

Sequential Optimization Based Feature Selection Algorithm For Efficient Cancer Classification And Prediction

A. W. Banjoko¹ & W. B. Yahya²

Department of Statistics

University of Ilorin, P.M.B. 1515

Ilorin, Nigeria

Email: banjokoalabi@gmail.com¹, dr.yah2009@gmail.com²

ABSTRACT

This study proposes an efficient method for optimal selection of feature subsets to enhance the classification performance of Support Vector Machine (SVM) in a binary and multiclass response high-dimensional genomic microarray data using Multi-Objective Optimization (MOO) approach. In a Monte-Carlo experiment, a pre-selection of the features was performed with the filter method based on Sidak alpha value to reduce the number of false positive features in the data. The optimal values of the tuning parameters for both the SVM cost and Radial Basis Function (RBF) kernel were determined by grid search in a 10-fold cross-validation. The SVM with RBF kernel was then fitted sequentially to select the set of near optimal genes that are correlated with the response class. The proposed algorithm was compared with the following four machine learning methods: Naïve bayes (NB), Random Forest (RF), Random Forest with variable selection (RFVS) and LASSO. The Misclassification Error Rate (MER) of the proposed method on simulated data was 1.1% with a sensitivity of 97.8% using four (near) optimal selected genes. In contrast, the MERs of NB, RF, RFVS and LASSO classifiers with 10, 10, 9 and 37 genes were 4.28%, 5.03%, 4.98% and 0.00% respectively using the data. Application of the proposed method on published Leukaemia data yielded an MER of 0.03% with a sensitivity of 99.95% based on three (3) optimally selected genes. On the other hand, the MERs of NB, RF, RFVS and LASSO classifiers for the Leukaemia data were 1.0%, 3.0%, 5.67% and 0.00% based on 93, 93, 2 and 31 genes respectively. These same fits of performance were achieved by all the methods considered on multiclass response DNA data set. The results generally showed that the proposed algorithm is more parsimonious and achieved better predictive performance than some of the existing methods considered. The sets of optimally selected gene subsets in the data employed here can be further investigated by molecular biologist to establish the pathology of these genes with respect to their respective tumour classes.

Keywords: Support Vector machines, Feature selection, Multi-Objective Optimization.

iSTEAMS Conference Proceedings Paper Citation Format

Banjoko, A. W. & Yahya, W. B. (2018): Sequential Optimization Based Feature Selection Algorithm for Efficient Cancer Classification and Prediction. Proceedings of the 14th iSTEAMS International Multidisciplinary Conference, AlHikmah University, Ilorin, Nigeria, Vol. 14, No 2 Pp 265-274

INTRODUCTION

Identification and selection of informative gene biomarkers that possess good biological relationship with the specific type of cancer tumour has eased the processes of non-clinical cancer tumour diagnosis and classification in the recent time [1,2]. This has been particularly possible due to the advent of microarray technology that allows for simultaneous monitoring of the expression profiles of several thousand of gene signatures in biological samples [3,4]. In any high-dimensional microarray data, a major problem is how to efficiently select the few gene signatures that are biologically correlated with the clinical response group of the tissue samples. The candidate gene subsets so selected must not only be statistically correlated with the tumour class of interest, but must also be clinically relevant to the cancer tumour being diagnosed [5,6]. In this study, a novel multi-objective feature selection algorithm that optimally selects the best gene biomarkers in high-dimensional genomic data for tissue sample classification with Support Vector Machine (SVM) [7, 8, 9] is presented

2. MATERIALS AND METHODS

2.1 Simulation Scheme

Binary response microarray data were simulated following the scheme adopted in [10]. In all, $n = 100$ observations with $p = 1000$ features (genes) were simulated which yielded a $n \times p$ data matrix of 100 samples by 1000 genes. The first 50 samples represent the first sample group, the tumour group (n_1) each coded (-1) and the second 50 samples represent the normal (tumour-free) sample group (n_2) coded (+1) such that $n_1 + n_2 = n$.

Of the 1000 features simulated on all samples, 10 of them were simulated from the mixture of two multivariate normal densities having covariance matrices Σ_1 and Σ_2 with respective mean vectors μ_1 and μ_2 and the mixing parameter π was taken to be 0.5. These 10 features were those whose expression levels were strongly related to the response groups and were thus labelled $g_1, g_2, \dots, g_{10}$. The remaining 990 features ($v_{11}, v_{12}, \dots, v_{1000}$) were the features with relatively low expression levels and were simulated from multivariate normal densities with covariance matrix Σ and mean vector μ . Generally, the covariance matrix Σ defined as $\Sigma = \{\sigma_{ij}\}$, has a block structure such that

$$\sigma_{ij} = \begin{cases} 0.2, & \text{if } |j - i| \leq 5 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

2.2 Real life dataset

A published microarray (Leukaemia and DNA) datasets by [4, 10, 11, 13] was also used to validate the result from simulation study. Details about the datasets are available online at <https://github.com/ramhiser/datamicroarray>. The Leukaemia dataset is a binary response microarray data with ...genes and two groups of mRNA samples. Whereas, the DNA dataset is a multiclass response microarray data havinggenes and three biological response groups.

3. METHODS

The proposed sequential algorithm used the Bayesian t-test by [11] for the binary response where $k = 2$ and the Welch's F – test by [12] for multiclass response where $k > 2$ at the pre-selection stage. This was used to prune the number of features in the dataset before the sequential optimization was performed. The number of false positive features were controlled by the Sidak alpha (α_s) value [14] at a chosen Family-Wise-Error Rate (FWER) of 5% such that;

$$\alpha_s = 1 - (1 - \alpha)^{\frac{1}{q}} \quad (2)$$

Where

α is the chosen FWER and q is the number of the tests performed.

Since the objective is to maximize accuracy, using the minimum gene subset possible, a subset $\underline{g}_h$ is said to dominate another subset $\underline{g}_h^*$ provided $\underline{g}_h$ is better in performance than $\underline{g}_h^*$ such that $\underline{g}_h > \underline{g}_h^*$ if

$$\forall i \in (1, 2, \dots, m) f_i(\underline{g}_h) \geq f_i(\underline{g}_h^*) \wedge \exists j \neq i \in (1, 2, \dots, m) f_j(\underline{g}_h) > f_j(\underline{g}_h^*) \quad (3)$$

The goal of the proposed method is to identify a non-dominated gene subset Pareto front, then the gene subset will be evaluated by its accuracy on the test data and the number of genes selected at the near optimal Pareto front. Specifically, for this study

$$f_1(\underline{g}_h) \geq f_1(\underline{g}_h^*) \wedge f_2(\underline{g}_h) < f_2(\underline{g}_h^*) \quad (4)$$

These criteria represented by f_1 and f_2 respectively and are use in solving the multi-objective fitness function below. Hence, the fitness function is obtained as;

$$f_1 = A(g) \quad (5)$$

$$f_2 = n(g) \quad (6)$$

$$fitness(F) \equiv f_1 \cap f_2 \quad (7)$$

Where

$A(g) \in [0,1]$ is the accuracy of the test data using only the expression values of the selected genes in a subset while $n(g) \geq 1$ is cardinality of the gene subset. The above multi-objective optimization will be solved using the interactive method.

The traditional SVM with the RBF kernel was fitted on each of the selected features by the Bayesian t-test and the MER of the respective features was computed. The feature with the minimum MER was the first best gene subset to enter into the algorithm. This feature was added pair-wisely to each of the remaining features among the selected ones. Again, the SVM was fitted on the pair-wise subsets and the gene subset with the minimum MER was also obtained. A test of equality of proportions was performed at 5% significance level to compare the successive pairs of MERs yielded by SVM classifiers that employed the first and second gene subsets. The algorithm continues until the test of proportion is not significance at a particular features' combination. At such point, the feature subset with the least MER becomes the final best gene subset for the data.

The test of proportion performed employed 1000 iterations from the Mote-Carlo Cross-Validation (MCCV) as the sample size for the test. At each iteration, the situations where all the observations in the test samples were correctly classified and when at least one of the observations in the test samples was wrongly classified (i.e. $MER_i = 0$ and $MER_i \neq 0$, for $i = 1, 2, \dots, 1000$) were considered. The flow chart of the proposed algorithm is presented as Fig 1. All analyses were performed with R statistical software (www.cran.r-project.org/).

4. ANALYSIS AND RESULTS

The Bayesian t-test (for $k = 2$) and the Welch's F-statistic (for $k > 2$) were employed at the pre-selection stage of this algorithm to selected the 10 features ($g_1, g_2, \dots, g_{10}$) that were simulated to have strong biological relationship with the response class at a FWER of 5%. Thereafter, the optimal SVM cost (C) and the value of the RBF kernel hyper-parameters (γ) were determined by the grid search method using a 10-fold cross validation. This was also done for every gene subset at each iteration before the SVM classifier was fitted. The data were partitioned into the train-test sets ratio of 95:5 respectively [10]. The SVM was fitted on the train data and the performance of each subset was evaluated on the test set. This was done to obtain the out-of-bag (OOB) performance of the algorithm. The MCCV approach was used to obtain the OOB performance at 1000 iterations. The proposed sequential feature selection method finally identified and selected the best 4 optimal gene subsets from the 10 pre-selected genes by Bayesian t-test for classification of the tissue samples with SVM. The classification results at each feature selection steps are shown in Fig 2 by the plot of the estimated MERs against the features selected.

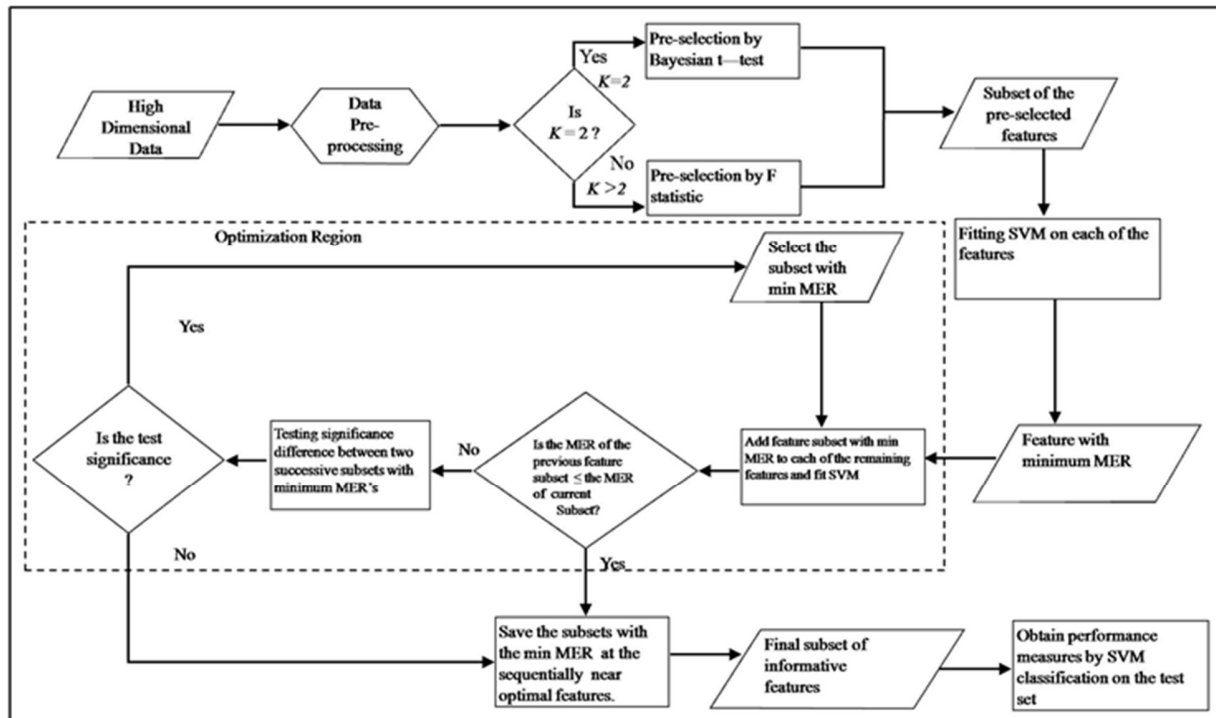


Fig. 1: The Flow chart of the proposed optimal sequential feature selection algorithm for binary and multiclass classification.

The performance indices of the proposed algorithm as employed using the four optimal selected features are presented in Table 1. Similarly, the performance indices of the proposed algorithm utilizing the three optimal selected features are presented in Table 2. The result of the Hierarchical Clustering with Single Linkage Search (HCSLS) using the four optimally selected features (genes) by the proposed method is presented in the dendrogram of Fig 3. In order to assess the goodness of the four selected optimal genes, each of the gene sets were employed to fit Principal Component (PC) models and the first constructed principal component (PC1) from each model was used to classify the two response groups in the data. Fig 4 presents the plot of the PC1 against all the 100 biological samples based on the four (4) selected optimal genes by sequential method proposed here.

Table 1: The estimated performance indices of the of the proposed method using four (4) optimal features selected in the simulated data.

Performance Indices of the proposed algorithm using the four selected features	Estimated Values (in %)	Standard Deviations of Estimates
Misclassification Error Rate (MER)	1.1	0.0513
Sensitivity	97.8	0.1026
Specificity	100	0.0001
Positive Predictive Value (PPV)	100	0.0001
Negative Predictive Value (NPV)	98.5	0.0684
Jaccard Index (JI)	97.8	0.1026
Cross-validated AUC	99.9	0.0176

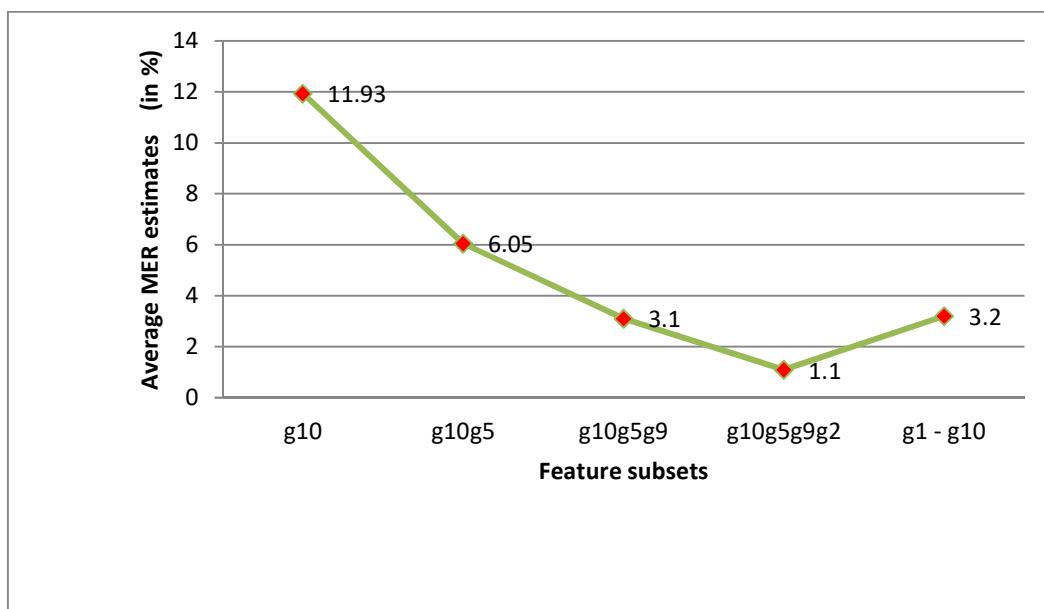


Fig. 2: The graph of successive estimated average MER at each selection step and the MER yielded by the first 10 genes (g1-g10) pre-selected by the Bayesian t-test

Table 2: The estimated performance indices of the of the proposed method using three (3) optimal features selected in the Leukaemia data.

Performance Indices of the proposed algorithm using the Three selected features	Estimated Values (in %)	Standard Deviations of Estimates
Misclassification Error Rate (MER)	0.03	0.0105
Sensitivity	99.95	0.0124
Specificity	100	0.0001
Positive Predictive Value (PPV)	100	0.0001
Negative Predictive Value (NPV)	99.90	0.0324
Jaccard Index (JI)	99.95	0.0124
Cross-validated AUC	100	0.0001

Table 3: The estimated performance indices of the proposed method using two (2) optimal features selected in the DNA Methylation data.

Performance Indices of the proposed algorithm	Estimated Values (in %)	Standard Deviations of Estimates
Misclassification Error Rate (MER)	0.02	0.0045
Sensitivity Blood	100	0.0001
Sensitivity Placenta	100	0.0001
Sensitivity Other	99.97	0.0069
Specificity Blood	100	0.0001
Specificity Placenta	99.98	0.0047
Specificity Other	100	0.0001
Positive Predictive Value (PPV) Blood	100	0.0001
Positive Predictive Value (PPV) Placenta	99.98	0.0453
Positive Predictive Value (PPV) Other	100	0.0001
Negative Predictive Value (NPV) Blood	100	0.0001
Negative Predictive Value (NPV) Placenta	100	0.0001
Negative Predictive Value (NPV) Other	99.96	0.0101

5. DISCUSSION OF RESULTS

A new algorithm that sequentially selects optimal gene signatures for binary tumour classification with SVM is proposed in this study. The result plotted in Fig. 2 shows a monotonically decreasing trend in the average MERs yielded by successive SVM classifiers as more features were selected by the sequential algorithm. The best optimal gene subset with the following four genes g_2 , g_5 , g_9 , g_{10} with the least MER of 1.1% was selected by the algorithm. This result was optimal as it was better than the result of the SVM that employed all the 10 selected features by the Bayesian t-test with average MER of 3.2% as shown in Fig 2. This shows that even when a pre-selection of features have are done, there are still large chances that the selected features can be optimized so as to obtain the minimum possible feature subsets that will maximize the classification accuracy.

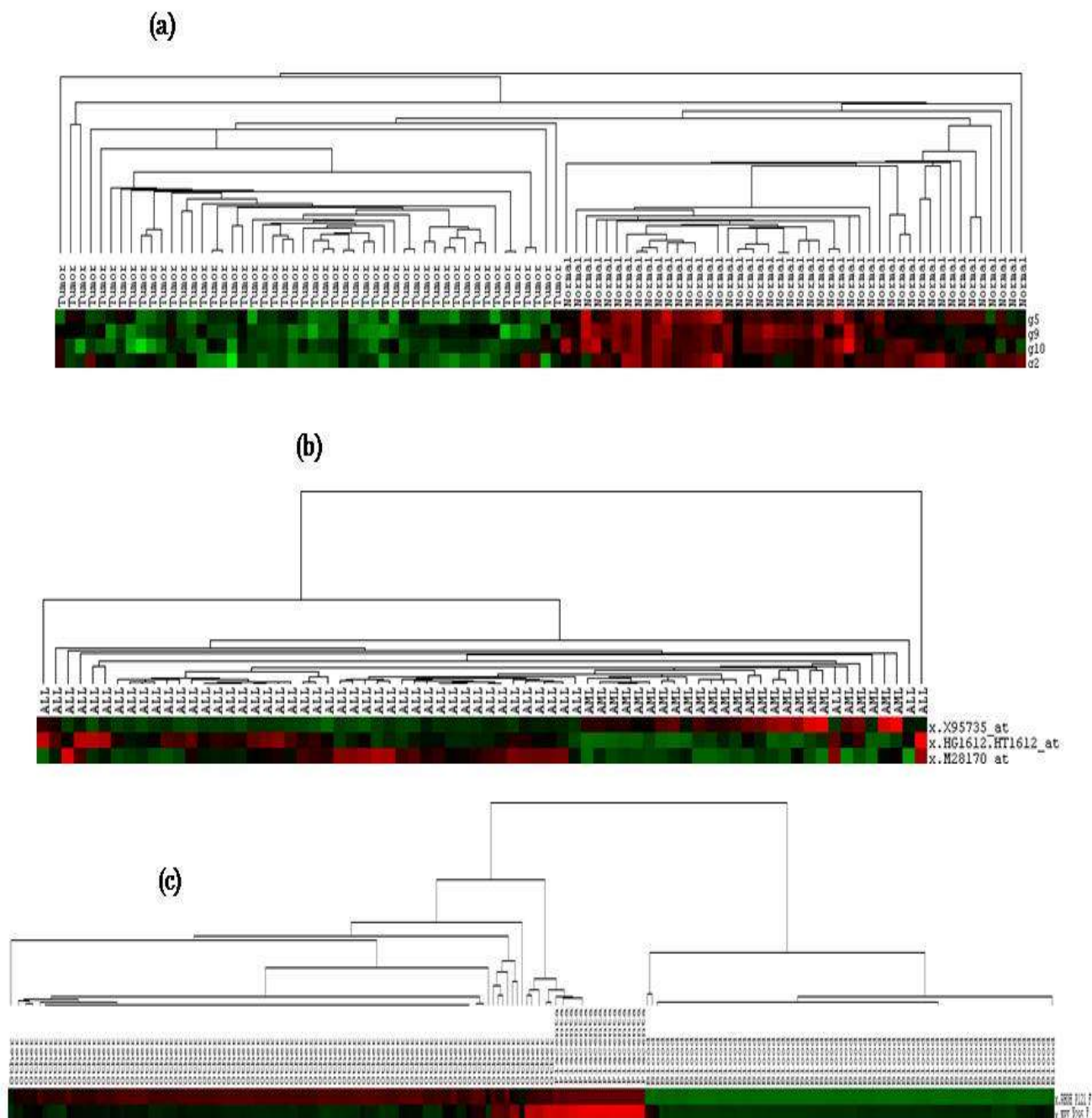


Fig. 3: The dendrogram of the clustering (HCSLS) result using the optimal sequentially selected features for the simulated (a), leukaemia (b) and DNA (c) data sets respectively

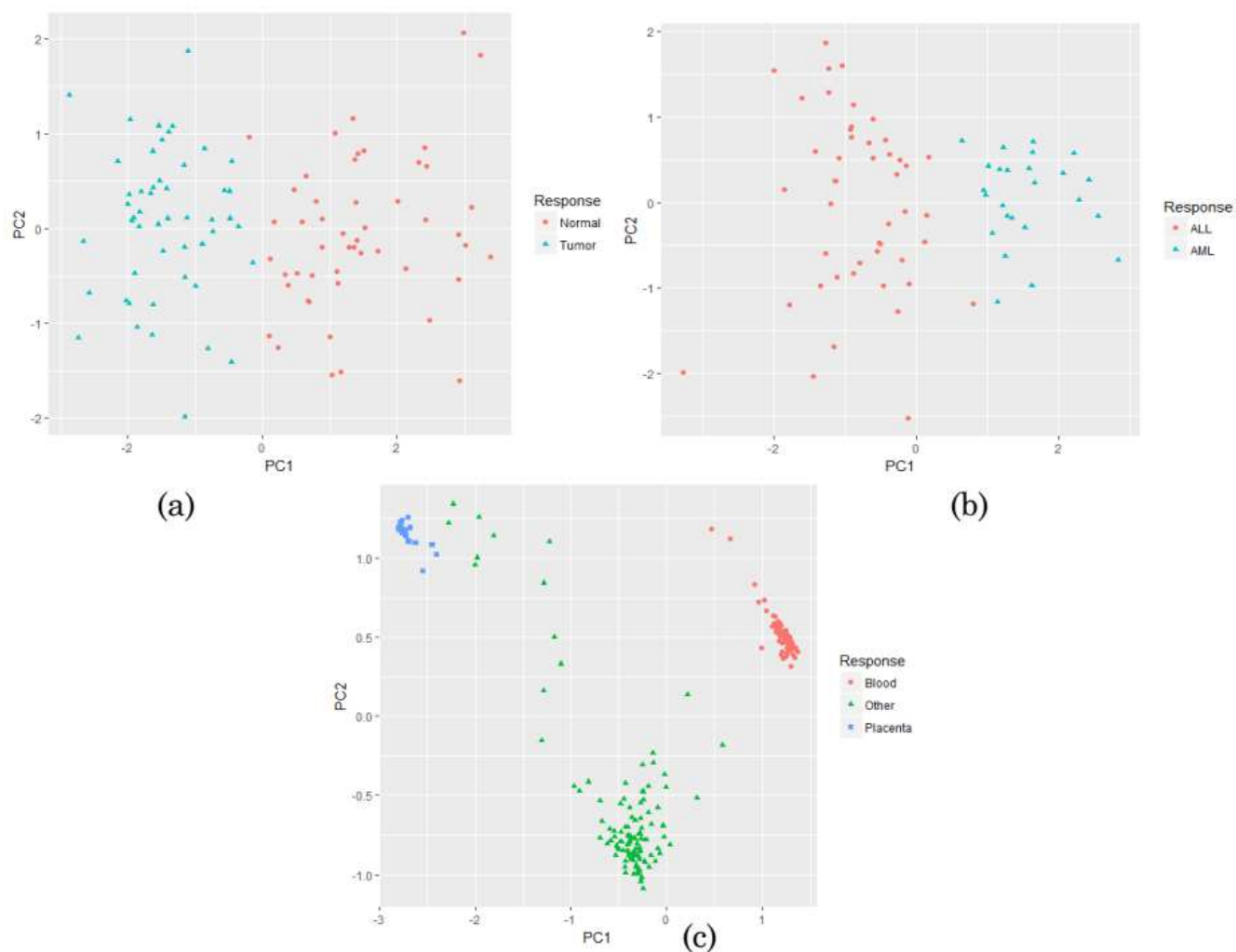


Fig. 4: Plot of the first PC (PC1) against the second PC (PC2). The PCs were constructed using the selected optimal genes subsets for the simulated (a), Leukaemia (b) and DNA (c) data respectively.

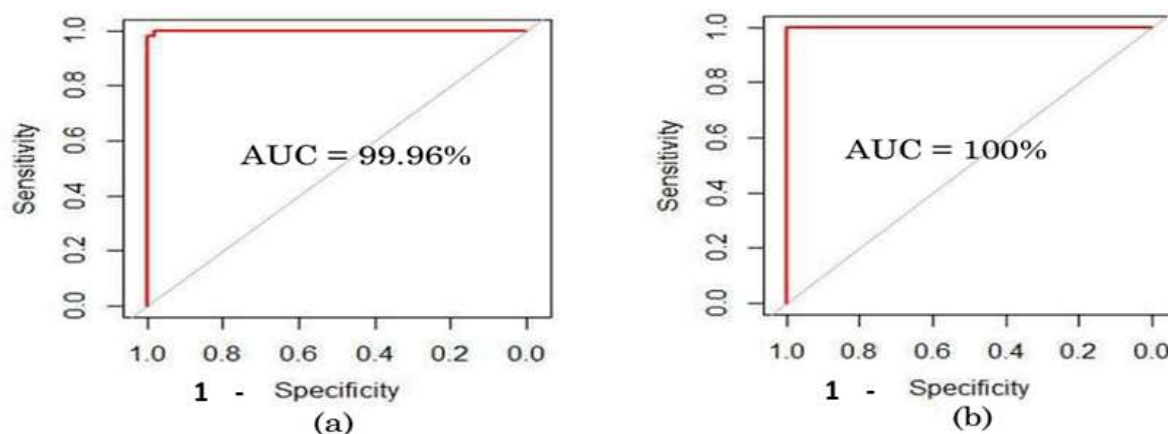


Fig. 5: Plot of ROC with Cross-Validated AUC for the simulated (a), Leukaemia (b) data respectively.

Figs. 3 presents the results of the Hierarchical Clustering with Single Linkage Search (HCSLS) using the near optimally selected gene subsets. Also, the strength of the optimal gene subsets at discriminating the tumour groups (types) in the sample was demonstrated by the plot of the first Principal Components (PC1) against the second Principal Components (PC2) for the datasets considered as shown in Fig 4. Finally, Table 3 presents the performance of the proposed algorithm compare to some of the existing ones. The result shows that the proposed algorithm competes favourably with some existing methods and is more parsimonious.

Table 3: The correct classification rates (CCR) of the proposed algorithm and those of the three existing methods selected. Number of genes used for classification by each method are in parenthesis.

Data set	Proposed 1-	NB	RF	FWC 1-	LASSO
Simulated	98.90% (4)	95.72% (10)	94.97% (10)	95.03% (9)	100% (37)
Leukaemia	99.97% (3)	99.00% (93)	97.00% (93)	91.33% (2)	100% (31)
DNA	99.98% (2)	100% (1055)	100% (1055)	100% (21)	100% (13)

6. CONCLUSION

This work has presented a novel algorithm for efficient feature selection and classification of biological sample in binary and multiclass response high-dimensional (microarray) data. The goodness of the features selected by this new technique has been demonstrated with SVM classification method using simulated high-dimensional, published Leukaemia and DNA data sets. The results from this work have again underscored the possibility of non-clinical classification of various cancer tumours using the gene expression profiles of biological samples.



REFERENCES

- [1] Cooper C.S (2001): Applications of microarray technology in breast cancer research. *Breast Cancer Research*, **3**: 158-171.
- [2] Hapfelmeier A, Yahya WB, Robert R, Ulm K. (2011): Predictive modeling of gene expression data. *Chapter 1, Hand book of Statistics in Clinical Oncology, 3rd ed., Chapman & Hall/CRC, New York.*
- [3] Rápolti E, Szigeti A, Farkas R, Bellyei S, Boronkai A, Papp A, Gömöri É, Horváth ÖP, Mangel L. (2009) Neoadjuvant radiochemotherapy in the treatment of locally advanced rectal tumors. *Hungarian Oncology*. **53**:345–349.
- [4] Yahya W. B, Aremu G. T and Garba M. K (2015): Multiclass Sequential Feature Selection and Classification Method for Genomic Data. *Journal of Applied Science and Technology*, **20** (1&2): 50-60.
- [5] Ting-Lee M.-L. (2004): *Analysis of microarray gene expression data*. Springer, New York.
- [6] Banjoko A. W., Yahya W.B., Garba M. K., Olaniran O. R., Olorede K. O. and Dauda K. A. (2015): Efficient Support Vector Machine Classification of Diffuse large B-Cell Lymphoma and Follicular Lymphoma mRNA Tissue Samples. *Annals. Computer Science Series*, **13** (2): 69 – 79.
- [8] Cristianini N., Shawe -Taylor J. (2012): *An introduction to support vector machines*. Cambridge University Press, UK.
- [9] James G., Witten D., Hastie T., Tibshirani R. (2013): *An Introduction to Statistical Learning with Applications in R*. Springer Science + Business Media New York.
- [10] Yahya WB, Ulm K, Fahrmeir L, Hapfelmeier (2011): *k*-SS: a sequential feature selection and prediction method in Microarray Study, *International Journal of Artificial Intelligence*, **6**(S11):19-47.
- [11] Oyebayo O. R. and W. B. Yahya (2017): Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique. *Journal of Modern and Applied statistical methods*. **16** (2): 185-196.
- [12] Welch B. L (1951): On the comparison of several mean values: An alternative approach. *Biometrika*, **38**, 330–336.
- [13] 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, Bloomfield C.D & Lander E. S. (1999): *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*. *Science*, **286**.5439: 531-537.