to each gene coefficient.

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

$$\hat{\beta}_{LASSO} = argmin_{\beta}\left[\sum_{i=1}^{n}\{\boldsymbol{y}_i\, ln\, \pi\,(\boldsymbol{x}_{ij}) + (1-\boldsymbol{y}_i)\, ln(\,1-\pi(\boldsymbol{x}_{ij})\} + \lambda \sum_{j=1}^{p}|\beta_j|\right], \quad (9)$$

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 method was developed as a means of mitigating the bias inherent in the Lasso approach. The degree of bias in the model estimate is determined by the value of The Alasso method employs adaptive weights to penalize distinct coefficients in the penalty (Zou, 2006). The utilization of the weight penalty is a technique employed to decrease the bias of Lasso. This is achieved by assigning relatively lower weights to variables with larger coefficients and higher weights to variables with smaller coefficients. Conversely, variables with large coefficients are assigned lower weights. It is imperative to maintain the sparsity property of Lasso while reducing its selection bias. The regularized logistic regression employing the adaptive LASSO is mathematically defined as follows:

$$\hat{\beta}_{ALASSO} = argmin_{\beta}\left[\sum_{i=1}^{n}\{\boldsymbol{y}_i\, ln\, \pi\,(\boldsymbol{x}_{ij}) + (1-\boldsymbol{y}_i)\, ln(\,1-\pi(\boldsymbol{x}_{ij})\} + \lambda \sum_{j=1}^{p} w_j|\beta_j|\right], \qquad (10)$$

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

3.3. The Proposed Method with New Weight

The goal of gene or variable selection is to improve the performance of classification, speed up and increase accuracy on the gene selection process, and learn more about the basic classification

problem. Looking at a regularized logistic regression technique that involves ranking variables and choosing the right ones is the main goal of this study. This will improve the accuracy of classification. The study introduces a new weight formula for the adaptive Lasso technique, incorporating FS, CS, and IG filters to penalize the model effectively. It compares this method with other techniques and clarifies the weight for utilization. A desirable weight is nonnegative and equal to or greater than the weight value at the j[th] position, denoted as w_j.

The utilization of resampling methodology is employed in order to calculate the weight below:

$$\boldsymbol{w_j = \frac{average\ Regression\ Coefficient}{average\ Standard\ Erorr}}$$

where $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots\ldots, \hat{\beta}_p)^T$ represented the vector of coefficients regression estimated also this vector are consistent estimators (Asar and Genç,2016). $\hat{S} = (S_{\hat{\beta}_1}, S_{\hat{\beta}_2}, \ldots\ldots, S_{\hat{\beta}_P})^T$ is vector represents the standard error of the regression, then $w_j = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}}$ with $w_{ratio} = \left(w_1, w_2, \ldots\ldots, w_p\right)^T$ represented the ratio vector of weights, and *j=1,2,3,….,p.*

Algorithm 1: The Proposed Method with New Weight of ALASSO

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: Get a filtered training dataset that includes only the true significant variables based on the sure independence screening (SIS) approach $\frac{s}{\log s}$ using RVS methods.

Step 3: A logistic regression model is built, the regression coefficient and standard error are found, and the average regression coefficient and standard error are calculated. This gives us the new weight value, which we call $\boldsymbol{w_j = \frac{average\ Regression\ Coefficient}{average\ Standard\ Erorr}}$

Step 4: Use the new w_j in the Adaptive Lasso with filter methods.

Step 5: Utilize a testing dataset and assess different performance metrics.

The workflow of The Proposed Method with New Weight of ALASSO framework is shown in Figure 1.

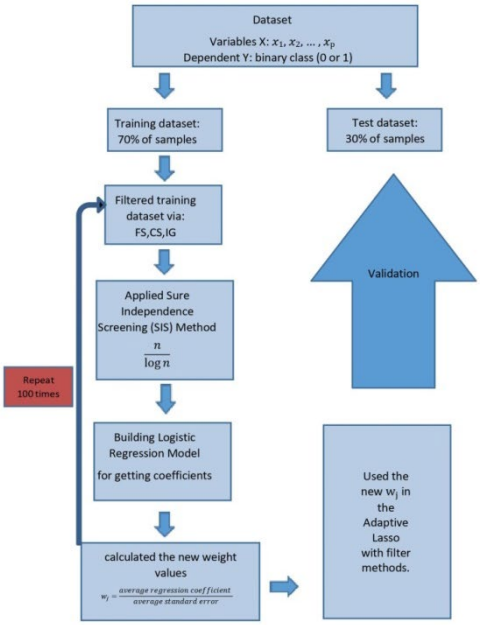

Fig.1: The workflow of the proposed method with the new weight of ALASSO

3.4. 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 (Patil et al., 2019). 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) (Sokolova et al., 2006).

$$Accuracy = \frac{TP + TN}{N}$$

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad (11)$$

Also we used geometric mean is the joint performance was evaluated by utilizing the geometric mean of sensitivity and specificity(Algamal and Lee,2015).

$$\text{G-Mean}=\sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (12)$$

And 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 calculation of the area under a curve involves the use of the following formula (Sokolova et al.,2006) :

$$A_{ROC} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} = \frac{1}{PN}\int_0^N TP * dFP \quad (13)$$

4. Result and Discussion

4.1. Data Description and Experimental Setting

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 (Alon et al., 1999), leukemia data (Golub et al., 1999), and prostate cancer (Singh et al.,2002). The datasets under consideration include a response variable that is 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. All the implementations of the study on real-data applications are carried out using R and Origin 2023b.

Table 1 shows the type of data used in the study with their sizes and classification for each case.

Table 1: 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 dataset is divided into two subsets, with 70% of the samples allocated for training and the remaining 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 and the proposed adaptive LASSO method. The model-building process uses a 10-fold cross-validation (CV) approach with 100 iterations. Finally, the testing dataset is utilized to estimate different performance metrics. The average accuracy, area under the receiver operating characteristic curve (AUROC), and geometric mean (G-mean) were estimated using the aforementioned methods. The provided dataset is preprocessed, ensuring that the gene expressions have been normalized and that there are no missing values.

The tuning parameter value, denoted as λ, was set to a fixed value between 0 and 100 for each method. Subsequently, the 10-fold cross-validation technique was used to determine the optimal value for each λ parameter. Out of the 10 folds, one fold was designated as the validation dataset, while the other nine folds were used as the training dataset for fitting the employed methods with a certain λ value. We ran the cross-validation (CV) process 100 times. Then, we found the best tuning value by looking at the smallest misclassification error that CV could find.

The histogram in Figure 2 provides a summary of the pairwise correlation. The average correlation coefficient among genes in colon cancer is 0.44. The evidence suggests a strong link between genes. The mean correlation between genes (variables) in cases of leukemia and 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. From this, we conclude that the high-dimensional data of the genetic matrix have high correlations between them, and therefore these data must be processed to extract important variables that have a direct impact on the disease and use them in the diagnosis of affected patients.

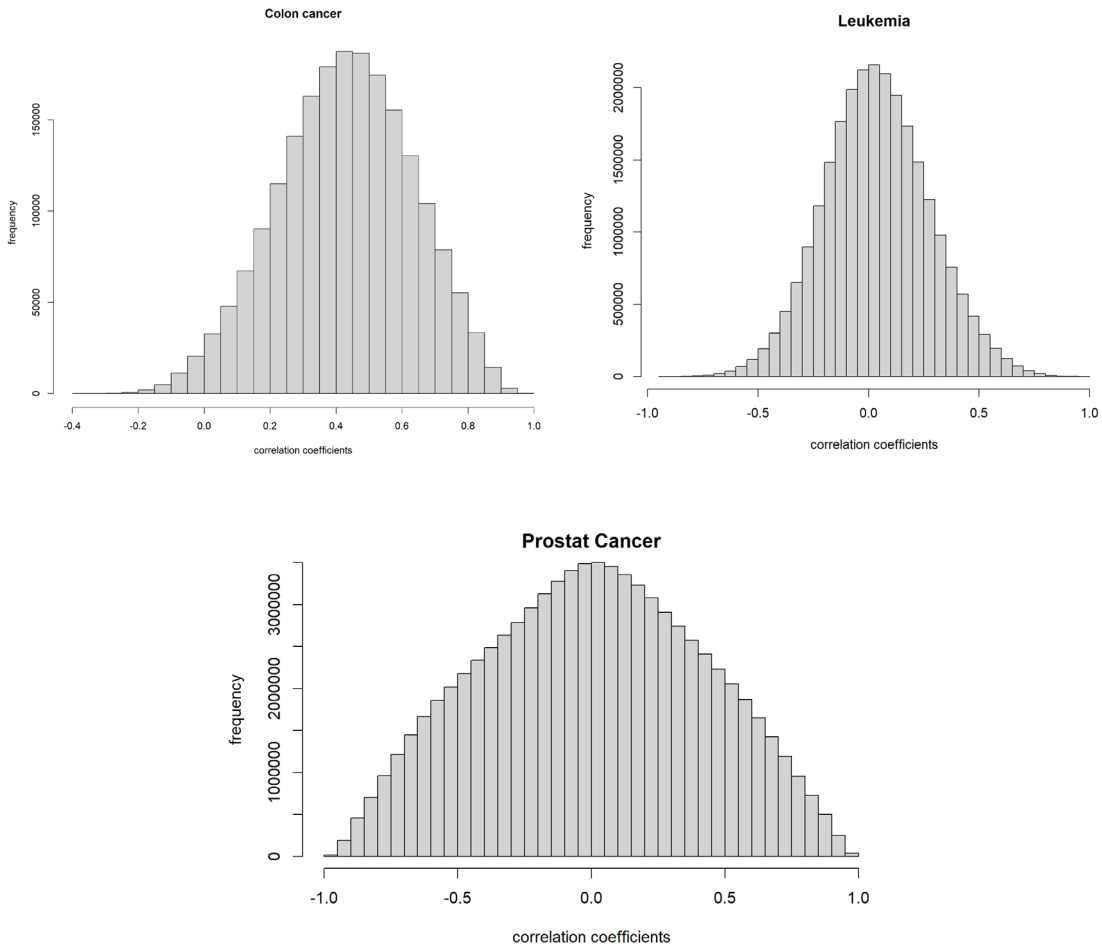

Fig.2: The histogram of pairwise correlation coefficients of a real dataset

4.2. Results of Statistical Significance Test

A paired, two-tailed t-test was used to compare the proposed method with each competing method one at a time. This was done to make sure that the proposed method worked to find highly relevant genes and make accurate classifications. The experiment was carried out by evaluating the area under the curve (AUC) of the training dataset. The adaptive Lasso classifier was used in our study with three filters, as well as the LASSO and ALASSO without filters. Also, ALASSO has weights of ridge and weights of marginal maximum likelihood. The results indicated that the FS filter produced the most advantageous outputs.

Figure 3 presents a boxplot that illustrates the area under the curve (AUC) for each method used across all three datasets. The results imply that the proposed approach's area under the curve (AUC) compares favorably to that of proposed adaptive lasso with IG (PALIG), proposed adaptive lasso with CS (PALCS), Lasso non-filtered (LNF), adaptive lasso non-filtered (ALNF), adaptive lasso with ridge weights (ALR), and adaptive lasso with marginal maximum likelihood weights (ALML).

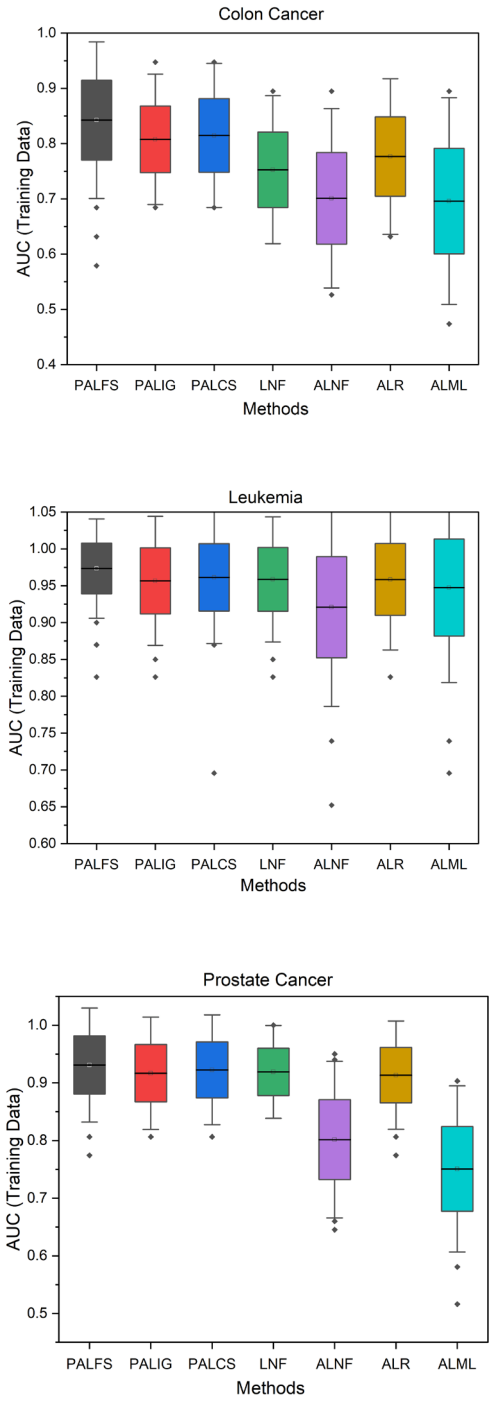

Fig.3: Boxplot of the AUC for the three datasets achieved by the seven used methods of study.

The paired two-tailed t-test findings at a specified significance level $\alpha = 0.05$ are shown in Table 2.

Table 2: P-values for the paired t-test of our suggested method against six other methods on three sets of real data.

Dataset	**PALFS Vs. PALIG**	**PALFS Vs. PALCS**	**PALFS Vs. LNF**	**PALFS Vs. ALNF**	**PALFS Vs. ALR**	**PALFS Vs. ALML**

Colon	0.017 (*)	0.015 (*)	0.0018 (*)	0.0012(*)	0.023(*)	0.0013(*)
Leukemia	0.018 (*)	0.016 (*)	0.017 (*)	0.0093(*)	0.016(*)	0.0095(*)
Prostate	0.035 (*)	0.028 (*)	0.019 (*)	0.0042(*)	0.032(*)	0.0072(*)

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

According to the results shown in Table 2, the proposed method demonstrates superior statistical performance in terms of the area under the curve (AUC) when compared to PALIG, PALCS, LNF, ALNF, ALR, and ALML across all datasets. The method described in this study demonstrates a statistically significant increase in the area under the curve (AUC) compared to all other methods used in the training dataset.

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

The term "molecular diagnosis" emerged subsequent to the identification of numerous genes responsible for a wide range of diseases and the discovery of the human genome. This term is more accurate than clinical diagnosis when it comes to figuring out what disease someone has because it doesn't depend on symptoms. Instead, it looks into and tests the genes that have the disease-causing mutations before the symptoms show up (Cheung et al., 2019).

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 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 SIS-ALASSO with new weights, was implemented for each route. Tables 3, 4, and 5 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 (Yang et al., 2018; Shukir, 2012; Chen et al., 2016; Mao and Shao, 2013; Han et al., 2011; Liang et al., 2013; Wang et al., 2010).

Table 3: The top significant genes selected for the colon dataset by the proposed method with filters

Variable selection with no. of appeared of proposed ALASSO with filters( )	Gene symbol
X66(3)	AJUBA
X377(3)	TEAD4
X249(2)	TCF7
X1772(3)	IL8
X1870(2)	CXCL2
X765(3)	CXCL1
X493(2)	CD44
X1423(3)	MMP3

*( ) The shared genes (variables) present in all three filter methods.

Table 4: The top significant genes selected for the leukemia dataset by the proposed method with filters

Variable selection with no. of appeared of proposed ALASSO with filters( )	Gene symbol
X5039(3)	KIT
X461 (3)	RFXAP
X3320(3)	CD19
X6515(2)	CD3E
X3847(2)	ITGA2B
X4847(3)	LCK
X2020(2)	CD8
X3208(2)	DNTT

Table 5: The top significant genes selected for the prostate dataset by the proposed method with filters

Variable selection with no. of appeared of proposed ALASSO with filters( )	Gene symbol
X6185(3)	ALDH3A2
X10234(3)	GUSB
X8965(2)	GSTP1
X5890(3)	UGP2
X7623(3)	CFD
X9172(2)	XYLB
X11858(2)	ALDH2
X9850(2)	ODC1

Figure 4 displays the correlation matrix of significant genes that were chosen in this investigation using the suggested methodology across the three real data sets examined in this study. The observed correlations between genes had mostly high values, with a subset displaying moderate values. This study of the chosen genes builds a strong scientific foundation in the medical field, which makes it easier to look into how these important genes are correlated.

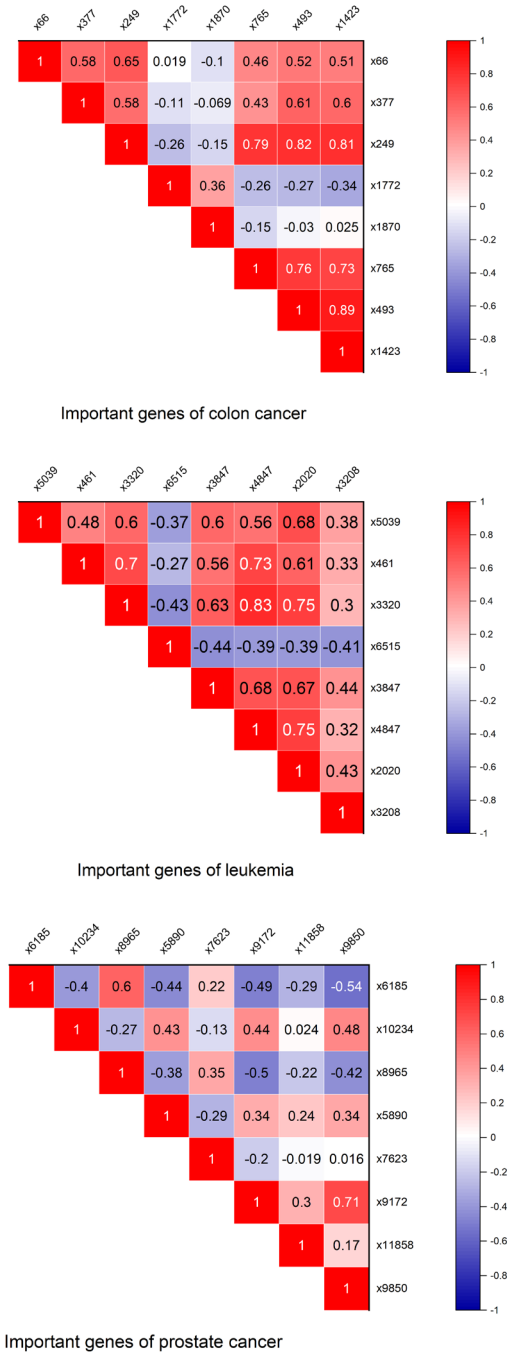

Fig.4: The correlation matrix among the most important selected genes of the real dataset

4.4. Performance of Classification

Table 6 presents the computed values of accuracy, area under the curve (AUC), and geometric mean (GM-mean) for all the methods. We conducted an analysis using real data comprising three types of cancer, depending on gene expression data.

Table 2: Classifiers evaluation a real testing dataset with an average of 100 iterations.

Cancer data	Methods	Accuracy(SD)	AUROC(SD)	G-Mean(SD))
Colon	PALFS	**0.821(0.071)**	**0.782(0.091)**	**0.725(0.102)**
	PALIG	0.788(0.084)	0.743(0.101)	0.666(0.190)
	PALCS	0.801(0.081)	0.744(0.102)	0.669(0.210)
	LNF	0.743(0.101)	0.675(0.113)	0.554(0.253)
	ALNF	0.693(0.095)	0.657(0.116)	0.570(0.244)
	ALR	0.769(0.092)	0.721(0.104)	0.672(0.157)
	ALML	0.688(0.104)	0.654(0.116)	0.574(0.216)
Leukemia	PALFS	**0.982(0.040)**	**0.967(0.051)**	**0.961(0.057)**
	PALIG	0.959(0.058)	0.948(0.083)	0.942(0.098)
	PALCS	0.970(0.048)	0.962(0.057)	0.957(0.065)
	LNF	0.961(0.057)	0.957(0.072)	0.953(0.083)
	ALNF	0.940(0.080)	0.918(0.114)	0.903(0.156)
	ALR	0.967(0.051)	0.947(0.083)	0.941(0.095)
	ALML	0.952(0.067)	0.946(0.085)	0.941(0.097)
Prostate	PALFS	**0.923(0.044)**	**0.918(0.045)**	**0.915(0.046)**
	PALIG	0.914(0.048)	0.904(0.045)	0.902(0.047)
	PALCS	0.919(0.048)	0.915(0.046)	0.905(0.047)
	LNF	0.901(0.050)	0.909(0.047)	0.907(0.048)
	ALNF	0.791(0.082)	0.793(0.078)	0.778(0.130)
	ALR	0.905(0.048)	0.906(0.048)	0.904(0.049)
	ALML	0.744(0.079)	0.741(0.079)	0.739(0.080)

* PALFS: Proposed adaptive lasso with FS, PALIG: Proposed adaptive lasso with IG, PALCS: Proposed adaptive lasso with CS, LNF: Lasso non-filtered, ALNF: Adaptive lasso non-filtered, ALR: adaptive lasso with weights of ridge, ALML: adaptive lasso with weights of marginal maximum likelihood, (SD) standard deviation.

It's clear from Table 2 that the classifiers work much better on all three real datasets when the suggested new weight of ALASSO is added to the FS, CS, and IG filters. This is shown by the accuracy, AUC, and G-mean. For colon cancer data, it is apparent that the application of different methodologies in adaptive lasso, such as the utilization of new weight and filter techniques, consistently produced outcomes that demonstrated superior performance in terms of accuracy, G-Mean, and AUC. We found that the proposed SIS-ALASSO worked better than other methods in terms of accuracy (82.1%), G-mean (72.5%), and AUROC (78.2%). On the other hand, for leukemia, it is clear that the methods combining a filter consistently exhibit superior performance in terms of accuracy, G-mean, and AUC. The proposed SIS-ALASSO with new weights worked better than other methods in terms of accuracy (98.2%), G-Mean (96.1%), and AUROC (96.7%). As well as, the PALFS method has high performance in classifying prostate cancer across all 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.

The boxplot in figure 5 shows the average accuracy of the testing dataset to the real data used in this study.

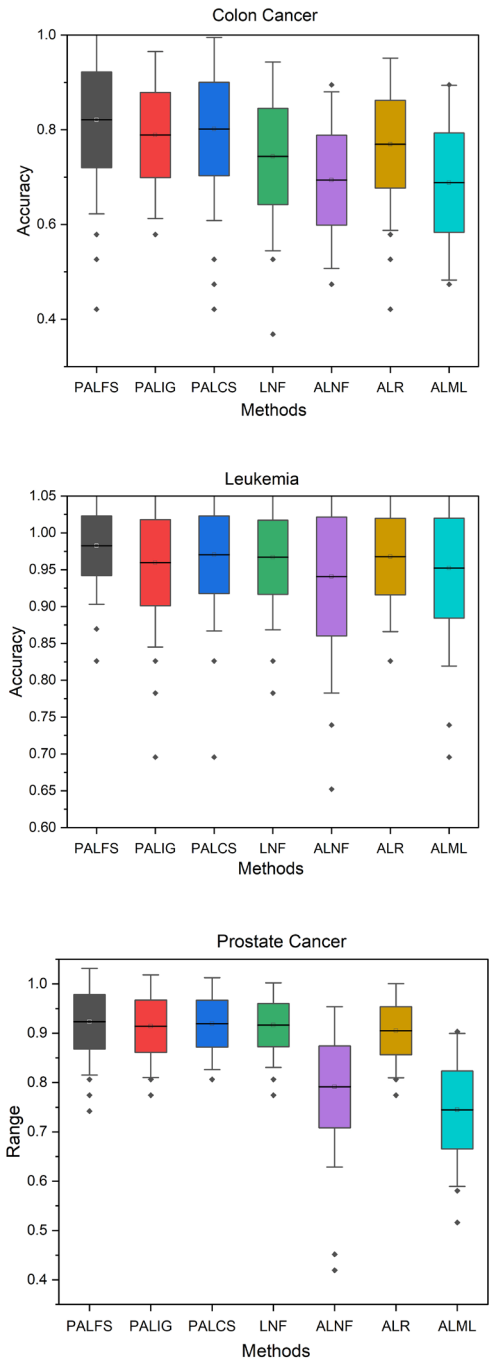

Fig.5: The boxplot of the average accuracy of the different RVS methods with classifiers on the real dataset.

5. Conclusion

This study proposed an integrated two-step approach using SIS filtering and weighted ALASSO for gene selection and cancer classification with high-dimensional gene expression data. The method leverages the strengths of both filter-based screening and regularized regression for gene selection. Experimental results on three cancer datasets demonstrate superior and robust classification performance compared to standard LASSO and adaptive LASSO variants. The SIS-ALASSO model with the Fisher score filter achieved best results, with up to 98.2% accuracy for leukemia data. The top genes identified also showed biological relevance to their cancer types. The study provides a promising framework for utilizing regularized regression and filtering techniques together for informative gene

selection from high-dimensional data. Future work can validate the approach on larger gene expression datasets and for other cancer prediction tasks.

Acknowledgements

We thank the editor and the referees for reading the article very carefully and making a number of valuable and kind comments that improve the presentation of the article.

References

Abd Algafore, H. A. & Hashem, S. H. (2016). Spam filtering based on naïve Bayesian with information gain and ant colony system. *Iraqi Journal of Science*, 719-727.

Algamal, Z. Y. & Lee, M. H. (2015). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, 42, 9326-9332.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. & Levine, A. J. (1999). 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*, 96, 6745-6750.

Asar, Y. & Genç, A. (2016). New Shrinkage Parameters For The Liu-Type Logistic Estimators. *Communications In Statistics-Simulation And Computation*, 45, 1094-1103.

Bhola, A. & Singh, S. (2018). Gene selection using high dimensional gene expression data: an appraisal. *Current Bioinformatics*, 13, 225-233.

Bunea, F. (2009). Honest variable selection in linear and logistic regression models via L1 and L1+L2 penalization. *Electronic Journal of Statistics*. 2, 1153 – 1194.

Chatterjee, A., Gupta, S. and Lahiri, S. N. (2015). On the residual empirical process based on the ALASSO in high dimensions and its functional oracle property. *Journal of Econometrics*. 186, 317 – 324.

Chen, W. & Angelia, S. (2013). Classification consistency analysis for bootstrapping gene selection. *Journal of Statistics*, 39, 7270-7280.

Chen, Y., Wang, L., Li, L., Zhang, H. & Yuan, Z. (2016). Informative gene selection and the direct classification of tumors based on relative simplicity. *BMC bioinformatics*, 17, 1-16.

Cheung, P. K., Ma, M. H., Tse, H. F., Yeung, K. F., Tsang, H. F., Chu, M. K. M., Kan, C. M., Cho, W. C. S., Ng, L. B. W. & Chan, L. W. C. (2019). The applications of metabolomics in the molecular diagnostics of cancer. *Expert review of molecular diagnostics,* 19, 785-793.

Dong, Y., Song, L., Wang, M. and Xu, Y. (2013). Combined-penalized likelihood estimations with a diverging number of parameters. *Journal of Applied Statistics*.41, 1274 – 1285.

Elyasigomari, V., Lee, D., Screen, H. R. & Shaheed, M. H. (2017). Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. *Journal of biomedical informatics*, 67, 11-20.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70, 849-911.

Fan, J. & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality, Ann Stat ,38,3567–3604.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R. & Caligiuri, M. A. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286**,** 531-537.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3**,** 1157-1182.

Han, B., Li, L., Chen, Y., Zhu, L. & Dai, Q. (2011). A two step method to identify clinical outcome relevant genes with microarray data. Journal of Biomedical Informatics, 44**,** 229-238.

Hart, P. E., Stork, D. G. & Duda, R. O. (2000). *Pattern classification*, Wiley Hoboken.

Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction, Springer.

Honrado, E., Osorio, A., Palacios, J. & Benítez, J. (2006). Pathology and gene expression of hereditary breast tumors associated with BRCA1, BRCA2 and CHEK2 gene mutations. *Oncogene*, 25**,** 5837-5845.

Huang, J., Ma, S. & Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica***,** 1603-1618.

Ing, C.-K. & Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica***,** 1473-1513.

Jiang, Y., He, Y. and Zhang, H. (2015). Variable Selection with Prior Information for Generalized Linear Models via the Prior LASSO Method. *Journal of the American Statistical Association*, Published online.

Kim, S. & Kim, J.-M. (2019). Two-stage classification with sis using a new filter ranking method in high throughput data. Mathematics, 7**,** 493.

Kock, A. B. (2016). Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. *Econometric Theory,* 32**,** 243-259.

Korkmaz, S., Zararsiz, G. & Goksuluk, D. (2014). Drug/nondrug classification using support vector machines with various feature selection strategies. *Computer methods and programs in biomedicine,* 117**,** 51-60.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50**,** 1-45.

Liang, Y., Liu, C., Luan, X.-Z., Leung, K.-S., Chan, T.-M., Xu, Z.-B. & Zhang, H. (2013). Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. BMC bioinformatics, 14**,** 1-12.

Luo, S. and Chen, Z. (2014). Sequential lasso cum EBIC for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*. 109, 1229 − 1240.

Mao, G. (2015). Efficient penalized estimation for linear regression model. *Communications in Statistics - Theory and Methods*. 44, 1436 − 1449.

Mao, Z., Cai, W. & Shao, X. (2013). Selecting significant genes by randomization test for cancer classification using gene expression data. *Journal of biomedical informatics*, 46**,** 594-601.

Marquardt, D. W. & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, 29**,** 3-20.

Mitra, P., Murthy, C. & Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, 24**,** 301-312.

Noor Alhuda Khalid, H. & Basad, A.-S. (2022). Deep Learning and Machine Learning via a Genetic Algorithm to Classify Breast Cancer DNA Data*. Iraqi Journal of Science* , 63**,** 3153-3168.

Okeh, U. & Oyeka, I. (2013). Estimating the fisher's scoring matrix formula from logistic model. *Am. J. Theor. Appl. Stat,* 2**,** 221-227.

Park, M. Y. & Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics,* 9**,** 30-50.

Patil, A. R., Chang, J., Leung, M.-Y. & Kim, S. (2019). Analyzing high dimensional correlated data using feature ranking and classifiers. *Computational and Mathematical Biophysics*, 7**,** 98-120.

Piao, Y., Piao, M., Park, K. & Ryu, K. H. (2012). An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. *Bioinformatics*, 28**,** 3306-3315.

Qin, X. Zhang, S. Yin, D. Chen, D. Dong,X..(2022). Two-stage feature selection for classification of gene expression data based on an improved Salp Swarm Algorithm[J]. *Mathematical Biosciences and Engineering*, 19(12): 13747-13781.

Ross, Q. J. (1993). C4. 5: programs for machine learning. San Mateo, CA.

Salesi, S., Cosma, G. & Mavrovouniotis, M. (2021). TAGA: Tabu Asexual Genetic Algorithm embedded in a filter/filter feature selection approach for high-dimensional data. *Information Sciences*, 565**,** 105-127.

Shukir, F. S. (2012). Class Prediction Methods Applied to Microarray Data for Classification. *Iraqi Journal of Science*, 53**,** 1193-1206.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'amico, A. V. & Richie, J. P. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer cell, 1**,** 203-209.

Sokolova, M., Japkowicz, N. & Szpakowicz, S.(2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. *Australasian joint conference on artificial intelligence, Springer*, 1015-1021.

Song, L., Smola, A., Gretton, A., Borgwardt, K. M. & Bedo, J.(2007) Supervised feature selection via dependence estimation. *Proceedings of the 24th international conference on Machine learning*, 2007. 823-830.

Song, Q. & Liang, F. (2015). High-dimensional variable selection with reciprocal l 1-regularization. *Journal of the American Statistical Association,* 110**,** 1607-1620.

Sun, J., Wu, Q., Shen, D., Wen, Y., Liu, F., Gao, Y., Ding, J. & Zhang, J. (2019). TSLRF: two-stage algorithm based on least angle regression and random forest in genome-wide association studies. *Scientific reports*, 9**,** 18034.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58**,** 267-288.

Wang, L., You, Y. & Lian, H. (2015). Convergence and sparsity of Lasso and group Lasso in high-dimensional generalized linear models. *Statistical Papers,* 56**,** 819-828.

Wang, M., Song, L. and Wang, X. (2010). Bridge estimation for generalized linear models with a diverging number of parameters. *Statistics & Probability Letters*.80, 1584 – 1596.

Wang, S.-L., Li, X., Zhang, S., Gui, J. & Huang, D.-S. (2010). Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. *Computers in Biology and Medicine*, 40**,** 179-189.

Yang, W., Ma, J., Zhou, W., Li, Z., Zhou, X., Cao, B., Zhang, Y., Liu, J., Yang, Z. & Zhang, H. (2018). Identification of hub genes and outcome in colon cancer based on bioinformatics analysis. Cancer Management and Research**,** 323-338.

Zeng, L. and Xie, J. (2012). Group variable selection via SCAD-L2. *Statistics.* 48,49 ‑ 66.

Zhang, L., Qian, L., Ding, C., Zhou, W. & Li, F. (2015). Similarity-balanced discriminant neighbor embedding and its application to cancer classification based on gene expression data. *Computers in biology and medicine,* 64**,** 236-245.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101**,** 1418-1429.