Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 2, April 2021, pp. 1071~1079
ISSN: 2302-9285, DOI: 10.11591/eei.v10i2.2769

A genetic algorithm for prediction of RNA-seq malaria vector gene expression data classification using SVM kernels

Marion, O. Adebiyi[1,2], Micheal, O. Arowolo[2], Oludayo Olugbara[3]
[1,3]Computer Science and Information Technology, Durban University of Technology, Durban 4001, South Africa
[2]Department of Computer Science, Landmark University, Omu-Aran, Kwara State Nigeria

Article Info

Article history:

Received Oct 13, 2020
Revised Jan 23, 2021
Accepted Feb 14, 2021

Keywords:

Genetic algorithm
Machine learning
Malaria
RNA-seq
SVMs

ABSTRACT

Malaria larvae embrace unpredictable variable life periods as they spread across many stratospheres of the mosquito vectors. There are transcriptomes of a thousand distinct species. Ribonucleic acid sequencing (RNA-seq) is a ubiquitous gene expression strategy that contributes to the improvement of genetic survey recognition. RNA-seq measures gene expression transcripts data, including methodological enhancements to machine learning procedures. Scientists have suggested many addressed learning for the study of biological evidence. An enhanced optimized Genetic Algorithm feature selection technique is used in this analysis to obtain relevant information from a high-dimensional *Anopheles gambiae* dataset and test its classification using SVM-Kernel algorithms. The efficacy of this assay is tested, and the outcome of the experiment obtained an accuracy metric of 93% and 96% respectively.

Corresponding Author:

Micheal, O. Arowolo
Department of Computer Science
Landmark University
Omu-Aran, Kwara State Nigeria
Email: arowolo.olaolu@lmu.edu.ng

1. INTRODUCTION

Next-generation high-throughput sequencing technology has created great wide-ranging data sets. This gigantic data expanse helps biologists to analyze and conduct daunting gene transcripts, such as disease-associated and RNA such as diseases (malaria), cancer, inherited, genomic, physiological, among others [1]. Blood-sucking mosquitoes such as mosquito anopheles with main *Plasmodium falciparum* malaria vectors are found mostly in Africa. Anopheles mosquito is a lethal malaria parasite which is responsible for thousands of deaths. When a fight against blowouts in antimalarial suppositories, state-of-the-art care for antimalarials improves, looking for groundbreaking drugs requires a greater knowledge of these species. Why anopheles mosquito parasite tolerates precise gene expression parameters, it has been a major interrogation with the need for an enhanced, thorough extrapolation model for its transcriptions [2, 3].

In the RNA-seq analysis, approachable disclosing genetic inquiries were made by designing a careful purposeful biological technique by improving the sequencing sample. RNA-seq data includes eliminating the high-dimensional curse, such as; sounds, illnesses, inconsistency, irrelevance, duplication, unfitting data, among others [4]. Advanced capabilities have strengthened solutions to the development of groundbreaking treatment frameworks such as effective public wellbeing nursing systems, advanced treatments, and other medical diagnosis and disorders [5]. In the last decade, numerous machine learning methods have been established with eloquent novelties to investigate the enormous volume of next-generation sequencing of RNA gene expression data analysis by learning the biologically applicable

backgrounds [6]. Quite a lot of scientists have used machine learning approaches with high-performance levels for the RNA-Seq gene expression data [7, 8]. The problem of curse dimensionality in high dimensional data has generated limitations in several traditional machines learning approaches, working with an efficient approach is of the essence.

Gene expression data needs significant developments for diagnosis, predictions of ailments and classifications, due to the fact that they are met with challenges such as irresistible numbers of genes relative to numbers of samples, comprising of irrelevant discrepancies. It is required to fetch optimal genes in the given data, to provide better accuracy and performance, previous works in literature have shown the importance of algorithms such as genetic algorithm, due to its strength of fetching range of medium relevant features, by identifying the small subset of genes to improve the gene analysis [5]. This study proposes an enhanced optimized Genetic algorithm approach to achieve the high dimensionality in the gene expression data, SVM kernel classification approaches are utilized to assess discrete genetic structures with classifications that may be suggested as useful techniques in the prediction and finding new genes for malaria infection.

The remainder of the study is organized as follows: Section 2 discusses the literature reviews. Section 3 discusses the research materials and method. Section 4 provides experimental research results. Section 5 discusses the conclusion of the study.

2. LITERATURE REVIEW

Computational approaches are effective on a large genomic dataset, and genes can be found which are responsible for the existence of ailments. Numerous methods are used to identify differentially expressed genes (DEG). Measures for machine learning (ML) are important in identifying differences between genes obtained from the human genome. Numerous approaches to machine learning are rivaled when analyzing and classifying patterns of gene expression from many diseases. The importance of unfolding gene expression data and its approaches was bestowed through several machine learning. Numerous academic findings in this area are being discussed. Current violations of work in analyzing gene expressions are known [5].

Oh *et al.* [9] suggested estimation of autism variation ailment with blood-based gene expression signs and machine learning to identify effective transcripts in classification. For machine learning algorithms, RNA information from the Gene expression compilation database is used on the R computer set. Rated cluster review found a fairly well-discriminated autism variation ailment from panels existed. SVM and KNN classifiers where data acceptance is used lead in a full classification class accuracy of 93.8%. Ren *et al.* [10] suggested a clustering and classification of RNA-seq utilizing a cumulative assessment, emphasizing the methods using clusters and classifier approaches as prevailing anomalies in recent times, non and linear scRNA-seq dimensional reduction approaches, incorporated and recorded scRNA-seq data.

Rating broad collections of genes calculated with RNA-seq focused on supervised learning methods for collecting RNA-seq genes was proposed, using variable range measurements generated through random forest classifier and defined extreme pseudo-sample channels with autoencoder variations and regressions extracting ranks from 12 RNA-seq cancer datasets containing about 1,200 samples. Results proved latent of the supervised learning-based selection of features in RNA-seq training and addressed the need for gene selection approaches to gene expression analysis [11].

A supervised approach to research was proposed [12], on RNA-seq data classification. By incorporating unbiased function collection from a simplified dimensional space inference method by introducing a generalizable method with a greatly detailed classification of single cells. They added scPred to the from mononucleate cells, pancreas tissue, biopsies of colorectal tumours and dendric cells that circulated. Proving scPred is highly effective in classifying different cells. A machine learning RNA-DNA analysis was proposed [13] specifying low gene expressed data that can mutually be inclined by PAH disease. We suggested a groundbreaking collection of features and advanced methods for classifying a trivial range of incredibly useful genes in machine learning algorithms. Studies revealed that clusters of genes with limited expression reveal modified types of PAH when forecasting and discriminating.

Characterization of data on gene expression using CNN for stomach cancer was proposed [14], they established a classification method focused on deep learning on patients with stomach cancer to demonstrate its application to data communication. PCA, heatmaps, and CNN algorithm were used to test 60000 genes of data from 300 patients. Researchers joined the scientific review of clinical evidence and RNA-seq gene analysis, and CNN to test these. They had 95.96% and 50.51% accuracy. RNA-seq discovery of secret transcripts in malaria parasites was proposed [15], by explaining the distinctions of RNA-seq technique to deconvolute transcript differences for approximately 500 different rodents and malaria parasites for human beings; they found distinct transcript signatures tucked inside.

An ensemble machine learning algorithm was proposed [16], to identify data on the expression of the cancer genes, based on the C4.5 decision tree, and improved ensemble decision tree classifiers supervised cancer classification methods, seven freely obtainable malignant microarray data relating to the classification methods and perform better than the independent decision trees. The design of a classification method for the gene expression of cancer information by the analytical ensemble was proposed [17] using combinatory recursive feature elimination was done through adaboost algorithm for appropriate classification features and reported changes. Tarek *et al.* [18] focused on classifying cancer for evidence on gene expression. We suggested an approach to the classification of the operational Ensemble that improves the introduction of the description and the poise of the performance. The results of the Ensemble are less dependent on the originalities of a particular range of instruction.

Duval and Hao [19], summarized current advances of metaheuristic-based approaches an embedded feature selection method, developed a metaheuristics method for selecting genes and classifying RNA/DNA data, highlighted the usefulness and importance of mixing problem-specific data into the search operators of such a process. The worked in what way linear classifier constants like SVM can be used lucratively in successful local experimentation for the collection and classification of elements. Shukla *et al.* [20] focused on a genetic algorithm-based hybrid system by implementing a groundbreaking hybrid feature selection algorithm using a filter-wrapper-based feature selection method to identify problems and resolve shortcomings of existing approaches. Five UCI biological datasets with several instances and dimensionality were proposed for the study. The findings demonstrate that the proposed method offers adequate support for major feature reduction and beats the state-of-the-art with the lowest classification accuracy of 40.04% and the highest precision of 99.32% using k-NN and SVM.

To improve tree model classification in selecting features of ensemble classifier, [21] employed an ensemble classification function collection with random trees and wrapper method. Future classification technique knowledge of an ensemble creates subdivision through the bagging, wrapper, and random tree methods. Potential strategy removes unnecessary features and uses a likelihood weighting method to select the best features for classification. Potential function selection method is tested using SVM, RF, and NB tests with its output correlating with the proposed techniques. The procedure reaches a ranking accuracy of 92%. Ching *et al.* [22] reported on the study of multiple function extractions gene expression analysis, such as the PCA, ICA, PLS, and LLE. Discussions and software purpose was discussed in the method.

3. MATERIALS AND METHODS

A lot of methods have been suggested in the literature for the investigation of high dimensional data. Genetic algorithm and SVM classification algorithm are considered in this analysis to minimize RNA-seq data tremendously in terms of dimensionality. Two thousand four hundred fifty-seven instances with seven gene attributes are used, data from western Kenya containing mosquito genes, comprising of significant genes of resistant and susceptible mosquitoes [4]. A descriptive overview of the dataset is shown in Table 1, where the dataset comprises of the attributes of the sample genes and the instances of the feature samples.

Table 1. Dataset

Dataset	Attributes	Instances
Mosquito *Anopheles gambiae*	7	2457

3.1. Methods

MATLAB was used to evaluate the data obtained from [19] as an experimental tool, and optimized GA was utilized to select features from the high dimensional data, to fetch a relevant subset of features. The fetched features were classified using the SVM classification [20]. Figure 1 shows the experimental workflow of this study. In this study the high dimensional data is passed into the genetic algorithm to fetch for a relevant subset of the data, the reduced data is sent to the SVM classifier for evaluating the performance of the experiment in terms of accuracy and other performance metrics.

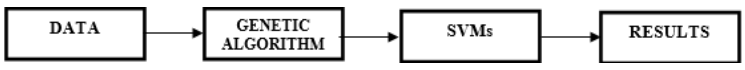

Figure 1. Experimental workflow

3.1.1. Genetic algorithm

Genetic algorithm (GA) is a versatile tool used to analyze the correct functionality of high-dimensional datasets. GA leading in function collection is wrapper-based approaches. There are several usages of parameters for genetic algorithms where mutation and crossover operatives are generally linked to the principles of binary parameters. Genetic algorithms are used to recognize appropriate features [21]. The RNA takes N numbers of features correspondingly representing structures with values 1 and 0 as picked and unselected. Addressing the value of functions, GA is utilized to consider the optimum subclass of features with the designated function number for dynamic presentation of classification. In Algorithm 1 below the general structure of the GA is defined by adopting [20]:

```
Algorithm 1. Genetic algorithm
Require: Initialize the parameters nPop = m, tmax, t = 0;
Check: Optimum feature subclass with the maximum fitness value.
while (t<=tmax) do
        Create pop m, tmax;
        For k = 1 to m do
                Parents [m1, m2] = system selection (m, nPop)
                Child = Xor [m1, m2]
                M u = mutation [Child}
        End for
        Replace m with Child1, Child2, …, Childm
        t = t+ 1;
        End while
        Store the Highest fitness value;
```

M is a population dimension, r is an arbitrary number flanked by 0 to 1, chromium corresponds to the designated or undesignated function by 0.5, and α is the maximum number of listed functions. The key problems of the particular method are the identification of the highest appropriate functionality from the known datasets. In this study, the genetic algorithm uses the 0.5 thresholds with an optimized iteration in the mutation ranging from 0 s to 1 s.

3.1.2. Support vector machine

Support vector machine (SVM) is a machine learning system which Vapnik presented in 1992 [23]. SVM works to find the best hyperplane in input space which isolates between groups. SVM is a linear classifier; it is generated by combining the kernel ideas into high-dimensional workspaces to deal with non-linear problems. For non-linear problems, SVM uses a kernel to train the data to spread the dimension narrowly. When tweaking the proportions, SVM should search for the ideal hyperplane, which can distinguish a class from other classes [23]. The method for finding the strongest hyperplane using SVM, as shown by the adoption of Aydadenta and Adiwijaya [23]:

i. Let $y_i \in \{y_1, y_2, \ldots, y_n\}$, where y_i are the p-attributes and target class $z_i \in \{+1, -1\}$
ii. Assuming the classes +1 and -1 divided totally by a hyperplane, as defined in (2) and (3):

$$v.y+c=0 \quad (1)$$

From (1), can get (2) and (3):

$$v.y + c \geq +1, \text{ for class } +1 \quad (2)$$

$$v.b + c \leq -1, \text{ for class } -1 \quad (3)$$

SVM is a machine learning system which Vapnik proposed in 1992 [23]. SVM works to find the best hyperplane in the input space which isolates between groups. SVM is a linear classifier; it is generated by combining the kernel ideas in high-dimensional workspaces to deal with non-linear problems. For non-linear problems, SVM uses a kernel to train the data to spread the dimension narrowly. If tweaking the proportions, SVM can look for the ideal hyperplane and can distinguish a class from other classes [23]. The technique to find the best hyperplane using SVM, as shown by the adoption of Aydadenta and Adiwijaya [23]:

– SVM-Gaussian kernel

Gaussian kernel [24] is related to a general supposition of smoothness in all subordinates of the kth order. Kernels that manage a certain prior data recurrence material can be built to represent earlier learning problems. Every input vector x is translated to an infinite-dimensional vector with all the polynomial extensions of the x components [25-27].

3.2. Performance evaluation

Assessing machine-learning algorithm efficiency needs certain validation metrics. The uncertainty matrix is often used to evaluate four characteristics of classification models; true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Determining the examples categorized appropriately and inaccurately from the data set sample specified to check the model [5]. Performance metrics are presented below with its formula [26].

3.3. Applications

Analysis of gene expression provides an enhanced route for the detection of RNA-seq results. The necessity to explore specific genes is beneficial in creating various applications such as modified treatment, cancer detection, gene and drug development, tumour recognition, illnesses such as malaria and typhoid. Machine learning knowledge in discovering designs and data inconsistency, it holds excellent procedures as instruments that apply to diverse areas.

Program development simplicity for designers, physicists, academics, among others, matrix laboratory (MATLAB) is used to experiment. MATLAB is an arithmetical processing environment with multi-worldview and a limited programming language documented by MathWorks. It allows application controls, tasks and knowledge visualization, algorithm execution, user interface development, written in C, C++, C #, Fortran, Java, and Python languages [16]. The key idea of this analysis is to predict Malaria infection, using the RNA-seq data technology on the MATLAB method. The computer conformation used as the executing tool for determining this study is iCore2 processor, 64-bit System, 4 GB RAM size, and MATLAB 2015a.

4. RESULTS AND ANALYSIS

This research explores the RNA-seq innovation of vulnerable and tolerant genes, carrying 2457 instances of *Anopheles gambiae* mosquitoes. Optimized genetic algorithm to diminish the burden of dimensionality was applied to the results. GA selection feature dimensionality reduction captures the optimal data sub-set and eliminates uncorrelated attributes to determine the maximum variance with a lesser number of mutable subset features. GA is optimized and used on the Anopheles mosquito data in this analysis, which offers important gene detail that is valuable for further research. MATLAB tool uses the SVM classification kernel algorithms to execute the pattern. With 0.5 thresholds, 708 significant optimal subset features of genes were using optimized GA as a feature selection method.

SVM classification algorithms, 10-fold cross-validation and 0.05 parameter holdout were utilized to evaluate the performance implementation of classification models, and training data uses 75% and 25% testing to verify the classification accuracy. The classifier uses a learning valuation procedure for sampling bias eradication, by training and testing estimated segments. Using MATLAB, this procedure is implemented. The measurement outcome described is based on the quantitative time and efficiency parameters (accuracy, specificity, sensitivity, precision, f-score and recall) [26]. This analysis measures the model classification efficiency with 93.3% and 95% accuracy, respectively, using L-SVM and RBF-SVM classifiers. The result performance and the matrix for uncertainty are revealed in Figure 2.

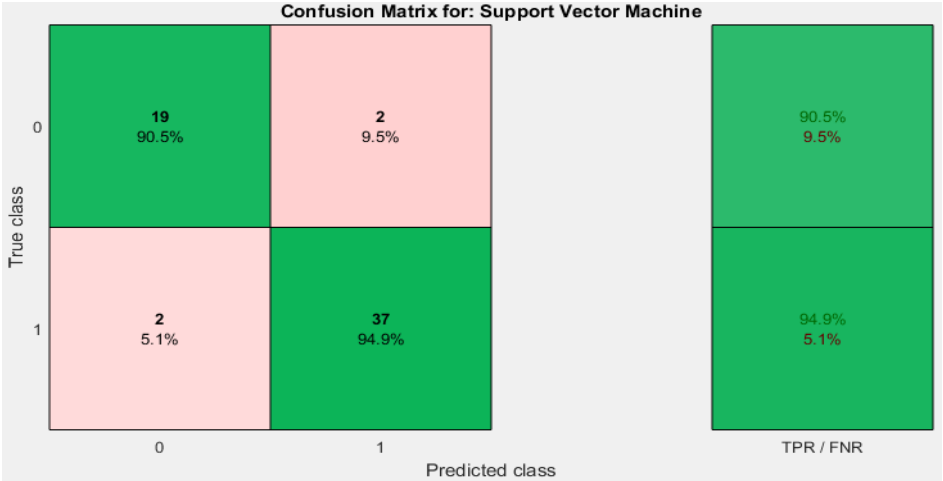

Figure 2. Confusion matrix for the classification data using L-SVM, TP=37; TN=19; FP=2; FN=2

This learning uses GA to gather relevant components from the loaded data. The chosen features are passed to the SVM classification, and the outcome is seen in the following Figures 2 and 3. The uncertainty matrix gives quality metrics a solution. The L-SVM classification kernel analysis achieved 93.3% accurate, the RBF-SVM kernel classification system is 95% accurate; other efficiency metrics are shown in tabulated form in Table 2.

In this study, the classification of the experiment was performed. It yielded the confusion matrices used to calculate the evaluation performances, Figure 2 and Figure 3 shows the confusion matrices that depicts the True positive. This outcome shows the correctly predicted positive classes: the true negative shows the outcomes that are correctly predicted but negative class. The false-positive shows the outcome of the model that is incorrectly predicted with negative classes. In comparison, the false-negative shows the outcomes of the model that are incorrectly predicted with negative classes.

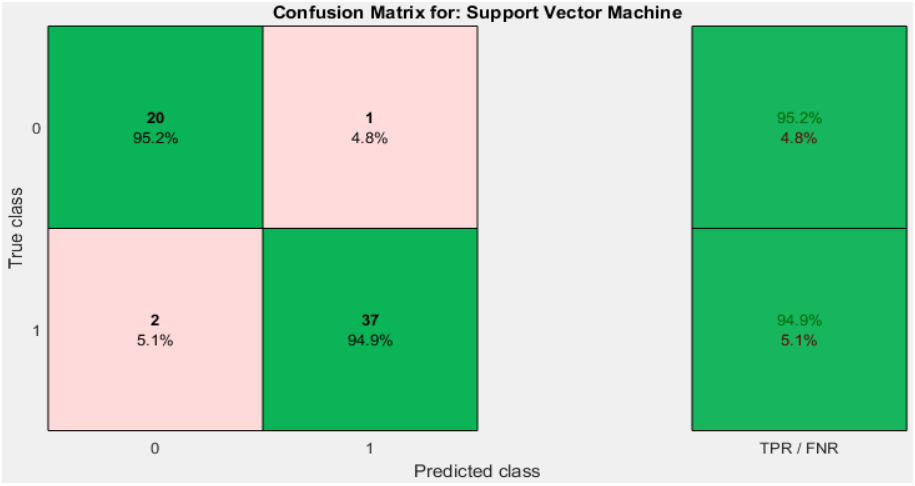

Figure 3. Confusion matrix for the classification data using RBF-SVM, TP=37; TN=20; FP=1; FN=2

RNA-seq results for *Anopheles gambiae* mosquito [28]. Two thousand four hundred fifty-seven gene features were obtained, GA was utilized as a guide for the lessening of dimensionality, 708 features were chosen as a subset of the results. Then these components are categorized using the SVM classification to forecast their performance. The outcome demonstrates the machine-learning technology's success in embryos. The success findings are shown and compared in Table 2 below for confirmation of the method. The analysis reveals that RBF-SVM outperforms L-SVM in terms of less training time and output accuracy. Table 2 shows the comparative analysis of the classification results of the proposed experiment using two types of SVM kernels which are the linear and radial basis function SVMs.

Table 2. The performance metrics

Performance metrics	GA-L-SVM classification	GA-RBF-SVM classification
Accuracy (%)	93.3	95.0
Sensitivity (%)	94.9	94.9
Specificity (%)	90.5	95.2
Precision (%)	94.9	97.4
Recall (%)	94.9	94.9
F-Score (%)	95.0	96.13

The output of the processed malaria vector data using the proposed model is evaluated and validated by clinicians. The results have to be validated for observations and automated machine learning-based approach in terms of the percentage of gene characterizations. It is evident that one of the major causes of the poor performance of classification is either overfitting or underfitting the data, this study trained and tested the reduced data with an approximate target function, in order to positively impact the performance of the model, by fetching out noises from the model. Proper validation is of the essence before clinical usage and testing, in order to provide more accurate services to clinicians and patients. Machine learning procedures

have proven to provide an accurate percentage of genes when compared with other methods and can be monitored for accurate prediction. This study is a better approach than traditional ways of determining observations and can provide a better assessment for malaria infection and transmissions in human. Table 3 shows the comparison with other state-of-the-art results. This study will help clinicians in decision making of prediction, detection and designs of efficient drugs as well as better ways of eradicating malaria infections in Africa. This work is limited to malaria infections and its computational analysis for clinicians, which can be extended in future to other ailments and introducing other approaches. Table 3 shows the comparison of this study with other techniques in literature.

Table 3. Comparative approaches

Methods	Accuracy (%)
PSO+SVM [29]	89
Mutual information+KNN [30]	95
GA+MLP [31]	89
RF [32]	94
Bayesian [33]	91

5. CONCLUSION

This study can be useful in human malaria ailment prognosis and diagnosis. The theoretical solution uses machine learning methods such as model and classification procedures for the reduction of dimensionality. Dimensionality reduction approach follows the GA filtering function model, which uses the SVM classification. This study carried out the success analysis and assessment and showed the findings obtained the SVM classification algorithm. This study evaluated and enhanced the classification of malaria vector data. Multiple studies have suggested evaluations by investigators using performance metrics, the findings have shown that dimensionality reduction model utilizing feature extraction methods such as GA can boost classification efficiency such as SVM. It will be important to explore how the recently proposed research can strengthen the feature selection models and algorithms. Future work proposes to use hybridized dimensionality reduction approaches. In future, this study can optimize the genetic algorithm for better fitness iteration and integrating the approach with other dimensionality reduction methods such as the ant colony optimizer, as well as introducing other beneficial classifiers such as the KNN, then compare and fetch for better efficiency of classification of the genes.

REFERENCES

[1] Sun S., Wang C., Ding H., and Zou Q., "Machine learning and its applications in plant molecular studies. briefings in functional genomics Oxford academic," *Briefings in Functional Genomicsvol*, 19, no. 1, pp. 40-48. 2019, doi:10.1093/bfgp/elz036.

[2] D. F. Read, K. Cook, Y. Y. Lu, K. G. Le Roch, and W. S. Noble, "Predicting gene expression in the human malaria parasite *Plasmodium falciparum* using histone modification, nucleosome positioning, and 3D localization features," *PLOS Computational Biology*, vol. 15, no. 9, p. e1007329, 2019, doi.org/10.1371/journal.pcbi.1007329.

[3] Anopheles gambiae 1000 Genomes Consortium, "Genetic diversity of the African malaria vector *Anopheles gambiae*," *Nature*, vol. 552, no. 7683, pp. 96-100, 2017, doi:10.1038/nature24995.

[4] M. O. Arowolo, M. Adebiyi, and A. A. Adebiyi, "A dimensional reduced model for the classification of RNA-seq *Anopheles gambiae* data," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 23, pp. 3487-3496, 2019.

[5] S. Karthik and M. Sudha, "A survey on machine learning approaches in gene expression classification in modelling computational diagnostic system for complex diseases," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 2, pp. 182-191, 2018.

[6] 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, 2018, doi:10.1261/rna.062802.117.

[7] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321-332, 2015.

[8] Z. Jagga and D. Gupta, "Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms," *BMC Proceedings*, vol. 8, no. S6, pp. 1-7, 2014.

[9] D. H. Oh, I. B. Kim, S. H. Kim, and D. H. Ahn, "Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning," *Clin Psychopharmacology Neuroscience*, vol. 15, no. 1, pp. 47-52, 2017, doi:10.9758/cpn.2017.15.1.47.

[10] Q. Ren, M. Anjun, Q. Ma, and Q. Zou, "Clustering and classification methods for single-cell RNA-seq data," *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1196-1208, 2020.

[11] S. Wenric and R. Shemirani, "Using supervised learning methods for gene selection in RNA-seq case-control studies," *Frontiers in Genetic*, vol. 9, p. 297, 2018, doi.org/10.3389/fgene.2018.00297.
[12] J. Alquicira-Hernandez, A. Sathe, H. P. Ji, Q. Nquyen, and J. E. Powell, "scPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data," *Genome Biology*, vol. 20, no. 1, pp. 1-17, 2019, doi:10.1186/s13059-019-1862-5.
[13] S. Cui, Q. Wu, J. West, and J. Bai, "Machine learning-based microarray analyses indicate low-expression genes might collectively influence PAH disease," *PLOS Computational Biology*, vol. 15, no. 8, p. e1007264, 2019, doi: org/10.1371/journal.pcbi.1007264.
[14] H. S. Shon, Y. G. Yi, K. O. Kim, E. J. Cha, and K. A. Kim, "Classification of stomach canacer gene expression data using CNN algorithm of deep learning," *Journal of Biomedical Translation Research*, vol. 20, no. 1, pp.15-20, 2019, doi: org/10.12729/jbtr.2019.20.1.015.
[15] A. J. Reid, A. M. Talman, H. M. Bennett, A. R. Gomes, M. J. Sanders, C. J. R. Illingworth, O. Billker, M. Berriman, and M. K. N. Lawniczak, "Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites," *Elife*, vol. 7, p. e33105, 2018, doi:10.7554/eLife.33105.
[16] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Applied Bioinformatics*, vol. 2, no. 3, pp. 75-83, 2003.
[17] N. Song, K. Wang, M. Xu, X. Xie, G. Chen, and Y. Wang, "Design and analysis of ensemble classifier for gene expression data of cancer," *Advancement in Genetic Engineering*, vol. 5, no. 1, pp. 1-7, 2016, doi:10.4172/2169-0111.1000152.
[18] S. Tarek, R. A. Elwahab, and M. Shoman, "Gene expression based cancer classification," *Egyptian Informatics Journal*, vol. 18, no. 3, pp. 151-159, 2017, doi: 10.1016/j.eij.2016.12.001.
[19] B. Duval and J-K. Hao, "Advances in metaheuristics for gene selectio and classification of microarray data," *Briefings in Bioinformatics*, vol. 11, no. 1, pp. 127-141, 2010.
[20] A. K. Shukla, P. Singh, and M. Vardhan, "A new hybrid feature subset selection framework based on binary genetic algorithm and information theory," *International Journal of Computational Intelligence and Applications*, vol. 18, no. 3, p. 1950020, 2019.
[21] M. O. Arowolo, M. Adebiyi, A. Adebiyi, and O. Okesola, "PCA Model For RNA-Seq Malaria Vector Data Classification Using KNN And Decision Tree Algorithm," *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS) IEEE*, pp. 1-8, 2020. 0.1109/ICMCECS47690.2020.240881
[22] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. E. Barnes, and D. E. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.
[23] H. Aydadenta and Adiwijaya, "On the classification techniques in data mining for microarray data classification," *Journal of Physics Conference Series*, vol. 971, no. 1, p. 012004, 2018, doi: 10.1088/1742-6596/971/1/012004.
[24] D. A. Vanitha C., Devaraj D., and Venkatesulu M., "Gene expression data classification using support vector machine and mutual information-based gene selection," *Procedia Computer Science*, vol. 47, pp. 13-21, 2015.
[25] 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 and Vector*, vol. 8, no. 1, pp. 1-13, 2015, doi: https://doi.org/10.1186/s13071-015-1083-z.
[26] M. O. Arowolo, S. O. Abdulsalam, R. M. Isiaka, and K. A. Gbolagade, "A comparative analysis of feature selection and feature extraction models for classifying microarray dataset," *Computing and Information System*, vol. 22, no. 2, pp. 29-38, 2018.
[27] M. O. Arowolo, S. O. Abdulsalam, R. M. Isiaka, and K. A. Gbolagade, "A hybrid dimensionality reduction model for classification of microarray dataset," *International Journal of Information Technology and Computer Science*, vol. 9, no. 11, pp. 57-63, 2017.
[28] M. O. Arowolo, M. O. Adebiyi, A.A. Adebiyi, and O. J. Okesola, "A Hybrid Heuristic Dimensionality Reduction Methods for Classifying Malaria Vector Gene Expression Data," *IEEE Access*, vol. 8, pp. 182422 -182430, 2020, 10.1109/ACCESS.2020.3029234.
[29] A. K. Shukla, "Identification of cancerous gene groups from microarray data by employing adaptive genetic and support vector machine technique," *Computational Intelligence*, vol. 36, no. 1, pp. 102-131, 2019.
[30] 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," *International Journal of Engineering Research and Technology*, vol. 13, no. 1, pp. 163, 2020, doi: 10.37624/ijert/13.1.2020.163-169 .
[31] S. Karthik and M. Sudha, "A survey on machine learning approaches in gene expression classification in modelling computational diagnostic system for complex diseases," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 2, pp. 182-191, 2017.
[32] C. Chakraborty, "Computational approach for chronic wound tissue characterization," *Informatics in Medicine Unlocke*, vol. 17, p. 100162, 2019.
[33] C. Chakraborty, B. Gupta, and S. K. Ghosh, "Chronic wound characterization using bayesian classifier under telemedicine framework," *International Journal of E-Health and Medical Communications*, vol. 7, no. 1, pp. 78-96, 2016.

BIOGRAPHIES OF AUTHORS

Dr Marion Olubunmi Adebiyi, is a faculty of the Department of Computer Science at Landmark University, Omu-Aran, Nigeria. She holds a BSc Degree from University of Ilorin, Ilorin Nigeria. She had her MSc and PhD Degree in Computer Science from Covenant University, Nigeria respectively. Her research interests include Bioinformatics of Infectious (African) Diseases/ Population, Organism's Inter-pathway analysis, High throughput data analytics, Homology modelling and Artificial Intelligence. She has published widely in local and international reputable journals. She is a member of the Nigerian Computer Society (NCS), the Computer Registration Council of Nigeria (CPN) and IEEE member.

Arowolo Micheal Olaolu is a Lecturer at the Department of Computer Science, Landmark University, Omu-Aran Nigeria. He holds his Bachelor Degree from Al-Hikmah University, Ilorin, Nigeria and his Master's Degree from Kwara State University, Malete Nigeria. He is presently a PhD Student. His area of research interest includes Machine Learning, Bioinformatics, Datamining, Artificial Intelligence, Cyber Security and Computer Arithmetic. He has published widely in local and international reputable journals. He is a member of IAENG, APISE, SDIWC, and an Oracle Certified Expert.

Prof Oludayo Olugbara graduated with a first-class Bachelor of Science (Hons) in Mathematics from the University of Ilorin in 1991, he was a junior research fellow in at the University of Ilorin, after completing the national youth service corps. In 1993 he commenced his Master's Degree in Mathematics with specialization in Computer Science at the University of Ilorin and completed the degree in 1995. He holds a PhD degree in Computer Science from the University of Zululand in South Africa. He is a Professor of Information Technology at the Durban University of Technology in South Africa. He is a holder of academic awards and scholarships, including the International Federation of Information Processing (IFIP) TC2 sponsored by Microsoft Research Cambridge in 2007 and respected research paper award at International Conference on Machine Learning and Data Analysis, organized by the IAENG International Association of Engineers, San Francisco, the USA in 2012. He is a University Scholar at the University of Ilorin, Member of Marquis Whos' Who in the World (USA), Member of the Association for Computing Machinery (ACM, USA), Member of Computer Society of South Africa (CSSA) and other academic associations. He was awarded honorary referee of the Maejo International Journal of Science and Technology, Thailand in 2007-2010 and 2011. In December 2015, He was awarded an outstanding scientist by the Center for Advanced Research and Design of Venus International Foundation in India. He became an established researcher courtesy of the National Research Foundation (NRF) of South Africa rating in 2017. He has examined several postgraduate theses, dissertations and assessed research publications for professorial appointments both nationally and internationally. He has published widely, and he is a reviewer for many reputable journals. oludayoo@dut.ac.za.