Simulated Annealing and Partial Least Squares Regression for Optimized Feature Selection in Gene Expression Data

Ashwini C.[1] Mahesha D.M ,[2] Bhavya D.N,[3] Sharath Kumar Y.H [4]

[1,2,3] Department of Studies and Research in Computer Science, Karnataka State Open University, Mysore-570006, India.
[4] Maharaja Institute of Technology Mysore, Department of ISE, MIT Mysore, India.

Abstract

The selection of relevant feature subset helps make machine learning models simpler and improves their performance. If the number of original features is less, the most relevant subset can be found by exhaustively evaluating each possible subset. This strategy is computationally prohibitive for gene expression datasets with thousands of features. Stochastic optimization algorithms help in selecting the next feature subset to evaluate. The selection is based on minimization of error or maximization of performance accuracy /cross-validation score of an estimator algorithm. Thus, the feature-selection task can be formulated as an optimization task with the score of an estimator as its objective function. In this chapter, we have proposed an optimization-based feature selection method for gene expression datasets. We use simulated annealing (SA) for selection of optimal features and optimal number of components for Partial least squares regression PLSR. The features selected by simulated annealing are used to fit a partial least squares regression model with the number of components also selected from simulated annealing.

1. Introduction

Cancer has become one of the deadliest diseases globally, with an estimated 9.7 million deaths out of 20 million new cancer cases in 2022, according to the World Health Organization (WHO) [1]. Cancer results from the unconstrained growth of some anomalous cells, which divide and disperse to other cells, increasing malignant offspring cells. Lung, prostate, colorectal, and stomach cancers are the most common types of cancer that occur in men. Additionally, colorectal, lung, cervical, and breast cancers are the most common types among females. Acute lymphoblastic leukaemia and brain tumours are the most common cancers among children, except in Africa [2]. Cancer is prevalent worldwide and affects social life economically, in addition to affecting individuals. Public and government budgets in the health sector are being threatened due to the high cost of medical treatments. Premature deaths reduce the social workforce. Proper cancer

identification at an early stage can restrain the death rate and retain human resources at working ages. Manual diagnosis systems may lead to errors due to insufficient relevant resources. DNA microarray-based gene expression profiling can be a promising technique for detecting cancer at an early stage. Early cancer detection raises the chance of survival, reducing personal, societal, and economic costs. The gene expression datasets in the literature typically have a small number of samples. In contrast, the number of genes (the dimension of the features) per sample is significantly large, causing an overfitting concern in the respective machine learning model as it may perform worse on the test data after training. Therefore, the gene selection technique must be applied to the gene expression datasets [3]. AbdElNabi et al. [4] also mentioned the overfitting problem caused by the high dimensionality of genes compared to the instance size. Therefore, the gain of information was employed to reduce the number of irrelevant genes and eliminate the high dimensionality problem. Dimensionality reduction is also applicable for reducing computing time, constructing a robust model, and increasing the model's prediction quality [5]. This paper introduces an effective cancer classification method based on a machine learning technique using high-dimensional gene expression data that can be suitable for precise cancer detection and contribute to reducing the above-mentioned impacts on society and individuals. Specifically, an ensemble learning technique using feature dimensionality reduction in gene expression data is utilized for precise classification. To cope with the high dimensionality of genes in the limited training data, which may influence the accuracy of any machine learning model, the mutual information (MI) algorithm is used to select the most significant genes instead of using all of them. With the chosen influential genes, an ensemble technique comprising the bagging method, where base classifiers are Multilayer Perceptrons (MLPs), is applied, and performance metrics for various datasets are evaluated and compared.

2. Literature Review and Problem Statement

Cancer is one of the most severe diseases leading to death worldwide. Approximately 1 in 5 people have cancer at some point in their lives, and 1 in 9 men and 1 in 12 women pass away from cancer [1]. It is found that approximately 17.6% of women, as well as 26.3% of men, develop cancer before the age of 75 years old in Japan [6]. A total of 380,500 cancer deaths were projected in 2022, of which 219,300 were male and 161,200 were female in Japan, respectively [7]. **Figure 1** represents the cancer death statistics of males and females in Japan in 2021 based on the types of cancer. The recovery of a person depends on how early the disease is detected. Early detection lessens the chances of death. Additionally, cancer treatment at the initial stage is much simpler

than in its outbursts. The traditional investigation includes the physical investigation of the infected parts. Such physical medical investigations threaten the health of the examiner with infection, radiation, and so on. The results of ultrasonography depend on the quality of the images, which several factors can impact. On the contrary, gene expression data collected from DNA microarrays can effectively solve these issues [4].

Figure 1. Number of cancer deaths of males and females in Japan in 2021 [7].

Several techniques for cancer classification using gene expression have been investigated in recent years. Gene selection and accuracy prediction on chosen genes were two crucial criteria utilized by Salem et al. [8] for the performance evaluation of their suggested approach. Information gain was employed for feature selection, followed by a genetic algorithm for feature reduction and Genetic Programming for categorizing cancer types. Seven cancer gene expression datasets were considered to verify the suggested framework. The Modified K-Nearest Neighbor approach described in [9] was trained on derived features via information gain using microarray data. Yeganeh et al. [10] examined the problem of ovarian cancer with selected genes. They experimented with five machine learning models, named Random Forest (RF), Generalized Linear Model (GLM), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree (CART). The Random Forest classifier outperformed the other classifiers with 89% accuracy. Dey et al. [11] classified leukemia cancer into acute myeloid leukemia (AML) and acute lymphocytic leukemia (ALL). Principal component analysis (PCA) was used for dimension reduction, and XGBoost, Random Forest, and Artificial Neural Networks (ANNs) were implemented next. XGBoost and ANN became the victorious classifiers, obtaining the same accuracy of 92.3%. Akhand et al. [12] used minimum Redundancy Maximum Relevance (mRMR) as a feature selection technique and then employed ANN on four benchmark datasets for cancer classification. Souza et al. [5] attempted to compare and contrast two reduction methods—attribute selection and principal component analysis—to provide the most comprehensive comparison while analyzing

gene expression datasets. Data collected from the Mendeley data repository were analyzed for five different types of cancer in [13]. They showed the implementation of eight deep-learning models for cancer classification. CNN obtained the best outcomes among them all. Another finding was that a 70–30 split produced the best classifier performance. Erkal et al. [14] selected 137 prominent features out of 7129 genes and then used several machine learning classifiers. The Multilayer Perceptron performed the best, acquiring an accuracy of 97.61%, while the J48 method had the lowest accuracy at 73% for multi-class brain cancer classification. Almutairi et al. [15] classified breast cancer using three datasets—the WBCD, WDBC, and WPBC datasets—collected from the UCI repository. Each dataset consisted of two classes: benign or cancer. The gorilla troop optimization (GTO) method was applied as a feature selection technique. The classification was performed using Deep Q-learning (DQL), which is based on deep reinforcement learning, and the anticipated result was explained using LIME. Their proposed model achieved >98.50% accuracy for each of the datasets. Mallick et al. [16] designed a five-layer DNN model to classify acute lymphocyte leukemia (ALL) and acute myelocytic leukemia (AML). They compared their model with other traditional machine learning models: SVM, KNN, and Naive Bayes. Notably, they managed to gain 98.2% accuracy, 97.9% specificity, and 96.59% sensitivity, which was better than the compared model's performance. Joshi et al. [17] applied a deep learning approach to classify brain tumors with the help of gene expression data. An accuracy of 98.7% was obtained through the introduction of PSCS with deep learning. In the above research, the developed models showed promising results in cancer detection using advanced non-contact-based examination techniques using various machine learning approaches. However, there are still opportunities to improve the classification by taking accuracy to the maximum level. Accurate cancer identification is expected to reduce mortality and personal, societal, and economic costs. Therefore, this study proposes a more effective cancer classification method using a machine learning model for cancer detection with higher accuracy. In the proposed method, an ensemble learning technique through dimensionality reduction using only selected gene expression data is utilized for precise classification after selecting the influential genes using the mutual information (MI) algorithm.

3. Proposed Work

In this work, our focus is on selecting an optimal subset of genes for classification of gene-expression profiles. To achieve this, we have proposed a two-stage method. In stage one, we have used simulated annealing for identification of optimum genes and optimum number of components. These genes are then used to fit partial least squares regression (PLSR), with

the optimum number of components identified by simulated annealing. The genes with maximum PLSR coefficients are the final selected genes.

3.1 Data Pre-Processing

All input dataset is scaled to have a mean of 0 and variance 1. We used bootstrap sampling to increase the number of samples. A bootstrap sample is drawn randomly from the input gene expression samples with replacement. Genes in the bootstrap sample are used as training data and the samples not in bootstrap are used as test data. This step is repeated N times to generate the N bootstrap samples. The class distribution in gene expression data is imbalanced. The number of samples of different classes is often different. E.g., the number of samples in the normal class is large as compared to the number of samples in the Cancerous class. This class imbalance can lead to one class over influencing the training algorithm. To balance the class distribution, we have used the synthetic minority oversampling technique.

3.2 Selection of Features and Optimal Number of Latent Variables for PLS using Simulated Annealing

Simulated annealing (SA) is a nature-inspired algorithm for function optimization. It is inspired by the mechanism of annealing in physical solids. We have used simulated annealing to select genes from input cancer gene expression datasets. This is because simulated annealing has been demonstrated to converge to global optima of the objective function with a sufficient number of iterations and it reduces the collinearity among the input genes. Thus, we use SA to select genes that minimize the root mean square error in PLS and to decide the optimum number of components for PLS. Then these selected genes are used to fit a regression model in PLS. The genes with the highest PLS coefficients are treated as the final subset of the most relevant genes. This subset of the genes has been used to train the classifiers Support vector machine classifier, AdaBoost Classifier, Voting Classifier, Random Forest Classifier, and Multilayer perceptron classifier.

3.3 General simulated annealing algorithm is defined as follows

When solids are heated to very high temperatures, they melt. From a hot and molten state, they cool gradually with a reduction in the temperature. During this cooling process, the particles in the matter obtain an equilibrium at temperature T. Particles tend to be in

equilibrium or low-energy states or ground states. The probability of a system being in thermal equilibrium is the probability of a system being in a state with energy E. This probability is given by Boltzmann distribution as in equation 1.

$$P(E = E) = \frac{1}{Z(T)} e^{\frac{-E}{kT}} \quad (1)$$

Where k is the Boltzmann constant and $Z(T)$ is the normalization factor. As temperature decreases, states with lower energy have a higher probability of existence. The process of gene selection is modeled after this process of annealing, such that each successive subset of genes considered is one that minimizes the energy of the system. The energy of the system in case of feature selection problem is the mean squared error of the PLS method in our work. Starting with a random set of genes, at each iteration we choose a slightly different set of genes and compute the difference in the root mean square error of two subsets of genes. A different set of genes is selected randomly. If the difference of new subset with an older subset, (E), is negative then- new subset is better. The probability of accepting a new subset of genes is given by the Metropolis criterion as given in equation 2.

$$X = TP^T + E \quad (2)$$

With more iterations, simulated annealing finds a subset of the gene with a minimum value of root mean square error of PLS in our case. In different cases, SA can optimize the classification accuracy or loss function of a classifier directly. Cost function or loss function is analogous to the energy of a physical system being annealed. SA allows locally non-optimal subsets of genes to be considered but finally selects the globally optimum subset. Hyper-parameters for simulated annealing are cooling parameters, we have chosen an initial value of 0.001 for the cooling parameter and it keeps on decreasing as the value of the objective function decreases.

3.4 Feature Reduction using Partial Least Squares Regression

“Partial least squares regression (PLSR)”, method finds linear transformations of input features. The features should be as have high covariance with the response variable and uncorrelated among themselves. Such linear transformations of mutually uncorrelated and with high covariance with response variables are known as latent com- ponents. PLS then uses regression to predict the response variable and reconstructs the original data matrix from latent variables. The objective function for the selection of latent variables in PLS is the

maximization of the covariance of the latent variable with the response variable. It is resistant to multi-collinearity, noise, high dimensionality, and the cases when the number-of-dimensions is much higher than the number-of-samples as is the typical case with gene-expression data. In this regression, the predictors' matrix X with dimensions $n * k$ and target matrix Y of dimensions $n * m$ can be modeled as follows

$$X = TP^T + E \tag{3}$$

$$Y = UQ^T + F \tag{4}$$

$$u_a = b_a t_a + h \tag{5}$$

$a = 1, 2, 3, ..A$

Where A is the number of latent variables, $T = (t1, t2, , tA)$, $U = (u1, u2, , uA)$ are latent variable scores of X and Y . T is a projection of X and U is a projection of Y with same dimensions as A and Y respectively. P and Q are orthogonal loading matrices calculated by the non-linear iterative partial least squares method. E and F are random errors with the normal distribution. Equations 4.3 and 4.4 are the outer relations between two score matrices, b_a is the regression coefficient of u_a. Equation 4.5 is the inner relation between U and T, E, F , and h is the error in X,Y and u_a. Cross-validation is used to minimize prediction error. The proportion of variance explained by each latent variable determines the total number of latent variables used in PLS regression and it is an important parameter that determines the accuracy of prediction. The contribution of each gene to the class label is determined by decomposing the sum of squares of the gene expression values, where the total sum of squares of latent variables has 2 parts: the sum-of-squares of regression; and the sum-of-squares of error. The Sum-of-squares (*SS*) is the square of the difference between the actual value of the predicted variable, y, and its mean. The Sum of squares regression (*SSR*) is the sum of differences between predicted values of y and its mean. The Sum-of-squares-error (*SSE*) is the difference between the actual-value of y and the predicted-value. Equation 6 gives the relation of *SS*, *SSR* and *SSE*.

$$SS = SSR + SSE \tag{6}$$

Where SS or $SS(Y)$ is the total sum of squares of latent variables, SSR is the sum-of-squares of regression, *SSE* is the sum-of-squares of error. The importance of each gene in the input gene expression matrix is calculated by PLS as given in equation 7.

$$GI_j = \frac{s \sum_{k} w_{ja}^2 b_a^2 t_a^T t_a}{\sum_{a}^{a} b_a^2 t_a^T t_a} \quad (7)$$

Where GIj is the gene importance of j^{th} gene in the gene expression data set, wja is the weight of j^{th} gene to the a^{th} latent variable. Figure 1 shows all the steps in the proposed method for feature selection.

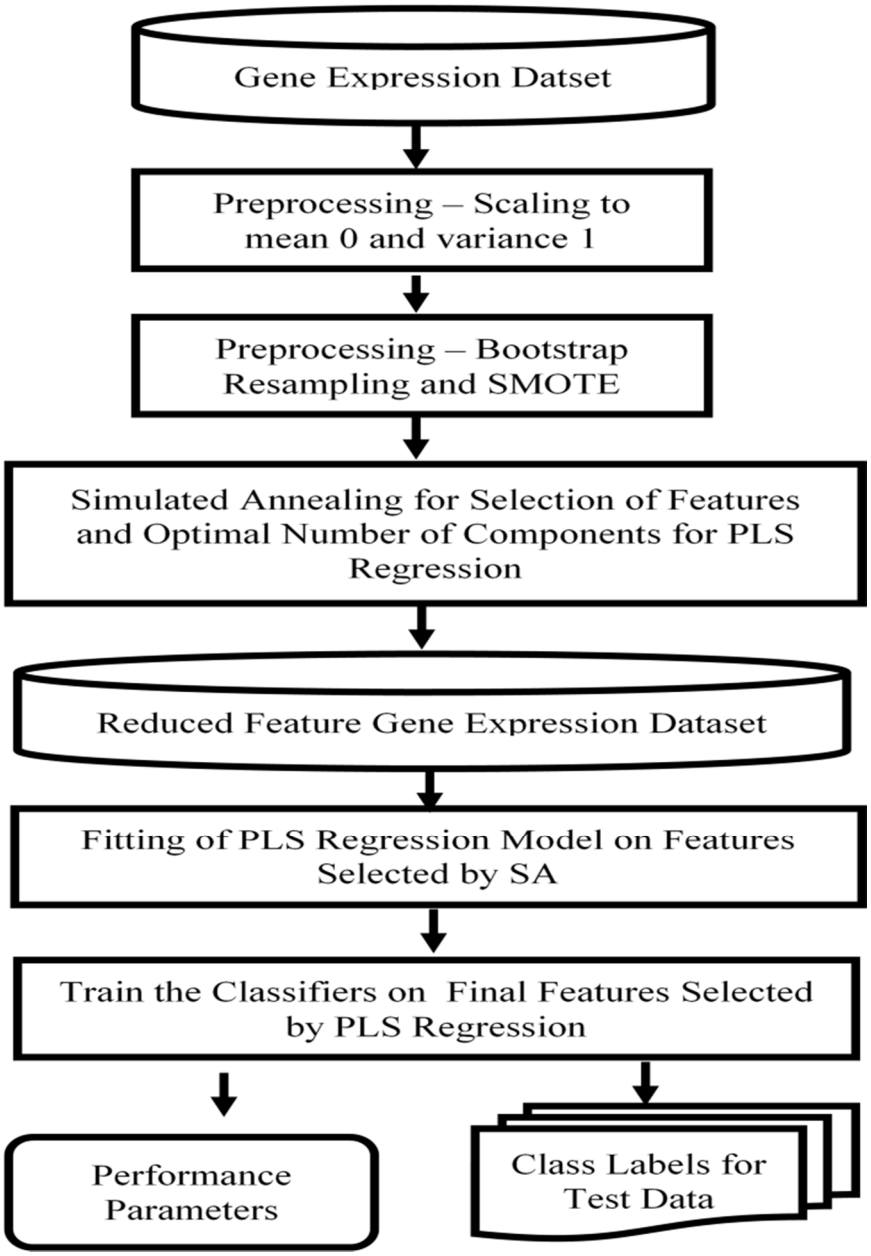

Figure 2: Proposed Method for Feature Selection using Simulated Annealing and PLS Regression

4. Experimental Setup

All the experiments were done on a Windows 10 Pro Laptop with Intel i7 processor and 8GB RAM. Programming was done using Anaconda Python 3.7. Proposed feature selection scheme was tested on 6 microarray cancer gene expression data sets. Of the 6 datasets, 3 are multiclass and 3 have 2 classes. All values in the datasets were scaled to lie between 0 and 1. For obtaining classification performance parameters, train-test ratio of 70:30 was used. For obtaining cross validation score, 5-fold cross-validation was used. We used 6 cancer microarray data sets for evaluation of the proposed scheme. Details about the datasets are given in Table 1.

Dataset	No. of Genes	No. of Samples	No. of Classes
Leukemia	3571	72	2
SRBCT	2308	83	4
Colon	2000	62	2
Lymphoma	4026	62	3
Prostate	6033	102	2
Brain	5597	42	5

Table 1: Datasets Used for Evaluation

Bootstrap resampling was applied to each dataset to increase the number of samples to more than 100 in each dataset. To these datasets with increased samples, we applied synthetic minority oversampling to increase the samples of the minority class. Finally, after bootstrap resampling and minority oversampling the datasets are used for feature selection with the simulated annealing algorithm. The selected features/genes and the optimal number of components are obtained from the simulated annealing algorithm. These genes are used to fit the PLS regression for the decided number of components. The features with the highest coefficients in the PLS regression model are taken as the final selected features. These genes were used to train classifiers - SVM, Adaboost, Voting-Classifier, Random Forest, and Multi-Layer Perceptron. The performance of these classifiers is reported as precision, re-call, classification-accuracy, and F1 score.

4.1 Results and Discussion

In this section, we present and interpret the findings of the experiment. We have shown the results of applying simulated annealing and the final number of genes chosen by PLSR for

each dataset. We present the classification performance for the small sub- set of selected genes with 6 classifier algorithms and comparison of classification- performance of our method with existing combinations of feature selection methods and classifiers.

4.2 Number of Selected Features.

On running the proposed feature selection method on the 6 pre-processed data sets, we obtained the optimal number of components for each dataset using simulated annealing and the final number of selected features using PLSR, as shown in the Table 2.

Dataset	Optimal Number of Components Selected by SA	Number of Features Selected by PLS
Colon	17	13
SRBCT	19	10
Prostate	12	12
Lymphoma	13	11
Brain	18	14
Leukemia	17	9

Table 2: Optimal Number of Components Selected by SA and Number of Features Selected by PLS

From the Table 2, we notice that application of simulated annealing helps in identification of a small optimal number of components that guide the second level of feature selection using PLSR model coefficients. This results in identification of a small number of relevant-genes for final classification. This small number of selected genes has helped in improving the classification-parameters and as shown in Table 2.

4.3 F-Test and Mutual Information of Selected Features with Class Labels

Figures 3,4 and 5 show the statistical significance of the features selected using the proposed method from Leukemia, Prostate and Brain datasets.

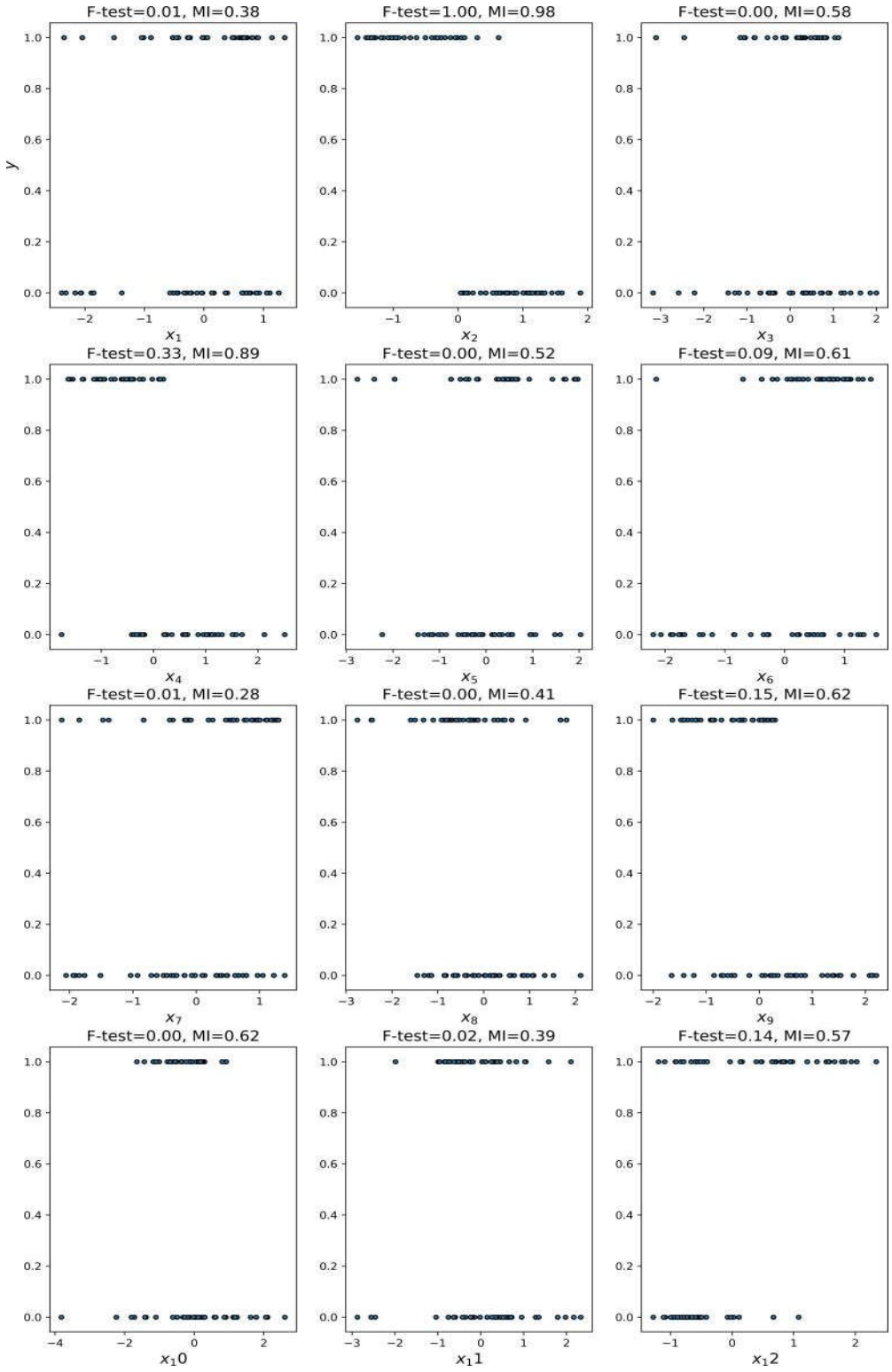

Figure 3: F-Test and Mutual Information of Features Selected with Class Labels by the proposed SA-PLS For Leukemi

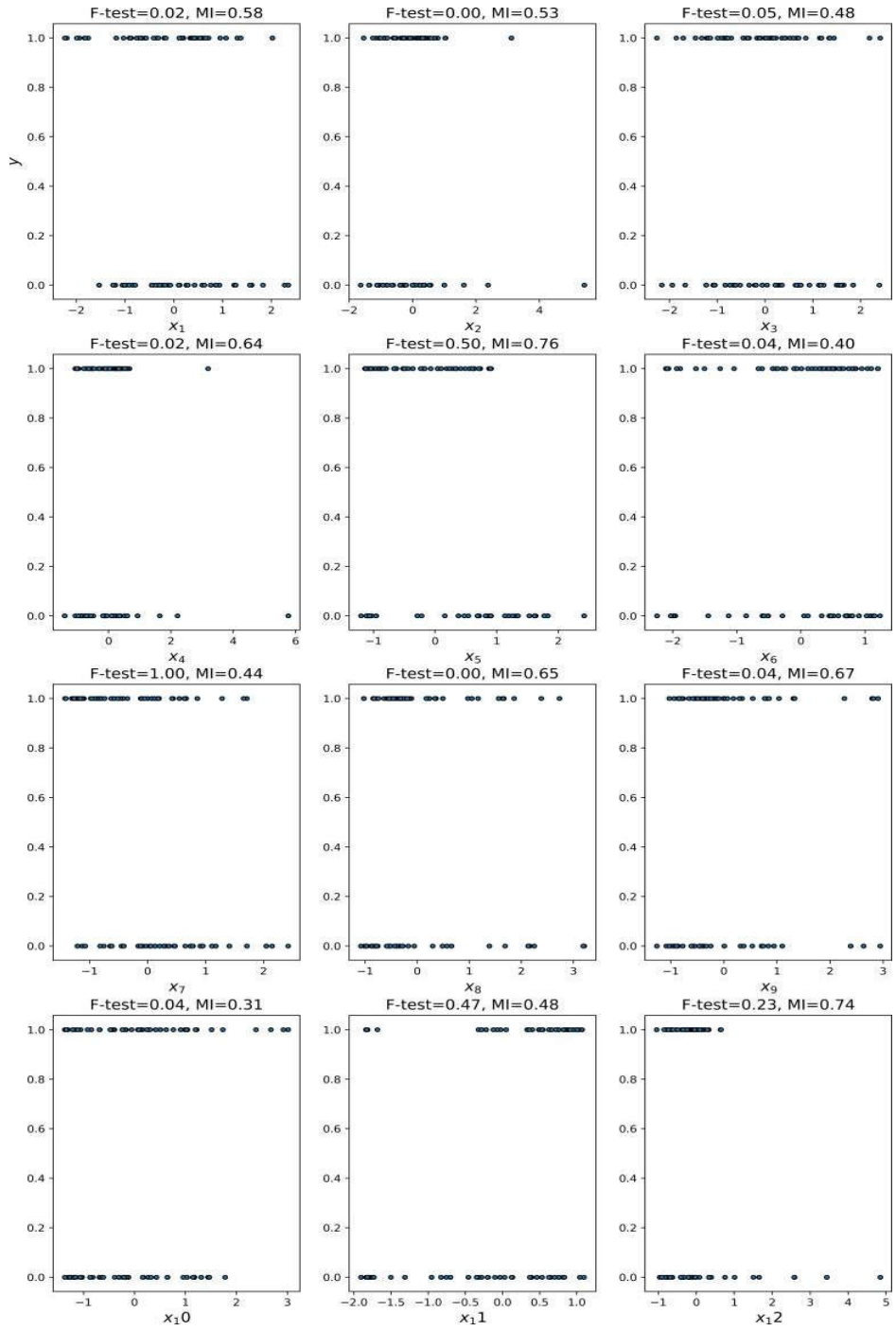

Figure 4: F-Test and Mutual Information of Features Selected with Class Labels by the proposed SA-PLS For Prostate Dataset

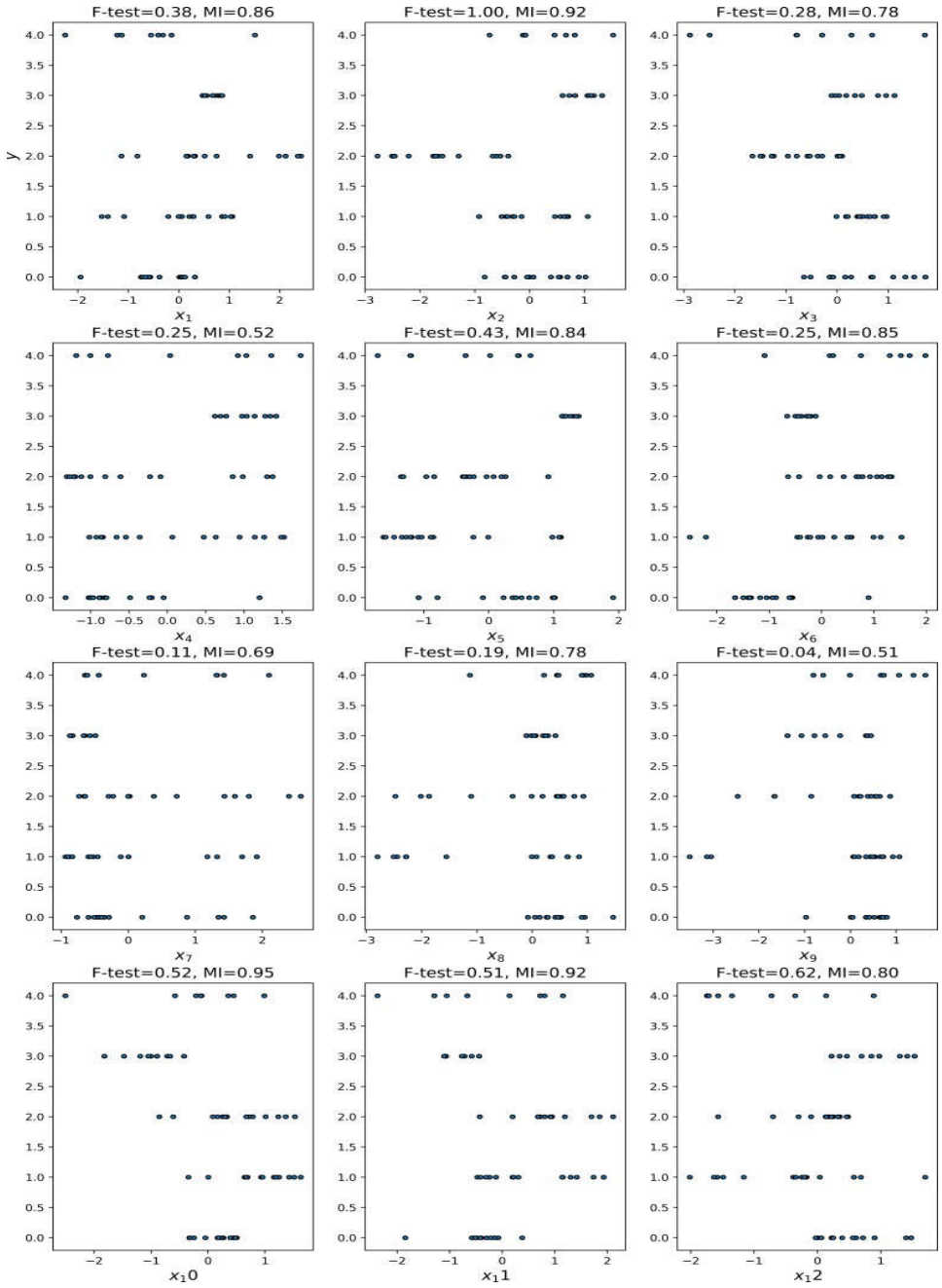

Figure 5: F-Test and Mutual Information of Features Selected with Class Labels by the proposed SA-PLS For Brain Dataset

4.4 Classification Accuracy and F1 Score with Selected Features

On training the classifiers with the above-selected genes from all the datasets, we observed very good performance metrics for all the classifiers. Only for Adaboost, the performance was less as compared to other classifiers. We also observed that for the brain cancer dataset, the performance metrics are less as compared to other datasets. Table 3 shows the metrics Precision, Recall, Classification accuracy, and F1-score the classifiers Linear SVM, RBF SVM, Adaboost, Voting Classifier, Random Forest Classifier, and Multilayer perceptron classifier. We observe from the Table 3, that the genes selected by the proposed method give a high predictive accuracy for 5 out of 6 datasets.

4.5 Comparison with Other Existing Methods of Feature Selection

The performance of the proposed feature-selection approach was compared with exist- ing heuristic methods of feature selection and classifier combinations to evaluate the performance for all 6 datasets. We compared the following combinations of feature-selection method and classifiers Genetic algorithm (GA) feature selection and lin- ear Support-vector-classifier (SVC), GA and Random-Forest Classifier (RFC), GA and Multilayer Perceptron (MLP), Binary Particle Swarm Optimization (BPSO) and SVC, BPSO and MLP, Proposed Method (SA-PLS) with SVC, SA-PLS and RFC, SA-PLS, and MLP. The classification accuracy obtained with each of the 6 pre-processed datasets is reported in Table 4. From the comparison, we see that the selected features give a 100% classification accuracy for all datasets with an SVC classifier. With RFC and MLP also the accuracy is 100% for 4 out of 6 data sets. With the brain cancer dataset, we see that the accuracy is 88% with RFC and 96% with MLP. While it is 100% with BPSO. The reason for this is that BPSO selected 1321 features for the brain data set.

Dataset	Classification Algorithm	Precision	Recall	Accuracy	F1 Score
Colon	Linear SVM	1.0	1.0	100	1.0
	RBF SVM	1.0	1.0	100	1.0
	AdaBoost	0.96	0.95	95.45	0.95
	Voting Classifier	1.0	1.0	100	1.0
	Random Forest Classifier	1.0	1.0	100	1.0
	Multilayer Perceptron	1.0	1.0	100	1.0
SRBCT	Linear SVM	1.0	1.0	100	1.0
	RBF SVM	1.0	1.0	100	1.0
	AdaBoost	0.91	0.91	91	0.91
	Voting Classifier	1.0	1.0	100	1.0
	Random Forest Classifier	1.0	1.0	100	1.0
	Multilayer Perceptron	0.97	0.97	97.11	0.97
Prostate	Linear SVM	0.87	0.88	87.53	0.87
	RBF SVM	0.87	0.88	87.53	0.87
	AdaBoost	0.96	0.97	97.61	0.97
	Voting Classifier	0.86	0.87	87.33	0.87
	Random Forest Classifier	0.96	0.97	97.58	0.97
	Multilayer Perceptron	0.96	0.97	97.58	0.97
Lymphoma	Linear SVM	1.0	1.0	100	1.0
	RBF SVM	1.0	1.0	100	1.0
	AdaBoost	1.0	1.0	100	1.0
	Voting Classifier	1.0	1.0	100	1.0
	Random Forest Classifier	1.0	1.0	100	1.0
	Multilayer Perceptron	1.0	1.0	100	1.0
Brain	Linear SVM	0.91	0.93	93.33	0.93
	RBF SVM	0.91	0.93	93.33	0.93
	AdaBoost	0.64	0.61	61.11	0.61
	Voting Classifier	0.91	0.93	93.33	0.93
	Random Forest Classifier	0.94	0.88	88.88	0.88
	Multilayer Perceptron	0.96	0.96	0.96	0.96
Leukemia	Linear SVM	1.0	1.0	100	1.0
	RBF SVM	1.0	1.0	100	1.0
	AdaBoost	0.97	0.97	97.0	0.97
	Voting Classifier	1.0	1.0	100	1.0
	Random Forest Classifier	1.0	1.0	100	1.0
	Multilayer Perceptron	1.0	1.0	100	1.0

Table 3: Classification Metrics Obtained with the Selected Features on 6 Classifiers

Method / Dataset	GA-SVC	GA-RFC	GA-MLP	BPSO-SVC	BPSO-RFC	BPSO-MLP	Proposed-SVC	Proposec-RFC	Proposed-MLP
Colon	90.9	95.45	95.45	100	98.3	100	100	100	97.1
SRBCT	96	92.3	100	100	100	100	100	100	100
Prostate	84.3	87.5	98.39	100	99.5	98.5	199	97.3	97.58
Lymphoma	100	98.99	100	100	100	100	100	100	100
Brain	88.88	83.33	99.98	100	100	100	100	88	96
Leukemia	100	100	100	100	98.9	99.1	100	100	100

Table 4: Comparison of Classification Accuracy Obtained with the Proposed Method with other Feature Selection Techniques

5. Summary

Application of machine learning algorithms to high dimensional cancer gene expression datasets requires addressing the issues of a smaller number of samples, selection of the subset of relevant genes, class imbalance, multi-collinearity in features, and in-stability in the results. We have addressed these challenges in this chapter. With scaling, we brought the data to mean 0 and variance 1. This prevents any particular feature with high variance across samples from dominating the results. By bootstrap re-sampling, we increased the number of samples to more than 100 in each data set. By applying synthetic minority oversampling, we balanced the class distribution in the datasets. Then, by applying a combination of simulated annealing and partial least squares regression we selected very few relevant genes from the original data sets with thousands of genes. Selected genes were used to train 5 classifiers including ensemble classifiers Voting classifier and Random Forest. We observed very good performance metrics with all classifiers with almost all the data sets. In the next chapter, we present a technique that can solve two problems - that of inference of gene-regulatory-networks and that of gene-selection. The proposed technique uses extra tree regression and network centrality to infer GRNs and identify gene subsets for classification.

References

[1] Global Cancer Burden Growing, Amidst Mounting Need for Services. Available online: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing–amidst-mounting-need-for-services (accessed on 19 February 2024).

[2] Cancer. Available online: https://en.wikipedia.org/wiki/Cancer (accessed on 19 February 2024).

[3] Alromema, N.; Syed, A.H.; Khan, T. A hybrid machine learning approach to screen optimal predictors for the classification of primary breast tumors from gene expression microarray data. *Diagnostics* 2023, *13*, 708. [Google Scholar] [CrossRef] [PubMed]

[4]AbdElNabi, M.L.R.; Wajeeh Jasim, M.; El-Bakry, H.M.; Taha, M.H.N.; Khalifa, N.E.M. Breast and colon cancer classification from gene expression profiles using data mining techniques. *Symmetry* 2020, *12*, 408. [Google Scholar] [CrossRef]

[5] De Souza, J.T.; De Francisco, A.C.; De Macedo, D.C. Dimensionality reduction in gene expression data sets. *IEEE Access* 2019, *7*, 61136–61144. [Google Scholar] [CrossRef]

[6]JapanCancerSurvivorshipCountryProfile.Availableonline: https://cancersurvivorship.eiu.com/countries/japan/ (accessed on 22 December 2023).

[7]CancerStatisticsinJapan.2023.Availableonline: https://ganjoho.jp/public/qa_links/report/statistics/2023_en.html (accessed on 20 February 2024).

[8] Salem, H.; Attiya, G.; El-Fishawy, N. Classification of human cancer diseases by gene expression profiles. *Appl. Soft Comput.* 2017, *50*, 124–134. [Google Scholar] [CrossRef]

[9] Ayyad, S.M.; Saleh, A.I.; Labib, L.M. Gene expression cancer classification using modified K-Nearest Neighbors technique. *Biosystems* 2019, *176*, 41–51. [Google Scholar] [CrossRef] [PubMed]

[10] Yeganeh, P.N.; Mostafavi, M.T. Use of machine learning for diagnosis of cancer in ovarian tissues with a selected mRNA panel. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 3–6 June 2018; pp. 2429–2434. [Google Scholar]

[11] Dey, U.K.; Islam, M.S. Genetic expression analysis to detect type of leukemia using machine learning. In Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, 3–5 May 2019; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6. [Google Scholar]

[12] Akhand, M.A.H.; Miah, M.A.; Kabir, M.H.; Rahman, M.M.H. Cancer Classification from DNA Microarray Data using mRMR and Artificial Neural Network. *Int. J. Adv. Comput. Sci. Appl.* 2019, *10*, 106–111. [Google Scholar] [CrossRef][Green Version]

[13] Rukhsar, L.; Bangyal, W.H.; Ali Khan, M.S.; Ag Ibrahim, A.A.; Nisar, K.; Rawat, D.B. Analyzing RNA-seq gene expression data using deep learning approaches for cancer classification. *Appl. Sci.* 2022, *12*, 1850. [Google Scholar] [CrossRef]

[14] Erkal, B.; Başak, S.; Çiloğlu, A.; Şener, D.D. Multiclass classification of brain cancer with machine learning algorithms. In Proceedings of the 2020 Medical Technologies Congress (TIPTEKNO), Antalya, Turkey, 19–20 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–4. [Google Scholar]

[15] Almutairi, S.; Manimurugan, S.; Kim, B.G.; Aborokbah, M.M.; Narmatha, C. Breast cancer classification using Deep Q Learning (DQL) and gorilla troops optimization (GTO). *Appl. Soft Comput.* 2023, *142*, 110292. [Google Scholar] [CrossRef]

[16] Mallick, P.K.; Mohapatra, S.K.; Chae, G.S.; Mohanty, M.N. Convergent learning–based model for leukemia classification from gene expression. *Pers. Ubiquitous Comput.* 2023, *27*, 1103–1110. [Google Scholar] [CrossRef] [PubMed]

[17] Joshi, A.A.; Aziz, R.M. Deep learning approach for brain tumor classification using metaheuristic optimization with gene expression data. *Int. J. Imaging Syst. Technol.* 2023, *34*, e23007. [Google Scholar] [CrossRef].

[18]LeukemiaData.Availableonline://https://hastie.su.domains/CASI_files/DATA/leukemia.html(accessed on 9 January 2024).

[19] Feltes, B.C.; Chandelier, E.B.; Grisci, B.I.; Dorn, M. CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. *J. Comput. Biol.* 2019, *26*, 376–386. [Google Scholar] [CrossRef] [PubMed]

[20] Srividya-Sundaravadivelu/Cancer-Classification-Using Machine Learning. Available online: https://github.com/Srividya-sundaravadivelu/cancer-classification-Using-Machine-Learning(accessed on 7 January 2024).