An Additive Sparse Logistic Regularization Method for Cancer Classification in Microarray Data

Vijay Suresh Gollamandala[1] and Lavanya Kampa[1,2]
[1]Department of Computer Science and Engineering, Lakireddy Bali Reddy College of Engineering, India
[2]Department of Information Technology, Lakireddy Bali Reddy College of Engineering, India

Abstract: *Now a day's cancer has become a deathly disease due to the abnormal growth of the cell. Many researchers are working in this area for the early prediction of cancer. For the proper classification of cancer data, demands for the identification of proper set of genes by analyzing the genomic data. Most of the researchers used microarrays to identify the cancerous genomes. However, such kind of data is high dimensional where number of genes are more compared to samples. Also the data consists of many irrelevant features and noisy data. The classification technique deal with such kind of data influences the performance of algorithm. A popular classification algorithm (i.e., Logistic Regression) is considered in this work for gene classification. Regularization techniques like Lasso with* L_1 *penalty, Ridge with* L_2 *penalty, and hybrid Lasso with* $L_{1/2}+2$ *penalty used to minimize irrelevant features and avoid overfitting. However, these methods are of sparse parametric and limits to linear data. Also methods have not produced promising performance when applied to high dimensional genome data. For solving these problems, this paper presents an Additive Sparse Logistic Regression with Additive Regularization (ASLR) method to discriminate linear and non-linear variables in gene classification. The results depicted that the proposed method proved to be the best-regularized method for classifying microarray data compared to standard methods.*

Keywords: *Microarray data, sparse regularization, feature selection, logistic regression, and lasso.*

Received April 30, 2020; accepted September 17, 2020
https://doi.org/10.34028/iajit/18/2/10

1. Introduction

In the area of genome research, the most important task is the classification of cancer based on the gene expression. The most popular classification technique is Logistic Regression provides an emphatic statistical depiction on cancer data. In the gene expression research, the number of genes under this study is more than the sample size. This is known as High dimensional-low sample size problem. To handle such kind of problem, applied technique which is regularization method with penalty. A popular regularization method is the penalty [15] which is, Least Absolute Shrinkage and Selection Operator (LASSO). This method can do reduction of features and gene selection at the same time. The other method related to this Adaptive Lasso (ALASSO) [26], assigns dynamic weights to the coefficients in the penalty. In the Logistic regression [10], may lead to bias with absence of future selection at predicting parameters. Another method [21] introduced penalty which could be considered as a symbol penalties. This is advantageous with respect to its sparsity and also computationally more efficient. The main characteristics of penalty are unbiasedness and oracle properties [23]. The penalty fails in the process of dealing with data, which consists of dependent features. It cannot identify the correlation between the features. Another drawback is penalty can select a single variable from a group of variables in which the pair wise correlations are high. Due to this, task relevant necessary data may lose which causes the improper classification. Many authors published their work on sparse regularization methods like Group Lasso [17], and Elastic Net [25]. But all the above stated techniques were restricted to the dimensionality reduction and feature selection based on the parametric methods. This paper focuses on regularization penalty with additive models. This model helps to smoothen the regression parameters. It could also discriminate the linear and nonlinear variables for removing the irrelevant variables. The proposed method tested over real microarray data sets and results are promising compared to the standard regularized methods. The rest of the article is organized as follows. In section 2, described about related work, section 3, defined the approach and presented an efficient algorithm for solving the logistic regression model with the penalty. In sections 4 and 5, we evaluated the performance of our proposed approach on the simulated data and five public gene expression datasets. We presented a conclusion of the paper in section 6.

2. Related Work

The popular Regularized logistic regression is applied to cancer classification and is support both large and small features [3, 12]. However, such methods fails to reach oracle and smoothing properties. An additive model [16] with regularized method for feature selection produces the properties of both the sparsity and smoothness. [6]. Introduced the Component Selection and Smoothing Operator (COSSO) method, used feature selection on basis of nonparametric regression models. The Wang *et al*. [18] make use of both Group Lasso and Smoothly Clipped Absolute Deviation (SCAD) methods for model selection. GLMNET (Lasso and elastic-net regularized generalized linear models) with L_1 penalized regression introduced [22] resolve problems of standard penalized methods. [1, 24] Designed methods for gene classification using Logistic Regression with $L_{1/2}$ penalty [11, 19]. A new algorithm with Group Lasso using multinomial logistic regression solves multi-class classification.

3. Regularized Methods and Model Framework

3.1. Lasso Regularization

Consider a dataset *X* with *n* samples where x={(x_1,y_1),(x_2,y_2),…,(x_n,y_n)}.. These *n* samples are labelled with two class names. x_i={x_{i1},x_{i2},…x_{in}} indicates the i^{th} instance with p dimensions in X. In the case of genome data, x_{ij} represents the value of i^{th} sample j^{th} gene. The class label of i^{th} instance or i^{th} sample is Y_i which holds a value of 0 or 1. As there are two classes in the dataset, if Y_i=0 then i^{th} instance represents class 1 otherwise Y_i then i^{th} instance represents class 2. The Lasso model is shown as:

$$\Gamma(\lambda,\beta)=\arg\min\frac{1}{n}\sum_{i=1}^{n}(y-X'\beta)^2+\lambda P(\beta) \qquad (1)$$

Where P(β)-Regularization term. The term can be defined as:

$$P(\beta)=\sum_{j=1}^{p}|\beta_j|^1$$

In the recent study, many regularization methods were proposed L_1 regularization is the one identified as a well-organized method by many researchers. In this extension methods are SCAD regularization, the adaptive Lasso, Elastic Net, and Elastic SCAD. But in the analysis of gene expression data, the L_1 type regularization methods are not sufficient for drawing the conclusions. The real dataset which is collected through microarray-sequence consists of many predicates. But all these are not informative genes. There is a necessity of selecting required number of informative genes from the dataset. Other than this, L_1 regression is asymptotically biased [6, 10]. To enhance the accuracy of classification technique, use the Lq (0<q<1) regularization technique [21] which is more accurate than L_1 and L_0 regularizations. Consider the $L_{1/2}$ regularization penalty to be represented as Lq (0<q<1). In the consideration of the high dimensional problem $L_{1/2}$ penalty with logistic regression works as an efficient method [2]. The unbiased and oracle properties are the main characteristics of $L_{1/2}$ penalty [20]. However, this method selects only one variable from the all highly correlated data and form corresponding group. This was the main problem with $L_{1/2}$ regularization method. Next, extension is Lq (0<q<1) regularization with robustness which can offer better theoretical characteristics and also employed high performance in maximum applications [5, 7]. The sparse logistic regression technique with the combination of Least Absolute Deviation (LAD) and Lq (0 < q < 1) is to be considered as

$$\beta=\arg\min\left\{\sum_{i=1}^{n}\left|Y_i - x_i^T\beta\right|+\sum_{j=1}^{p}\lambda\left|\beta_j\right|^q\right\} \qquad (2)$$

The above combination offers tremendous properties like sparsity, unbiasedness, oracle properties and uniformity for the selection of a variable.

3.2. Regularized Lasso to Additive Models

In the process of analyzing the high dimensional datasets, the most preferred statistical tool is the additive nonparametric regression. This model is stated as (2):

$$Y=f_o+\sum_{j=1}^{p}f_j(x_i)+\varepsilon \qquad (3)$$

From Equation (3), it is known that X_j- predictor variable, ε -errors. These errors are not depending on the predictor variables. The other predictor variables are E(ε)=0, Var (ε)=2, f_j-univariate smoothening function, f_o -Constant and Y-response variable. the flexibility and interoperability of this additive nonparametric regression model made more popular. But it suffers with curse of dimensionality. To overcome this problem [7, 8, 14, 21], the Lasso technique is associated to this Additive model. Due to this adoption, the additive model could handle high dimensional data. The advantage of this method is to generate low dimensional data with equivalent high dimension. Consider the solution for constrained optimization problem which is represented in Equation (4):

$$\arg\min_{\beta_1,\beta_2,\ldots,\beta_p}\left\|Y-\sum_{k=1}^{p}x_j\alpha_j\right\|_2^2 \text{ subject to } \sum_{k=1}^{p}|\beta_k|\le\tau \qquad (4)$$

Where is a τ -predefined value. First apply the Cubic Smoothing spline on nonlinear data by extending the above equation in the following way:

$$\arg\min_{\beta_1,\beta_2,\ldots\ldots\beta_p}\left\|Y-\sum_{k=1}^{p}f_k(x_k)\right\|_2^2+\sum_{k=1}^{p}\lambda_k\int_{ak}^{bk}f_k''(x)^2dx \qquad (5)$$

The value of f_K is estimated by considering the values ranging from a_i to b_k. The interval between a_i and b_k is completely unreliable for the data. This equation consists of a smoothing parameter λ_k which is used to smoothing the errors. The second smoothing spline is B-spline. The equation for B-spline is:

$$\min_{\beta_1,\beta_2,\dots\dots,\beta_p} \left\| Y - \sum_{k=1}^{p} \mathcal{N}_k \beta_k \right\|_2^2 + \sum_{k=1}^{p} \lambda_k \beta_k^t \Omega_k \beta_k \\ \text{subject to } \sum_{k=1}^{p} \frac{1}{\lambda_k} = \frac{p}{\lambda} \quad \lambda_k > 0 \tag{6}$$

In Equation (6), $\mathcal{N}_k$ is bias which is evaluated at x_{ik}. This is applied as a function $f_k(x_k)$ in Equation (5). The other function in Equation (5) is $f_k(x_k)^2$ which is defined as $\beta_k^t \Omega_k \beta_k$. In Ω_k, a matrix which is in the size of $(n+2)^2$. β_k acts as the coefficient of the function. It is known that the given approach is an advancement to Lasso regression with the Additive models using Cubic Splines. The Equation (6) could be optimized and improved by applying a root square function to it. The optimized Equation is (7) given below:

$$\min_{\beta_1,\beta_2,\dots\dots,\beta_p} \left\| Y - \sum_{k=1}^{p} \mathcal{N}_k \beta_k \right\|_2^2 + \frac{\lambda}{p} \left(\sum_{k=1}^{p} \sqrt{\beta_k^t \Omega_k \beta_k} \right)^2 \tag{7}$$

4. Proposed Work

4.1. Additive L1/2+L2 Regularization (ALR)

The regularization is hybridized with the combination of $L_{1/2}$ regularization and L_2 regularizations [3]. The equation helps to handle correlation data. The equation for this is as follows:

$$L(\lambda_1,\lambda_2,\beta) = \arg\min \frac{1}{n} \sum\nolimits_{i=1}^{n} \left(Y - X'\beta \right)^2 + \lambda_1 |\beta|_{1/2} + \lambda_2 |\beta|^2 \tag{8}$$

Where β-set of coefficients i.e., $\beta=(\beta_1,\beta_2,\dots\beta_p)$, $\lambda_1 - L_{1/2}$ Regularization, $\lambda_2 - L_2$ Regularization, $|\beta|_{1/2} = \sum\nolimits_{k=1}^{p} |\beta_k|^{1/2}$ and $|\beta|^2 = \sum\nolimits_{k=1}^{p} |\beta_k|^2$ .To minimize the above stated equation the $L_{1/2+}2$ Regularization LR proposes an estimator $\hat{\beta}$ which is stated as

$$\hat{\beta} = \arg\min_{\beta} \{ L(\lambda_1, \lambda_2, \beta) \} \tag{9}$$

Further to minimize (9), consider another parameter α which is defined as $\alpha = \lambda_1 / (1+\lambda_2)$.It can handle the values of $\lambda_1 \,\&\, \lambda_2$ to represent $L_{1/2}$ Regularization and L_2 regularization. Apply this α value in the Equation (10) and then it turns the equation as follows:

$$\hat{\beta} = \arg\min_{\beta} \left\{ |Y - X'\beta|^2 + \lambda(\alpha |\beta|_{1/2} + (1-\alpha)|\beta|^2) \right\} \tag{10}$$

The value of α decides the regularization. If α = 0 then the above equation turns into L_2 regularization. If α = 1 then the above equation turns into $L_{1/2}$ Regularization. The combination of these two penalties makes efficient to produce a concise result. To fit the Equation (8) to non-parametric model the equation is modified and mentioned

$$\min_{\beta_1,\beta_2,\dots\dots,\beta_p} \left\| Y - \sum_{k=1}^{p} \mathcal{N}_k \beta_k \right\|_2^2 + \frac{\lambda_1}{p} \left(\sum_{k=1}^{p} \sqrt{ |\beta_k^t|_{1/2}\, \Omega^1{}_k \, |\beta_k|_{1/2} } \right)^2 \text{To} \\ + \frac{\lambda_2}{p} \left(\sum_{k=1}^{p} \sqrt{ |\beta_k^t|_2 \, \Omega_k^2 \, |\beta_k|_2 } \right)^2 \tag{11}$$

Optimize the Equation (11), parameter α is used

$$\min_{\beta_1,\beta_2,\dots\dots,\beta_p} \left\| Y - \sum_{k=1}^{p} \mathcal{N}_k \beta_k \right\|_2^2 + \lambda \left\{ \begin{aligned} & \left(\alpha \sum_{k=1}^{p} \sqrt{ |\beta_k^t|_{½}\, \Omega^1{}_k \, |\beta_k|_{½} } \right)^2 \\ & + (1-\alpha) \left(\sum_{k=1}^{p} \sqrt{ |\beta_k^t|_2 \, \Omega_k^2 \, |\beta_k|_2 } \right)^2 \end{aligned} \right\} \tag{12}$$

Similarly mentioned in above α=0 then Equation (13) to be considered as additive L_2 Regularization and in case of α=1 will be considered as Additive $L_{1/2}$ Regularization, respectively. Other cases this will be treated as ALR.

4.2. Additive Sparse Logistic Regression with ALR method (ASLR)

Let's consider a dataset $\mathcal{M}$ with *n* instances $\mathcal{M} = \{ (x_1,y_1), (x_2,y_2), \dots, (x_n,y_n) \}$. In this X_i represents X_i $\{X_{i1}, X_{i2}, \dots, X_{ip}\}$ the ith instance with p attributes (genes) and target variable to be mentioned as Y_i and value is 0 or 1. Next, to perform classification popular approach (i.e., The Logistic regression) is used and is represented as:

$$P(Y_i = 1/X_i) = P(X_i'\beta) = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)} \tag{13}$$

β-set of calculated coefficients $\beta=(\beta_1,\beta_2,\dots\beta_p)$ representation for p attributes. The regression technique represents the estimated coefficients using a simple algebra which is give below:

$$\Gamma(\beta) = -\sum\nolimits_{i=1}^{n} \{ Y_i \log[P(X_i'\beta)] + (1-Y_i) \log[1-(X_i'\beta)] \} \tag{14}$$

Now apply the Lasso Regularization (LR) technique with logistic regression model. It offers an optimized solution. To handle the fixed nonnegative λ and α, apply the sparse logistic regression model which is based on the ALR technique is defined as:

$$\Gamma(\lambda,\alpha,\beta) = -\sum\nolimits_{i=1}^{n} \{ Y_i \log[P(X_i'\beta)] + (1-Y_i) \log[1-P(X_i'\beta)] \} \\ + \lambda(\alpha |\beta|_{1/2} + (1-\alpha)|\beta|^2) \tag{15}$$

At end classification to the gene data performed by integrating proposed ALR technique with logistic regression model. It offers an optimized solution called the ALR technique is defined as:

$$\Gamma(\lambda,\alpha,\beta) = -\sum\nolimits_{i=1}^{n} \{ Y_i \log[P(X_i'\beta)] + (1-Y_i) \log[1-P(X_i'\beta)] \} + \\ \lambda \left\{ \left(\alpha \sum_{k=1}^{p} \sqrt{ |\beta_k^t|_{½}\, \Omega_k^1 \, |\beta_k|_{1/2} } \right)^2 + \left((1-\alpha) \sum_{k=1}^{p} \sqrt{ |\beta_k^t|_2 \, \Omega_k^2 \, |\beta_k|_2 } \right)^2 \right\} \tag{16}$$

5. Results and Discussion

The performances of regularized methods shown in Table 2, are assessed by applying on various real microarray datasets [13, 16, 18]. The short comprehensive information about each Regularized methods used in this work shown in Table 2. This paper considers Prostate, Diffuse Large B-cell Lymphomas (DLBCL) and Lung cancer gene data sets. Table 1 shows the comprehensive information associated to these datasets.

Table 1. Gene data set for classification.

Short	Datasets	Samples	Genes	Classes
D1	Prostate Cancer	102	12600	2
D2	Lymphoma Cancer	77	7129	2
D3	Lung Cancer	164	22401	2

The dataset Prostate is ready with enormous assortment of 12,600 genes expression profiles. The normal (i.e., 50) and abnormal prostate tissues (i.e., 52) are present in this dataset. Next, one is Lymphoma, which has 77 attributes corresponds to gene expression in Microarray Data. From the dataset, identified two classes labelled one is DLBCL and other is Follicular Lymphomas (FL). The complete gene data sampled with 7,129 expression profiles. The last dataset considered in this paper is Lung cancer dataset which comprises 164 instances and 22,401 genes expression profiles. This dataset consists of two classes labeled as lung adenocarcinomas which holds 87 samples and adjacent normal tissues holds 77 samples [4].

Table 2. Regularized methods.

Short	Methods
M1	Lasso
M2	SCAD L_2
M3	Elastic Net
M4	$L_{1/2}$
M5	$L_{1/2}$+L_2
M6	ASLR

Every classification task divides the given dataset into two parts called training dataset and testing dataset. The proposed method applied on the above discussed datasets.

Table 3. Optimized genes selected by applying regularizations methods.

Dataset	M1	M2	M3	M4	M5	M6
D1	13	20	14	8	14	17
D2	13	26	13	11	14	16
D3	13	28	20	18	16	14

Before applying classification method, derived optimal genes of each data sets with regularized methods and results shown in Table 3. This model divided the dataset in the ratio of 70:30 for training and testing. In addition, 10-fold cross-validation to extract the optimal tuning parameters on the training dataset. Use the estimated tuning parameters in the sparse logistic regression to develop a classification model. After building the classification model, use this model on test dataset to check the classification accuracy.

Table 4. Accuracy of training data by different regularizations methods.

Dataset	M1	M2	M3	M4	M5	M6
D1	89.4	92.7	93.4	96.2	97.3	97.6
D2	90.0	92.9	93.9	96.7	97.7	98.0
D3	86.7	91.4	92.8	94.7	96.8	97.2

Table 5. Accuracy of testing data by different regularizations methods.

Dataset	M1	M2	M3	M4	M5	M6
D1	82.2	90.4	91.0	93.4	94.4	95.2
D2	82.6	90.5	91.9	93.8	94.9	96.1
D3	81.6	89.6	90.0	94.9	95.4	96.9

The complete processes done at 500 times with random partition [9]. The part of the results shown in Tables 4 and 5, where Training and Testing accuracy of the classifier modeled from M6 to M1 namely, ASLR, $L_{1/2}$ + L_2, $L_{1/2}$, Elastic Net, SCAD L_2 and Lasso approaches with average 10-fold Cross Validation (CV) on three microarray dataset D1, D2 and D3 presented. ASLR (i.e., M6) model nominated 16 genes with an average accuracy rate of 98.0% and the average test accuracy of 96.1%. The classifier modeled through $L_{1/2}$ + L_2 (i.e., M5) with average 10-fold CV on Lymphoma dataset nominated 14 genes with an average accuracy rate of 97.7% and the average test accuracy of 94.9% (shown in Tables 4 and 5). The classifier modeled through $L_{1/2}$ (i.e., M4) on Lymphoma dataset nominated 11 genes with an average accuracy rate of 96.7% and the average test accuracy of 93.8%. Elastic Net (i.e., M3), SCAD L2 (i.e., M2) and Lasso(i.e., M1) approaches with average 10-fold CV on Lymphoma dataset nominated 13, 26 and 13 genes with an average accuracy rate of 93.9%, 92.9% and 90.0 % and the average test accuracy of 91.9%, 90.5%, and 82.6%.The complete accuracy results shown in Figures 1 and 2. After examining all the methods on Lymphoma dataset, the ASLR (i.e., M6) has proven as the best with respect to both training and testing accuracy. The classifiers modeled through ASLR with an average 10-fold CV on Prostate dataset and Lung cancer dataset were offering nearly equal accuracy rate as ASLR on both training and testing datasets.

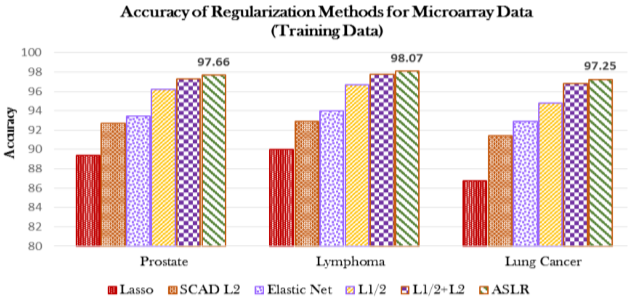


Figure 1. Accuracy of Training Data by different Regularizations Methods on Three data sets.

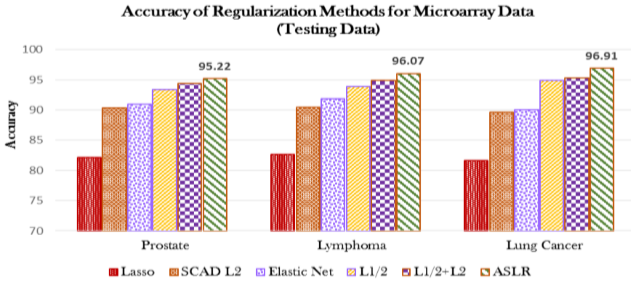
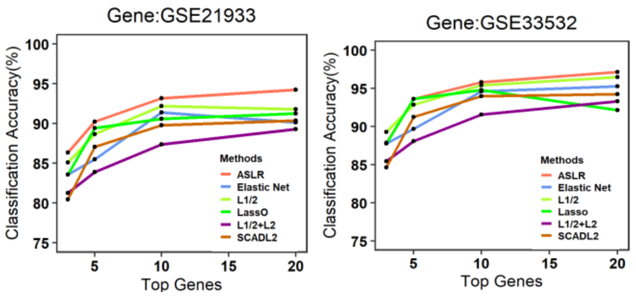


Figure 2. Accuracy of testing data by different regularizations methods on three data sets.

Though, the proposed ASLR (i.e., M6) technique attained its accuracy 97.2 in training and 95.4 in test case using only 14 predictors (genes), compared to 18 genes for the $L_{1/2}$ (i.e., M4) method and $L_{1/2}$ +L_2 (i.e., M5) method with 16 genes with training accuracy of 94.7 and 96.8. Even though the Lasso or $L_{1/2}$ approaches achieved the sparsest results, the performance with respect to these two classification approaches were inferior to $L_{1/2}$ +L_2 (i.e., M5) and ASLR (i.e., M6) techniques. To regulate the cost, choose only few features while conducting a precise test for screening and diagnostic claims.

5.1. Analysis on Gene Expression Data

Further analysis extended to select top rank genes from the individual gene profile data sets. Classification accuracy is resulted to be best by extending the number of genes which is shown in Tables 6, 7, and 8. In case of Lung cancer data set gene expression GSE21933, compared accuracy result with top ranks genes range from 5 to 20.Among proposed ASLR (i.e., M5), produced best accuracy using 20 genes compared to standard regularized methods $L_{1/2}$ +L_2 (i.e., M6), $L_{1/2}$ (i.e., M4) and Lasso (i.e., M1).



a) Gene Data GSE21933. b) Gene Data GSE33532.

Figure 3. Classification accuracy performance on lymphoma gene profiles using proposed method.

The results of classification method over Lung cancer with corresponding genes shown in Table 6. The gene profile, GSE21933 produce 94.2 accuracy with 20 genes compare to standard methods results shown in Figure 3-a). The similar kind of results observed in other gene expression data GSE33532 from the Lung cancer produces better accuracy value 97.1 using top 20 genes compared to non-sparse additive regularized methods and results are shown in Figure 3-b).

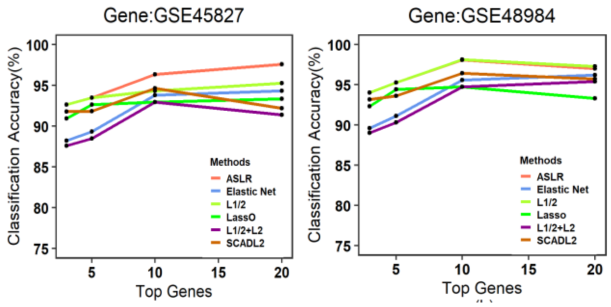
Table 6. Performance analysis of classifier methods on lung cancer gene data using FS method.

Genes	GSE21933					
	M1	M2	M3	M4	M5	M6
3	83.5	80.4	83.5	81.2	85.1	86.3
5	89.4	87.0	85.4	83.9	88.6	90.2
10	90.6	89.7	91.4	87.3	92.2	93.1
20	91.2	90.3	90.1	89.2	91.7	94.2
Genes	GSE33532					
	M1	M2	M3	M4	M5	M6
3	87.7	84.6	87.7	85.4	89.3	87.8
5	93.6	91.2	89.6	88.1	92.8	93.6
10	94.8	93.9	94.6	91.5	95.4	95.8
20	92.1	94.2	95.2	93.2	96.4	97.1

From Lymphoma data, it is observed that gene data GSE45827 produces better accuracy 97.6 with 20 genes and in other gene GSE48984 only with 10 genes will get accuracy 98.0 and which is not promising compared to $L_{1/2}$ +L_2 and it produced 98.2 with 10 genes. The classification results of gene GSE45827 are shown in Figure 4-a). In other gene data GSE48984, proposed method results shown in Figure 4-b).

Table 7. Performance analysis of classifier methods on lymphoma cancer gene data using fs method.

Genes	GSE45827					
	M1	M2	M3	M4	M5	M6
3	90.9	91.7	88.2	87.6	92.6	92.6
5	92.6	91.8	89.3	88.5	93.4	93.4
10	92.9	94.6	93.8	92.9	94.3	96.3
20	93.3	92.2	94.3	91.3	95.2	97.6
Genes	GSE48984					
	M1	M2	M3	M4	M5	M6
3	92.3	93.1	89.6	89.0	94.0	94.0
5	94.4	93.6	91.1	90.3	95.2	95.2
10	94.7	96.4	95.6	94.7	98.2	98.0
20	93.2	95.7	96.2	95.3	97.2	97.0



a) Gene data GSE45827. b) Gene data GSE48984.

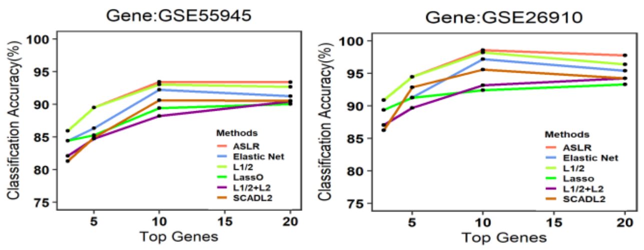
Figure 4. Classification accuracy performance on lymphoma gene profiles using proposed method.

The Prostate cancer, gene data GSE55945 with best accuracy 93.4 produced by proposed ASLR (i.e., M6) with only 10 genes and results shown in Figure 5-a). Moreover, in case of gene data GSE26910 retained 98.5 with 10 genes using ASLR (i.e., M6).

Table 8. Performance analysis of classifier methods on prostate cancer gene data using FS method.

Genes	GSE55945					
	M1	**M2**	**M3**	**M4**	**M5**	**M6**
3	84.4	81.3	84.4	82.0	85.9	85.9
5	85.2	84.9	86.3	84.7	89.5	89.5
10	89.4	90.6	92.2	88.2	93.0	93.4
20	90.0	90.5	91.2	90.4	92.6	93.3
Genes	**GSE26910**					
	M1	**M2**	**M3**	**M4**	**M5**	**M6**
3	89.3	86.2	89.3	87.0	90.9	90.9
5	91.2	92.8	91.2	89.7	94.4	94.4
10	92.4	95.5	97.2	93.1	98.2	98.5
20	93.2	94.2	95.3	94.2	96.3	97.7

The results of GSE26910 genes shown in Figure 5-b). The comparison of classification results from non-sparse regularized to sparse regularized over prostrate gene profiles data with optimized genes shows in Table 8.



a) Gene Data GSE55945. b) Gene Data GSE26910.

Figure 5. Classification accuracy performance on prostate cancer gene profiles using proposed method.

From the Table 9, it clear that the proposed ASLR (i.e., M6) method shows better accuracy, sensitivity and specificity values than the standard hybrid method $L_{1/2}$ +L_2 (i.e., M5) in all three real gene datasets (D1 to D3).

Table 9. Performance analysis of classifier on different datasets with FS methods.

	Metric (%)	**M5:$L_{1/2}$+L_2**		**M6:ASLR**	
D1	**Features**	15	32	15	32
	Accuracy	97.01	96.59	97.98	98.06
	Sensitivity	81.14	77.65	85.71	86.47
	Specificity	92.25	90.03	81.14	77.65
D2	**Features**	16	42	16	42
	Accuracy	97.13	97.09	98.13	98.41
	Sensitivity	81.03	81.16	86.44	89.09
	Specificity	94.19	94.45	93.55	95.49
D3	**Features**	13	42	13	42
	Accuracy	96.86	96.49	97.26	97.18
	Sensitivity	86.95	88.37	74.44	76.16
	Specificity	91.37	95.62	93.09	94.25

Moreover, the ASLR method produced better sensitivity from suitable genes compared to the non-additive method. In the case of lung cancer (i.e., D3), the proposed method retained promising results in terms of accuracy: 97.18% associated on 42 genes and with good specificity: 94.25% and Sensitivity: 76.16%. In lymphoma data (i.e., D2), the proposed method retained promising results in terms of accuracy: 98.41% associated on 42 genes and with good specificity: 95.49% and Sensitivity: 89.09%. Similarly, in the case of lung cancer (i.e., D1), proposed method also produced impressive results compared to the standard method shown in Table 9.

6. Conclusions

This paper examined a new feature selection approach termed ASLR method. It is designed by considering the best features from $L_{1/2}$ and L_2 penalties. An innovative method is projected in this paper for additive function which is applied to the combination of $L_{1/2}$ + L_2 Regularization. It helps to optimize the calculated coefficients. It has established feature selection Additive Lasso Regularization and then applied classification using Sparse Logistic method. In the end, observed results of the proposed method were promising when compared to the standard methods including Lasso, $L_{1/2}$, SCAD-L_2 and Elastic Net models. Hence, empirical results also show proposed one is a best feature selection method for gene data.

References

[1] Algamal Z. and Lee M., "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.

[2] Bashir K., Li T., and Yahaya M., "A Novel Feature Selection Method Based on Maximum Likelihood Logistic Regression for Imbalanced Learning in Software Defect Prediction," *The International Arab Journal of Information Technology*, vol. 17, no. 5, pp. 721-730, 2020.

[3] Becker N., Toedt G., Lichter P., and Benner A., "Elastic SCAD as A Novel Penalization Method for SVM Classification Tasks in High-Dimensional Data," *BMC Bioinformatics*, vol. 12, no. 138, pp. 1-13, 2011.

[4] Goh L., Song Q., and Kasabov N., "A Novel Feature Selection Method to Improve Classification of Gene Expression Data," *in Proceedings of the 2nd Conference on Asia-Pacific Bioinformatics*, Dunedin, pp. 161-166, 2004.

[5] Hu Y. and Kasabov N., "Ontology-Based Framework for Personalized Diagnosis and Prognosis of Cancer Based on Gene Expression Data," *in Proceedings of International Conference on Neural Information Processing*, Kitakyushu, pp. 846-855, 2008.

[6] Knight K. and Fu W., "Asymptotics for LASSO-Type Estimators," *The Annals of Statistics*, vol. 28, no. 5, pp. 1356-1378, 2000.

[7] Lavanya K., Reddy L., and Reddy B., *Computational Intelligence in Data Mining*, Springer, 2019.

[8] Lavanya K., Reddy L., and Reddy B., "Modelling of Missing Data Imputation using Additive LASSO Regression Model in Microsoft Azure," *Journal of Engineering and Applied Sciences*, vol. 13, no. 8, pp. 6324-6334, 2018.

[9] Lin Y. and Zhang H., "Component Selection and Smoothing in Multivariate Nonparametric Regression," *Annals of Statistics*, vol. 34, no.5, pp. 2272-2297, 2006.

[10] Malioutov D., Cetin M., and Willsky A., "Sparse Signal Reconstruction Perspective for Source Localization with Sensor Arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010-3022, 2005.

[11] Meier L., Geer S., and Bühlmann P., "The Group LASSO for Logistic Regression," *Journal of the Royal Statistical Society Series B*, vol. 70, pp. 53-71, 2008.

[12] Meinshausen N. and Yu B., "Lasso-Type Recovery of Sparse Representations For High-Dimensional Data," *Institute of Mathematical Statistics*, vol. 37, no. 1, pp. 246-270, 2009.

[13] Singh D., Febbo P., Ross K., Jackson D., Manola J., Ladd C., Tamayo P., Renshaw A., D'Amico A., Richie J., Lander E., Loda M., Kantoff P., Golub T., and Sellers W., "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203-209, 2002.

[14] Taylan P. and Weber G., *Data Science and Digital Business*, Springer, 2019.

[15] Tibshirani R., "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267-288, 1996.

[16] Van-de-Geer S., "High-Dimensional Generalized Linear Models and the Lasso," *Institute of Mathematical Statistics*, vol. 36, no. 2, pp. 614-645, 2008.

[17] Vincent M. and Hansen N., "Sparse Group Lasso and High Dimensional Multinomial Classification," *Computational Statistics and Data Analysis*, vol. 71, pp. 771-786, 2014.

[18] Wang L., Chen G., and Li H., "Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data," *Bioinformatics*, vol. 23, no. 12, pp. 1486-1494, 2007.

[19] Wu S., Jiang H., Shen H., and Yang Z., "Gene Selection in Cancer Classification Using Sparse Logistic Regression with L1/2 Regularization," *Applied Sciences*, vol. 8, no. 9, pp. 1569, 2018.

[20] Xu Z., Chang X., Xu F., and Zhang H., "L1/2 Regularization: A Thresholding Representation Theory and A Fast Solver," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 7, pp. 1013-27, 2012.

[21] Xu Z., Zhang H., Wang Y., Chang X., and Liang Y., "L1/2 Regularization," *Science China Information Sciences*, vol. 53, no. 6, pp. 1159-1169, 2010.

[22] Yuan G., Ho C., and Lin C., "An Improved GLMNET for L1-Regularized Logistic Regression," *Journal of Machine Learning Research*, vol. 13, pp. 1999-2030, 2012.

[23] Zeng J., Lin S., Wang Y., and Xu Z., "L1/2 Regularization: Convergence of Iterative Half Thresholding Algorithm," *IEEE Transactions on Signal Processing*, vol. 62, no. 9, pp. 2317-2329, 2014.

[24] Zhu J. and Hastie H., "Classification of Gene Microarrays by Penalized Logistic Regression," *Biostatistics*, vol. 5, no. 3, pp. 427-443, 2002.

[25] Zou H. and Hastie T., "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 3, pp. 301-320, 2005.

[26] Zou H., "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418-1429, 2006.

Vijay Suresh Gollamandala working as an Associate Professor in Department of CSE in Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India. He is pursing Ph.D degree in Computer Science from J.N.T. University Hyderabad, He has published more than 12 research papers in various international journals.

Lavanya Kampa working as an Associate Professor in Department of IT in Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India. She earned Ph.D degree in Computer Science from J.N.T. University Anantapur, She has published more than 15 research papers in various international journals. Also she worked as Reviewer for various international journals.