TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 19, No. 1, February 2021, pp. 310~316
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v19i1.16381

A genetic algorithm approach for predicting ribonucleic acid sequencing data classification using KNN and decision tree

Micheal Olaolu Arowolo, Marion Olubunmi Adebiyi, Ayodele Ariyo Adebiyi
Department of Computer Science, Landmark University, Omu-Aran, Kwara State, Nigeria

Article Info

Article history:

Received Apr 14, 2020
Revised Jun 9, 2020
Accepted Sep 24, 2020

Keywords:

Decision tree
Genetic algorithm
KNN
Mosquito anopheles
Ribonucleic acid sequencing

ABSTRACT

Malaria larvae accept explosive variable lifecycle as they spread across numerous mosquito vector stratosphere. Transcriptomes arise in thousands of diverse parasites. Ribonucleic acid sequencing (RNA-seq) is a prevalent gene expression that has led to enhanced understanding of genetic queries. RNA-seq tests transcript of gene expression, and provides methodological enhancements to machine learning procedures. Researchers have proposed several methods in evaluating and learning biological data. Genetic algorithm (GA) as a feature selection process is used in this study to fetch relevant information from the RNA-Seq Mosquito Anopheles gambiae malaria vector dataset, and evaluates the results using kth nearest neighbor (KNN) and decision tree classification algorithms. The experimental results obtained a classification accuracy of 88.3 and 98.3 percents respectively.

Corresponding Author:

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

1. INTRODUCTION

Next-generation high-throughput sequencing technology has created profuse wide-ranging datasets, this enormous data expanse helps biologists to analyze and perform daunting gene transcripts, such as disease related and RNA such as infections (malaria), cancer, inherited, genetics, physiology, among others [1]. Blood-sucking mosquitoes such as Mosquito Anopheles with vectors of malaria plasmodium falciparum are found in Africa. Mosquito Anopheles is a deadly malaria parasite, responsible for demises of thousands of humans daily. Antimalaria combat suppositories blowouts, state-of-the-art antimalarials treatment upsurges, fetching for ground-breaking medications requires improved biotic studies of this infenctions. The parasite tolerates precise parameter of gene expression query enormously and necessitates making enhanced thorough extrapolative model transcriptions of vectors [2].

Approachable revealing genetic inquiries have been made in ribonucleic acid sequencing (RNA-seq) study by unfolding a cautious purposeful biological strategy for enhancement of the learning. RNA-Seq data requires removal of expletive high-dimension, such as; noises, complaints, repetition, irrelevant, inactivity, unfitting data, and others [3]. New capabilities strengthen solutions to the development of ground-breaking healthcare frameworks such as effective public wellbeing nursing systems, advanced interventions and medical diagnosis and disorders [4].

Machine learning means have been established with convincing uniqueness to investigate the enormous amount of cutting-edge RNA-Seq knowledge by studying the naturally material structures [5]. Scientists have used machine learning algorithms with relevant achievement for gene expression data results

of RNA-Seq [6-8]. In this study, a genetic algorithm (GA) pre-processor, to obtain reduced dimensionality of data with kth nearest neighbor (KNN) and decision tree classifiers are proposed to classify discrete genetic structures and obtain advances that are suitable system for predicting and detecting innovative genes for malaria ailments in human.

2. REVIEWS

Computational procedures are based on enormous samples of individuals genes with or without diseases, mutations may be found accountable for the precense of diseases. Differential expressed genes (DEG) are defined through some methods. Machine Learning measures are important for spotting the variation between genes found from human genome. Machine learning techniques have been emulated severally in investigating and classifying various profiles of diseases gene expressions. Various machine learning approaches are reported and reviewed, using recent trends in the evaluations [4].

Machine learning for predicting Autism spectrum ailment was experimented and classify transcripts, using RNA data from gene omnibus expression data. This study ranked cluster analysis and relatively discriminated, using SVM and KNN classifiers, an estimate accuracy of 94% was achieved [9]. Clustering and classification of RNA-Seq data was carried out by performing a mutual valuation, and emphasizing the expertise and ploys of methods occurring in recent time as predominant shifts, uing nonlinear and linear dimension reduction systems, by combining scRNA-seq data [10]. Group of RNA-Seq genes for ranking genes set of huge ensembles using a supervised learning approach was carried out using random forests classification method, on 1210 samples of tumor RNA-Seq datasets showed hidden supervised learning selection approaches necessity on analysis [11]. A supervised single-cell RNA-Seq data classification model was proposed using a comprehensive approach by combining independent feature selection approaches. scPred RNA-seq datasets showed high accuracy [12]. RNA-DNA machine learning analysis was proposed to indicate small genome expression to influence PAH ailment, feature selection algorithm was proposed to classify relevant genes with an outcome that reveals unique PAH [13].

Stomach tumor gene expression data using CNN classification procedure was developed based on deep learning approach, 60,000 data made up of stomach tumor genes were evaluated using PCA), heatmaps, and CNN algorithms with an accuracy of 96% and 51% [14]. RNA-Seq hidden transcripts in malaria parasites was proposed by relating variations of procedures to deconvolute transcriptional differences for distinct mosquitos and revealed hidden distinct transcriptional signatures [15].

An ensemble classification algorithm for cancer dataset was developed using decision tree, ensemble decision trees algorithms on available cancerous microarray, the results enhances than the decision trees classification [16]. An investigative cancer gene expression ensemble classification method was proposed using a hybrid RFE-Adaboost algorithm to fetch significant features for enhancing classification performancet [17]. Classification of cancer data was carried out using an effective ensemble classification method by increasing the classification, the result were less contingent [18]. A metaheuristics system for fetching relevant RNA/DNA data genes for classification was proposed by briefing recent developments of metaheuristic-based methods in embedded feature selection methods, useful data for operatives for ranking coefficients of SVM classifier is used [19]. A GA presenting a state-of-the-art approach was proposed using filter-wrapper based feature selection on five biological datasets, the results showed an important reduction of features for classification [20]. An enhanced ensemble classification for certain features was proposed for learning an ensemble-based feature selection approach with random trees using a subset, the method removes the unfitting structures and picks the best structures by means of a probability weighing value for classification evaluation using RF, SVM, and NB [21]. Review of several feature extraction algorithms for gene expression investigation, such as the PCA, ICA, PLS, and LLE was carried out and discussed for the purpose of the machine learning applications [22].

3. MATERIALS AND METHODS

Several high-dimensional data enhancement methods are in place, this paper carries out a feature selection using GA technique and ensemble classification algorithm for fetching relevant information in a huge dimensional data and classification. A western Kenya RNA-Seq data mosquitos' genes with 2457 instances with 7 gene attributes [23], MATLAB environment tool is used carry out the experiment using GA to select relevant subset of features from the dataset as shown in Table 1, Ensemble algorithm approach is used as a classifier on the selected features [24].

Table 1. Features of the data

Dataset	Attributes	Instances
Mosquito Anopheles Gambiae	7	2457

3.1. GA

GA is a wrapper-based feature selection approach that examines suitable features from a given high dimensional datasets, with numerous parameter procedures, where mutation and crossover operators are associated with relevant recognized binary constraints features [19]. RNA features with *N* number denotes a feature having selected and unselected values 1 and 0 respectively. GA is important and helps to find feature subset models with selected figure features for composite classifications. GA structure is adopted and well-defined in Algorithm 1 [20]:

```
Algorithm 1. Genetic Algorithm
Necessitate: Set parameters nPop = m, t_max, t = 0;
Confirm: Optimal feature subset with the maximum suitable value.
        1: while (t<=t_max) do
        2:      Create pop m, t_max;
        3:      For k = 1 to m do
        4:              Parents [m1, m2] = system selection (m, nPop)
        5:              Child = Xor [m1, m2]
        6:              M u = mutation [Child}
        7:      End for
        8:      Replace m with Child1, Child2, …, Childm
        9:      t = t+ 1;
        10:     End while
        11:     Store the Highest fitness value;
```

m = population size, *r* = random number 0 to 1, chrome = certain or non-certain feature through threshold δ, set value = 0.5, and α = threshold number of picked features. Selecting maximum fit features from the predictable datasets is the main problem of the GA technique.

3.2. KNN

A supervised learning K-nearest neighbor classification technique for gene datasets, performs neighborhood classification evaluation value of innovative application occurrence. KNN algorithm classifies innovative entity developed on instances, attributes as well as training models. KNN classifiers do not train models to fit but built on retention. The features selected are assumed as input to segments. The K value of nearest neighbors are selected nearest to the query spot. Detachment between query-instance and training models are considered and sorted based on the K^{th} minimum determined distance. Group Y of the nearest neighbors is fetched. The unassuming common of the group of nearest neighbors as the estimate amount of the query instance is used. Bonds can fragment randomly [25].

3.3. Decision trees

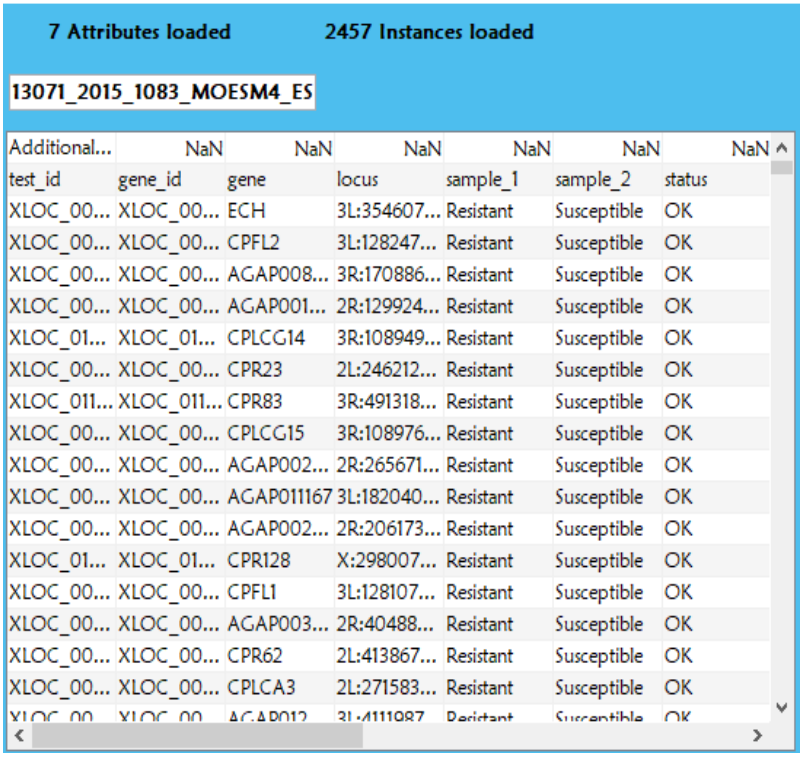
Decision tree classification algorithm divides recursively instance spacing with hyperplanes orthogonally. Decision tree model assembles derivative nodes signifying attributes, based on instance space attribute value roles selected inversely for algorithms, using its values. Advanced data sub-space iteratively divides till end principle is determined and terminal nodes (leaf nodes) are allocated to class labels characterizing the classification. Accurate conventional end procedure is a significant tree with too huge, overfitted and trivial trees, underfitted and suffers loss in accuracy. Algorithms have assembled overfitting strategies, labelled trimming, classifying new instances by leading the tree basis down a leaf, with respect to the examination result along the pathway [26]. Competent models are discovered using decision tree classifiers and ensembles, with unbalanced varying trained datasets, with resultant models totally unalike.

3.4. Performance evaluation and applications

Machine learning model need evaluation and validation of performance metrics using a confusion matrix and its formula [4, 27]. Expression of gene analysis suggest enhanced RNA-Seq data path identification, to learn applicable helpful genes in advancing applications such as treatment modifications, diseases diagnosis, drugs and gene discoveries, classification of cancers, typhoid, malaria, among other ailments. Designs and inconsistency findings between machine learning data has discovered great algorithms applicable to many fields such as engineering, banking, health sectors among others. MATLAB 2015A is proposed as an experimental and executing tool for the prognosis of malaria infections on an iCore2 processor, 4GB RAM size, 64-bit System.

4. RESULTS AND ANALYSIS

In this study, 2457 instances of RNA-Seq dataset "Mosquitoes Anopheles Gambiae" containing resistants and susceptibles of genes is used on a GA to draw optimal reduced number of subsets in the data, taking away uncorrelated attributes to pick maximum variance features. The result shows important gene evidence suitable for KNN and decision tree classification algorithm study on MATLAB environment for the model experiment. Genetic algorithm makes use of 0.5 threshold and achieves 708 optimal subset features of significant genes. Classifiers used 10-folds cross validation was used on KNN and decision tree classifiers, to implement evaluations of the model's performance with 0.05 holdout training data parameter and classifier accuracy tests the data with 25%. A learning classification procedure evaluation train and test evaluates using 10-fold cross validation to remove the partiality in sampling. The performance metrics and time computation is evaluated [27] and relates the model classification performance, by means of KNN (bagging) and decision tree, with 98.3% and 88.3% accuracy separately using confusion matrix and result outpus as shown in Figure 1. Related components were fetched by GA from the full data shown in Figure 1, the subset data features pass into KNN as well as decision tree and shows the Confusion matix result in the Figure 2 and Figure 3 to derive the solution to the performance metrics. KNN classification algorithm achieves an accuracy of 88.3%, while the decision tree classification algorithm achieved an accuracy of 98.3%, metrics of other performance are shown in Table 1.



7 Attributes loaded 2457 Instances loaded

13071_2015_1083_MOESM4_ES

Additional...	NaN	NaN	NaN	NaN	NaN	NaN
test_id	gene_id	gene	locus	sample_1	sample_2	status
XLOC_00...	XLOC_00...	ECH	3L:354607...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	CPFL2	3L:128247...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	AGAP008...	3R:170886...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	AGAP001...	2R:129924...	Resistant	Susceptible	OK
XLOC_01...	XLOC_01...	CPLCG14	3R:108949...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	CPR23	2L:246212...	Resistant	Susceptible	OK
XLOC_011...	XLOC_011...	CPR83	3R:491318...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	CPLCG15	3R:108976...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	AGAP002...	2R:265671...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	AGAP011167	3L:182040...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	AGAP002...	2R:206173...	Resistant	Susceptible	OK
XLOC_01...	XLOC_01...	CPR128	X:298007...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	CPFL1	3L:128107...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	AGAP003...	2R:40488...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	CPR62	2L:413867...	Resistant	Susceptible	OK
XLOC_00...	XLOC_00...	CPLCA3	2L:271583...	Resistant	Susceptible	OK
XLOC_00	XLOC_00	AGAP012	3L:4111987	Resistant	Susceptible	OK

Figure 1. Loaded mosquito anopheles gambiae on MATLAB environment

In this study, RNA-Seq data uses a mosquito anopheles gambiae dataset [28], to test the machine learning method performance. Genetic algorithm dimensionality reduction model selects 708 subset features from 2457 features of genes form the data. The selected components were classified using classification algorithms (KNN and decision tree) performance evaluation. The efficiency of machine learning approach in genes are shown in the results to confirm the method, the outcomes are revealed and related in Table 2 showing GA-decision tree outperforms GA-KNN terms of accuracy. In this study, an improved classification of malaria vector data is analyzed using GA with decision tree and KNN algorithms respectivel, numerous works have been reviewed, the results prove that GA enhances classification yield for KNN and decision tree.

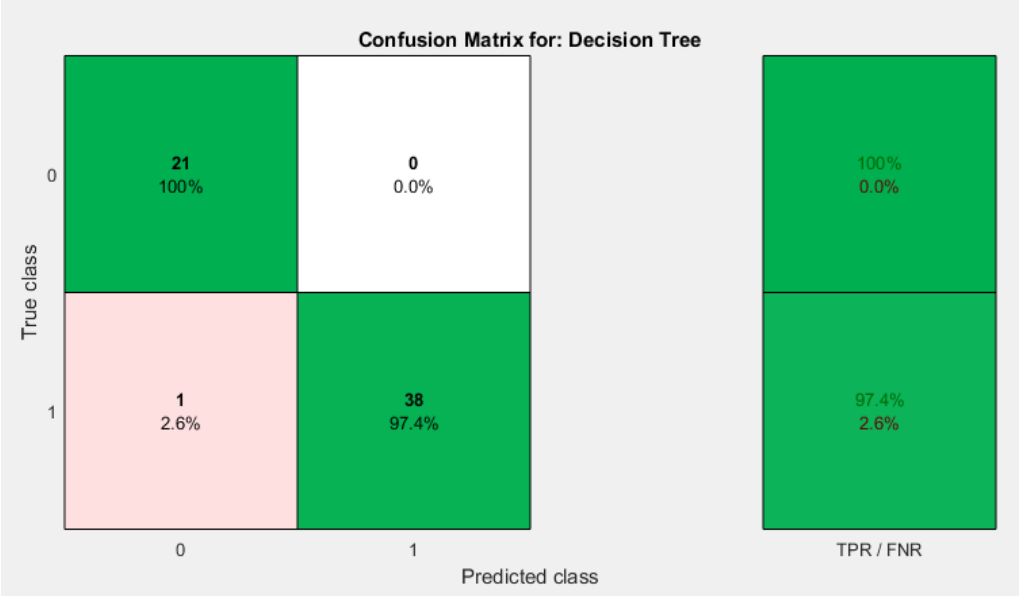


Figure 2. RNA-Seq confusion matrix using decision tree algorithm TP=38; TN=21; FP=0; FN=1

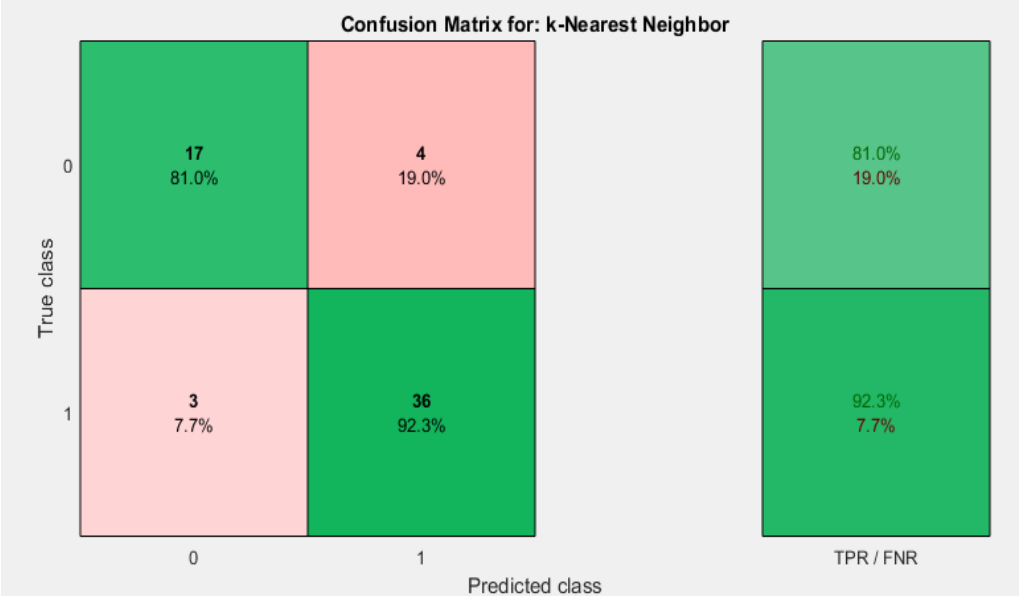


Figure 3. RNA-Seq confusion matrix using KNN TP=36; TN=17; FP=4; FN=3

Table 2. Performance metrics table for the confusion matrix

Performance Metrics	GA-Decision Tree Classification	GA-KNN Classification
Accuracy (%)	98.3	88.3
Sensitivity (%)	97.4	92.3
Specificity (%)	100	81.0
Precision (%)	100	90.0
Recall (%)	97.4	92.3
F-Score (%)	98.7	91.1

5. CONCLUSION

In this study, improvements efficience for predicting and detecting malaria ailments in human are proposed using machine learning dimensionality reduction and classification techniques. GA feature selection dimensionality reduction and KNN and decision tree classifiers were employed by performing evaluating and

analysing the performance results obtained. This study enhanced malaria vector data classification, and compared with quite a lot of proposed works in reviews by numerous researchers, the outcomes demonstrates that, GA dimensionality reduction model helps to develop classification output such as decision tree. Investigating current works proposed in literature can improve feature selection models and algorithms and compared with recent other state-of-the-art classification algorithm.

REFERENCES

[1] S. Shanwen, *et al.,* "Machine Learning and Its Applications in Plant Molecular Studies," *Briefings in Functional Genomics*, vol. 19, no. 1, pp. 40-48, 2019.

[2] F. David., *et al.,* "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, 2019.

[3] M. Arowolo, *et al.,* "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. 3487-96, 2019.

[4] S. Karthik, *et al.,* "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.

[5] N. Johnson, *et al.,* "Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers?," *Cold Spring Harbor Laboratory Press for the RNA Society*, vol. 24, no. 9, pp. 1119-1132, 2018.

[6] M. Libbrecht, *et al.,* "Machine learning applications in genetics and genomics," *Nat Rev Genetics*, vol. 16, pp. 321-332, 2015.

[7] Z. Jagga, *et al.,* "Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms," *BMC Proceedings*, vol. 8, no. 2, 2014.

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

[9] D. Oh, *et al*., "Predicting Autism Spectrum Disorder Using Blood-based Gene Expression Signatures and Machine Learning," *Clin Psychopharmacology Neuroscience*, vol. 15, no. 1, pp. 47-52, 2017.

[10] Q. Ren, *et al.,* "Clustering and Classification Methods for Single-cell RNA-Seq Data," *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1-13, 2019.

[11] W. Stephen, *et al*., "Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies," *Frontiers in Genetic*, vol. 9, no. 297, pp. 1-6, 2018.

[12] J. Alquicira-Hernandez, *et al.,* "scPred: Accurate Supervised Method for Cell-type Classification from Single-cell RNA-seq Data," *Genome Biology*, vol. 20, no. 264, 2019.

[13] S. Cui, *et al.,* "Machine Learning-based Microarray Analyses Indicate Low-Expression Genes Might Collectively Influence PAH Disease," *PLOS Computational Biology*, vol. 15, no. 8, 2019.

[14] H. Shon*, et al.,* "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.

[15] J. R. Adam, *et al.,* "Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites," *eLife Sciences*, vol. 7, pp. 1-29, 2018.

[16] A. Tan, *et al.,* "Ensemble Machine Learning on Gene Expression Data for Cancer Classification," *Applied Bioinformatics,* vol. 2, no. 3, pp. 75-83, 2003.

[17] N. Song, *et al.,* "Design and Analysis of Ensemble Classifier for Gene Expression Data of Cancer," *Advancement in Genetic Engineering*, vol. 5, no. 1, pp. 1-7, 2016.

[18] S. Tarek, *et al.,* "Gene Expression Based Cancer Classification," *Egyptian Informatics Journal*, vol. 18, no. 3, pp. 151-159, 2017.

[19] Duval, *et al.,* "Advances in Metaheuristics for Gene Selectio and Classification of Microarray Data," *Briefings in Bioinformatics*, vol. 11, no. 1, pp. 127-141, 2010.

[20] A. Shukla, *et al.,* "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, pp. 1-11, 2019.

[21] A. Tan, *et al.,* "Ensemble Machine Learning on Gene Expression Data for Cancer Classification," *Applied Bioinformatics,* vol. 3, pp. 1-10, 2003.

[22] K. Kamran, Kiana J. M., Mojtaba H., Sanjana M., Laura B., Donald B., "Text Classification Algorithms: A Survey," *Information MDPI*, vol. 10, no. 150, pp. 2-68, 2019.

[23] B. Mariangela, *et al.,* "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. 474, pp. 1-13, 2015.

[24] G. James, *et al.,* "An introduction to statistical learning with application in R," *New York (NY): Springer*; 2013.

[25] J. Bose, *et al.,* "A Hybrid GA/KNN/SVM Algorithm for Classification of Data," *BioHouse Journal of Computer science*, vol. 2, no. 2, pp. 5-11, 2016.

[26] I. Polaka, *et al.*, "Decision Tree Classifiers in Bioinformatics," *Scientific Journal of Riga Technical University,* pp. 110-123, 2010.

[27] M. Arowolo, *et al.,* "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.

[28] Mariangela Bonizzoni, *et al.,* "Additional file 4: of 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," 2015. 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

BIOGRAPHIES OF AUTHORS

Micheal Olaolu Arowolo, is a faculty of the Department of Computer Science at Landmark University, Omu-Aran Nigeria. He holds a Bachelor Degree from Al-Hikmah University, Ilorin, Nigeria and a Masters Degree from Kwara State University, Malete Nigeria, he is presently a PhD Student of Landmark University, Omu-Aran Nigeria. His area of research interest includes Machine Learning, Bioinformatics, Datamining, 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.

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 modellin and Artificial Intelligence. She has published widely in local and international reputable journals She is a member of Nigerian Computer Society (NCS), the Computer Registration Council of Nigeria (CPN) and IEEE member.

Professor Ayodele Ariyo Adebiyi, is a faculty and former Head of Department of Computer and Information Sciences, Covenant University, Ota Nigeria. He is currently the Head of Department of Computer Science at Landmark University, Omu-Aran, Nigeria, a sister University to Covenant University. He holds a BSc degree in Computer Science and MBA degree from University of Ilorin, Ilorin Nigeria. He had his MSc and PhD degree in Management Information System (MIS) from Covenant University, Nigeria respectively. His research interests include, application of soft computing techniques in solving real life problems, software engineering and information system research. He has successfully mentored and supervised several postgraduate students at Masters and PhD level. He has published widely in local and international reputable journals. He is a member of Nigerian Computer Society (NCS), the Computer Registration Council of Nigeria (CPN) and IEEE member.