ICA Learning Approach for Predicting RNA-Seq Data Using KNN and Decision Tree Classifiers

Marion Olubunmi Adebiyi[1], Ayodele A. Adebiyi[1], Olatunji Okesola[2] and Micheal Olaolu Arowolo[1],

Department of Computer Science, Landmark University, Omu-Aran, Nigeria.[1]
Department of Computer Science, First Technical University, Ibadan, Nigeria.[2]
Arowolo.micheal@lmu.edu.ng, marion.adebiyi@lmu.edu.ng,
ayo.adebiyi@lmu.edu.ng and olatunjiokesola@tech-u.edu.ng

Abstract

Parasites such as malaria accept disturbed variation of life unit's growth through uncountable stratospheres of mosquito vector. There are transcriptomes with distinct species, Ribonucleic acid sequencing (RNA-seq) is a predominant protein sequence expression process leading to increase genetic query identification. RNA-seq calculate records of gene expressions and imposes machine learning diagnostic enhancements techniques. Biological data learning approaches for analyses have been suggested by numerous researchers. An Independent Component Analysis (ICA) feature extraction dimensionality reduction algorithm is proposed to draw hidden components from RNA-seq dataset with high dimension, utilizing KNN and Decision Tree classification procedures to evaluate the performances, the study attained a performance metrics of 81.7.7% and 73.3% classification accuracy respectively.

Keywords: Mosquito Anopheles, RNA-Seq, ICA, KNN, Decision Tree.

1. Introduction

The next-generation sequencing proficiency high-throughput has generated abundant datasets widely, with enormous volume of data that allow biologists to carry out investigations and determine challenging transcriptions of genes, for instance RNA and ailment relationships such as melanomas, contagions (malaria, fevers), heredities, transmissible, biological, and so on [1].

An African malaria blood-sucking parasite (*Anopheles gambiae*) having significant vectors of Plasmodium falciparum is very deadly. Antimalarial state-of-the-art and medicines has increased, the need for advanced medicines requires improved genetic organism understandings. Anopheles Mosquitoes specific endurance parameter of expression of genes has remained a need for building enhanced comprehensive predictive model for transcriptions of malaria vector [2] [8].

RNA-seq produces receptive revealing genetic inquiries that defines faltering efficient biological design by enhancement of sequencing learning. RNA-Seq information entails removal of high-dimensionalities [3]. Technologies have enhanced approaches in developing state-of-the-art health care models for example smart health nursing systems, enhanced treatments, diseases and complaints analyses [4]. Quite a lot of machine learning approaches are established with suggestively evaluating enormous sum of RNA-Seq data developments with mutable achievable rates for learning related biological frameworks [5-7].

In this study, ICA technique is utilized to draw out hidden components in the given data, KNN and Decision Tree classification approaches are carried out to determine different genetic outlines respectively by delivering a developed classification accuracies methods for predicting and detecting new genes for malaria.

2. Reviews

Huge genetic dataset of individuals possessing or not possessing diseases applies computational approaches for genes accountable for presence of distinguishing diseases. Several procedures are used for recognizing Differentially Expressed Genes (DEG), Machine Learning (ML) is an important process for finding differences between human genomic features. Numerous approaches have been proposed in investigating and classifying of numerous diseases exploiting machine learning technology. Gene expression outlining approaches using innumerable machine learning approaches and current study work gaps predictable in examining gene expressions are discussed [4].

An evaluation of Autism field disease using blood-based gene expression profiling RNA-Seq data to detect and classify using ranked clustering investigation, SVM and KNN classifiers were proposed with an evaluation accuracy of 93.8% [9]. An RNA-Seq data using clustering and classification was carried out on an integrated valuation, by highlighting methods occurring recently as predominant scRNA-seq data variations for providing data descriptions 10]. A supervised learning technique for 12 cancer RNA-Seq gene grouping using ranking large ensembles of genes restrained was proposed using variable rank measures on random forests classification algorithm. There results showed a supervised gene selection learning-based approaches in trainings data and spot the requirement of utilizing feature selection methods on analysis of genes [11]. Classification of RNA-Seq data uses a supervised approach to predict a comprehensive procedure on a mononuclear cell, pancreatic material, colorectal growth tissue removal, and mixing dendritic cells datasets with single cells classification, by merging impartial feature selection reduced dimensional space method and displayed scPred classifies distinct cells with better accuracy [12].

An RNA-DNA investigation indicating low expressed genomes jointly inclined PAH ailment was proposed using a state-of-the-art feature selection and improved classification algorithms on an irrelevant set of genes, their result shows the small-expression clustered genes are close-fitting at expecting and characterizing transformation of PAH [13]. Stomach cancer data classification using CNN deep learning approach was proposed, its application to data expression was demonstrated on 60,000 genes of data, heatmaps, PCA, and CNN classification algorithm associates experimental data, it studied the protein sequence and evaluated them with 95.96% and 50.51% accuracy respectively. RNA-Seq disclosure of concealed records in malaria vermin was proposed by relating the distinction of RNA-seq technique to deconvolute transcriptional difference for 500 vermin, and revealed hidden distinct transcriptional signs [15].

An ensemble classification algorithm for cancerous data was proposed with C4.5 decision tree, as a supervised machine learning approach on seven microarray data and correlated these approaches in this experiment, bagged and boosted ensemble learning was noticed to advance better as a classifier [16]. Designing an investigative ensemble classification method was proposed for cancer gene expression data, using a mutual Recursive Feature Elimination with Ada-boost ensemble enhanced classification algorithm evaluation [17]. Gene expression cancer data classification was carried out using an effective ensemble classification method for increasing classification performance results, which shows less dependent ensemble results on originalities of individual training set [18]. An enhanced ensemble classification selected feature model tree was proposed using an ensemble-based by means of random tree and feature selection wrapper-based procedure to develop classification that originates a subclass using bagging-ensemble, wrapper-based system, and random trees. The feature selection removes the unrelated features and selects optimum features for classification and evaluated the performance, with a classification accuracy of 92% [23]. Numerous dimensionality reduction approach for gene expression investigation was proposed, an instantaneous and bases of the software were stated in the preceding segment for feature extraction process [24].

3. Materials and Methods

Investigation of high dimensional data have been carried out by several authors, this learning proposes an Independent Component Analysis (ICA) feature extraction for RNA-Seq huge data and an Ensemble classification algorithm for an improved performance on genes of 2457 instances and 7 attributes from a western Kenya data, containing transcriptions of mosquito genes with

disparities in transcripts of deltamethrin- resistant and susceptible Anopheles gambiae mosquitoes and shown in table1 [19].

Table 1. Dataset Features

Dataset	Attributes	Instances
Mosquito Anopheles Gambiae	7	2457

3.2 Methods

In this study experimental tool using MATLAB was proposed to assess the provided data [19], ICA was proposed to extract latent component features and perform the ensemble classification algorithm approach [20].

3.2.1 Independent Component Analysis (ICA)

ICA is an esteemed extended conservative of PCA since the separation visor of independent bases from linear grouping [24]. The original fact of ICA possesses uncorrelation of universal PCA. Built n x p on data matrix X, with rows ri (j=1…, n) reckon to observed variables and columns c_j (j=1…, p) are entities matching variables. ICA model of X is as follows:

$$X = AS \qquad\qquad (1)$$

With complete overview, A is a n x n fusion matrix, where S is a n x p is a basis matrix with the need of statistically independently conceivable. Independent components are original variables kept in rows of S, with variables detected linearly composed independent components. The independent components achieved by learning the precise linear groupings of the detected variables, then union can be reversed as:

$$U = S = A^{-1}X = WX \qquad\qquad (2)$$

3.2.2 Kth Nearest Neighbours

K-nearest neighbor (KNN) algorithm is a supervised knowledge classifier approach for gene data, the result of innovative occurrence request is classified built on mutual KNN cluster. KNN algorithm uses neighborhood arrangement for value evaluation of innovative, it classify novel entity built on features and training models. KNN uses no model to appropriate retention. Particular features are specified as an involvement to this segment. The Kth principles that are contiguous to the query opinion are carefully chosen, distance among query-instance and training models are considered. The distance is fixed and nearest neighbors created on the Kth least distance is resolute, group Y of the nearest neighbors is grouped and uses unassuming common group of nearest neighbors as prediction value of the query instance, by fragmenting any draws arbitrarily [21].

3.2.3 Decision Trees

Decision tree is a classifier that recursively divides instances space by means of hyperplanes orthogonally to its axes, a derivation node representing a feature model is developed, the instance fragment space is built on attribute values functions. Individually new data sub-space is divided to original reiterative sub-spaces till finish condition is encountered and leaf nodes are allocated respectively to a class label denoting class of common instances checked in the sub-space. It is very important to set the right end condition for overfitting and underfitting that can suffer loss in accuracy. Most algorithms have built in mechanisms dealing with overfitting known as pruning. New instances are classified by directing after tree-root to down layer of the leaf, rendering the test result path [22]. Decision tree yields unstable effective models, if the given training datasets

slightly vary, the ensuing models can be dissimilar totally, decision trees are used in classifier ensembles frequently.

3.3 Performance Evaluation

Machine learning model performance assessment entails some proof metrics such as the confusion matrix which is frequently used in analyzing classification model illustrations from the experiment sampling model [4], [25].

3.4 Applications

Analyzing genes suggests an enhanced pathway to recognize RNA-Seq data, determining appropriate genes are supportive in developing many applications like improved treatments, analyzing diseases, realizing protein sequence and medications, cancer classification, complaints such as typhoid, malaria, and so on. Machine learning technology for discovering strategies and inconsistency amongst data possesses great algorithmic tools useful on several fields. MATLAB is employed to achieve the aim of this study [16]. This study predicts RNA-Seq Malaria database knowledge using the MATLAB tool with computer system configuration having processor-iCore, RAM-8GB, 64-bit and 2015a MATLAB implementing tool.

4. Results

In this study, Mosquitoes Anopheles Gambiae RNA-Seq data with 2457 uses an ICA algorithm to reduce the curse of dimensionality by fetching the latent components, spots and remove uncorrelated Features that adopt determined modifications with a lesser number of Key Components, this study gives important genetic factor information beneficial for additional learning. KNN and Decision Tree classifiers were carried out on the MATLAB environment tool for the proposed model.

In this study, ICA was used as a feature extraction technique generated 25 signification latent Component features of genes, KNN and Decision Tree classifiers uses a 10-folds cross validation in evaluating the classification model with 0.05 holdout was used as a data parameter for training and 25% for testing and removing the sampling biases of the classifier knowledge evaluation procedure, the evaluation results are built on time computation and performance metrics [25].

In this study, ICA-KNN and ICA-Decision Tree were compared with 81.7% and 73.3% accuracy results individually and procedures of the confusion matrix result output are revealed below.

Figure 1. Experiment Data Display.

In this study, ICA algorithm was used to extract hidden components from the dataset shown in figure 1, and passed the features extracted into an ensemble classification algorithm, the reduced data is also passed into a decision tree classifier, the results and scattered plot of the reduced data are revealed in the figures beneath. Confusion matrix shows a resolution of performance metrics and shown in figure 4 and 5 below respectively.
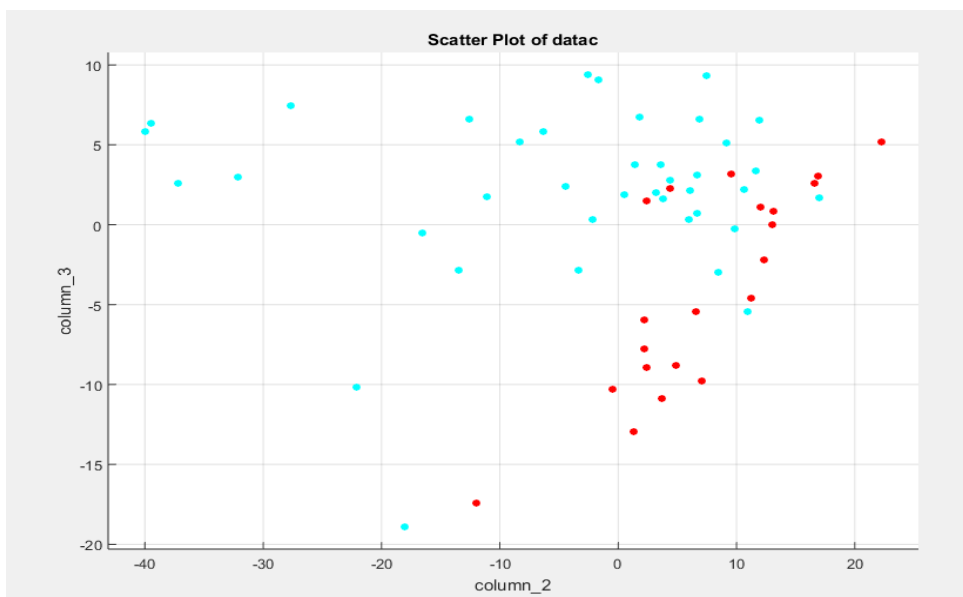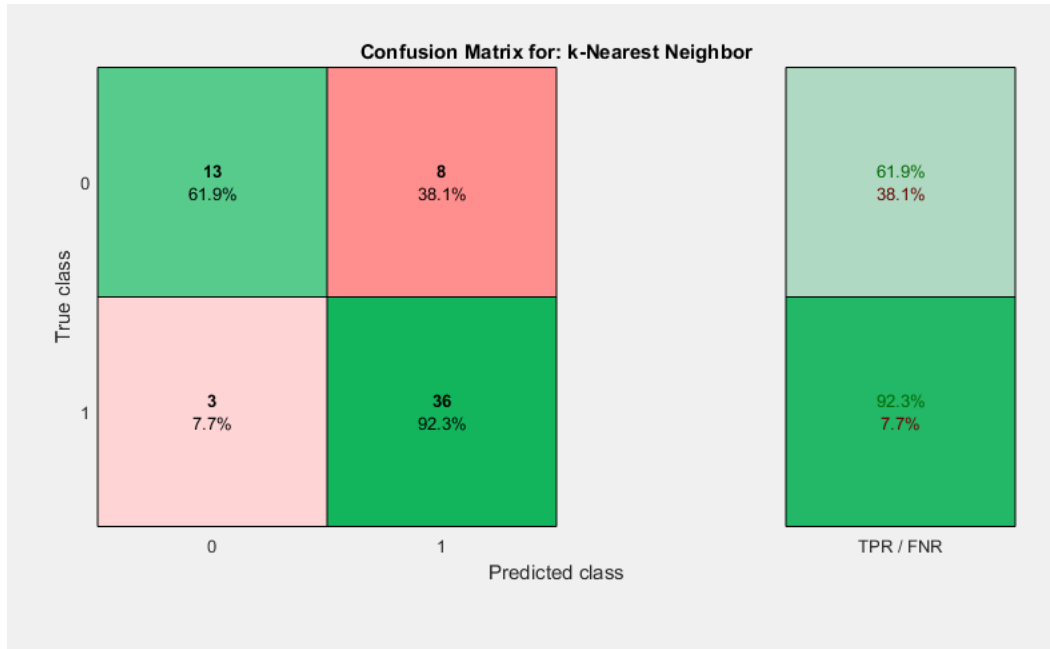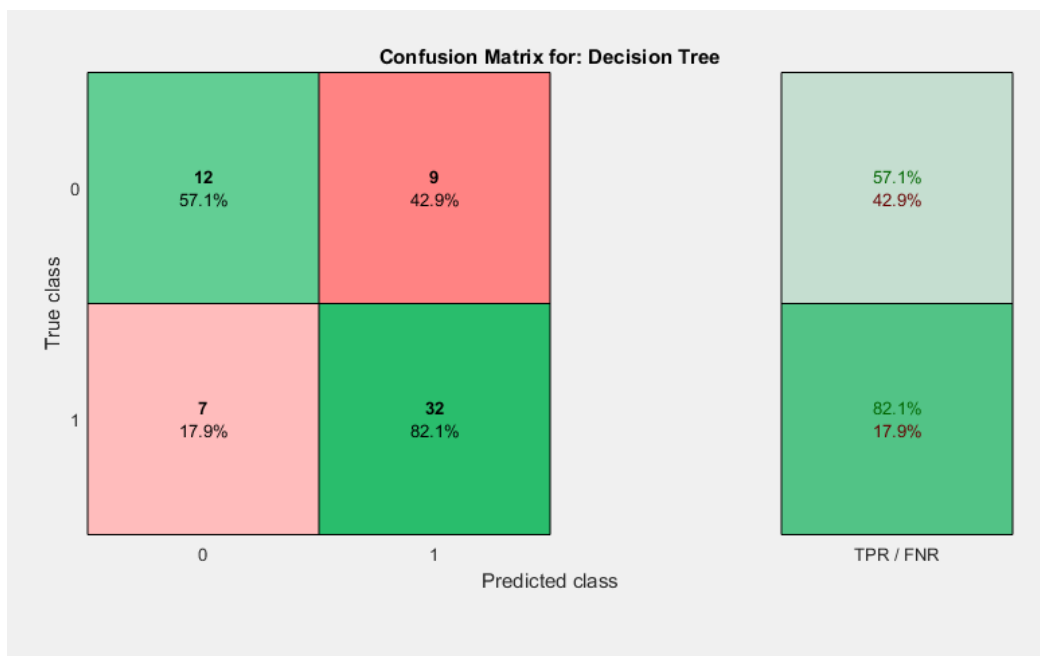

Figure 2. Scattered Plot of the Reduced Data.

12277

KNN	
Medium KNN	81.7%
Tree	
Simple Tree	73.3%

Fig. 3. Classfiication Results Using ICA-KNN and ICA-Decision Tree Algorithms

**Figure 4. Confusion Matrix for RNA-Seq Data Classification Using ICA-KNN
TP=36; TN=13; FP=8; FN=3**

**Figure 5. Confusion Matrix for the RNA-Seq Data Classification Using ICA-Decision Tree
TP=32; TN=12; FP=9; FN=7**

In this study, Mosquito Anopheles Gambiae RNA-Seq data was copied [33] containing 2457 gene features, ICA was utilized as a feature extraction algorithm, with relevant hidden components for classification utilizing Ensemble classifier, decision tree was also

12278

used and the performances were compared. The outcome displays the efficiency of this approach. The result displays ICA-KNN performs better than ICA-Decision Tree in terms of a reduced amount of training time and accuracy performance, the evaluation is revealed and compared in table 2 below.

Table 2. Performance Metrics Table for the Confusion Matrix

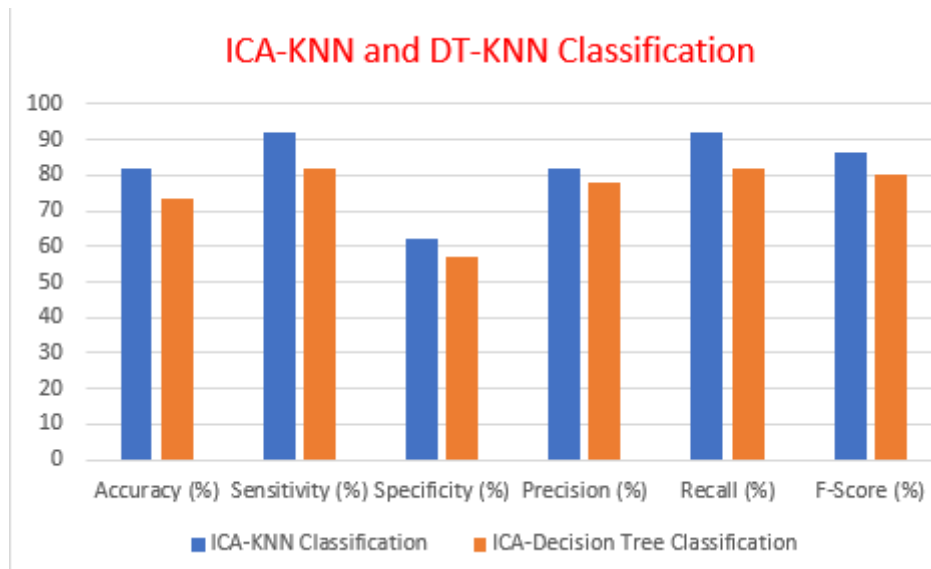| Performance Metrics	ICA-KNN Classification	ICA-Decision Tree Classification
Accuracy (%)	**81.7**	**73.3**
Sensitivity (%)	**92.3**	**82.1**
Specificity (%)	**62.0**	**57.1**
Precision (%)	**81.8**	**78.1**
Recall (%)	**92.3**	**82.1**
F-Score (%)	**86.7**	**80.1**

Fig .6. Result Classification Chart

In this study, classification of mosquito anopheles data was investigated and enhanced, quite a lot of studies have been carried out by different authors utilizing the performance metrics, the result chart is displayed in figure 6 above, the results of this study proves that ICA can enhance classification productivity for KNN and Decision tree classifiers.

5. Conclusion

In this study, prediction and detection of malaria is carried out using machine learning procedures like dimensionality reduction (ICA) and KNN and Decision Tree classification procedures. Investigation and assessment of the performance carried out showed KNN performs better than Decision Tree Classification algorithm.
Classification of malaria vector data was improved; the outcomes shows ICA as a helpful algorithm to advance classification productivities.
Future works can carry out hybridized ICA for classification of RNA-Seq dataset.

References

[1] S. Shanwen, W. Chunyu, D. Hui, Z. Quan, "Machine Learning and its Applications in Plant Molecular Studies," Briefings in Functional Genomics Oxford Academic, 2019, pp.1-9. doi:10.1093/bfgp/elz036

[2] F.R. David, C. Kate, Y.L. Yank, G. Karine, L. Roch. "Predicting Gene Expression in the Human Malaria Parasite Plasmodium Falciparum Using Histone Modification, Nucleosome Positioning, and 3D Localization Features" PLOS Computational Biology, 2019 doi.org/10.1371/journal.pcbi.1007329

[3] M.O. Arowolo, M. Adebiyi, A.A. Adebiyi. "A Dimensional Reduced Model for the Classification of RNA-Seq Anopheles Gambiae Data", Journal of Theoretical and Applied Information Technology. 2019, 97(23) pp.3487-96.

[4] S. Karthik, 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. 2018, 8(2), pp.182-191

[5] N.T. Johnson, A. Dhroso, K.J. Hughes, D. Korkin. "Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers?". RNA. 2018, 24(9), pp.1119–1132. doi:10.1261/rna.062802.117.

[6] M.W. Libbrecht, W.S. Noble. "Machine learning applications in genetics and genomics" Nat Rev Genetics. 2015, 16, pp.321–332.

[7] Z. Jagga, D. Gupta. "Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms". BMC Proceedings. 2014:8(2).

[8] Anopheles gambiae 1000 Genomes Consortium; Data analysis group; Partner working group; Genetic diversity of the African malaria vector Anopheles gambiae. Nature.2017;552(7683) pp.96–100. doi:10.1038/nature24995

[9] D.H. Oh, I.B. Kim, S.H. Kim, D.H. Ahn. "Predicting Autism Spectrum Disorder Using Blood-based Gene Expression Signatures and Machine Learning". Clin Psychopharmacology Neuroscience. 2017;15(1): pp.47–52. doi:10.9758/cpn.2017.15.1.47

[10] Q. Ren, M. Anjun, M. Qin, Z. Quan. "Clustering and Classification Methods for Single-cell RNA-Seq Data". Briefings in Bioinformatics. 2019: pp.1-13

[11] W. Stephen, S. Ruhollah. "Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies. Frontiers in Genetic". Bioinformatics and Computational Biology. 2018:9(297); pp.1-6. doi.org/10.3389/fgene.2018.00297

[12] J. Alquicira-Hernandez, A. Sathe, H.O. Ji, Q. Nquyen, J.E. Powell. "scPred: Accurate Supervised Method for Cell-type Classification from Single-cell RNA-seq Data". Genome Biology. 2019:20(264) doi:10.1186/s13059-019-1862-5

[13] S. Cui, Q. Wu, J. West, J. Bai. "Machine Learning-based Microarray Analyses Indicate Low-Expression Genes Might Collectively Influence PAH Disease". PLOS Computational Biology. 2019. doi.org/10.1371/journal.pcbi.1007264

[14] H.S. Shon, Y.G. Yi, K.O. Kim, E.J. Cha, K.A. Kim. "Classification of Stomach Canacer Gene Expression Data Using CNN Algorithm of Deep Learning". Journal of Biomedical Translation Research. 2019:20(1); pp.15-20. doi.org/10.12729/jbtr.2019.20.1.015

[15] J.R. Adam, M.T. Arthur, M.B. Hayley, R.G. Ana, J.S. Mandy, J.R.I. Christopher, Oliver B, Matthew B, Mara KNL. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *Elife*. 2018;7. doi:10.7554/eLife.33105

[16] A.C. Tan, Gilbert D. Ensemble Machine Learning on Gene Expression Data for Cancer Classification. 2003:2(3);75-83.

[17] N. Song, Wang k, Xu M, Xie X, Chen G, Wang Y. Design and Analysis of Ensemble Classifier for Gene Expression Data of Cancer. Advancement in Genetic Engineering. 2016:5(1);1-7.
doi:10.4172/2169-0111.1000152

[18] S. Tarek, Elwahab RA, Shoman M. Gene Expression Based Cancer Classification. Egyptian Informatics Journal. 2017:18(3);151-159. Doi:10.1016/j.eij.2016.12.001.

[19] B. Mariangela, Eric O, William AD, Monica B, Yaw A, Guofa Z, Joshua H, Ming L, Jiabao X, Andrew G, Joseph F, Guiyun Y. 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. 2015:8(474);1-13. https://doi.org/10.1186/s13071-015-1083-z

[20] G. James, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with application in R. New York (NY): Springer; 2013.

[21] J.S.C. Bose, S.B., Changalesetty, A.S., Badawy, W. Ghribi, J. Baili, and H. Bangali. "A Hybrid GA/KNN/SVM Algorithm for Classification of Data". BioHouse Journal of Computer science. 2:(2), 2016, pp.5-11

[22] I. Polaka, I. Tom, and A, Borisov. "Decision Tree Classifiers in Bioinformatics". Scientific Journal of Riga Technical University. 2010, pp. 110-123.

[23] A.C. Tan, Gilbert D. Ensemble Machine Learning on Gene Expression Data for Cancer Classification. Applied Bioinformatics. 2003:3;1-10.

[24] K. Kamran, Kiana JM, Mojtaba H, Sanjana M, Laura B, Donald B. Text Classification Algorithms: A Survey. Information MDPI. 2019:10(150);2-68

[25] M.O. Arowolo, Abdulsalam SO, Isiaka RM, Gbolagasde KA. A Comparative Analysis of Feature Selection and Feature Extraction Models for Classifying Microarray Dataset. Computing and Information System.2018:22(2);29-38.

[26] E. Guzman, El-halaby M, Bruegge B. Ensemble Methods for App Review Classification : An Approach for Software Evolution, in: 30th IEEE/ACM Int. Conference of Automative Software Engineering. 2015: pp;771–776. doi:10.1109/ASE.2015.88.

[27] Y. Ren, Suganthan PN, Srikanth N. Ensemble methods for wind and solar power forecasting : A state-ofthe-art review, Reneweable Sustainable Energy Revolution.2015:50(4);:82-91. doi:10.1016/j.rser.2015.04.081.

[28] S. Flennerhag. Machine Learning Ensemble, (2017). doi:10.5281/zenodo.1042144.

[29] C.F. Tsai, Y.F. Hsu, D.C. Yen. "A comparative study of classifier ensembles for bankruptcy prediction", Application Soft Computing Journal. 2014:24; pp. 977–984. doi:10.1016/j.asoc.2014.08.047

[30] A. Mayr, A. Binder, O. Gefeller, M. Schmid. "The Evolution of Boosting Algorithms From Machine Learning to Statistical Modelling", Methods Informatics and Medicine. 2014:53; pp.419–427.

[31] A. Nisioti, A. Mylonas, P.D. Yoo, S. Member, V. Katos. "From Intrusion Detection to Attacker Attribution : A Comprehensive Survey of Unsupervised Methods". IEEE Commun Surv Tutorials. 2018; pp. 1-11.

[32] S. Hafizah, S. Ariffin, N. Muazzah, A. Latiff, M.H.H. Khairi, S.H.S. Ariffin,. "A Review of Anomaly Detection Techniques and Distributed Denial of Service (DDoS) on Software Defined Network (SDN)". Technol Appl Sci Res [Internet]. 2018;8(2) pp. 2724–30.

[33] 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

Authors

Arowolo Micheal Olaolu, is a staff 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 Master's 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 in local and international reputable journals, he is a member of IAENG, APISE, SDIWC, and an Oracle Certified Expert.

Dr. Marion O, Adebiyi, is a faculty of the Department of Computer Science at Landmark University, Omu-Aran, Nigeria. She holds a B.Sc Degree from University of Ilorin, Ilorin Nigeria. She had her M.Sc and Ph.D 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 modeling 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.

Prof. Ayodele, A. 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 B.Sc degree in Computer Science and MBA degree from University of Ilorin, Ilorin Nigeria. He had his M.Sc and Ph.D 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 Ph.D 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.

Olatunji Julius Okesola, is a Professor of Cybersecurity at the 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 member of Information System Audit and Control Association (ISACA), Computer Professionals of Nigeria (CPN), and a fellow of Nigerian Computer Society (NCS). Okesola is a scholar, an Information Security expert and a seasoned banker. Until November 2016, he was the Group Head, for Information Systems Control and Revenue Assurance at Keystone Bank (Nig.) Ltd, Lagos. An alumnus of University of South Africa. His research interests include Cyber security, biometrics, and Software engineering. He has several publications in scholarly journals and conference proceedings both local and international.