# A Comparative Analysis of Feature Selection and Feature Extraction Models for Classifying Microarray Dataset

Arowolo M. Olaolu, Sulaiman O. Abdulsalam, Isiaka R. Mope, and Gbolagade A. Kazeem
Department of Computer Science, College of Information and Communication Technology, Kwara State University, Malete, Nigeria.
micheal.arowolo14@kwasu.edu.ng, abdulsalamny@gmail.com, imabdulrafiu@yahoo.com and kazeem.gbolagade@kwasu.edu.ng

**ABSTRACT**

**Purpose:** The purpose of this research is to apply dimensionality reduction methods to fetch out the smallest set of genes that contributes to the efficient performance of classification algorithms in microarray data.
**Design/Methodology/Approach:** Using colon cancer microarray dataset, One-Way- Analysis of Variance is used as a feature selection dimensionality reduction technique, due to its robustness and efficiency to select relevant information in a high-dimension of colon cancer microarray dataset. Principal Component Analysis (PCA) and Partial Least Square (PLS) are used as feature extraction techniques, by projecting the reduced high-dimensional data into efficient low-dimensional space. The classification capability of colon cancer datasets is carried out using a good classifier such as Support Vector Machine (SVM). The study is analyzed using MATLAB 2015.
**Findings:** The study obtained high accuracies and the performances of the dimension reduction techniques used are compared. The PLS-Based attained 95% accuracy having edge over the other dimension reduction methods (One-Way- ANOVA and PCA).
**Practical Implications:** The major implication of this research is getting the local dataset in the environments which lead to the usage of an open resource dataset.
**Originality:** This study gives an insight and implications of high dimensional data in microarray gene analysis. The application of dimensionality reduction helps in fetching out irrelevant information that halts the performance of a microarray data technology.

***Keywords:*** *Dimension Reduction, One-Way-ANOVA, PCA, PLS, Classification.*
***Paper Type:*** *Research work*

## 1 Introduction

Recently, procedures of machine learning in data mining have improved (Shengyan and Iagnemma, 2010; Techo, Nattee and Theeramunkong, 2008), microarray technology has become known as a proficient technique for cancer diagnosis, prognosis and treatment (Mukesh, Nitish, Amitav, and Satanu, 2015). Deoxyribonucleic Acid (DNA) microarray analysis have great effect by fetching informative genes that causes cancer (Flores, Hsiao, Chiu, Chuang, Huang and Chen, 2013), but its existing major drawback is caused by the curse of dimensionality, it hinders helpful information in a dataset which results into instability of computation. Therefore dimension reduction methods as a preprocessing step for retrieving similar data in a very high dimensional space is necessary, so as to select or extract relevant features which are very important in the analysis of microarray data of cancer (Mukesh, Nitish, Amitav, and Satanu, 2015).

To enhance the development of microarray technology (Mukesh, Nitish, Amitav, and Satanu, 2015), which has been supportive in studying thousands and millions of genes concurrently and also generating huge amount of data, several dimension reduction techniques and classifiers based on machine learning approach have been proposed in literature by various researchers and practitioners (Wang, Chu and Xie, 2007). Several algorithms and techniques have been proposed for dimensionality reduction. In feature selection, distinctive features are done by selection and

removal of redundant and irrelevant features (Cecille, Dana and Otman, 2013; Wald, Khoshgoftaar and Napolitano, 2013). Due to high dimensionality of original dataset, features are selected before classification, by using some feature selection approaches such Filter, wrapper and embedded methods as a dimension reduction method. Feature extraction, is one of the most important technique used in analysis and interpretation of microarray data, it is sectioned into supervised and unsupervised machine learning methods. Among the most effective methods are PCA, PLS and Independent Component Analysis (ICA) (Maryam and Mohammad, 2016), by considering the important components for classification.

Classification is an important aspect of cancer diagnosis and treatment; microarrays offer hope that cancer classification can be objective and highly accurate, providing clinicians with the information to choose the most appropriate forms of treatment. Classification operation performs intelligent discrimination by means of features obtained from dimensionality reduction methods. Many classifiers have been used for microarray analysis task, such as Fisher Linear Discrimination Analysis, Support Vector Machine, K-Nearest Neighbor (KNN) and aggregated classifiers (Atiyeh and Mohammad, 2016 ; Smitarani and Pratishya, 2014). In this paper, SVM is adopted based on its constructive learning procedures, it is used for classification tasks, and it uses linear models in implementing non-linear class boundaries by transforming input space using a non-linear mapping into a new space. SVM produces an accurate classifier with less over fitting and it is robust to noise (Vapnik, 1998).

In this paper, simple yet very efficient methods for cancer classification using dimension reduction methods are proposed. One-Way-ANOVA is used as a feature selection method to select a subset of the dataset and analyze the performance, PCA and PLS is used as feature extraction methods for dimensionality reduction, to project the dataset into a low-dimensional space and constructs a new dimension by analyzing relationships hidden in the dataset. Classification is carried out using SVM to classify the predetermined data. The experiment shows PLS outperforms PCA and One-Way-ANOVA by reducing the dimension and attaining a better performance.

## 2 Related Works

In literature, several studies have been carried out on the classification of the microarray data, using dimensionality reduction techniques and with the use of different classifiers. Vijayarani and Maria Sylviaa (2016) worked on two divisions of dimension reduction, Feature extraction (PCA, LDA) and feature selection (FA). Feature extraction techniques performed more adequate than the feature selection. Reduction was done to the larger medical dataset (Thyroid, Oesophagal) to decrease the curse of dimensionality. Ali, Paul and Mahdhu, (2011), Analyed microarraydata in the field of diagnosis and treatment of patients, a very high dimensional dataset containing some noise, non-useful information and a small number of relevant features for disease, proposes a non-linear dimensionality reduction algorithm Local Principal Component (LPC) which aims to maps high dimensional data to a lower dimensional space.

The reduced data represents the most important variables underlying the original data. Experimental results and comparisons were presented to show the quality of the proposed algorithm.

Research is still on for innovative techniques to select unique attributes or features so that the classification accuracy can be improved and the processing time can be reduced. From the above studies, there are limitations regarding to the problem of improving, developing efficiency and effectiveness of classification. Although a number of comparative studies have been made on feature selection, feature extraction and classification methods of gene expression of microarray data, they were all conducted on different gene expression data sets.

## 3 Methodology

### 3.1 Data Set

Colon cancer dataset was used in this experiment, it contains an expression of 2000 genes with highest minimal intensity across 62 tissues, derived from 40 tumor and 22 normal colon tissue samples (Alon, Barkai and Notterman, 1999). The gene expression was analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. The gene intensity has been derived from about 20 feature pairs that correspond to the gene on the DNA microarray chip by using a filtering process. Details for data collection methods and procedures are described in (Alon, Barkai and Notterman, 1999), and the data set is available from the website http://microarray.princeton.edu/oncology/.

The computer configuration exploited for the purpose of comparative study uses iCore2 processor, 4GB RAM size, 64-bit System and MATLAB 2015a as the implementing tools.

### 3.2 Methods

This study presents an approach for classification of microarray data, which consists of two phases:

i. Input colon cancer dataset (Alon, Barkai and Notterman, 1999) and preprocess using dimension reduction methods such as One-Way-ANOVA as a feature selection technique, PCA and PLS as feature extraction techniques.
ii. After selecting the relevant features, SVM is applied to classify the microarray dataset in terms of accuracy, sensitivity, specificity and precisions (Fig.1).

To compute the importance of relevant gene in a dataset, One-Way-ANOVA based on p-value was used as a technique to analyze the colon cancer data, in which one or more response variables are measured under various conditions identified by one or more classification variables. In an analysis of variance, the variation in the response is separated into variation attributable to differentiate between the classifications variables and variation attributable to random error. An analysis of variance constructs tests to determine the significance of the classification effects. A typical goal in an analysis of variance is to compare means of the response variable for various combinations of the classification variables (Bharathi and Natatjan, 2010).

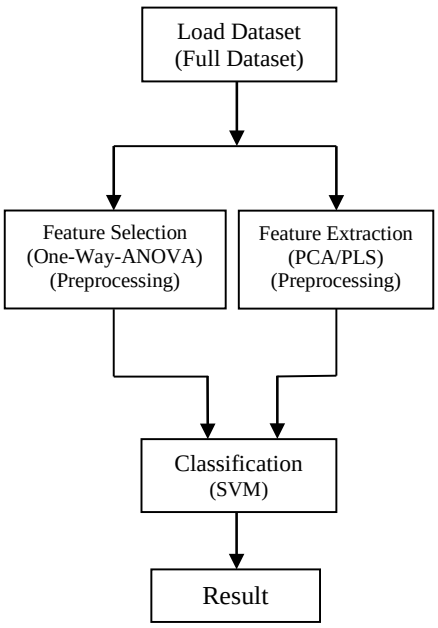


**Figure 1: Technique Workflow**

### 3.2 Dimensionality Reduction Methods

#### 3.3.1 Feature Selection Method

The motivation is to carry out a one-way ANOVA feature with 0.05 p-values for selection of responsive data; it tests whether or not all the different classes of Y contain the equivalent mean as X, by adopting (Nadir, Othman and Ahmed, 2013). The following features apply:

$N_j = The\ number\ of\ cases\ with\ Y = j$
$X_j = The\ sample\ mean\ of\ predictor\ X\ for\ target\ class\ Y = j$

$S_j^2 = The\ sample\ variance\ of\ predictor\ X\ for\ target\ class\ Y = j$:

$$S_j^2 = \sum_{i=1}^{N_j} (X_{ij} - \bar{X}_j)^2 / (N_j - 1) \quad (1)$$

$\bar{X}$ : $The\ grand\ mean\ of\ predictor\ X$:

$$\bar{X} = \sum_{j=1}^{J} N_j X_j / N \quad (2)$$

The notations above are based on non-missing pairs of the sample and attribute of the dataset used in terms of (X, Y). The p-value is calculated by p value = Prob {F (J-1, N-J)>F}:

Where,

$$F = \frac{\sum_{j=1}^{J} N_j (x_j - \bar{x})^2 / (J-1)}{\sum_{j=1}^{J} (N_j - 1) x_j^2 / (N-1)} \quad (3)$$

F (J-1, N-1) is an indiscriminate variable that follows an F distribution with level of freedom J-1 and N-J. If the denominator for a predictor is zero, set the p-value = 0.5 for the predictor. Predictor is classed by sorting according to the p value in ascending order. If there is tie, sort by F in descending order and if it still ties sort by N in descending order. Classification of features shows that out of 2001 features, 416 features were the most significant features correlated to the microarray data analysis. It uses the p-values to rank the significant features with small values and the sorted numbers of features are further processed.

### 3.3.2 Principal Component Analysis (PCA)

PCA is a widely used unsupervised feature extraction technique; it works by replacing the original variables in a data with numerical variables called principal component by capturing the most descriptive features with respect to the most relevant ones (Arif, 2009). PCA mathematically transforms data by referring them to a different coordinate system in order to obtain the greatest variance. A number of correlated variables into a smaller number of uncorrelated variables called principal components (Smitarani and Pratishya, 2014).

PCA identifies patterns of similarities and differences in a data, these patterns are determined and can be compressed by reducing the numbers of dimensions without much loss of information.

In order to conduct the PCA analysis for the input data, the following steps are performed by adopting [16]:

1) Create N x d data matrix with one row vector $\boldsymbol{x_n}$ per data input
2) Subtract mean from each row vector $\boldsymbol{x_n}$ in X
3) Calculate the covariance matrix of X
4) Find Eigen Vectors and Eigen values of $\sum$
5) Fetch the Eigen vector with the largest Eigen values

The PCA are uncorrelated and the components explain the largest percentage in the dimensional dataset with results in extracting 10 components which are considered relevant in the colon cancer dataset used.

### 3.3.3 Partial Least Square (PLS)

Partial Least Square (PLS) is a supervised feature extraction technique, which is widely used as a procedure in modeling associations linking blocks of experimental variables by means of latent variable, it tries finding uncorrelated linear transformations (latent components) of the original predictor variables which have high covariance with the response variables (Xue, and Guo, 2014). The goal of PLS is to find the linear relationship between the response and explanatory variables y and X:

$$X = TP^T + E_x \quad (4)$$

$$y = TC^T + E_y \quad (5)$$

Where T represents the scores (latent variables) P and C are loadings, and $\boldsymbol{E}x$ and $\boldsymbol{E}y$ are the residual matrices obtained the original **X** and **y** variables.

Feature extraction using PCA ignores the response variable and its equivalence. PLS integrates the response variable during the dimensionality reduction procedure. PLS outperforms PCA in the case of microarray gene expression, PLS only consists of indicating the amount of gene components whereas PCA necessitates choosing the essential gene

components (Nebu, Vinjay, Gayathri and Jaisankra, 2012).

### 3.4 Classification

In this step, the results for classification are computed using SVM for classification. SVM is a constructive learning procedure based on statistical knowledge theory (Vapnik, 1998), it is used for classification tasks, and it uses linear models in implementing non-linear class boundaries by transforming input space using a non-linear mapping into a new space. SVM produces an accurate classifier with less over fitting and it is robust to noise.

Assuming $\{(x_1, y_1), \dots, (x_n, y_n)\}$ be a training set with $x_{1i} \epsilon R^d$ and $y_i$ is the corresponding target class. SVM can be reformulated as:

Maximize:

$$\mathbf{J} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T, x_j) \quad \textbf{(6)}$$

**Subject to:**

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \ and \ \alpha_i \geq 0, i = 1, 2, \dots, n \quad \textbf{(7)}$$

This is the weighted average of the training features. Here, αi is a Lagrange multiplier of the optimization task and αi is a rank label. Values of α′si are non-zero for all the points lying inside the margin and on the correct side of the classifier. The kernel function is used to solve the problem. The Kernel function analyses the relationship among the data and it creates a complex divisions in the space (ML, 2015).

## 4 Results, Data Analysis and Discussion

To assess the performance of the proposed approach, several experiments have been conducted on a publicly available dataset (Alon, Barkai, and Notterman, 1999). A brief description of the dataset (the salient features of each dataset is summarized in Table I).

**Table I: Result Evaluations**

| **Dataset** | **Features Selected (one-way ANOVA)** | **Feature Extracted (PCA)** | **Feature Extracted (PLS)** |
|---|---|---|---|
| Colon Cancer (2001x62) | 416 | 10 Components | 20 Components |

In the first experiment, feature selection methods using One-Way-ANOVA and SVM as the classifier for the function of the number of genes retained: 416 out of 2001. The average performance of the approach across the dataset (the performance for the dataset is reported in a confusion matrix).

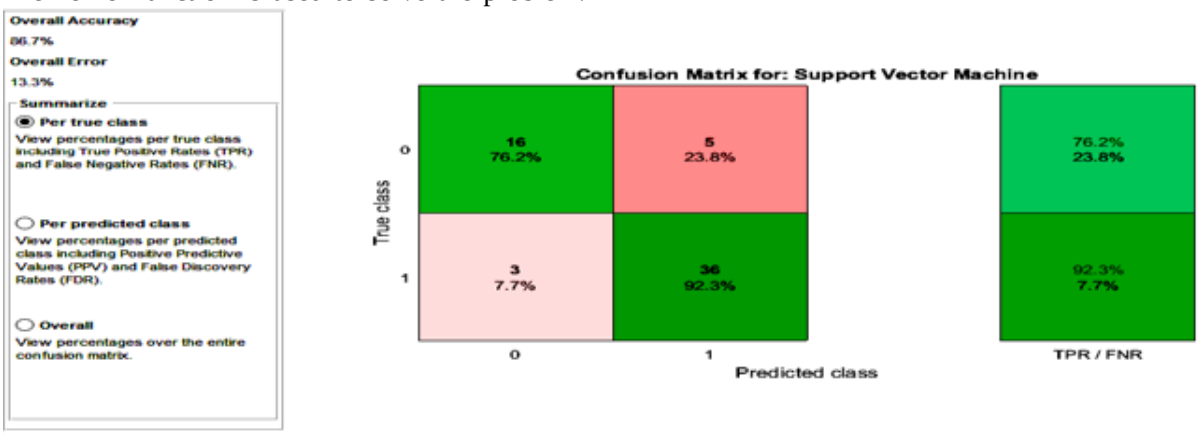


**Figure 2: Confusion Matrix of Proposed Classification, using One-Way-ANOVA for Classification**

True Positive Rate 76.2% and False Negative Rate yields 92.3%. TP=36 FP=5 FN=3 TN=16

**ACCURACY:** *(TP + TN) / (TP+TN+FP+FN) = 86.67%*

**SENSITIVITY:** *TP/ (TP+FN) = 92.31*

**SPECIFICITY:** *TN/ (FP+TN) = 76.19*

**PRECISION:** *TP/ (TP+FP) = 87.81*

In the second experiment, feature extraction method is compared using PCA to reduce the high-dimension and SVM is used as the classifier. PCA is used to de-correlate the data and 10 components was achieved, in Fig. 3, the overall accuracy on all the datasets obtained using PCA as feature extraction to transform and extract the dataset is reported in a confusion matrix.

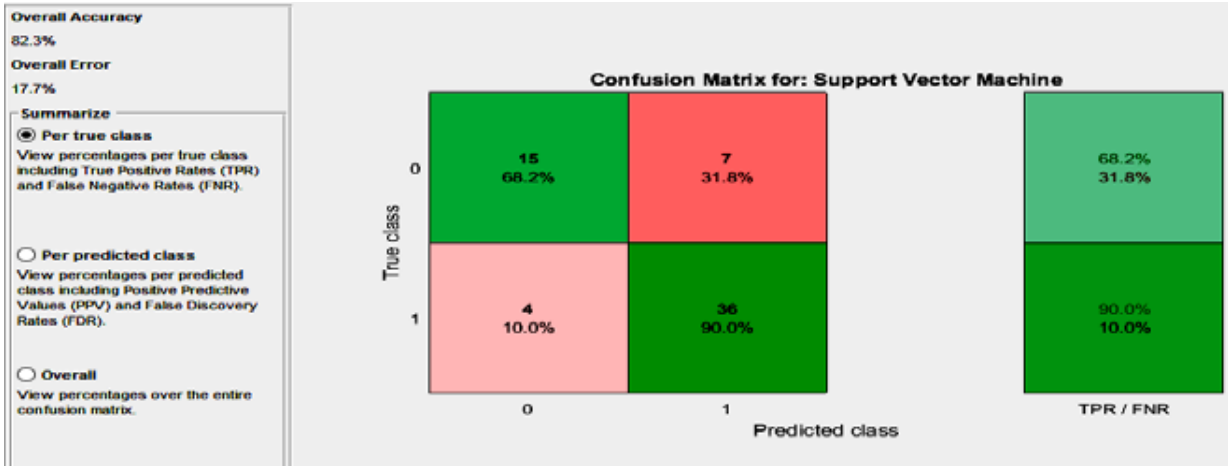


**Figure 3: Confusion Matrix of Proposed Classification, using PCA-Based for Classification**

True Positive Rate 68.2% and False Negative Rate yields 90.0%. TP=36 FP=7 FN=4 TN=15

**ACCURACY:** *(TP + TN) / (TP+TN+FP+FN) = 82.3%*

**SENSITIVITY:** *TP/ (TP+FN) = 90.00*

**SPECIFICITY:** *TN/ (FP+TN) = 68.18*

**PRECISION:** *TP/(TP+FP) = 83.72*

In the third experiment, this paper evaluates the performance of colon cancer dataset using PLS as a feature extraction method with SVM as a classifier. It obtained a better overall accuracy compared to the first and second experiment as shown in the confusion matrix of Fig. 4, it de-correlated the features of colon cancer dataset into 20-components which are considered relevant.

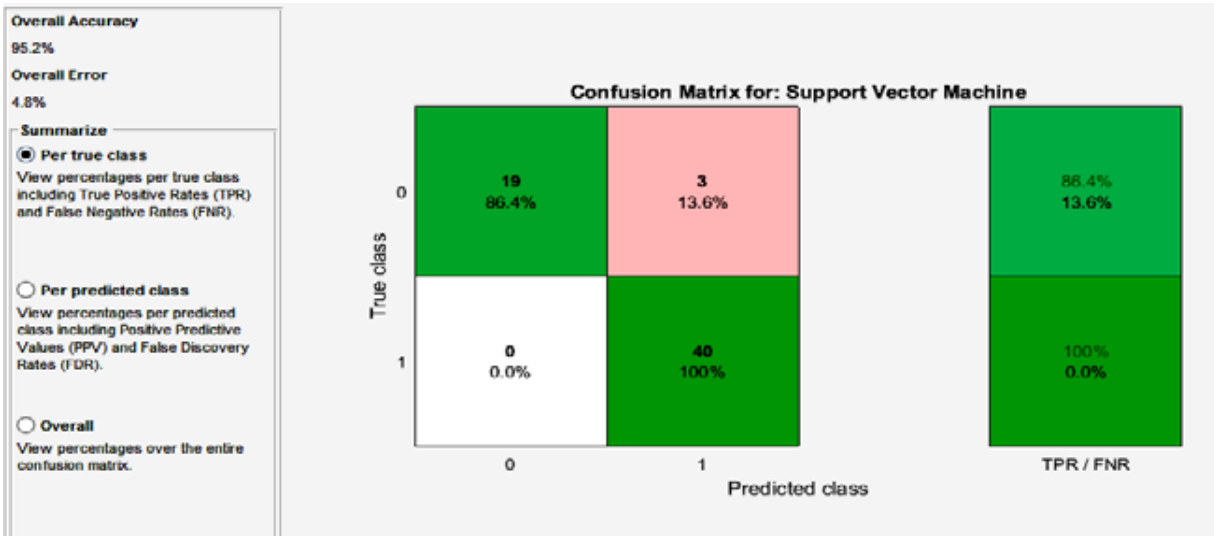


**Figure 4: Confusion Matrix of Proposed Classification, using PLS for Classification**

True Positive Rate 86.4% and False Negative Rate yields 100.0%. TP=40 FP=3 FN=0 TN=19

**ACCURACY:** *(TP + TN) / (TP+TN+FP+FN) = 95.16%*

**SENSITIVITY:** *TP/ (TP+FN) = 100*

**SPECIFICITY:** *TN/ (FP+TN) = 86.36*

**PRECISION:** *TP/ (TP+FP) = 93.02*

**Table II: Performance Evaluation of Proposed One-Way-ANOVA, PCA and PLS Methods**

| S/No | Performance Metrics | One-Way-ANOVA Method | PCA Method | PLS Method |
|---|---|---|---|---|
| 1 | Training Time | 71.2777 | 47.2686 | 5.0088 |
| 2 | Accuracy (%) | 86.67 | 82.30 | 95.16 |
| 3 | Sensitivity (%) | 92.31 | 90.00 | 100 |
| 4 | Specificity (%) | 76.19 | 68.18 | 86.36 |
| 5 | Precision (%) | 87.81 | 83.72 | 93.02 |
| 6 | Area Under Curve | 0.907204 | 0.848864 | 0.994318 |
| 7 | Error (%) | 13.3 | 17.7 | 4.8 |

Table II illustrates a comparative chart between the three methods used in terms of several performance measures such as accuracy, sensitivity, specificity, precision, error, time and area under curve. This comparison shows the integrity of the proposed approach with respect to the state of the art. The colon cancer dataset used to generate our result achieved its best on PLS for feature extraction; it makes this method suitable for practitioners.

## 5 Conclusion

In this paper, a widely used datasets was used in the literature for the evaluation of the algorithms used. The dimension reduction algorithms used to eliminate high dimensional data were One-Way-

ANOVA, PCA and PLS, it uses SVM as its classifier, and it was successfully implemented on MATLAB. For the purpose of finding the smallest gene subsets for accurate cancer classification, PLS method is highly effective compared to One-Way-ANOVA and PCA.

**References**


Shengyan, Z., & Iagnemma, K. (2010). Self-supervised learning method for unstructured road detection using Fuzzy Support Vector Machines, *International Conference on Intelligent Robots and Systems*, IEEE, pp. 1183–1189.

Techo, J., Nattee, C., & Theeramunkong, T. (2008). A Corpus-Based Approach For Keyword Identification Using Supervised Learning Techniques, *5th International Conference On Electrical Engineering/Electronics, Computer, Telecommunications And Information Technology*, Ecti-Con, Vol 1, pp. 33–36.

Mukesh, K., nitish, K.R., Amitav, S., & Santanu, K.R. (2015). Feature Selection and Classification of Microarray Data Using Map Reduce Based ANOVA and K-Nearest Neighbor, *11th International Multi-Conference on Information Processing, Procedia Computer Science*, Vol 54, pp. 301-310.

Flores, M., Hsiao, T., Chiu, Y., Chuang, E., Huang, Y., & Chen, Y. (2013). Gene Regulation, Modulation, and their Applications in Gene Expression Data Analysis, *Advances in Bioinformatics* 2013, pp. 360678–360678.

Wang, L., Chu, F., & Xie, W. (2007). Accurate Cancer Classification using Expressions of very Few Genes, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (TCBB), Vol 4, No 1, pp. 40–53.

Cecille, F., Dana, K., & Otman, B. (2013). Feature Selected Tree-Based Classification, *IEEE Transactions on Cybernetics*, Vol 43, No. 6, pp. 199-204.

Wald, R., Khoshgoftaar, T.M., & Napolitano A. (2013). Stability of Filter- and Wrapper-Based Feature Subset Selection, *IEEE 25th International Conference On Tools With Artificial Intelligence*, pp. 374–380.

Maryam, M., and Mohammad, H.M. (2016). *A Novel Feature Extraction Approach Based on Ensemble feature Selection and Modified Discriminant Independent Component Analysis for Microarray Data Classification*, Nałęcz Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences. Published by Elsevier, Vol 36, pp. 521-529.

Atiyeh, M., & Mohammad, H.M., (2016), Robust Feature Selection from Microarray data Based on Cooperative Game Theory and Qualitative Mutual Information. *Advances in Bioinformatics*, Hindawi, Article ID 1058305, 16 pages.

Smitarani, S., & Pratikshya, M., 2014 “Microarray Classification Using Intelligent Techniques”, International Journal of Scientific and Engineering Research, Vol 5, No. 7, pp. 1-5

Vapnik, V.N. (1998). Statistical Learning Theory. New York: John Wiley & Sons.

Alon, U., Barkai, N., & Notterman, D.A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceeding of Nationall myAcademic Sciences USA 1999, Vol 96, pp. 6745-50

Bharathi, A., & Natatjan, A.M. (2010), Cancer Classification of Bioinformatics using ANOVA, International Journal of Computer Theory and Engineering, Vol 2, No. 3, pp. 369-373.

Nadir, O.E., Othman, I., & Ahmed, H.O. (2013). A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. Research Journal of Applied Sciences, Engineering and Technology Vol 7, No. 3, pp. 625-638.

Arif, M.W. (2009), Microarray Classification with Hybrid Approaches, *Canadian Journal of Pure and Applied Sciences*, Vol 3, No. 1, pp. 759-763.

Vijayarani, S., & Sylviaa, S.M., (2016), Comparative Analysis of Dimensionality Reduction Techniques, *International Journal*

*of Innovative Research in Computer and Communication Engineering*, Vol 4, No 1, pp. 23-29.

Xue-Qjang, Z., & Guo-Zheng, L., (2014), Dimension Reduction for Protein Recognition by Using Incremental Partial Least Squares, *IEEE*. pp. 73-79.

Nebu, V., Vinay, V., Gayathri, P., & Jaisankra, N. (2012). A Survey of Dimensionaliity Reduction and Classification Methods. *IJCSES*, Vol 3, No. 3, pp. 45-54.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: John Wiley & Sons.

Vijayarani, S. & Maria, S. S (2016). Comparative Analysis of Dimensionality Reduction Techniques, *International Journal of Innnovative Research in Computer and Communication Engineering*, Vol 4, No 1, pp. 23-29.

Ali, A., Paul J.K., & Madhu, G. (2011). Dimension Reduction of Microarray Data Based on Local Principal Component, *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* Vol 5, No 5.

**About the authors**

**Micheal O. Arowolo** Holds his B.Sc and M.Sc degree in Computer Science from Al-Hikmah University, Ilorin, and Kwara State University, Malete, Nigeria respectively. He is an Oracle Certified Expert, a member of the IAENG, EAI and SDIWC. His research interests are Bio-informatics, Datamining, Machine Learning and Software Engineering.

Email: micheal.arowolo14@kwasu.edu.ng

**Sulaiman O. Abdulsalam** He holds a Bachelor of Science degree and Master of Science Degree in Computer Science, both from University of Ilorin, Nigeria. He is currently a lecturer at the Department of Computer Science, Kwara State University, Malete, Nigeria. He is a member of Nigeria Computer Society. His research intrests include Data Mining, Machine Learning and Software Engineering.

Email: abdulsalamny@gmail.com

**Rafiu M. Isiaka** he has his Ph.D. in Computer science, he is a lecturer in the Department of Computer Science, Kwara State University Malete since 2009. His research interest includes soft computing, e-learning, data mining and information security.
Email: imabdulrafiu@yahoo.com

**Kazeem A. Gbolagade** is a Professor and Provost at the College of Computer in Information Science, Kwara State University, Malete, Nigeria. He was born in Iwo (Osun State), Nigeria, on the 27th of August, 1974. He received his B.Sc degree in 2000 in Computer Science from the University of Ilorin, Kwara State, Nigeria. In 2004, he obtained his Masters degree from the University of Ibadan, Nigeria. In April 2007, he joined the Computer Engineering Laboratory group at the Delft University of Technology (TU Delft), The Netherlands. In TU Delft, he pursued a PhD degree under the supervision of Prof. Sorin Cotofana. He is a member of the IEEE. His research interests include Digital Logic Design, Computer Arithmetic, Residue Number Systems, VLSI Design, and Numerical Computing.
Email: kazeem.gbolagade@kwasu.edu.ng