EFFICIENT SUPPORT VECTOR MACHINE CLASSIFICATION OF DIFFUSE LARGE B-CELL LYMPHOMA AND FOLLICULAR LYMPHOMA MRNA TISSUE SAMPLES

A. W. Banjoko[1], W. B. Yahya[1], M. K. Garba[1], O. R. Olaniran[1], K. A. Dauda[2], K. O. Olorede[2]

[1] Department of Statistics, University of Ilorin, P.M.B. 1515, Ilorin, Nigeria
[2] Department of Statistics and Mathematical Sciences, Kwara State University, Malete, P.M.B 1530, Ilorin, Nigeria

Corresponding author: W. B. Yahya (dr.yah2009@gmail.com; wbyahya@unilorin.edu.ng)

ABSTRACT: In this study, an efficient Support Vector Machine (SVM) algorithm that incorporates feature selection procedure for efficient identification and selection of gene biomarkers that are predictive of Diffuse Large B–Cell Lymphoma (DLBCL) and Follicular Lymphoma (FL) cancer tumor samples is presented. The data employed were published real life microarray cancer data that contained 7,129 gene expression profiles measured on 77 biological samples that comprised 58 DLBCL and 19 FL tissue samples. The dimension reduction approach of the Welch statistic was employed at the feature selection phase of the SVM algorithm. The cost and kernel parameters of the SVM model were tuned over a 10–fold cross-validation to improve the efficiency of the SVM classifier. The entire sample was randomly partitioned into 95% training and 5% test samples. The SVM classifier was trained using Monte Carlo Cross-validation approach with 1000 replications. The performance of this classifier was assessed on the test samples using misclassification error rate (MER) and other performance measures. The results showed that the SVM classifier is quite efficient by yielding very high prediction accuracy of the tumor samples with fewer differentially expressed genes. The selected gene biomarkers in this work can be subjected to further clinical screening for proper determination of their biological relationship with DLBCL and FL tumour sub-groups. However, more studies with large samples might be needed in future to validate the results from this work.
KEYWORDS: SVM, Diffuse Large B-Cell Lymphoma, Follicular Lymphoma, 10-fold cross-validation, Welch Statistic.

1. INTRODUCTION

Early identification of cancer tumour in patients has been known to be of great help for efficient management and treatment of such cancer types. However, the management and treatment of cancer tumour largely depends on the specific types of cancer ([Sch05]). Most of the clinical diagnostic methods often require thorough examination of the tumour cell in a microscope before proper diagnoses can take place which obviously might take considerable longer time before the cancer status of the ailing patients can be determined ([YRU14]). A major consequence of this is the risk of having the cancer tumour (if present in the patient) metastasizes to the neighboring tissue cell areas due to prolonged period of incubation.

One of the most common lymphoid malignancies in adults is Diffuse Large B–Cell Lymphoma (DLBCL) and is curable in less than 50% of patients (Shipp et al. 2002). However, like other forms of cancer tumours, clinical diagnosis of this type of tumor has been reported to be very slow before the true nature of the tumour can be determined. But studies have shown that cancer tumour might be discovered faster with microarray analysis than with clinical methods ([NCG03]; [YOJ12]; [YRU14]).

The advent of microarray technology that enables simultaneous investigation of several thousands of gene expression profiles that interact with the tissue samples to produce clinical responses has thrown up additional challenges in the analysis of microarray cancer data. It is a known fact that only few of among the thousands of gene expression signatures are actually biomarkers of the patient's clinical outcome ([YRU14]). Hence, the need for efficient non-clinical techniques for proper identification and selection of such few biomarker genes becomes necessary.

As reported by Ting-Lee ([Tin04]), identification of biomarkers involves the selection of genes that are correlated with clinical responses of the tissue samples. Such exercise will give rise to the following advantages; (i) Reduce the dimension of the data space from $n \times p$ to $n \times k$ with $k \ll p$ (ii) Remove irrelevant and noisy genes from the data (iii) Yield a small subsets of genes that might be both statistically significant and biologically relevant (iv) Reduce the computational time (v) Improve the classification accuracy (vi) Reduce cost of investigation in clinical settings.

The traditional support vector machine (SVM) algorithm ([Yah09]; [Ya12]; [Vap95]) has been shown to be very efficient machine learning tool for classification of response groups. SVM has become a state-of-the-art performance in real-world applications such as text categorization, hand written character recognition, image categorization, bio-sequence analysis ([CS12]).

The basic idea behind the SVM method is to construct an optimal separating hyperplane for the two sample groups by mapping the gene expression data to a higher-dimensional space. This involves finding a hyperplane defined by a weight vector ***w***and a bias ***b*** such that the separation of the two groups is maximized in a specific sense. Using kernel representations, linear separation in the higher-dimensional space corresponds to a nonlinear decision boundary in the original space.

In this present work, a thorough review of the SVM method for cancer tumour classification in a binary response microarray data problem is presented. Further techniques to improve the classification performance of the original SVM method by incorporating variable selection approaches using different levels of *family-wise error rate* (FWER) α_F were proposed. The efficiency of this improved SVM algorithm was examined on a published Affymetrix microarray cancer dataset for feature selection and tissue sample classification in a binary response data structure.

2. MATERIAL AND METHODS

2.1 Data Description

The data used in this work are published data set ([S+02]) that contained 7,129 gene expressions profiling on 77 biological subjects. Of these 77 subjects, 58 were *diffuse large B-cell lymphoma* (DLBCL) tissue samples while the remaining 19 subjects were *follicular lymphoma* (FL) tissue samples. The data are freely available and can be accessed at www.genome.wi.mit.edu/MPR/lymphoma.

2.2 Flow-Chat of the Conceptual SVM Feature Selection and Classification Method

The flow-chat of the conceptual SVM feature selected and classification method as employed in this work is presented by Fig 1.

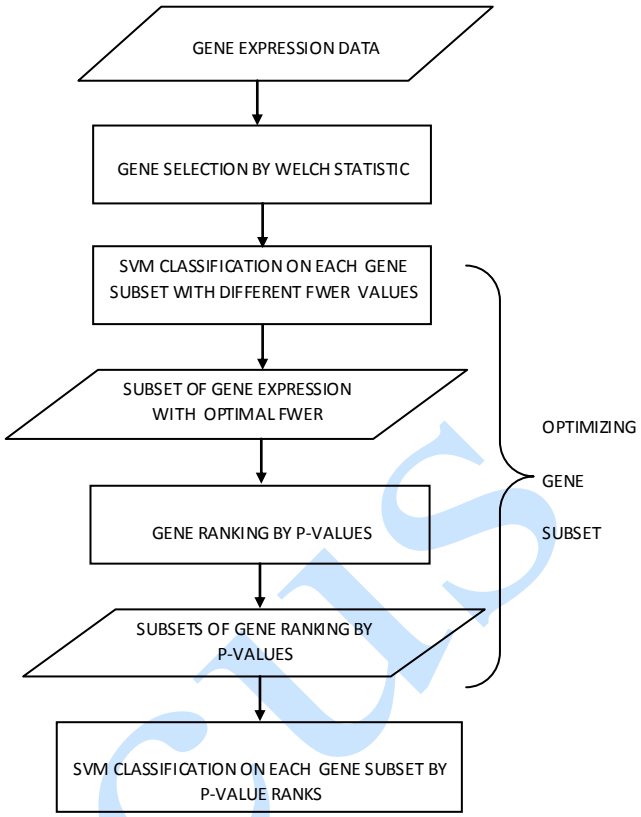

Fig 1: Conceptual SVM algorithm for gene selection and classification of tissue samples

2.3 The Mathematics of The SVM Method

In this section, overview of the concepts of SVM method, its kernel functions, and choice of kernel are presented.

2.3.1 *Overview of SVM for hard margin*

Given n training points such that each input x_i has D attributes and is in one of the two classes $y_i = \pm 1$ .Thus, the training dataset is of the form $\{x_i, y_i\}$, where $i = 1,2, \dots, n$ and $y_i \in \{-1, +1\}, x_i \in \Re^D$. Here, we assume that the data is linearly separable into two classes by drawing a line for D = 2 and a hyperplane for D > 2. Let x_i be the nearest data point to the hyperplane and w be the weight vector orthogonal to the hyperplane, then we have the condition that:

$$w'x = 0 \quad (1)$$

For any bias b , the equation of the hyperplane can therefore be described by

$$w'x + b = 0 \quad (2)$$

To normalize w with minimum x, it is required that

$$|w'x_i + b| = 1 \quad (3)$$

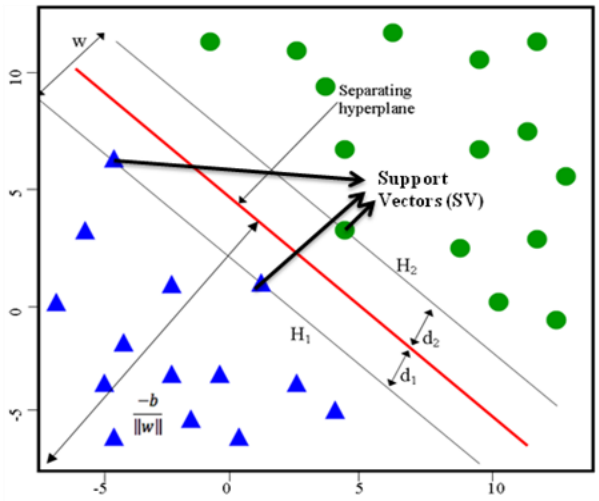

Fig 2: The graph showing a typical separating hyperplane and the maximal margin hyperplane (H_1 and H_2) for the linearly separable subjects with two distinct subject groups. The support vectors are indicated as shown on the graphs

The implementation of SVM boils down to selecting w and b so that the training data $\{x_i.y_i\}$ can be describe by

$$x_i.w + b \geq +1, for\ y_i = 1 \quad (4)$$
$$x_i.w + b \leq -1, for\ y_i = -1 \quad (5)$$

Equations (4) and (5) above reduced to

$$y_i(x_i.w + b) - 1 \geq 0 \quad (6)$$

A typical separating hyperplane and the maximal margin hyperplanes for the linearly separable biological subjects with two distinct subject groups are presented by Fig 2.

From Fig 2, the SVM margins (distances) that separated the two biological sample groups from the separating hyperplane to the two maximal margin hyperplanes H_1 and H_2 are denoted by d_1 and d_2 respectively with $d_1 = d_2 = d$. Hence, the main objective of the SVM is to maximize the distance d that separated the two sample groups. This is stated by the objective function:

$$max_{w,b}\, d \quad (7)$$

subject to inequality statement in (6).

Let a unit vector $\hat{w}$ of w be obtain by

$$\hat{w} = \frac{w}{\|w\|} \quad (8)$$

Thus, distance d can be computed by

$$d = \hat{w}'(x_i - x) \quad (9)$$

Since d is non-negative, therefore

$$d = |\hat{w}'(x_i - x)| \quad (10)$$

Using (8) in (10), we have that

$$d = \frac{|w'x_i - w'x|}{\|w\|}$$
$$\rightarrow$$
$$d = \frac{|(w'x_i + b) - (w'x + b)|}{\|w\|}$$

and by (2) and (3), the above becomes

$$d = \frac{1}{\|w\|} \quad (11)$$

Hence, maximizing d subject to (6) is equivalent to minimizing $\|w\|$. Thus, the objective function becomes

$$\min\|w\| \quad (12)$$

subject to (6).

2.3.2 *The optimization problem*

As reported by Fletcher ([Fle09]) and Yahya ([Yah12]) among others, optimizing (12) is equivalent to $\frac{\min \square 1}{2}\|w\|^2$ which makes it possible to perform the Quadratic Programming (QP) optimization. Hence, the constrained optimization is obtain as

$$\frac{\min \square 1}{2}\|w\|^2 \quad (13)$$

subject to (6).

The solution is then obtain by introducing a Lagrange multiplier α_i in the constraint (6) and adding to the objective function as

$$\alpha_i[y_i(w'x_i + b) - 1] = 0 \quad (14)$$

By adding (14) to the objective function (13) we have

$$\frac{\min 1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(w'x_i+b)-1] \quad (15)$$

which is equivalent to

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i y_i(w'x_i+b) + \sum_{i=1}^{n}\alpha_i \quad (16)$$

The objective here is to find the values of w that minimizes (16) and b that maximizes α for all $\alpha_i \geq 0$. This is achieve by differentiating $\mathcal{L}(w,b,\alpha)$ with respect to w and b and equating them to zero. Thus, we have;

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n}\alpha_i y_i x_i \quad (17)$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \quad (18)$$

Therefore, we have the function

$$\mathcal{L}(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=i}^{n}\alpha_i\alpha_j y_i y_j x_i x_j \quad (19)$$

with $\alpha_i \geq 0, \forall i$.
Maximizing the objective function $\mathcal{L}(\alpha)$ in (19) using QP method is equivalent to minimizing $-\mathcal{L}(\alpha)$. That is;

$$\min_{\alpha}\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=i}^{n}\alpha_i\alpha_j y_i y_j x_i x_j - \sum_{i=1}^{n}\alpha_i\right) \quad (20)$$

By Karush-Kuhn-Tucker (KKT) condition ([Kar39]; [KT52]), some of the $\alpha_i's$ would be zero at the optimal solution level and they are those $\alpha_i's$ for which the margin

$$\alpha_i[y_i(w'x_i+b)-1] = 0 \quad (21)$$

Therefore, it is either $\alpha_i = 0$ or $[y_i(w'x_i+b)-1] = 0$. Since α_i cannot be zero, it shows that equation $y_i(w'x_i+b) = 1$ has to be solved for w and b.
If $\alpha_i > 0$ then x_i is a support vector (SV). The SVs are the points that defines the plane and margin for which the α 's are positive. Thus, the value of w is obtained by

$$w = \sum_{x_i \in SV}\alpha_i y_i x_i \quad (22)$$

And the value of bias b can be obtain by solving the equation $y_i(w'x_i+b) = 1$ using any of the support vectors.

2.3.3 SVM *for soft margin*

Sometimes, the data are not perfectly linearly separable and it might be necessary to consider a classifier based on a hyperplane that does not perfectly separate the two classes in the interest of greater robustness to individual observations and better classification of most of the training observations James et al. ([J+13]).A typical non-separable hyperplane with some misclassified samples in the two groups are shown by Fig 3.

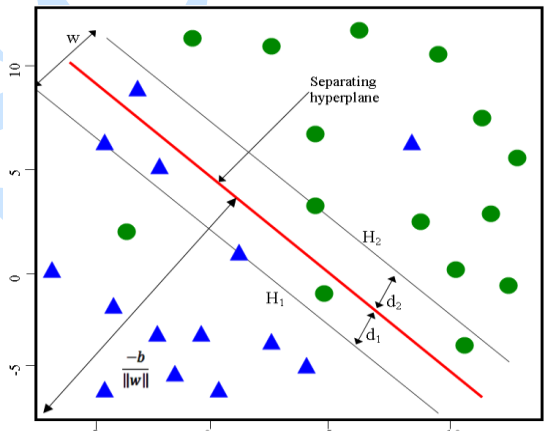

Fig 3: The graph showing a typical separating hyperplane and the maximal margin hyperplanes (H_1 and H_2) for the linearly non-separable subjects with two distinct subject groups. Some of the misclassified samples are indicated by the respective symbols on the graph.

In order to extend the SVM algorithm to handle data that is not fully linearly separable, we relax the constraints (4) and (5) slightly to allow for misclassified points. This is done by introducing a positive slack variable $\varepsilon_i, i = 1, \ldots, n$ as follows;

$$x_i.w + b \geq 1 - \varepsilon_i \quad for\ y_i = 1 \quad (23)$$

$$x_i.w + b \leq -1 + \varepsilon_i \quad for\ y_i = -1 \quad (24)$$

with $\varepsilon_i \geq 0$. Equations (23) and (24) can be combined to have

$$y_i(x_i.w + b) - 1 + \varepsilon_i \geq 0 \quad (25)$$

Each of the positive slack variable ε_i are called violations and total violations is $\sum_{i=1}^{n} \varepsilon_i$ ([J+ 13]). The new optimization problem then becomes

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\varepsilon_i \quad (26)$$

Subject to the constraint in(25).
By re-formulating and introducing a new Lagrange multiplier β_i as done in (14) through (19), we have

$$\mathcal{L}(w,b,\varepsilon,\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{n}\alpha_i\left(y_i(w'x_i+b)-1+\varepsilon_i\right) - \sum_{i=1}^{n}\beta_i\varepsilon_i \quad (27)$$

Minimizing $\mathcal{L}(w,b,\varepsilon,\alpha,\beta)$ w.r.t w, b and ε, and maximizing it w.r.t each $\alpha_i, \alpha_i \geq 0$ and $\beta_i \geq 0$ we have back the objective function (21) as

$$\min_{\alpha}\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=i}^{n}\alpha_i\alpha_j y_i y_j x_i x_j - \sum_{i=1}^{n}\alpha_i\right)$$

subject to the constraint that $\sum_{i=1}^{n}\varepsilon_i \leq C$ for $0 \leq \alpha_i \leq C, i = 1,2,\dots,n$. The values of w and b can then be obtained as before.

2.3.4 *The kernel function*

As a kernel based machine learning algorithm, the four most commonly known SVM kernels are: i.) The linear kernel: $K(x_i.x) = (x_i.x)$, ii.) The polynomial kernel: $K(x_i.x) = [(\gamma x_i.x + r)]^d$, iii.) The Radial Basis Function (RBF): $K(x_i.x) = \exp(-\gamma\|x_i - x\|^2)$ and iv.) The sigmoid: $K(x_i.x) = tanh(\gamma x_i.x + r)$, where γ, r and d are the kernel parameters ([Vap95]; [Yah12]).
The concept of kernel, as it relates to the feature space, is considered in order to address the non-linearity problem often encountered in many classification problems. It turns out that the non-linearity in a lower dimensional space is actually linear in a higher dimensional space using a suitable mapping $X \to \emptyset(X)$. The SVM algorithm described above only involved the *inner products* of the observations (as opposed to the observations themselves) within the kernel setting.
In practice, the optimal SVM and the kernel parameters are obtain through grid search and they are those that give the minimum misclassification error.
In the traditional SVM, the choice of kernel is by trial and error. This involves searching for the best parameters for the four kernels through grid search and applying each of the kernels for classification. The best kernel is the one with minimum misclassification error. But works by Keerthi and Lin ([KL03]) and James et al. ([J+13]) suggest the use of linear kernel when the number of gene chips p is larger than the sample size n ($p \gg n$).The RBF kernel performs better for $p \ll n$ ([H+10]; [KL03]; [Cic15]), while the polynomial kernel contained too many parameters and suffers from numerical difficulties Chih-Jen ([LL03b]). Also, the study by Lin and Lin ([LL03a]) and Chih-Jen ([LL03b]) showed that the sigmoid kernel is not always a valid kernel and does not in any way better than the RBF kernel whenever the RBF works.
As a result of the above reasons, our choice of kernels in this study shall be restricted to both the linear and RBF kernels depending on the number of features (p) and the sample size (n) in the data.

3. ANALYSIS

The first step to take based on our proposed improved SVM implementation as presented in the flow-chat in Fig 1 is to perform feature selection using the Welch statistic ([Wel47]). This is meant to filter out the numerous noisy and non-informative genes (in univariate sense) in the data. Using the scheme adopted by Mahata and Mahata ([MM07]) for features' optimization, the remaining few selected informative genes were ranked by their respective p-values as provided by the values of their Welch statistic.
For a binary response group $y_i \in \{-1, +1\}$ as the case here, $y_i = -1$ indicates that the ith sample is from DLBCL tumour group and $y_i = +1$ indicates that the ith sample is from the FL tumour group. Therefore, the Welch statistic that selects gene expression profile X_j that discriminates these two tumour groups is of the form;

$$t_w = \frac{\bar{X}_{+1j} - \bar{X}_{-1j}}{\sqrt{\frac{S_{+1}^2}{n_{+1}} + \frac{S_{-1}^2}{n_{-1}}}} \sim t_v$$

for $j = 1,2, \dots, G$ (number of gene expressions in the data) where v is a modified degrees of freedom of the form

$$v = \frac{\left(\frac{S_{+1}^2}{n_{+1}} + \frac{S_{-1}^2}{n_{-1}}\right)^2}{\left(\frac{S_{+1}^4}{n_{+1}^2(n_{+1}-1)} + \frac{S_{-1}^4}{n_{-1}^2(n_{-1}-1)}\right)}$$

In order to control for the number of false positives in the preliminary gene selection loop of the SVM algorithm, the p-values of the Welch statistic were compared with the Family-Wise-Error rate (FWER) α_F given by $\alpha_s = 1 - (1 - \alpha_F)^{1/G}$ at some specified nominal Type I error rate α_F in a multiple hypothesis testing for comparing G gene variables ([Sid67]). By this, gene X_j would be selected to be differentially expressed if its p-value from the Welch statistic, say p_j is less than α_s (i.e. if $p_j < \alpha_s$).
Various values of α_F (1%, 5%, 10%, 15%, 20% and 100%) were employed in this work for optimal search of differentially expressed genes. The best kernel and SVM parameters were determined through grid search with 10-fold cross-validation for each of the gene subsets selected at each chose α_F values.
The entire sample size was partitioned into 95% training sample which was used to build the SVM classification model and 5% test sample which was used to assess its performance. The SVM algorithm for classification was run over 1000 MCCV runs for each selected gene subsets with the appropriate kernel to ensure model stability.
The subset of genes with the best accuracy was selected as the combination of gene expressions that are most appropriate for the tumor sample classification. As adopted by Mohamad et al. ([M+07]), this subset of genes was ranked based on their respective p-values from the Welch statistic. In order to optimize this feature selection process, the first few genes (i.e. the first 5, first 10, first 15, first 20 and first 25 genes in ranks) from the ranked genes were again re-selected for classification of the two tumour groups for optimal results.
The improved SVM algorithm for feature selection and classification was implemented in R statistical package. The R library '**e1071**' was employed for the implementation of the SVM method.

4. RESULTS

The results of the improved SVM method for feature selection and classification of the DLBCL and FL tumour groups are presented in this section.

In Table 1, the numbers of gene signatures selected by the selection scheme of the Welch statistic at various chosen FWERs are presented. The SVM algorithm for tissue sample classification was implemented using the linear kernel. This choice of kernel is appropriate for the data since the number of features p (7,129 genes) is greater than the sample size n (77 samples).
The optimal value of the SVM parameter C at each chosen FWER was determined using 10-fold cross-validation as presented in Table 1. The cross-validation (CV) error rates over all the possible values of the SVM parameter C were computed. The value of C that yielded the least CV error rate becomes the optimal SVM parameter value C at each chosen FWER as indicated in Table 1.
For a clearer overview of the performance of the SVM while tuning the parameter of the SVM, we provide the graph that shows the plot of minimum cross-validation error (misclassification error) rate against the values of the FWER employed for feature selection in Fig 4. The number genes selected at each FWER level are provided in parenthesis on the graph.
Having determined the optimal values of the SVM parameter C at all the chosen levels of the FWER, these were then employed for proper SVM classification using the training data over 1000 Monte-Carlo cross-validation (MCCV) runs. Linear kernel was employed for the SVM implementation since $n \ll p$ in all the six gene subset cases considered (i.e. at the six chosen FWERs considered).
The results of the SVM classification performance over 1000 MCCV runs were presented in Table 2. The table reported the test sample percent correct classification rate (%CCR), misclassification error rate (MER), sensitivity, specificity, positive (+) predictive value, negative (-) predictive value and the Jaccard index of the SVM classifier.
The classification performance of the SVM in Table 2 showed that the best value of FWER for the data is 1%. This is the α_F value that yielded the highest prediction accuracy (98.5%CCR) at relatively fewer number of gene variables (173).All the 173 genes selected by SVM algorithm at this α_F level were ranked based on their respective p-values (from the Welch statistic results) in order to select the optimal (parsimonious) gene subset out of these number that would make biological sense. By this, gene variable with the least p-value was ranked first, followed by the one with the second least p-value and so on.
The determination of the optimal gene subset from the entire 173 genes begins by selecting the first 5, the first 10, … , the first 25 gene subsets from the 173 ranked genes as shown in Table 3.The SVM algorithms for tumour groups classification was run

separately on each of these gene subsets using RBF kernel function (since $n \gg p$ in all cases). Both the SVM and RBF parameters were tuned for optimality at10-fold cross-validation. The optimal SVM and RBF parameters values as determined from this implementation are provided in Table 3.The optimal parameters values as indicated in Table 3 are the respective parameter values that yielded the minimum CV error rate over 10-fold cross-validation.

The SVM method was implemented using each of the gene subsets and their predetermined respective optimal RBF and SVM parameter values as provided in Table 3 over 1000 MCCV runs. The classification performance of the improved SVM algorithm is presented in Table 4 for various performance indices. The graphs of these performance measures over the five selected gene subsets are presented in Fig 6.

Table 1: Table of selected gene expression profiles by the Welch statistic in the SVM implementation. The optimal SVM parameter values as determined by grid search as well as the minimum cross-validation error rate at each FWER rates of the SVM implementation using linear kernel are presented

	FWER (α_F) in %					
	1%	**5%**	**10%**	**15%**	**20%**	**100%**
No. of genes selected	173	248	289	313	342	7129
SVM Parameter C	0.1	0.01	0.01	0.01	0.01	0.01
Minimum Misclassification (Cross-validation) Error Rate	0.0125	0.0125	0.0125	0.0268	0.0268	0.0375

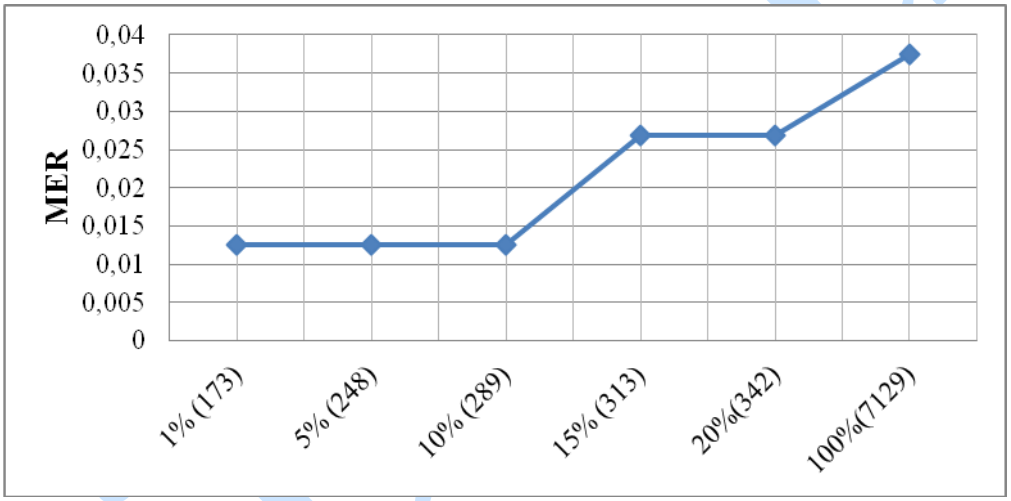

Fig 4: Graph showing the plot of MER at varying FWERs (in %). The number of genes that yielded the respective MER was indicated in parenthesis along with their respective FWER. The SVM algorithm was run at 10-fold cross-validation to tune the parameter of the SVM

Tables 2: Table of classification results of the SVM method for each gene subset using linear kernel. (*) indicates the FWER level that yielded the best prediction accuracy (highest % CCR of 98.5%)

Assessment Criteria	FWER (α_F) in %					
	***1%**	**5%**	**10%**	**15%**	**20%**	**100%**
No. of genes selected	173	248	289	313	342	7129
CCR (%)	**98.50**	98.53	97.40	97.23	97.47	95.67
MER	0.0150	0.0147	0.026	0.0277	0.0253	0.0433
Sensitivity	1.0000	1.0000	1.0000	1.0000	1.0000	0.8778
Specificity	0.9808	0.9813	0.9676	0.9654	0.9685	0.9819
Positive Predictive Value	0.9409	0.9412	0.8979	0.8926	0.9001	0.9328
Negative Predictive Value	1.0000	1.0000	1.0000	1.0000	1.0000	0.9609
Jaccard Index	0.9409	0.9412	0.8979	0.8926	0.9001	0.8278
No. of Support Vectors	24	31	28	27	28	45

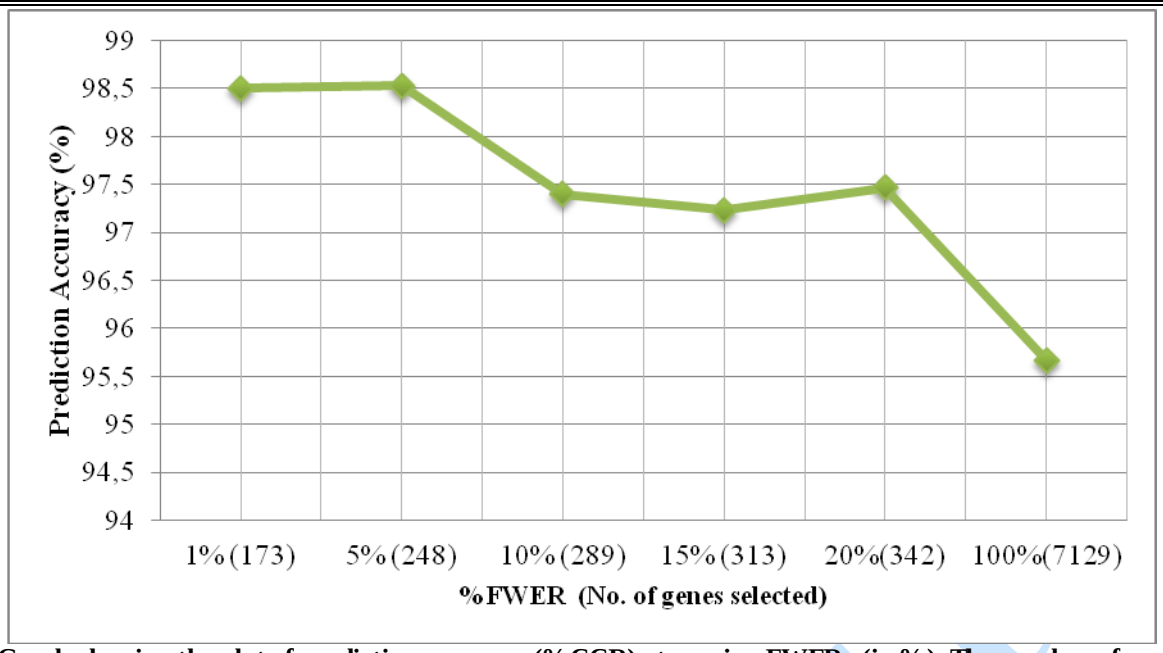

Fig 5: Graph showing the plot of prediction accuracy (%CCR) at varying FWERs (in %). The number of genes that yielded the respective CCR was indicated in parenthesis along with their respective FWER

Table 3: Results of the optimal values of the tuning parameters of SVM and RBF using the selected genes by p-value ranks. The optimal SVM and RBF parameter values were determined for each of the gene subset by grid search using the minimum cross-validation error

The first set of genes selected from the ranked genes	RBF kernel parameter	SVM parameter	Minimum CV error
	γ	C	
5	3	1	0.0893
10	1	1	0.0911
15	0.0001	10000	0.1018
20	0.01	100	0.0625
25	0.01	100	0.0750

Tables 4: Table of classification results of the SVM method using each selected gene subset by ranks based on RBF kernel

Assessment Criteria	No. of genes Employed by p-value ranks				
	5	**10**	**15**	**20**	**25**
CCR (%)	89.07	**91**	87.20	**92.50**	92.13
MER	0.1093	**0.0900**	0.1280	**0.0750**	0.0787
Sensitivity	0.8101	**0.8939**	0.8207	**0.9214**	0.9025
Specificity	0.9181	**0.9154**	0.8888	**0.9252**	0.9272
Positive Predictive Value	0.7610	**0.7766**	0.7061	**0.8108**	0.8033
Negative Predictive Value	0.9402	**0.9684**	0.9415	**0.9701**	0.9709
Jaccard Index	0.6419	**0.7056**	0.6134	**0.7594**	0.7332
No. of Support Vectors	70	**72**	21	21	21

Tibiscus

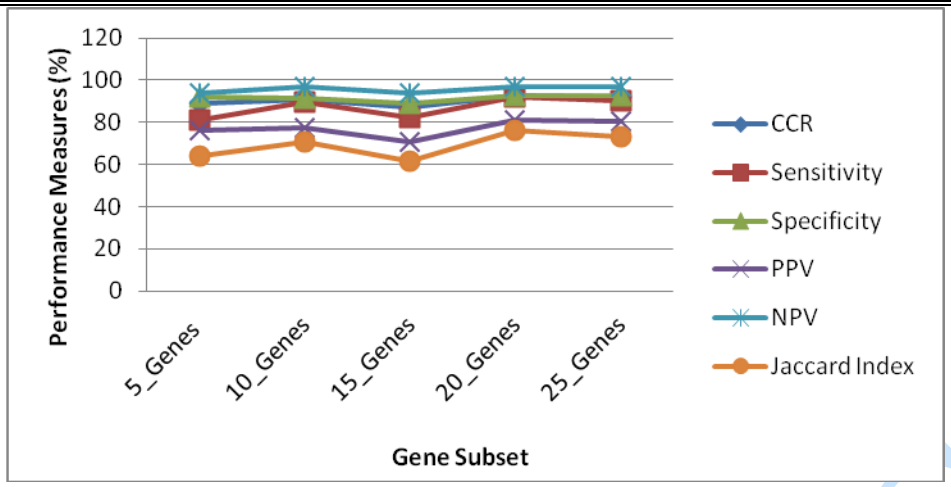

Fig 6: Graph of the various performance measures over the five selected gene subsets. The number of genes that yielded the respective CCR was indicated in parenthesis along with their respective FWER

5. DISCUSSIONS

This work presents an efficient non-clinical method for classification of Diffuse Large B–Cell Lymphoma and Follicular Lymphoma tumor samples using the expression levels of 7,129 gene signatures.

The fewer numbers of differentially expressed genes selected at each FWER levels indicated that the data contained appreciable number features that are complex at producing a binary response signal. However, at 1% FWER level, an optimal gene subset with 173 gene signatures was obtained with high prediction accuracy as provided in Table 1.

Also, optimal values of the kernel and SVM tuning parameters were determined for each of the gene subsets with linear kernels (for $n \ll p$ data structures) using 10-fold cross-validation approach as shown in Table 2. The results of the SVM classifier yielded very good prediction performances in terms of (high prediction accuracy, high sensitivity, high specificity, etc.) with fewer differentially expressed genes than when all the genes were employed for classification. This indicates that in the presence of noisy genes, a classifier might not perform as good as when only the most differentially expressed ones among the competing ones are employed by the classification method.

For final gene optimization in order to construct efficient parsimonious SVM classification model, the selected gene subsets at 1% FWER (the optimal value as determined for these data in Table 1) were ranked based on their respective p-values of the Welch statistics. Based on this ranking, the first few gene subsets were selected sequentially to optimize both the SVM and RBF kernel parameters as shown in Table 3.

Similarly, the SVM classification results using each gene subsets as provided in Table 4 indicated the presence of some false positive genes among those selected at 1% FWER. For example, the prediction accuracy of SVM classifier was about 89% using five genes as can be seen in Table 4. This increases to 91% when ten genes were employed by classifiers showing the marginal contribution of the five additional genes. Surprisingly, when fifteen genes were employed, the prediction accuracy of the SVM classifier dropped to around 87%. This is an indication of some false positive (noisy) genes among the new fifteen crop of genes used for classification. Hence, there is the need to further screen the selected gene variables in order to determine those that are actually correlated with the two tumour groups.

Various results of further genes screening as presented in Table 4 revealed the best prediction performance of the improved SVM classifier when twenty differentially expressed genes were employed for classification. In other words, twenty differentially expressed genes are quite sufficient to provide good discrimination between the two tumour groups of DLBCL and FL in the microarray cancer data analyzed in this work.

CONCLUSION

An improved SVM algorithm for tissue sample classification is presented in this work. Efficient scheme to optimize feature selection process for SVM algorithm was equally provided. Various results obtained generally showed that the improved SVM method is quite efficient.

The relative gains in the need to optimize the parameters of the SVM and the two kernel functions employed in the SVM algorithm evidently manifested in the improved prediction accuracies as shown in the relevant tables in Section 4. The optimal tuning parameters values for the SVM method with Linear and RBF kernels were

efficiently determined by grid search over 10-fold cross-validation using the minimum cross-validation errors criteria. This obviously improved the efficiency of the SVM classification results.

Selecting important genes for cancer classification using gene expression data has become inevitable in cancer research, especially when the prime goal is to seek quick diagnosis of the tumour subjects. The conceptual SVM algorithm for gene selection and classification presented in this work has immensely improve the classification and prediction accuracy of the standard SVM method in which all the entire features in the data were used for tumour classification as shown by the SVM results in Table 2. In that result, the best prediction results of the SVM were obtained at 1% FWER using 173 gene variables (CCR = 98.5%; Sensitivity = 100%). However, results of the standard SVM classifier that employed all the 7129 genes provided CCR of 95.67% with Sensitivity of 87.78%.

This work has created opportunity for further research, especially for molecular biologist in the area of investigating into the biological composition of the selected genes regarding their biological relationship to the two tumour groups of Diffuse Large B–Cell Lymphoma and Follicular Lymphoma tumor samples as contained in the data.

REFERENCES

[AY15] **Aremu G. T., Yahya W. B.** - *Competing Algorithms For Microarray-Based Multiclass Sequential Feature Selection and Classification*. Proceedings of 4th International Science, Technology, Education, Arts, Management & Social Sciences (iSTEAMS) Research Nexus Conference, Nigeria: 675 – 682, 2015.

[Cic15] **Cichosz P.** - *Data mining algorithms explained using R*. John Wiley & Sons, N.Y., 2015

[CS12] **Cristianini N., Shawe-Taylor J.** - *An introduction to support vector machines*. Cambridge University Press, UK, 2012.

[Fle09] **Fletcher T.** - *Support Vector Machines Explained*. www.cs.ucl.ac.uk/staff/T.Fletcher, 2009.

[HCL10] **Hsu C.-W., Chang C.-C., Lin C.-J.** - *A Practical Guide to Support Vector Classification*. Vítor https://www.scribd.com/vmangaravite Mangaravite, 2010.

[H+12] **Hapfelmeier A., Yahya W. B., Rosenberg R., Ulm K.** - *Predictive modelling of gene Expression data*. In: *Handbook of Statistics in Clinical Oncology, 3rd ed.*, Edited by Crowley, J. and A. Hoering. Chapman and Hall/CRC, New York, 2012, pp., 463-475.

[J+13] **James G., Witten D., Hastie T., Tibshirani R.** - *An Introduction to Statistical Learning with Applications* in R. Springer Science + Business Media New York, 2013.

[Kar39] **Karush W.** - *Minima of functions of several variables with inequalities as side constraints*.Master's thesis, Dept. of Mathematics, University of Chicago, 1939.

[KL03] **Keerthi S. S., Lin C.-J.** - *Asymptotic behaviors of support vector machineswith Gaussian kernel*. Neural Computation 15 (7), 1667-1689, 2003.

[KT52] **Kuhn H. W., Tucker A. W.** - *Nonlinear programming*. Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probabilities, 481–492, 1952.

[LL03a] **Lin H.-T., Lin C.-J.** - *A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods*. ww.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf, 2003.

[LL03b] **Lin K.-M., Lin C.-J.** - *A study on reduced support vector machines*.IEEE Transactions on Neural Networks. 12: 1449 – 1559, 2003.

[MM07] **Mahata P., Mahata K.** - *Selecting differentially expressed genes using minimum probability of classification error*. Journal of Biomedical informatics, 40, 775-789, 2007.

[M+07] **Mohamad M. S., Omatu S., Deris S., Hashim S. Z. M.** - *A model for gene selection and classification of gene expression data*. Artif. Life Robotics, 11: 219–222, 2007.

[NCG03] **O'Neill G, Catchpoole D. R., Golemis E. A.** - *From correlation to casualty: microarrays, cancer and cancer treatment*. In: Berrer DP, Dubitzky W, Granzow M, eds. 34:S64-S71, 2003.

[Sch05] **Schulz W. A.** - *Molecular Biology of Human Cancers: An Advanced Student's Textbook*. 1st ed. New York: Springer: 1-508, 2005.

[Sid67] **Sidak Z. K.** - *Rectangular Confidence Regions for the Means of Multivariate Normal Distributions*. Journal of the American Statistical Association, 62 (318): 626–633. doi:10.1080/01621459.1967.10482935, 1967.

[S+02] **Shipp M. A., Ross K. N, Tamayo P., Weng A. P., Kutok J. L., Aguiar R. C. T., Gaasenbeek M., Angelo M., Reich M., Pinkus G. S., Ray T. S., Koval M. A., Last K. W.., Norton A., Lister T. A., Mesirov J., Neuberg D. S., Lander E. S., Aster J. C., Golub T. R.** - *Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning*. Nature Medicine, 8, 68-74, 2002.

[Tin04] **Ting-Lee M.-L.** - *Analysis of microarray gene expression data*. Springer, New York, 2004.

[Vap95] **Vapnik V. N.** - *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[Wel47] **Welch B. L.** - *The generalization of "student's" problem when several different population variances are involved*. Biometrika 34: 28-35, 1947.

[Yah09] **Yahya W. B.** - *Sequential dimension reduction and prediction methods with high dimensional microarray data*. Universitätsbibliothek, Ludwig-Maximilians-Universität, München, Germany. Ph.D. Thesis. URL: http://edoc.ub.uni-muenchen.de/10254/, 2009.

[Yah12] **Yahya W. B.** - *Genes selection and Tumour Classification in Cancer Research: A new approach*. Säbruck, Germany: Lambert Academic Publishing, 2012.

[YAG15] **Yahya W. B., Aremu G. T., Garba M. K.** - *Multiclass Sequential Feature Selection and Classification Method for Gene Expression Data*. Journal of Applied Science and Technology. 20 (1 & 2): 50 – 61, 2015.

[YOJ12] **Yahya W. B., Oladiipo M. O., Jolayemi E. T.** - *A fast algorithm to construct neural networks classification models with high-dimensional genomic data*. Annals. Computer Science Series, 10 (1): 39-58, 2012.

[YRU14] **Yahya W. B., Rosenberg R., Ulm K.** - *Microarray-based Classification of Histopathologic Responses of Locally Advanced Rectal Carcinomas To Neoadjuvant Radiochemotherapy Treatment*. Turkiye Klinikleri Journal of Biostatistics, 6(1): 8-23, 2014.

[Y+11] **Yahya W. B., Ulm K., Ludwig F., Hapflemeir A.** - *K-SS: A sequential feature selection and prediction method in microarray study*. International Journal of Artificial Intelligence, Spring, 6, No. S11:19-47, 2011.

FIDISCUS