IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 24, 2020, accepted October 2, 2020, date of publication October 7, 2020, date of current version October 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029234

A Hybrid Heuristic Dimensionality Reduction Methods for Classifying Malaria Vector Gene Expression Data

MICHEAL O. AROWOLO[1], MARION OLUBUNMI ADEBIYI[1], (Member, IEEE), AYODELE ARIYO ADEBIYI[1], (Member, IEEE), AND OLATUNJI JULIUS OKESOLA[2]

[1]Department of Computer Science, Landmark University, Omu-Aran 24001, Nigeria
[2]Department of Computer Science, First Technical University, Ibadan 200284, Nigeria

Corresponding author: Micheal O. Arowolo (arowolo.olaolu@lmu.edu.ng)

This work was supported by the Landmark University, Omu-Aran, Kwara, Nigeria.

ABSTRACT Malaria is the world's leading cause of death, spread by Anopheles mosquitoes. Gene expression is a fundamental level where the effects of unseen vital revealing genes and developmental systems can be evident for detection of distinctions in malaria infections, to recognize the biological processes in human. Ribonucleic acid sequencing offers a large-scale measurable generated profiling transcriptional data results that help a variety of applications such as scientific and clinical condition studies. A fundamental limitation of ribonucleic acid sequencing consists of high dimensional, infrequent and noises, making classification of genes challenging. Several approaches have proposed enhancing the problem of the curse of dimensionality problem, requiring more improvement, yet it is critical to obtain accurate results. In this study, a hybrid dimensionality reduction technique proposes an optimized Genetic algorithm to pick pertinent subset features from the data. Features chosen is passed into principal component analysis and independent component analysis methods grounded on their class variants, to help transform the selected elements into a lower dimension separately. Support vector machine kernel classifiers used the reduced malaria vector dataset to assess the classification performance of the experiment.

INDEX TERMS Genetic algorithm, principal component analysis, independent component analysis, support vector machine, hybrid approach, RNA-sequencing, gene expression, malaria vector.

I. INTRODUCTION

Malaria has led to the death of millions of humans, and it requires a precise and fast method for identifying anopheles' mosquitoes, predicting and diagnosing malaria transmission is of the essence [1]. The massive volume of genomic and proteomic data is obtainable freely in several repositories. It is increasingly vital to develop enhanced skills of processing these data for beneficial health interpretability of prognosis and diagnosing health conditions and obtaining cost-effective results from the potential data [2], [3].

Gene expression exposes results of several genomic observable programs, over the years, determining transcripts of genetic factors has credibly transferred from microarray technologies to sequencing. Ribonucleic acid sequencing (RNA-seq) runs a quantifiable transcriptional result on large numbers of cells and enables a variety of clinical and scientific application.

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu.

A lot has proposed about the features of RNA-seq datasets, and numerous performances of developed methods [4]. RNA-seq advances have generated massive transcriptome datasets, with advanced cell variety understanding and its fundamental procedures in standardized populations. Nevertheless, RNA-seq data are high-dimensional, containing noises, insufficient. It requires several steps for its computational analysis, including dimensionality reduction, clustering, mapping, expressed genes, classification, among others [5]. The advent of computing technology has made rapid development and challenges in health care practices for diagnostic methods and keeping up-to-date records, resulting in mining massive medical records [3].

Owing to the high dimension and insignificant gene expression data sample-sizes, making RNA-Seq technology challenging [6]. Dimensionality reduction techniques lessen unwanted properties in RNA-Seq data for better performance feature. Dimensionality reduction approaches are used in selecting select [7] and extract [8] relevant and specific information in a massive amount of dataset,

which are known as the feature selection and the feature extraction [9].

Classification of diverse types of predominant gene expression RNA-Seq data in enumerating genes for biological questions has given beneficial information for identifying and discovering drugs [10], [11].

A hybrid dimensionality reduction technique is proposed, to classify malaria vector gene expression RNA-Seq data and identify significant features using genetic algorithm as a feature selection technique to fetch pertinent subset features from given dataset. The picked features are conceded into the Principal Component Analysis (PCA) and Independent Component Analysis (ICA) algorithms separately to extract additional optimal latent components on the RNA-Seq malaria vector dataset, Support Vector Machine (SVM) kernel classification algorithms used to evaluate the effectiveness of this experiment. The hybrid approaches presented a classification performance of 96.12% and accuracy, respectively, for prediction of drug classification for malaria infection.

II. RELATED WORKS

Ali *et al.* [9], worked on dimensionality reduction of bioinformatics data using PCA and Factor analysis, on leukemia gene expression dataset to reduce the number of attributes, and extract essential features, there experiment achieved an enhanced result compared to state of the art.

Pradhan [2], proposed a computational procedure, using Probabilistic-PCA and an evolutionary programmed improved supervised classifier using Artificial Neural Network (ANN) for cancer gene expression data, the enhance classifier trained using the backpropagation algorithm to minimize error. The proposed method improved ANN compared to other classification approaches.

Uma and Kirubakaran [6], proposed a hybrid investigative dimension reduction method for classifying gene expression data by combining Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO), the experimental technique correlates well with classification error after training, with the classification accuracy enhanced, compared to other predictable methods.

Jain and Singh [3], proposed a hybrid dimensionality reduction model, using Relief-F and PCA methods on various long-term disease datasets. There experiment determined the best value of threshold in picking relevant features; they obtained a reduced significant computational time with appropriate features selected, and 50% reduced irrelevancy.

Mei *et al.* [12], suggested a dimensionality reduction method for classifying tumor gene expression data. They used a discriminant hybrid structure to enhance data separability and improving tumor classification accuracy.

Karthik and Sudha [13], reviewed machine learning methods for classifying gene expression model computational analytical structure for complicated diseases, by identifying several differentially expressed gene techniques. The paper provided a thorough overview of several machine learning techniques utilized in investigating and classifying gene expression data for ailments.

Guo *et al.* [14], proposed a hybrid method using Information gain for filtering foreign genes. SVM s used to classify and to eliminate noises efficiently. The proposed IG-SVM used as an input for LIBSVM classifier. Their experiment showed higher accuracy related to state of the art.

Utami and Rustam [15], proposed a hybrid method for classifying cancer using PSO, Ant Bee Colony (ABC) and SVM, they selected revealing genes by removing ineffective genes in the high-dimensional data with a rank method. Their experiment picked revealing genes, SVM was used in removing disproportionate genes after filtration using PSO and ABC, there experiment presented a comparatively enhanced accuracy of 88%.

Li *et al.* [16], proposed dimensionality reduced approach for gene expressional data classification. It supervised principal component regressions on experimental gene datasets that proves effective outcomes and stability utilized.

Souza *et al.* [17] proposed a dimensionally reduced gene expression dataset. They analyzed dimensionality reduction methods for gene expression datasets by preprocessing and evaluating. They introduced a combined consistent based subset evaluation and minimum redundant maximum relevancy called the CSE-minimum Redundancy Maximum Relevance (mRMR) for improving and enhancing classification proficiency. Their outcome showed an enhance classification for both methods.

Jabeen *et al.* [18] suggested a machine learning technique for classifying RNA-Seq datasets by discussing numerous approaches and implementation. Progresses in bioinformatics, and machine learning offers prevailing toolboxes that rank information transcriptomes existing in RNA-Seq data.

Chiesa *et al.* [19], proposed a genetic algorithm approach for identifying robust subset features in high dimensional datasets for revealing certain features in high-dimensional datasets, by outperforming typical feature selection approaches with better classification accuracies in reasonable computational time.

Tsuyuzaki *et al.* [20], proposed a conventional PCA for large-scale Single-Cell-RNA-Seq (Sc) data. They reviewed existing fast with memory sufficient PCA algorithm implementations. They evaluated the applications which showed that PCA procedures grounded on Krylov subspace and random singular value breakdown are fast, retention adequate and precise.

Huynh *et al.* [21] proposed an innovative hybrid Deep Convolutional NN (DCNN)-SVM approach, classifying gene expression RNA-Seq data. They used DCNN to abstract hidden features from a cancer RNA-Seq gene expression data, SVM used as a classifier which the results were efficiently accurate compared to hi-tech.

Piao *et al.* [22], proposed an ensemble approach for predicting prostate cancer RNA-Seq data, by simulating non-parametric methods and submitted its application on other diseases.

Shon *et al.* [23], proposed a classification approach using an endways, cost-effective hybrid deep-learning method on a combined clinical kidney cancer gene data, by combining deep symmetric auto-encoder. They determined the optimal and estimated classification accuracy model, the experiment showed a better efficiency compared to hi-tech and can be functional for extracting features from a gene biomarker for prognosis, diagnosis and prevention of kidney cancer.

Varghese *et al.* [24], proposed a dimensionality reduction and classification methods. Their experiment offered a knowledge for diverse existing dimensionality reduction methods and applicability of techniques to gene expression dataset parameters with algorithms. They also presented procedures for selecting relevant algorithms for specified and determined instances when algorithms are appropriate for performing predefined tasks.

Arowolo *et al.* [25], proposed a reduced dimensional approach for classifying gene expression RNA-Seq data. They compared several SVM classification kernels. They used PCA to extract underlying components from a malaria vector Anopheles gambiae data. An efficient accuracy achieved related to hi-tech.

Compared to studies discussed in the literature, this study uses an optimized based GA feature selection approach to select the optimal subset of relevant genes which produces higher reliable accuracy to report the issue of dimensionality in the optimization for genetic algorithm. Further, the reduced data utilizes a class-based PCA or ICA feature extraction techniques based on its class variants to obtain a reduced latent component data of relevant genes for classification. The system framework is different from the existing ones in the way of hybridization by using two-dimension reduction techniques. Feature selection has become a requirement before feature extraction for hybridization in recent times [36]. The search for an optimal subset of relevant genes improves classification accuracy.

A metaheuristic study on RNA-Seq classification [55], was proposed by comparing several classifiers with and without transformations, their result revealed increasing sampling size.

An informative metaheuristic framework for gene expression data was proposed with GA-LV with 99% accuracy [56].

A GECKO approach was proposed for genetic algorithm and classification using a k-mer optimizer on RNA-Seq data.

III. MATERIALS AND METHODS

This section analyzes fundamental RNA-Seq gene expression data analysis and technology concepts by giving an overview and discussions on the RNA-Seq technology, dimensionality reduction and classification approaches.

A. RIBONUCLEIC ACID SEQUENCING GENE EXPRESSION

RNA-Seq is a broadly used diagnostic gene expression data technology. It realizes numerous facets of transcriptomes—a leading present-day DNA-microarray technology, used in carrying out a high-throughput investigation of gene expressions. Providing improved understanding into cell transcriptomes, giving unconventional treatments and better resolutions [26-27], it distinguishes early hidden variations happening in disease conditions by answering to therapeutics of diverse environments and other training, creating ample amount of sequencing data [21]. Gene expression RNA-Seq data classification has given beneficial evidence for identifying and determining germane drugs for ailments. Gene expression is a genetic factor for prevalent RNA-Seq method that quantifies and gain improved understandings to numerous biological questions [28]. A significant challenge of RNA-Seq is the problem of diagnostic challenges, and it gives unfitting outcome due to high dimensions of gene expression data. Numerous machine learning approaches proposed to enhance gene expression data classification such as; dimensional reduction, clustering, classification, among others [21].

B. DATASET

Gene expression RNA-Seq data for Anopheles gambiae larvae gathered around neighborhoods of Bungoma Western region of Kenya [25], [29]. It contains profiles of deltamethrin-resistant and susceptible mosquitoes for understanding resistance mechanisms; the dataset comprises of 7 attributes involving of the Test_ID, Gene_ID, Genes, Locus, Resistant, Susceptible and status with a predictor and 2457 instances [51].

C. DIMENSIONALITY REDUCTION

Dimensionality reduction is a well-known method for removing noises and redundant features. Gene expression datasets contain high dimensionality that results in high computational weight and deprivation of classification model algorithm performances. Dimensionality reduction methods are required, to remove redundancy and fetch irrelevant features that are disturbing the performance and operation by decreasing the feature ratios of the samples. This process helps in reducing the probability of overfitting [30]. Dimensionality reduction is two effective methods known as the feature selection [7] and feature extraction [31].

1) FEATURE SELECTION

Feature selection is a significant stage in building a machine learning classification, in technologies like RNA transcript to make useful selective identifier features for transcript sequences for training and testing models [18]. Feature selection helps in selecting appropriate elements to be executed in classification models and eliminate insignificant and unnecessary features, to diminish the curse of dimensionality. It helps in making learning procedure for classification period effective and increases the performance model [32]. Extensive data feature selection procedure, for example; RNA-Seq data requires supervised and unsupervised learning in making decisions. Rank features conferring to significance is essential for classification problems, and picking the best can advance the performance of the prediction model [33]. Feature selection is an effective method known as the Filter, Wrapper and Embedded approaches [18], [31].

2) GENETIC ALGORITHM

GA is an evolutionary wrapper-based feature selection algorithm used for explorations of engine optimization problems. GA can be reliant on real behaviors relating to human genes, in the survival of the fittest basis. GA consists of initial population generation, fitness evaluation, selection of parents, crossover and mutation [6], [31], [34].

GA is a heuristic exploration procedure with its most uncomplicated form with a population of arbitrarily created results (phenotypes or entities) having a suitable value giving to the objective purpose quantifying the solutions worthiness. Each chromosome or genotype has a set of properties usually characterized as 0s and 1s binary strings [7], [35].

GA has a limitation of optimality, although very sensitive to initial population, with its solution quality deteriorating while problem size increases, enhancing it for gene sampling has proven to generate reasonable quality solutions.

D. FEATURE EXTRACTION

Feature extraction is a method used in identifying significant features, traits or attributes present in a data. Examples of feature extraction technique include; pattern recognition and detecting common precedents in a group of reports. Data with stacks of dimensions requires feature extraction usage to generate a better brief of its characterization. Feature extraction makes innovative variables of selected features to diminish the curse of dimensionality present. Two large groups of feature extraction algorithms exist, namely: the linear (adopts data on a low-dimensional subspace, for example, PCA) and non-linear (works with a low-dimensional subspace represented on a high dimensional attribute for a non-linear connection between features can be originated, for example, ICA) [36].

1) PRINCIPAL COMPONENT ANALYSIS

PCA is a linear feature extraction algorithm technique, extensively utilized mainly in biological investigations [37]. PCA projects feature spaces from high to a lower dimension by reconstructing the k-dimensional unrelated features from the unique n-dimension feature of the area. PCA has recognized to be an essential tool for discovering high dimension gene expression data. It has frequently utilized on RNA-seq data [38].

PCA explores orthogonal alteration by converting a group of correlated variables to a group of uncorrelated variables [3]. PCA used for exploratory data analysis. PCA can apply for the examination of the relationships among a group of variables, and suitable for dimensionality reduction [39].

2) INDEPENDENT COMPONENT ANALYSIS

ICA [40] helps in obtaining concealed features from multi-dimensional information, by decomposing multi-variate indications into independent non-gaussian sections for the components to be statistically independent. ICA finds a correlation between data by decorrelating the data by exploiting or diminishing the specific information.

ICA adopts opinion X as a linear mixture of independent components S.

If A signifies the different matrix of a weighted matrix W, and columns of A characterize the source feature vectors of comment X.

$$S = W \times X, X = A \times S \tag{1}$$

ICA has been extensively utilized for biological information, recognitions and other grounds [4].

3) PRINCIPAL COMPONENT ANALYSIS VERSUS INDEPENDENT COMPONENT ANALYSIS

PCA is a linear transformation technique used in reducing the dimensionality and feature numbers. It is an ''unsupervised'' algorithm whereas ICA is ''non-linear'', ICA has been proven to work improvingly if a data is preprocessed [36].

E. CLASSIFICATION

Classification is said to be a supervised learning approach in data mining technique. It is a prevalent helpful task that assigns and predict class labels given to current data from a predefined class label. Classification building is carried out in two phases [7]:

- the learning phase, where the classification model built giving to set of training data, with a class label.
- The model is used to predict class labels for hidden data, while the accuracy of SVM classifier is measured.

SVM predicts the class of user-nominal purpose features, based on the level of the predictive features [40]. SVM allocates objects to prearranged classes in two steps called the training and testing phases.

Training includes the algorithm analyzing the learning data by generating a classification model.

Testing phase examines the accuracy model using added dataset. Examples of classifiers are Naive-Bayes, KNN, SVM, among other popular classification approaches for gene expression data [41].

1) SUPPORT VECTOR MACHINE

SVM a supervised machine learning classification algorithm, for insightful hyperplanes that optimally segments tuples of a class from another class. Hyperplanes initiated from margins and support vectors, calculated from directions defining the hyperplane [7].

SVM generates result boundaries between positive and negative groups by selecting the utmost pertinent instances intricate in the decision procedure. When data is linearly discrete, the hyperplane construction is always conceivable. SVM uses kernels where non-linearly maps into high dimension feature space for separating found hyperplanes [42], [43]. SVM kernels examples include; polynomial kernel, linear kernel, String Kernels, Radial basis function (RBF), Gaussian kernel, Sigmoid, among others [25].

F. HYBRID APPROACH DYMENSIONALITY REDUCTION

Increase in biological data dimensionality is a big issue to simple, predictable investigation methods. Using conventional methods for learning intricate designs at numerous layers

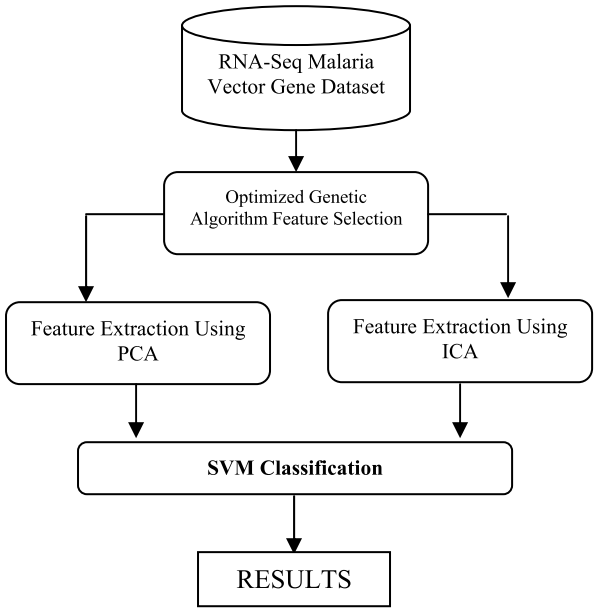


FIGURE 1. Proposed RNA-Seq gene classification framework.

stimulated from morphological operations receptive to processing is of the essence. Most standard techniques used in dealing with high dimensional data such as RNA-Seq data involves several complexities [18]. Combining different dimensionality reduction methods can be of the essence, exploiting specific benefits, where the gene subset attained from one process assisted as the input to alternative [6], [31], [34], [45]- [48], [50]. Generally, feature extraction procedures can be used to efficiently aid feature selection [35], [39], [49], [50], by using feature selection to select the original gene subset or benefit from eliminating redundant genes. Combination of numerous feature extraction techniques can be applied to extract the initial feature subsets [57]–[60].

G. PROPOSED MODEL

In this study, an efficient hybrid dimensionality reduction technique proposed for classifying malaria vector RNA-Seq gene expression data.

RNA-Seq comprises of great potentials to discover, identify and trace cell lineages, however, dimensionality reduction helps in the interpretation of the structures, yet data remains challenging, existing algorithms require the right formation to uncover relevant features, hybrid dimensionality approach has proven to be robust yet requires efficient algorithms to model, capture and visualize the low-dimensional structures in gene expression data.

The proposed classification technique comprises of three phases, namely:

- Feature selection
- Feature extraction
- Classification

The proposed hybrid framework for the classification of malaria gene expression dataset shows in Figure 1. The structure comprises of three subsystems, feature selection subsystem, class-based feature extraction subsystem and classification subsystem.

The feature selection subsystem uses an optimized GA by adapting algorithm one below for selecting an optimal subset by evaluating the fitness for chromosomes. The feature extraction subsystem utilized the PCA and ICA, due to its efficiency invariance projection of data along with orthogonal directions. The performance of the experiments is classified using SVM.

One of the significance of optimizing genetic algorithm is its evolutionary processing of the features of the algorithm, which in turn helps the multiple search point by exploring the optimal solution simultaneously and independently to generate high quality solution, this study uses an optimize genetic algorithm feature selection so as to decrease the number of features and retaining the discriminant features, feature extraction is suitable for transforming the reduced data into latent components, the richness of this is to reduce the thrive and suffers of both dimensionality reduction methods that can be used for malaria classification.

Algorithm 1:

Step 1: prepare parameters p and q and arbitrarily create the first population

Step 2: for i<popsize

Step 3: compute the tangent value $\tan(x_{in}/x_{in+1})$of the involved angle between the two vectors in adjacent dimensions for each pop(i)

Step 4: if step 3=0, then update the value of the nth dimension of the ith discrete to 0 then, do not update the value and continue in step 6

Step 4.1: if x>1

Step 4.2: calculate the compatible persistence with Euclidean distance D**between xi and xj**

Step 4.3: calculate the comparation principle within the distance L= |Xi-Xj| <D

Step 4.4: if no similarity,

i. Remove individual fitness with the parallel of biallelic loci (SD (Xi, Xj)) and regular parallel MSDi

ii. calculate the subpopulations M(t+1) ***;*** **otherwise** ***Merge*** N **entities in memory pool with subpopulation organized by capability in descending order**

Step 4.5: compute the subpopulation $*$ **threshold**

Step 5: judge the convergence condition

Step 6: compute the number of 0 basics in each dimension for the restructured pop; if it is above the critical value Q, delete this dimension

Step 7: get the updated population

Step 8: compute the fitness value F(i) of each discrete in the population

Step 9: set the new population

Step 10: pick two entities from the population according to the fitness with the relational selection algorithm

Step 11: if random (0, 1) < P_c, then move on to step 12; else, implement step 13

Step 12: apply the crossover operative rendering to the crossover probability P_c on the two entities

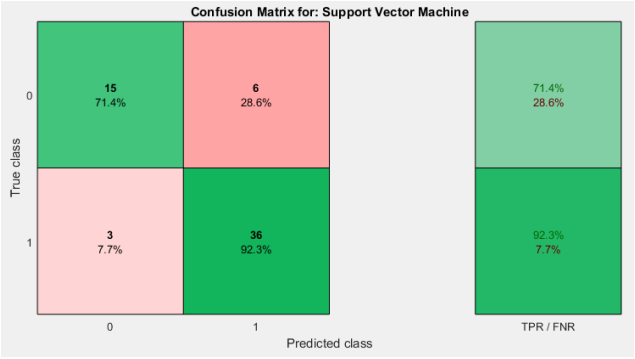


FIGURE 2. Confusion matrix for classifying mosquito anopheles RNA-Seq using GA+ICA+ medium gaussian-SVM TP= 39; TN= 16; FP= 5; FN= 0.

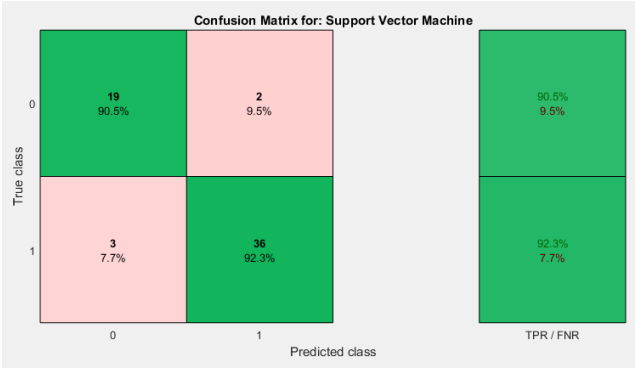


FIGURE 3. Confusion matrix for classifying mosquito anopheles RNA-Seq using GA+PCA+L-SVM TP= 36; TN= 19; FP= 2; FN= 3.

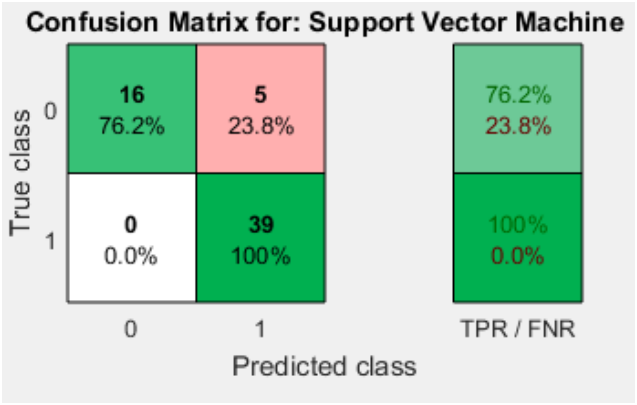


FIGURE 4. Confusion matrix for classifying mosquito anopheles RNA-Seq using GA+PCA+ medium gaussian-SVM TP= 36; TN= 15; FP= 6; FN= 3.

Step 13: if random (0, 1) < P_m, then move on to step 14

Step 14: apply the mutation operator according to the mutation probability P_m on the two individuals

Step 15: add the two new individuals into the new population

Step 16: Repeat this process until the N^{-th} generation produced; otherwise, return to step 4

Step 17: change the population with a new population

Step 18: Reiterate this procedure until the number of groups exceeds G; otherwise, return to step 8

Step 19: end

Figure 2 depicts the methodology and procedure using a malaria vector RNA-Seq data.

Genetic algorithm fetches the relevant features in data. The selected features are given to the feature extraction phases using PCA and ICA algorithms separately to carry for underlying components. SVM kernel classifiers used to calculate the performance metrics for the learning processes.

Step 1: Preprocess the imported data

Step 2: Apply a feature selection algorithm on the data using the Optimized Genetic algorithm

Step 3: Apply PCA feature extraction algorithm on the selected features on the outcome of step 2

Step 4: Apply SVM Classification kernels on the result of step 3

Step 5: Evaluate the performance

Step 6: Repeat step 3 using the ICA feature extraction algorithm

Step 7: Apply SVM kernel classification algorithms on the output of level 6.

Step 8: Evaluate the performance

Step 9: Compare the results of step 5 and step 6

All experiments executed using a computer system with Intel Core 5, 16GB RAM, 64-bit Operating system. All algorithms coded in C++ and MATLAB 2015. confusion matrices are generated as a classification investigation to ensure similar performances of the training and test sets evaluation metrics in terms of accuracy, sensitivity, specificity, precision and recall [51].

IV. RESULTS AND DISCUSSIONS

In this study, an experiment carried out on a malaria vector dataset collected from a publicly available repository with 2457 instances and seven attributes [25], [29], [52], using the MATLAB tool. The dataset experimented using a genetic algorithm to select relevant features in the data, with a threshold of 0.5, 708 subset features of genes were significant. The classifier proficiency related to the state-of-the-art for appropriate comparisons.

Seven hundred eight selected features using Genetic algorithm is firstly passed into the PCA algorithm and extracted ten latent variables in 1.4623 seconds. The extracted features are then given into the SVM classification algorithm using 10-folds cross-validation and the confusion matrix of L-SVM and Medium Gaussian SVM classification algorithms are evaluated.

The selected 708 features secondly passed into the ICA feature extraction algorithm; 25 latent variables extracted in 0.42794 seconds. The extracted features are then given into the SVM classification algorithm using 10-folds cross-validation and the confusion matrix of L-SVM and Medium Gaussian SVM classification algorithms are evaluated.

Classification of dimensionality reduced malaria vector, RNA-Seq data carried out, using GA+PCA+SVM and GA+ICA+SVM algorithms, using two relevant SVM kernels ''the Linear-SVM and Medium Gaussian SVM''.

GA+PCA+L-SVM achieved 85%, GA+PCA+ Medium Gaussian-SVM achieved 92%, GA+ICA+L-SVM achieve

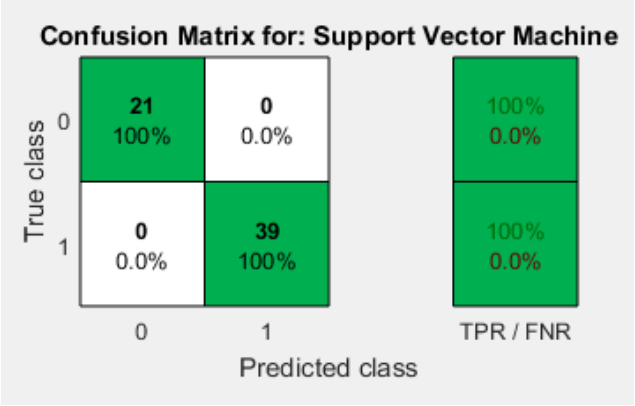


FIGURE 5. Confusion matrix for classifying mosquito anopheles RNA-Seq using GA+ICA+L-SVM TP= 39; TN= 21; FP= 0; FN= 0.

TABLE 1. Performance metrics table for the confusion matrix.

Performance Metrics (%)	GA+PCA+ L-SVM	GA+PCA+ MG-SVM	GA+ICA+ L-SVM	GA+ICA+ MG-SVM
Accuracy	91.7	85.0	100	91.7
Sensitivity	92.3	92.3	100	100
Specificity	90.5	71.4	76.2	76.2
Precision	94.7	85.7	100	88.6
Recall	92.3	92.3	100	100
F-score	93.5	88.9	100	94
Computational Time	1.4623	1.4623	0.4279	0.4279

100%, and GA+ICA+Medium-Gaussian-SVM achieved 92% accuracies respectively. Their performance evaluation tabulated below, comparing the two methods, GA+ICA+SVMs found to perform superior with a 100% accuracy, it is of the essence to note that even the PCA method lower compared to the ICA. Yet, it presented a higher average, compared to the state-of-the-art.

This study contributes with several important implications for gene expression analysis. The potential application of this study is to give insight into biological and technical considerations that can explain uncovered structures and interpretations for genes relevant for predictions, diagnosis of malaria and drug designs.

As reported in Table 1 above, the experiment achieved consistent performances using the applied algorithms relatively.

In this study, a hybrid dimensionality reduction was carried out using an optimized Genetic algorithm feature selection approach. PCA and ICA algorithms were used as feature extraction in the second phase. The third phase utilized an SVM classification algorithm, with 10-fold cross-validation parameter.

The result achieved an improved outcome, as shown in Table 1 above. The accuracies showed an improvement compared to the state-of-the-art.

In order to provide reliable detection and prediction approach for malaria transmission, several researchers have been studying the problem of malaria classification using machine learning algorithms, the results achieve in this study can be used for training required predominance of malaria

TABLE 2. Comparing the performance metrics with other methods.

Performance Metrics (%)	GLM+PCA [53]	GA+PCA+ NN [54]	GA+CCA+ NN [54]	GA+PCA& CCANN [54]
Accuracy	70	85.0	85	88

infection by clinicians, by using this procedure for compiling a pathologist curated dataset for training classifiers and augmentation methods for datasets in significantly increasing the dataset size, in light of the overfitting problems associated with training of datasets. The study of characterizing thousands of genes offers deep insight into malaria classification problems with abundant data explored, for drug discovery, prediction and diagnosis for malaria treatments and understanding functions of genes with the interaction between the genes in normal and abnormal conditions. The proposal of this study increases classification performance results and shows a less dependency of training set.

V. CONCLUSION

RNA-Seq data analysis gives valuable and precious benefits to the performance of the technology, with enormous contributions to advancing the gene expression profiling issues. Relevant applications of RNA-Seq technology includes the dimensionality reduction and classification approaches. It is a vital issue, due to the great curse of dimensionality bounded in the gene expression data. Several methods have proposed towards the enhancement of the technology, the prediction and detection of ailments derived from samples, dimensionality reduction has proven to solve these challenges. Yet, more investigations need to carried out. Hybrid approaches have also used in recent time for classification of gene expression data. These experiments carried out a dimensionality reduction approach using GA feature selection with ICA and PCA feature extraction algorithms separately, and tested their evaluation performance on SVM classification kernels, GA+ICA+SMV outperformed the GA+PCA+SVM based approach.

Hence in future work, this study aims to apply hybrid dimensionality reduction algorithms on classifiers like the KNN, NN to identify relevant genes for gene expression classification.

REFERENCES

[1] A. Ahmed and M. Ahmed, "Morphological identification of malaria vectors within anopheles species in parts of kano state, nigeria," *Bayero J. Pure Appl. Sci.*, vol. 4, no. 2, pp. 160–163, Apr. 2012.
[2] M. Pradhan, "Evolutionary computational algorithm by the blending of PPCA and EP-enhanced supervised classifier for microarray gene expression data," *IAES Int. J. Artif. Intell.*, vol. 7, no. 2, pp. 95–105, 2018.
[3] D. Jain and V. Singh, "An efficient hybrid feature selection model for dimensionality reduction," *Procedia Comput. Sci.*, vol. 132, pp. 333–341, Jan. 2018.
[4] C. Feng, S. Liu, H. Zhang, R. Guan, D. Li, F. Zhou, Y. Liang, and X. Feng, "Dimension reduction and clustering models for single-cell RNA-seq data: A comparative study," *Int. J. Mol. Sci.*, vol. 21, no. 2181, pp. 1–21, Mar. 2020.
[5] X. Guo, Y. Zhang, L. Zheng, C. Zheng, J. Song, Q. Zhang, B. Kang, Z. Liu, L. Jin, and R. Xing, "Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing," *Nature Med.*, vol. 24, pp. 978–985, 2018.

[6] S. M. Uma and E. Kirubakaran, "A hybrid heuristic dimensionality reduction technique for microarray gene expression data classification: A blending of GA, PSO and ACO," *Int. J. Data Mining, Model. Manage.*, vol. 8, no. 2, pp. 160–179, 2016.

[7] N. Almugren and H. Alshamlan, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," *IEEE Access*, vol. 7, pp. 78533–78548, 2019, doi: 10.1109/ACCESS.2019.2922987.

[8] R. Aziz, C. K. Verma, and N. Srivastava, "Dimension reduction methods for microarray data: A review," *AIMS Bioeng.*, vol. 4, no. 1, pp. 179–197, 2017, doi: 10.3934/bioeng.2017.1.179.

[9] M. Usman, S. Ahmed, J. Ferzund, A. Mehmood, and A. Rehman, "Using PCA and factor analysis for dimensionality reduction of bio-informatics data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 5, pp. 415–426, 2017.

[10] Y. Li, K. Kang, J. M. Krahn, N. Croutwater, K. Lee, D. M. Umbach, and L. Li, "A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data," *BMC Genomics*, vol. 18, no. 1, p. 508, Dec. 2017, doi: 10.1186/s12864-017-3906-0.

[11] P-H, Huynh, V-H, Nguyen and T-N, Do, "Novel hybrid DCNN-SVM model for classifying RNA-seq gene expression data," *J. Inf. Telecommun.*, vol. 3, no. 4, pp. 533–547, Oct. 2019, doi: 10.1080/24751839.2019.1660845.

[12] Q. Mei, H. Zhang, and C. Liang, "A discriminative feature extraction approach for tumor classification using gene expression data," *Current Bioinf.*, vol. 11, no. 5, pp. 561–570, Nov. 2016, doi: 10.2174/1574893611666160728114747.

[13] S. Karthik and M. Sudha, "A survey on machine learning approaches in gene expression classification in modelling computational diagnostic system for complex diseases," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 2, pp. 182–191, 2018.

[14] L. Gao, M. Ye, X. Lu, and D. Huang, "Hybrid method based on information gain and support vector machine for gene selection in cancer classification," *Genomics, Proteomics Bioinf.*, vol. 15, no. 6, pp. 389–395, Dec. 2017.

[15] D. A. Utami and Z. Rustam, "Gene selection in cancer classification using hybrid method based on particle swarm optimization (PSO), artificial bee colony (ABC) feature selection and support vector machine," in *Proc. AIP Conf.* 2019, Art. no. 020047, doi: 10.1063/1.5132474.

[16] J. Li, Z. Zhao, L. Zhou, and Y. Wang, "Y-SPCR: A new dimensionality reduction method for gene expression data classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Nov. 2019, pp. 401–408.

[17] J. T. Souza, A. C. Francisco, and D. C. Macedo, "Dimensionality reduction in gene expression data sets," *IEEE Access*, vol. 7, pp. 61136–61144, 2019.

[18] A. Jabeen, N. Ahmad, and K. Raza, "Machine learning-based state-of-the-art methods for the classification of RNA-seq data," in *Proc. Classification BioApps*, 2017, pp. 133–172, doi: 10.1101/120592.

[19] M. Chiesa, G. Maioli, G. I. Colombo, and L. Piacentini, "GARS: Genetic algorithm for the identification of a robust subset of features in high-dimensional datasets," *BMC Bioinf.*, vol. 21, no. 1, p. 54, Dec. 2020.

[20] K. Tsuyuzaki, H. Sato, K. Sato, and I. Nikaido, "Benchmarking principal component analysis for large-scale single-cell RNA-sequencing," *Genome Biol.*, vol. 21, no. 1, p. 9, Dec. 2020.

[21] P.-C. Huyng, V.-H. Nguyen, and T.-N. Do, "Novel hybrid DCNN-SVM model for classifying RNA-sequencing gene expression data," *J. Inf. Telecommun.*, vol. 3, no. 4, pp. 533–547, 2019, doi: 10.1080/24751839.2019.1660854.

[22] Y. Piao, N. H. Choi, M. Li, M. Piao, and K. H. Ryu, *Ensemble Method for Prediction of Prostate Cancer From RNA-Seq Data*. Science Technology, 2014, pp. 51–56.

[23] H. S. Shon, E. Batbaatar, K. O. Kim, E. J. Cha, and K.-A. Kim, "Classification of kidney cancer data using cost-sensitive hybrid deep learning approach," *Symmetry*, vol. 12, no. 1, p. 154, Jan. 2020, doi: 10.3390/sym12010154.

[24] N. Varghese, V. Verghese, P. Gayathri, and N. Jaisankar, "A survey of dimensionality reduction and classification methods," *Int. J. Comput. Sci. Eng. Surv.*, vol. 3, no. 3, pp. 45–54, Jun. 2012.

[25] M. O. Arowolo, M. O. Adebiyi, and A. A. Adebiyi, "A dimensional reduced model for classification of RNA-Seq anopheles gambiae data," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 23, pp. 3487–3496, 2019.

[26] S. Zhao, W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, "Comparison of RNA-seq and microarray in transcriptome profiling of activated t cells," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e78644, doi: 10.1371/journal.pone.0078644.

[27] K. R. Kukurba and S. B. Montgomery, "RNA-sequencing and analysis," *Cold Spring Harbor Protocols*, vol. 2015, no. 11, pp. 951–969, 2015, doi: 10.1101/pdb.top084970.

[28] N. T. Johnson, A. Dhroso, K. J. Hughes, and D. Korkin, "Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers?" *RNA*, vol. 24, no. 9, pp. 1119–1132, Sep. 2018, doi: 10.1261/rna.062802.117.

[29] M. Bonizzoni, E. Ochomo, W. A. Dunn, M. Britton, Y. Afrane, G. Zhou, J. Hartsel, M.-C. Lee, J. Xu, A. Githeko, J. Fass, and G. Yan, "RNA-seq analyses of changes in the anopheles gambiae transcriptome associated with resistance to pyrethroids in kenya: Identification of candidate-resistance genes and candidate-resistance SNPs," *Parasites Vectors*, vol. 8, no. 1, pp. 1–13, Dec. 2015.

[30] L. Shen, H. Jiang, M. He, and G. Liu, "Collaborative representation-based classification of microarray gene expression data," *PLoS ONE*, vol. 12, no. 12, Dec. 2017, Art. no. e0189533, doi: 10.1371/journal.pone.0189533.

[31] B. Sahu, S. Dehuri, and A. Jagadev, "A study on the relevance of feature selection methods in microarray data," *Open Bioinf. J.*, vol. 11, no. 1, pp. 117–139, Jul. 2018.

[32] S. C. Hoi, J. Wang, P, Zhao, and R. Jin, "Online feature selection for mining big data," in *Proc. 1st Int. Workshop Big Data, Streams Heterogeneous Source Mining: Algorithms, Syst., Program. Models Appl.*, 2012, pp. 93–100.

[33] M. Takahashi, H. Hayashi, Y. Watanabe, K. Sawamura, N. Fukui, J. Watanabe, T. Kitajima, Y. Yamanouchi, N. Iwata, K. Mizukami, T. Hori, K. Shimoda, H. Ujike, N. Ozaki, K. Iijima, K. Takemura, H. Aoshima, and T. Someya, "Diagnostic classification of schizophrenia by neural network analysis of blood-based gene expression signatures," *Schizophrenia Res.*, vol. 119, nos. 1–3, pp. 210–218, Jun. 2010.

[34] H. Motieghader, A. Najafi, B. Sadeghi, and A. Masoudi-Nejad, "A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata," *Informat. Med. Unlocked*, vol. 9, pp. 246–254, Jan. 2017.

[35] L. Wang, Y. Wang, and Q. Chang, "Feature selection methods for big data bioinformatics: A survey from the search perspective," *Methods*, vol. 111, pp. 21–31, Dec. 2016.

[36] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied in microarray data," *Adv. Bioinf.*, vol. 2015, pp. 1–13, Jan. 2015, doi: 10.1155/2015/198363.

[37] I. Jolliffe, "Principal component analysis," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Germany: Springer, 2011, pp. 1094–1096.

[38] F. Buettner, V. Moignard, B. Göttgens, and F. J. Theis, "Probabilistic PCA of censored data: Accounting for uncertainties in the visualization of high-throughput single-cell qPCR data," *Bioinformatics*, vol. 30, no. 13, pp. 1867–1875, Jul. 2014.

[39] R. G. Thippa, K. R. M. Praveen, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020, doi: 10.1109/ACCESS.2017.

[40] F. S. G. Hashemi, M. R. Ismail, M. R. Yusop, M. S. G. Hashemi, M. H. N. Shahraki, H. Rastegari, G. Miah, and F. Aslani, "Intelligent mining of large-scale bio-data: Bioinformatics applications," *Biotechnol. Biotechnol. Equip.*, vol. 32, no. 1, pp. 10–29, Jan. 2018.

[41] B. Yang, F. Liu, C. Ren, Z. Ouyang, Z. Xie, X. Bo, and W. Shu, "BiRen: Predicting enhancers with a deep-learning-based model using the DNA sequence alone," *Bioinformatics*, vol. 33, no. 13, pp. 1930–1936, Jul. 2017.

[42] F. Tan, X. Fu, Y. Zhang, and A. G. Bourgeois, "A genetic algorithm-based method for feature subset selection," *Soft Comput.*, vol. 12, no. 2, pp. 111–120, Sep. 2007, doi: 10.1007/s00500-007-0193-8.

[43] M. Loey, M. W. Jasim, H. M. E-Bakry, M. H. N. Taha, and N. E. M. Khalifa, "Breast and colon cancer classification from gene expression profiles using data mining techniques," *Symmetry*, vol. 12, no. 3, p. 408, 2020, doi: 10.3390/sym12030408.

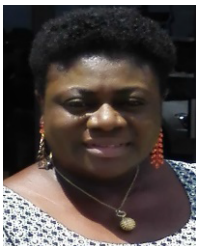
[44] A. Bhola and A. K. Tiwari, "Machine learning based approaches for cancer classification using gene expression data," *Mach. Learn. Appl., Int. J.*, vol. 2, pp. 1–12, Dec. 2015.

[45] L. Sun, X. Kong, J. Xu, Z. Xue, R. Zhai, and S. Zhang, "A hybrid gene selection method based on RefiefF and Ant colony optimization algorithm for tumor classification," *Nature Res. Acad.*, vol. 9, no. 8978, 2019.

[46] H. Alshamlan, G. Badr, and Y. Alohali, "MRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling," *BioMed Res. Int.*, vol. 2015, pp. 1–15, 2015, doi: 10.1155/2015/604910.

[47] E. Pamukçu, H. Bozdogan, and S. Çalık, "A novel hybrid dimension reduction technique for undersized high dimensional gene expression data sets using information complexity criterion for cancer classification," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–14, Jan. 2015.
[48] E. B. Huerta, B. Duval, and J.-K. Hao, "A hybrid LDA and genetic algorithm for gene selection and classification of microarray data," *Neurocomputing*, vol. 73, nos. 13–15, pp. 2375–2383, Aug. 2010.
[49] K. L. Lin, C. Y. Lin, C. D. Huang, H. M. Chang, C. Y. Yang, C. T. Lin, C. Y. Tang, and D. F. Hsu, "Feature selection and combination criteria for improving accuracy in protein structure prediction," *IEEE Trans. Nanobiosci.*, vol. 6, no. 2, pp. 186–196, Jun. 2007, doi: 10.1109/TNB.2007.897482.
[50] M. Veerabhadrappa and L. Rangarajan, "Bi-level dimensionality reduction methods using feature selection and feature extraction," *Int. J. Comput. Appl.*, vol. 4, no. 2, pp. 33–38, Jul. 2010.
[51] M. O. Arowolo, M. O. Adebiyi, and A. A. Adebiyi, "An efficient PCA Ensemble learning approach for prediction of RNA-Seq malaria vector gene expression data classification," *Int. J. Eng. Res. Technol.*, vol. 13, no. 1, pp. 163–169, 2020.
[52] [Online]. Available: https://figshare.com/articles/Additional_file_4_of_RNAseq_analyses_of_changes_in_the_Anopheles_gambiae_transcriptome_associated_with_resistance_to_pyrethroids_in_Kenya_identification_of_candidateresistance_genes_and_candidateresistance_SNPs/4346279/1
[53] C. Feng, S. Lu, H. Zhang, and X. Feng, "Dimension reduction and clustering models for single-cell RNA sequencing data: A comparative study," *Int. J. Mol. Sci.*, vol. 21, no. 6, p. 2181, 2020.
[54] S. J. Susmi and H. K. Nehimiah, "Hybrid dimensionality reduction techniques with genetic algorithm and neural network for classifying leukemia gene expression data," *Indian J. Sci. Technol.*, vol. 9, no. 1, pp. 1–8, Jan. 2016.
[55] G. Zararsiz and D. V. G. Eldem, "Comprehensive simulation study on classification of RNA-seq data," *PLoS ONE*, vol. 12, no. 8, 2017, Art. no. e0182507.
[56] M. Pyingkodi and R. Thangarajan, "Informative gene selection for cancer classification with microarray data using a metaheuristic framework," *Asian Pacific J. Cancer Prevention*, vol. 19, no. 2, pp. 560–564, 2018.
[57] A. Thomas, S. Barriere, L. Broseus, J. Brooke, and W. Ritchie, "GECKO is a genetic algorithms to classfiy gene and explore throughput seq data," *Commun. Biol.*, vol. 2, Jun. 2019, Art. no. 222.
[58] T. C. Nnodim, A. M. R. F. El-Bab, B. W. Ikua, and D. N. Sila, "Design and simulation of a tactile sensor for fruit ripeness detection," in *Proc. World Congr. Eng. Comput. Sci.*, 2019, pp. 390–395.
[59] T. C. Nnodim, A. M. R. F. El-Bab, B. W. Ikua, and D. N. Sila, "Estimation of the modulus of elasticity of mango for fruit sorting," *Int. J. Mech. Mechatron. Eng.*, vol. 19, no. 2, pp. 1–10, 2019.
[60] P. C. Okolie, E. C. Nwadike, J. L. Chukwuneke, and C. T. Nnodim, "The analysis of cigarate production using double exponential smoothing model," *Acad. J. Sci.*, vol. 7, no. 2, pp. 293–308, 2017.

MICHEAL O. AROWOLO received the bachelor's degree from Al-Hikmah University, Ilorin, Nigeria, and the master's degree from Kwara State University, Malete Nigeria. He is currently pursuing the Ph.D. degree with Landmark University, Omu-Aran Nigeria. He is also a Staff of the Department of Computer Science, Landmark University. He has published widely in local and international reputable journals. His research interests include machine learning, bioinformatics, data mining, cyber security and computer arithmetic. He is a member of IAENG, APISE, SDIWC, and an Oracle Certified Expert.



MARION OLUBUNMI ADEBIYI (Member, IEEE) received the B.Sc. degree from the University of Ilorin, Ilorin, Nigeria, and the M.Sc. and Ph.D. degrees in computer science from Covenant University, Nigeria. She is currently a Faculty Member of the Department of Computer Science, Landmark University, Omu-Aran, Nigeria. She has published widely in local and international reputable journals. Her research interests include bioinformatics of infectious (African) diseases/population, organism's inter-pathway analysis, high throughput data analytics, homology modeling, and artificial intelligence. She is a member of the Nigerian Computer Society (NCS) and the Computer Registration Council of Nigeria (CPN).

AYODELE ARIYO ADEBIYI (Member, IEEE) received the B.Sc. degree in computer science and the M.B.A. degree from the University of Ilorin, Ilorin, Nigeria, and the M.Sc. and Ph.D. degrees in management information system (MIS) from Covenant University, Nigeria. He is also a Faculty and a Former Head of the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. He is currently the Head of the Department of Computer Science, Landmark University, Omu-Aran, Nigeria, a sister University to Covenant University. He has successfully mentored and supervised several postgraduate students at master's and Ph.D. level. He has published widely in local and international reputable journals. His research interests include the application of soft computing techniques in solving real-life problems, software engineering, and information system research. He is a member of the Nigerian Computer Society (NCS) and the Computer Registration Council of Nigeria (CPN).

OLATUNJI JULIUS OKESOLA was the Group Head for Information Systems Control and Revenue Assurance with Keystone Bank (Nig.) Ltd., Lagos, in November 2016. He is currently a Professor of cybersecurity with First Technical University, Ibadan, Nigeria. He is a Certified Information Security Manager (CISM) and a Certified Information Systems Auditor (CISA) with a Ph.D. in computer sciences. He is a Scholar, an Information Security Expert, and a Seasoned Banker. He is an alumnus of the University of South Africa. He has several publications in scholarly journals and conference proceedings both local and international. His research interests include cybersecurity, biometrics, and software engineering. He is a member of the Information System Audit and Control Association (ISACA), the Computer Professionals of Nigeria (CPN), and a Fellow of the Nigerian Computer Society (NCS).