ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2020-77341

Data Mining Genome-Based Algorithm for Optimal Gene Selection and Prediction of Colorectal Carcinoma

Optimal Gen Seçimi İçin Veri Madenciliği Genom Tabanlı Algoritma ve Kolorektal Karsinomun Tahmini

Alabi Waheed BANJOKO[a]

[a]Department of Statistics, University of Ilorin, Ilorin, Kwara State, NIGERIA

ABSTRACT Objective: This study presents a method for optimal selection of gene subsets to enhance the non-clinical diagnostic classification and prediction of colorectal cancer using gene expression level of gene expression profiles obtained with an Affymetrix oligonucleotide array. **Material and Method:** A Hybrid multi-objective Support vector Machine (SVM) feature selection and classification algorithm was employed to determine the Biomarker gene subsets that are highly statistically and clinically relevant to the 62 (tumour or normal) responses of the gene expression levels. The genes selection was done in two stages with the first stage using the Bayesian t-test to prune the non-informative genes and the second stage employed the multi-objective optimization method that allows sequential addition of genes for optimal determination of the pre-selected gene subsets. The SVM with RBF kernel (SVM_{RBF}) was fitted sequentially to select the set of near-optimal genes that are correlated with the response class. **Results:** The optimally selected gene subset yielded an accuracy of 90.1% on the test data that were never used in the building process of the algorithm.Furthermore, the results obtained from the principal component analysis and the complete linkage hierarchical clustering indicated near-perfect discrimination of the two clinical response groups of the colorectal cancer status of the patients. **Conclusion:** This work has fully demonstrated that non-clinical colon cancer diagnosis and prediction of patients using their gene signatures from the gene microarray expression data is very possible when the appropriate data mining technique tools are used.

Keywords: Support vector machines; feature selection;
multi-objective optimization;
principal component analysis; clustering

ÖZET Amaç: Bu çalışma, afimetrik oligo-nükleotid dizisi ile elde edilen gen ekspresyon profillerinin gen ekspresyon seviyesini kullanarak kolorektal kanserin klinik olmayan tanı sınıflandırmasını ve tahminini geliştirmek amacıyla gen alt kümelerinin optimal seçimi için bir yöntem sunar. **Gereç ve Yöntemler:** Gen ekspresyon seviyelerinin 62 (tümör veya normal) yanıtları ile istatistiksel ve klinik olarak oldukça ilgili biyo-belirteç alt kümelerini belirlemek için hibrit çok amaçlı destek vektör makinesi (DVM) özelliği seçimi ve sınıflandırma algoritması kullanılmıştır. Gen seçimi iki aşamada yapılmıştır; ilk aşamada bilgi vermeyen genleri budamak için Bayesçi t-testi, ikinci aşamada önceden seçilen gen alt kümelerinin optimal belirlenmesi için genlerin için sekansiyel eklenmesine izin veren çok amaçlı optimizasyon yöntemi kullanılmıştır. RBF çekirdeğine sahip SVM (SVM_{REF}), yanıt sınıfı ile korele olan neredeyse optimal genler kümesini seçmek için sırayla yerleştirildi. **Bulgular:** Optimal olarak seçilen gen alt kümesi, algoritmanın oluşturulma sürecinde hiç kullanılmayan test verilerinde %90.1'lik bir doğruluk sağlamıştır. Ayrıca, ana bileşen analizinden ve tam bağlantı hiyerarşik kümelenmesinden elde edilen sonuçlar, kolorektal kanser durumunun iki klinik yanıt grubunun ayırt edilmesi için neredeyse mükemmele yakın ayrımını göstermiştir. **Sonuç**: Bu çalışma, klinik olmayan kolon kanseri teşhisinin ve hastaların gen mikro-dizi ekspresyon verilerinden kendi gen imzalarını kullanarak tahmin etmelerinin, uygun veri madenciliği teknik araçları kullanıldığında çok mümkün olduğunu tam olarak göstermiştir.

Anahtar kelimeler: Destek vektör makineleri; öznitelik seçimi;
çok amaçlı optimizasyon;
temel bileşenler analizi; kümeleme

Colorectal cancer has been reported to be one of the commonest cancers and a major cause of mortality.[1] It is estimated that about 600,000 deaths due to colorectal cancer are recorded annually.[2] Therefore, it is very important to determine biomarker gene subset from among the gene expression profiles that can be used

Correspondence: Alabi Waheed BANJOKO
Department of Statistics, University of Ilorin, P.M.B. 1515, Ilorin, Kwara State, NIGERIA
E-mail: banjokoalabi@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 12 Jun 2020 ***Received in revised form:*** 04 Sep 2020 ***Accepted:*** 07 Sep 2020 ***Available online:*** 21 Dec 2020

in the timely and accurate detection of the disease. However, because of various challenges, clinical diagnosis of cancer statuses have been reported to be inefficient.[3-7] This is due to the untimely and inaccurate prediction of the disease before the true status of the patient can be determined and also the risk associated with some of the methods adopted for the clinical diagnosis.[8] Therefore, it is necessary to employ a non-clinical diagnosis and prediction method for accurate prediction of these tissue samples arising from the gene expression profiles.[5,9-11]

Determining biomarker gene subsets for accurate prediction of cancer status using the non-clinical approach have gained a lot of attention in the literature[9-11] and have also eased the processes of cancer diagnosis classification and prediction.[12-15] This has been practically and experimentally possible due to the advent of microarray technology that allows simultaneousmonitoring of the expression profiles of the gene signatures in biological samples.[16,17]

Quite a several studies have utilized the gene expression profiles from the microarray experiment.[3,5,11,13-15,18-24] The major goal is to determine the biomarker gene(s) from among the several thousands of expression profiles of the gene signatures generated from the microarray experiment. This would then be used in the non-clinical diagnosis and prediction of the cancer status of the patient using some data mining techniques.

One of the ways to identify the biomarker genes is feature selection process discussed in several literatures.[5,9,11,19,25] The process selects all the genes that are assumed to be relevant to the cancer status of the subjects and those that are not biomarkers are filtered out. Then, the biomarkers genes are then used for classification of tumour samples.

In most cases when the feature selection methods are used, the number of biomarker genes q is usually large i.e. $(n \ll q)$, particularly when the combination of the genes in some of the genomic data are very complex.[9,11] The task is usually on how to optimize those set of biomarkers so as to select the most relevant among them that will be use for classification without compromising the accuracy in the classification task.[9,18,26]

One of the data mining technique that has been proven to be quite efficient in the field of pattern recognition and group response classification is the Support Vector Machine (SVM).[1,2,11,21,27] The SVM is a kernel-based learning algorithm that has been widely used in the literature due to its efficiency in classification task.[1] Adaptation and modification of the SVM have therefore also been reported.

The present study apply a sequential hybrid optimal feature selection method base on the SVM to determine the biomarker gene(s) that are both statistically important and clinically relevant to colorectal cancer.[28] The method uses a two-stage approach. The first stage prune or remove the noisy genes from the original gene expression microarray data using the Bayesian test of hypothesis while the second stage adopts the multi-objective optimization strategy to obtain the optimum gene subsets for better prediction of colorectal cancer.[25,29,30] The main objective is to maximize the prediction accuracy of colorectal cancer status using the minimum biomarker gene subsets possible.

MATERIAL AND METHODS

DATA RESOURCES

The colon cancer data is popularly known as it has been used in several works of literature.[18,19] The data was collected and first analysed by Alon et al.[31] The data set contains gene expression levels of 40 tumour and 22 normal colon tissues for 2,000 gene expression profiles obtained with an Affymetrix oligonucleotide array. The output variable is a binary response categorical defined as follows;

$$Y_i = \begin{cases} 1, if\ the\ i^{th}\ subject\ response\ is\ a\ normal\ colon\ tissue \\ -1, if\ the\ i^{th}\ subject\ response\ is\ a\ tumour\ colon\ cancer\ tissue \end{cases}$$

$$i = 1, 2, 3, \dots, n$$

The data is freely available in an online microarray data archive at https//github.com/ramhiser/datamicroarray.

The flow chart of the algorithm adopted in this work is proposed by Banjoko and Yahya[28]. The algorithm starts by preprocessing the data.

STATISTICAL ANALYSIS

Data Pre-processing

As noted, microarray gene expression readings will necessarily not have desirable statistical properties.[32] This is so because the laboratory process involved in the preparation of each biological sample $\boldsymbol{n}$ on a microarray glass slide introduces an arbitrary scale that is common to gene expression readings for all genes g. The usual or common practice is to correct the readings for the introduced scale through data normalization which allows for minimization of extraneous variation in the measured gene expression levels so that biological differences can be more easily distinguished. The data normalization method by was adopted in this study as[33]

Standardize each of the p genes in the gene expression data with n samples

$$g_{ij}^{*} = \frac{g_{ij} - \hat{\mu}_{g_i}}{\sqrt{\hat{\sigma}_{g_i}^{2}}} \qquad (1)$$

$\forall i = 1\,,2{,}3\,\ldots, p \text{and} j = 1, 2, 3, \ldots, n$

Preliminary Feature Selection

As earlier presented, microarray dataset is usually characterized by high dimension i.e. several thousands of $\boldsymbol{P}$ genes expression measured on relatively small subjects $\boldsymbol{n}$. Several microarray studies among others, have shown that among these several thousands of genes, only a few are differentially expressed or biomarkers and might be connected to the cancerous state of the biological samples.[9-11] It is necessary to determine from among the genes those that are biomarkers to the colorectal cancer status of the subjects.

The hypothesis testing technique proposed by was adopted because the response group of the colorectal cancer is binary.[25] The hypothesis is stated as.

H_0: Gene g_i is not differentially expressed to the colorectal cancer status

H_1: Gene g_i is differentially expressed to the colorectal cancer status

The test statistic for testing the hypothesis above as stated by is[25]

$$t_{BT} = \frac{\delta - \hat{\delta}_{BT}}{\sqrt{var(\hat{\mu}_{-1BT}) + var(\hat{\mu}_{+1BT})}} \sim t_v | H_0 \qquad (2)$$

Optimization

Following the method in, the objective of this study is to maximize the prediction accuracy of the colorectal cancer status of the patient using the minimum biomarker genes possible.[28] Therefore, the problem is viewed as a multi-objective optimization task. Hence, the optimization technique proposed in is adopted.[28]

SUPPORT VECTOR MACHINE (SVM) CLASSIFIER

SVM isa powerful machine learning tool that has been proven useful in classification problems encountered in working with microarray data as the most successful kernel-based learning method based on the recent advances in statistical learning theory (Figure 1).[27,34]

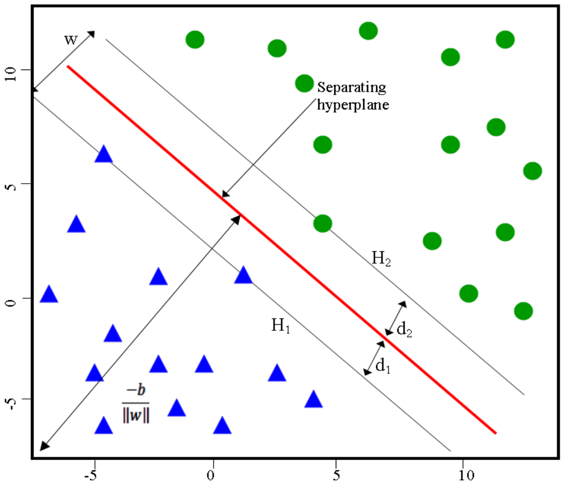

FIGURE 1: A separating maximal margin hyperplane for the linearly separable subjects with a binary response.

The idea of the SVM is to train$n$ data points such that each input $x_i$can be classified into one of the binary response group i.e. $y_i = \pm1$.Thus, the training dataset is of form$\{x_i, y_i\}$, where $i = 1,2, \dots, n$ and $y_i \in \{-1, +1\}, x_i \in \Re^D$ assuming the data is linearly separable into two classes by drawing a line for D = 2. A deep review of the SVM as reported in gives the objective function.[11]

$$\min_{\alpha}\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=i}^{n}\alpha_i\,\alpha_j y_i y_j x_i x_j - \sum_{i=1}^{n}\alpha_i\right) \tag{3}$$

Where $\alpha_i \geq 0$ is the langrage multiplier as equally reported [11].

In this paper, the SVM with Radial Basis Function (SVM_{RBF}) kernel was sequentially fitted on each of the genes declared as differentially expressed at the preliminary feature selection stage in a Monte-Carlo experiment with 1000 iterations using the flowchart already reported in and selecting the gene with the least misclassification error rate (MER).[28] The selected gene is the first gene to enter into the optimization stage and is then paired with each of the remaining genes. Similarly, the SVM_{RBF} is fitted on each pair and the pair with the least MER will be selected as the next optimum gene subset. As illustrated in, the process continues until a final Pareto fonts gene subset is obtained.[28]

For any two successive gene subsets in the optimization stage, the estimates of Misclassification Error Rates$(MERs)$ for the two gene subsets are obtained for the test sample over the Monte-Carlo cross-validation (MCCV) with 1000 iterations.

$MER_{1j}, MER_{2j}, \dots, MER_{1000j}andMER_{1j+1}, MER_{2j+1}, \dots, MER_{1000j+1}$

The success of any classifier is to classify the entire subject in the test sample correctly at each iteration. This requires that at each iteration, a success is recorded if $MER_i = 0$ and otherwise if $MER_i \neq 0$. Hence, the number of successes for the two successive subsets can be determined and the $Z -$ statistic is defined as;

$$Z = \frac{\hat{P}_j - \hat{P}_{j+1}}{\sqrt{pq\left(\frac{1}{n_j} + \frac{1}{n_{j+1}}\right)}} \sim N(0,1) \tag{4}$$

The final Pareto fonts gene subset is the used for classification of colorectal cancer and necessary performance indices were obtained.

All analyses were performed using R software (www.cran.r-project.org) version 3.4.4 using the e1071 package version 1.7-3.

ANALYSIS AND RESULTS

The goal is to optimally select the biomarker gene subsets that are highly informative of the response classes (tumourous or normal) and use same to predict any future (unseen) colon cancer response group when using gene expression data. As noted in Table 1 which is also shown in Figure 2, the preselected genes by the feature selection methods and their respective MERs in parenthesis are as follows; **Hsa.6814 (0.2277), Hsa.2928 (0.2173), Hsa.831 (0.1737), Hsa.601 (0.2403), Hsa.773 (0.2133), Hsa.8147 (0.1650).**

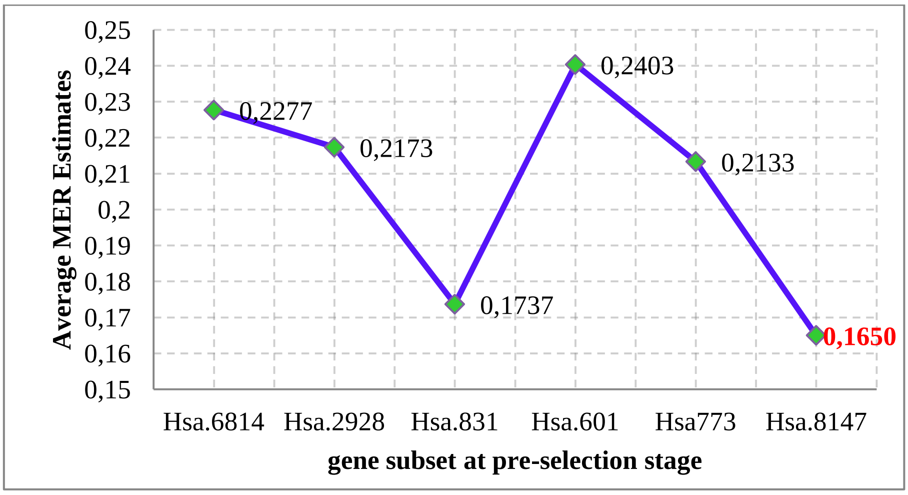

FIGURE 2: The graph of MER estimates of each of the pre-selected genes of the colon cancer data.

TABLE 1: Genes selected at the pre-selection stage.

Microarray data	Number of genes in the data	Number of genes declares as differentially expressed at pre-selection stage	Genes declare as differentially expressed
Colon cancer	2000	6	Hsa.6814, Hsa.2928, Hsa.831, Hsa.601, Hsa.773, Hsa.8147

The medical characteristics of the six preselected genes can be seen in Table 2. The results indicated that the selected genes have very strong relationship with the response. However, there is need to optimize the selected crop of genes as earlier explained so as to maximize the classification accuracy.

TABLE 2: Clinical characteristics of the preselected genes.

Probe-set Number	Gene Number	Sequence	Name
Hsa.6814	H08393	3'UTR	Collagen alpha 2(xi) chain (Homo sapiens)
Hsa.2928	X63629	Gene	H.sapiens mRNA for p cadherin.
Hsa.831	M22382	Gene	Mitochondrial matrix protein P1 precursor (human)
Hsa.601	J05032	Gene	Human aspartyl-tRNAsynthetase alpha-2 subunit mRNA, complete cds
Hsa.773	H40095	3'UTR	Macrophage Migration Inhibitory Factor (human)
Hsa.8147	M63391	Gene	Human desmin gene, complete cds

UTR: Untranslated region.

Base on the estimated means MERs displayed in Figure 1 above, the gene with probe-set Number-Has.8147 is the gene that provided the best prediction accuracy forhaving the least mean MER of 0.1650 among the 6 preliminarily selected genes. Hence, gene Has.8147 is the first gene to be selected by the method. The gene is then paired with each of the remaining 5 genes for the optimization process.

TABLE 3: Results of the proposed algorithm ateach step for the Colon cancer data.

Step	Enters	Gene subset	*MER*	Required test	Test Value	P-value	Decision
0	Hsa.8147	Hsa.8147	0.1650	-	-	-	Initialize
1	Hsa.601	Hsa.8147 Hsa.601	0.1317	Hsa.8147$vs$Hsa.8147 Hsa.601	-4.50	< 0.001	Continue
2	**Hsa.2928**	**Hsa.8147 Hsa.601 Hsa.2928**	**0.0990**	**Hsa.8147 Hsa.601vs Hsa.8147 Hsa.601 Hsa.2928**	$\mathbf{-3.36}$	**0.0004**	
3	Hsa831	Hsa.8147 Hsa.601 Hsa.2928 Hsa831	0.1093	X	X	X	Stop

From Table 3 above, optimal classification/prediction accuracy is achieved at the third step. As discussed in the flow chart of the method adopted, if the prediction accuracy of the previous subset is better than the current subset, then the algorithm terminates and the previous subsets are assumed to be the near-optimal gene subset that will be used for classification/prediction of the responses.

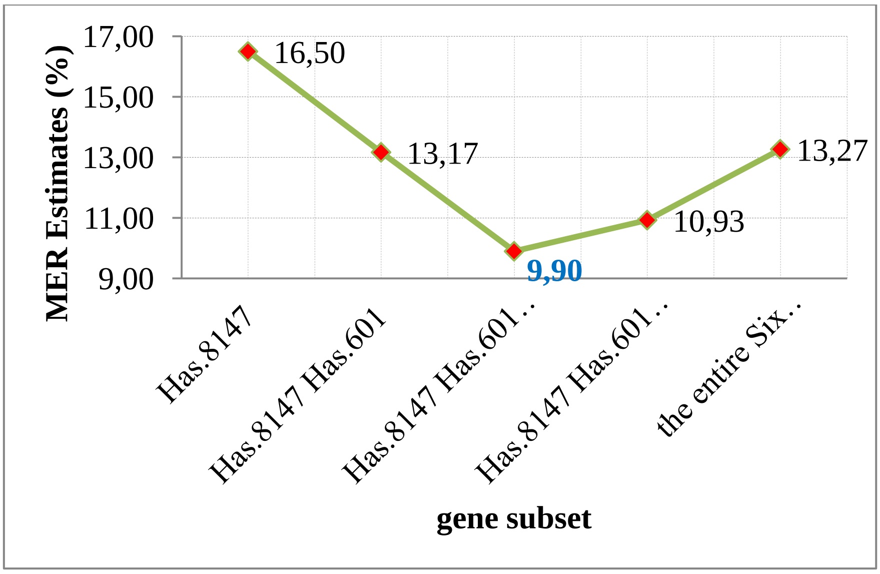

FIGURE 3: The graph of successive estimated MER of gene subsets at each sequential step.

The plot of the average MERs (in %) at each selection steps against the number of genes selected as presented in Figure 3. An indication of successive improvements in classification/prediction accuracy as additional genes are selected into the models is observed. It can also be observed that the minimum MER estimate was obtained when the number of gene subset is three. Addition of another gene to the three optimal genes worsen the result, hence the algorithm terminates at the optimally selected genes.

Also, it can be observed in Figure 3 that when the entire six preselected genes were used for classification, the MER became worse. This is a justification that using preselected genes alone for classification as usually done in the analysis of microarray data is not sufficient for optimal efficient prediction of cancer status. There is a need to optimize the preselected genes for better accuracy.

It can, therefore, be concluded that only three genes from the preselected genes in Table 2 are optimally needed for efficient classification of the colon cancer status. The three optimal gene subset is presented in Table 4 below.

TABLE 4: The three optimal gene subset.

Probe-set Number	Gene Number	Sequence	Name
Hsa.8147	M63391	Gene	Human desmin gene, complete cds
Hsa.601	J05032	Gene	Human aspartyl-tRNAsynthetase alpha-2 subunit mRNA, complete cds
Hsa.2928	X63629	Gene	H.sapiens mRNA for p cadherin.

The performance of the optimal gene subsets is evaluated via the test data which are not included in the optimization stage. The performance indices of these genes using the test set are reported in Table 5 and Figure 4. The indices show that the method achieves a very good performance.

TABLE 5: The estimated performance indices of the three (3) optimal gene subset selected.

Performance Indices of the proposed algorithm	Estimated Values (in %)	Standard Deviations of Estimates
Misclassification Error Rate (MER)	9.90	0.1735
Sensitivity	86.94	0.2976
Specificity	91.92	0.2093
Positive Predictive Value (PPV)	85.87	0.3072
Negative Predictive Value (NPV)	92.64	0.1986
Jaccard Index (JI)	76.05	0.3685
Cross-validated AUC	91.56	0.1831

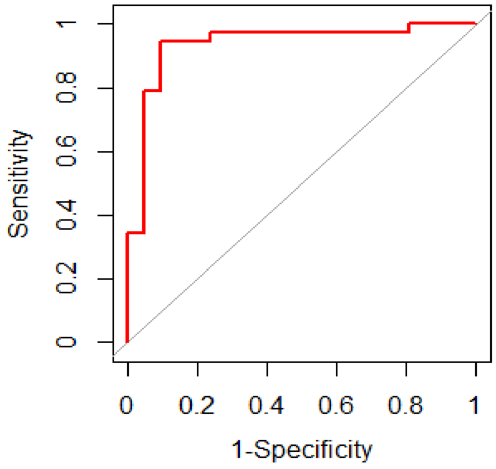

FIGURE 4: The CVROC curve for the selected three optimal gene subsets for the Colon cancer data. CVAUC=0.9156.

VISUALIZING THE EFFICACY OF THE SELECTED OPTIMAL GENE SUBSET

The unsupervised learning methods (Principal Component Analysis, PCA and the Clustering method) was applied to probe further into the efficacy of the near-optimal gene subsets by the method.

The discriminatory power of the selected genes by the proposed method is assessed via the principal component analysis (PCA). PCA refers to the process by which principal components are computed, and the subsequent use of these components in understanding the data. PCA is an unsupervised approach since it involves only a set of features and no associated response. The idea of the PCA in this study is to first fit a principal component regression (PCR) model using the optimal gene subsets selected by the method from each of the colon cancer data and obtain the graphical plots of the first two principal components (PC1 and PC2). If the selected gene subsets are good discriminators of their respective binary response classes, then the number of sub-groups in the response class must be separated on the PCA plots.

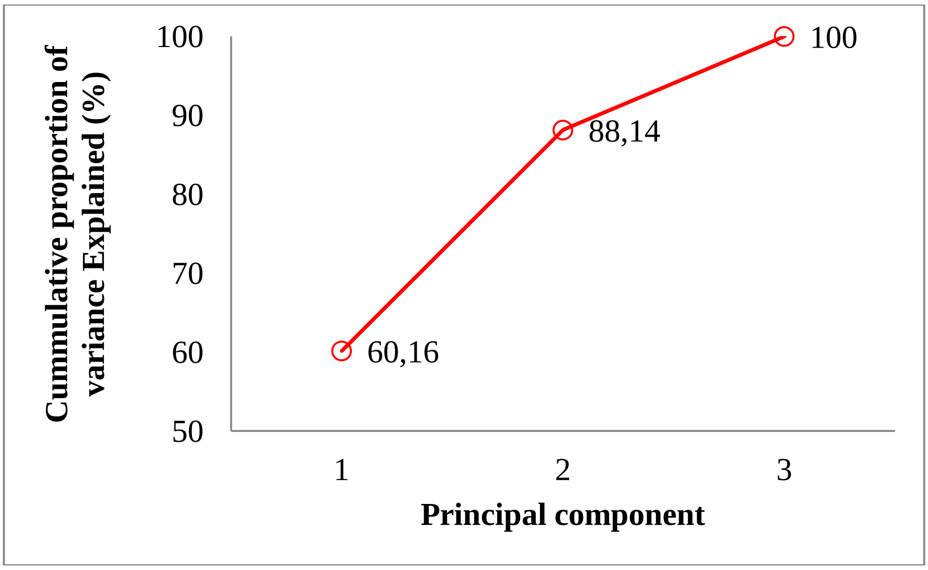

FIGURE 5: Plots of the cumulative proportion of variance explained for all colon cancer data.

From the plots in Figure 5 above, it can be observed that the first two principal components (PC1 and PC2) accounts for more than 88% of the variation in the colon cancer data is explained by PC1 and PC2. This means that the first two principal components determined by utilizing the three selected optimal gene subsets can provide a pretty accurate summary of the data. It is therefore evidenced that the selected gene subsets are good enough to provide good classification/prediction of colon cancer status of subjects.

Figure 6 below provides the plot of the first two principal components for the data sets. It can be observed that the different biological groups in the colon cancer data are separated on the PCA plots and this shows a very good summary of the data using two dimensions. This is an indication that the selected optimal genes are good predictors of the colon cancer status. The misclassifications noticed on the PCA plots is justified by the correct prediction (MER) estimated by using the near-optimal selected genes as reported in Table 5.

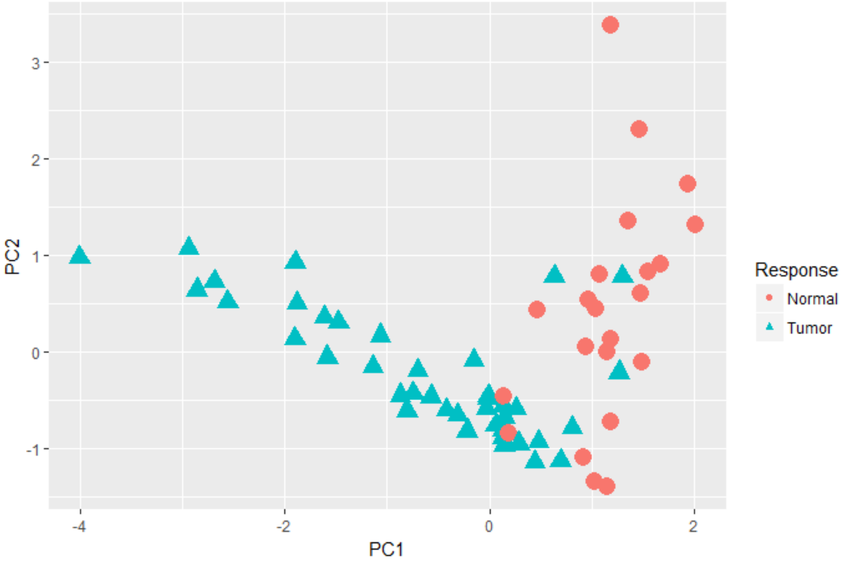

FIGURE 6: The plots of the first two principal components, PCA plots, constructed using the optimal genes.

Cluster analysis is also utilized to determine the efficacy of the selected gene subsets by directly and efficiently relating the existing biological sub-groupings of the colon cancer rather than utilizing all the genes. The *Complete-linkage hierarchical clustering* (CLHC) using Euclidean distance for the measure of similarity was adopted. Features or groups of features with similar expression pattern are adjacent to each other and the expression levels are shown in Figure 7 below.

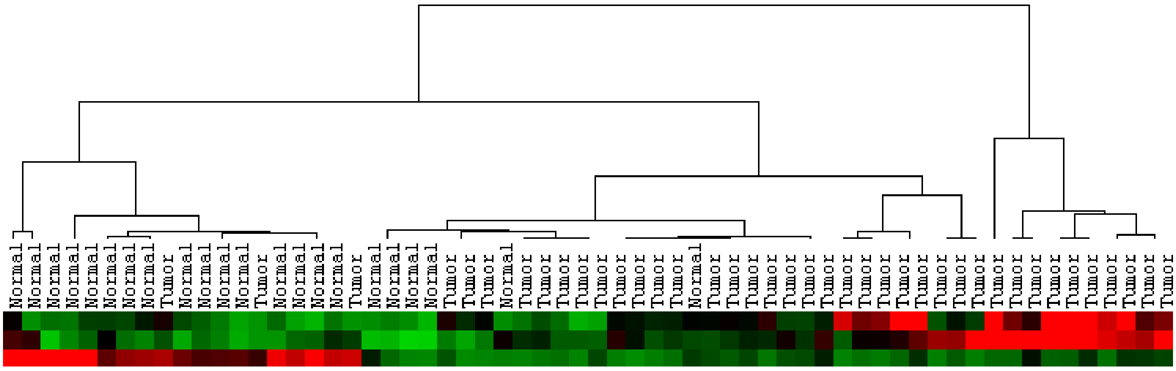

FIGURE 7: Dendrogram plots constructed using the selected optimal genes.

DISCUSSION

The clinical risks, time and cost of medical diagnosis involved when patients are been screened for colorectal cancer have made it worthy to device an efficient and reliable non-clinical colorectal cancer diagnosis and prediction through which early detection of the cancer status can be determined among patients to enable adequate and appropriate treatment. The alternative to the clinical diagnostic method is presented in this work which involves identifying and optimizing the core biomarker genes that are both statistically and medically related to colorectal cancer status of the patients using gene expression profiling data from microarray experiment.

The adopted method starts by preprocessing the data to minimize the extraneous variation in the measured gene expression levels so that biological differences can be more easily distinguished. Also, from the result in table 1, only six genes were determined to be differentially expressed among the two thousand gene expressions in the colon cancer data which is always often the case working with microarray gene expression data. The combinations of genes in the colon cancer are not complex in interacting with the sample to produce response signal. This argument is supported by the number of genes selected at the preliminary stage in this study.

Furthermore, it can be observed from figure three that sequential addition of genes leads to an appreciable increase in the prediction accuracy until at a point when the addition of an extra gene worsens the result earlier obtained. This has effectively justified the objective of the algorithm used in this study. The number of gene subsets is very few and the classification/prediction accuracy and other performance indices of the algorithm are very high.

Finally, the results of the principal component analysis and cluster method on the entire colon cancer data used in this research justify the efficacy of the selected near-optimal gene subsets by the proposed algorithm.

CONCLUSION

The study is an application of non-clinical approach to classification and prediction of colorectal carcinoma. It is an indication that the non-clinical approach is very possible. This work has opened up a new research activity for the molecular biologist to further examine and probe the optimally selected genes to establish the etiological pathology of these genes concerning their respective tumour classes. However, there is a need to validate the study with more clinical studies with colorectal carcinoma patients.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

This study is entirely author's own work and no other author contribution.

REFERENCES

1. Zhi J, Sun J, Wang Z, Ding W. Support vector machine classifier for prediction of the metastasis of colorectal cancer. Int J Mol Med. 2018;41(3):1419-26.[Crossref] [PubMed] [PMC]
2. Zhao D, Liu H, Zheng Y, He Y, Lu D, Lyu C. A reliable method for colorectal cancer prediction based on feature selection and support vector machine. Med Biol Eng Comput. 2019;57(4):901-12.[Crossref] [PubMed]
3. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7(6):673-9.[Crossref] [PubMed] [PMC]
4. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002;8(1):68-74.[Crossref] [PubMed]
5. Banjoko AW, Yahya WB, Garba MK. Multiclass response feature selection and cancer tumour classification with support vector machine. J Epidemiol Biostat. 2019;5:91-104.[Crossref]
6. Ge H, Yan Y, Wu D, Huang Y, Tian F. Potential role of LINC00996 in colorectal cancer: a study based on data mining and bioinformatics. Onco Targets Ther. 2018;11:4845-55.[Crossref] [PubMed] [PMC]
7. Ge H, Yan Y, Yue C, Liang C, Wu J. Long noncoding RNA LINC00265 targets EGFR and promotes deterioration of colorectal cancer: a comprehensive study based on data mining and in vitro validation. Onco Targets Ther. 2019;12:10681-92.[Crossref] [PubMed] [PMC]
8. Committee on Diagnostic Error in Health Care; Board on Health Care Services; Institute of Medicine; The National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care. Balogh EP, Miller BT, Ball JR, editors. Washington (DC): National Academies Press (US); 2015.[PubMed]
9. Yahya WB, Ulm K, Fahrmeir L, Hapfelmeier A. k-SS: a sequential feature selection and prediction method in microarray study. International Journal of Artificial Intelligence. 2011;6(11):19-47.[Link]
10. Hapfelmeier A, Yahya WB, Robert R, Ulm K. Predictive modeling of gene expression data. In: Crowley J, Hoering A, eds. Handbook of Statistics in Clinical Oncology. New York: Chapman & Hall/CRC; 2012. p.471-83.[Crossref]
11. Banjoko A, Yahya WB, Garba MK, Olaniran OR, Dauda KA, Olorede KO. Efficient support vector machine classification of diffuse large B-cell lymphoma and follicular lymphoma mRNA tissue samples. Annals Computer Science Series. 2015;13:69-79.[Link]
12. Kharrat N, Assidi M, Abu-Elmagd M, Pushparaj PN, Alkhaldy A, Arfaoui L, et al. Data mining analysis of human gut microbiota links Fusobacterium spp. with colorectal cancer onset. Bioinformation. 2019;15(6):372-9.[Crossref] [PubMed] [PMC]
13. Liang H, Yang L, Tao L, Shi L, Yang W, Bai J, et al. Data mining-based model and risk prediction of colorectal cancer by using secondary health data: a systematic review. Chin J Cancer Res. 2020;32(2):242-51.[Crossref] [PubMed] [PMC]
14. Pourhoseingholi MA, Kheirian S, Zali MR. Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients. Acta Inform Med. 2017;25(4):254-8.[Crossref] [PubMed] [PMC]
15. Wang T, Yang C, Zhao H. Prediction analysis for microbiome sequencing data. Biometrics. 2019;75(3):875-84.[Crossref] [PubMed]
16. Yahya WB, Aremu GT, Garba MK. Multiclass sequential feature selection and classification method for genomic data. JAST. 2016;20(1-2):50-61.[Link]
17. Rápolti E, Szigeti A, Farkas R, Bellyei S, Boronkai A, Papp A, et al. [Neoadjuvant radiochemotherapy in the treatment of locally advanced rectal tumors]. Magy Onkol. 2009;53(4):345-9.[Crossref] [PubMed]
18. Mohamad MS, Deris S, Illias RMD. A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray. IJCIA. 2005;5(1):91-107.[Crossref]
19. Mohamad MS, Omatu S, Deris S, Misman MF,Yoshioka M. Selecting informative genes from microarray data by using hybrid methods for cancer classification. Artif Life Robotics. 2009;13:414-7.[Crossref]
20. Michalewicz Z. Genetic Algorithms + Data Structures = Evolution Programs. New York: Springer Verlag; 1996.[Crossref]
21. Banjoko AW, Yahya WB, Garba MK, Abdulazeez KO. Weighted support vector machine algorithm for efficient classification and prediction of binary response data. J Phys Conf Ser. 2019;1366:012101.[Crossref]

22. Zagorecki A. Feature selection for naïve Bayesian network ensemble using evolutionary algorithms. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems. ACSIS. 2014. p.381-5.[Crossref]
23. Yang K, Zhou B, Yi F, Chen Y, Cheng Y. Correction to: colorectal cancer diagnostic algorithm based on sub-patch weight color histogram in combination of improved least squares support vector machine for pathological image. J Med Syst. 2019;43(12):333.[Crossref] [PubMed]
24. Zhao D, Liu H, Zheng Y, He Y, Lu D, Lyu C. Whale optimized mixed kernel function of support vector machine for colorectal cancer diagnosis. J Biomed Inform. 2019;92:103124.[Crossref] [PubMed]
25. Olaniran OR, Yahya WB. Bayesian hypothesis testing of two normal samples using bootstrap prior technique. JMASM. 2017;16(2):185-96.[Crossref]
26. Yahya WB. Gene Selection and Tumour Classification in Cancer Research. Lambert Academic Sabrick; 2012.
27. Vapnik VN. The Nature of Satistical Learning Theory. Springer-Verlag; 1995.[Crossref] [PubMed]
28. Banjoko AW, Yahya WB. Sequential optimization based feature selection algorithm for efficient cancer classification and prediction. In: 4th iSTEAMS International Multidisciplinary Conference, Vol. 14. p.265-74. (AlHikmah University, Ilorin, Nigeria, 2018).
29. Hwang CL, Masud ASM. Multiple Objective Decision Making, Methods and Applications: A State-Of-The-Art Survey. Berlin: Springer-Verlag; 1979.[Crossref]
30. Miettinen K, Ruiz F, Wierzbicki AP. Introduction to multiobjective optimization: interactive approaches. In: Branke J, Branke J, Deb K, Miettinen K, Słowiński R, eds. Multiobjective optimization: interactive and evolutionary approaches. Berlin, Heidelberg: Springer-Verlag; 2008. p.27-57.
31. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A. 1999;96(12):6745-50.[Crossref] [PubMed] [PMC]
32. Ting-Lee ML. Analysis of Microarray Gene Expression Data. New York: Kluwer Academics; 2004.
33. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7:142.[Crossref] [PubMed] [PMC]
34. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines: and Other Kernel-Based Learning Methods. Cambridge University Press; 1999.[Crossref] [PMC]