International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 1 (2020), pp. 163-169
 https://dx.doi.org/10.37624/IJERT/13.1.2020.163-169

An Efficient PCA Ensemble Learning Approach for Prediction of RNA-Seq Malaria Vector Gene Expression Data Classification

Micheal Olaolu Arowolo[1], Marion O. Adebiyi[2] and Ayodele A. Adebiyi[3]

[1,2,3] Department of Computer Science, Landmark University, Nigeria.

[1]ORCID: 0000-0002-9418-5346 and [2]ORCID:0000-0001-7713-956X

Abstract

Malaria parasites adopt outstanding variation of life phases as they evolve through manifold mosquito vector atmospheres. Transcriptomes of thousands of individual parasites exists. Ribonucleic acid sequencing (RNA-seq) is a widespread method for gene expression which has resulted into improved understandings of genetical queries. RNA-seq compute transcripts of gene expressions. RNA-seq data necessitates analytical improvements of machine learning techniques. Several learning approached have been proposed by researchers for analysing biological data. In this study, PCA feature extraction algorithm is used to fetch latent components out of a high dimensional malaria vector RNA-seq dataset, and evaluates it classification performance using an Ensemble classification algorithm. The effectiveness of this experiment is validated on aa mosquito anopheles gambiae RNA-Seq dataset. The experiment result achieved a relevant performance metrics with a classification accuracy of 93.3%.

Keywords: RNA-Seq, PCA, Ensemble Cassification, Malaria Vector.

I. INTRODUCTION

High-throughput next-generation sequencing technology has yielded numerous extensive data sets, this huge amount of data permit biologists to investigate and discover problematic transcripts of genes, such as relations amid RNA and diseases such as cancer, infections (malaria), genetics, hereditary, physiological, among others [1].

Anopheles gambiae are kind of blood-sucking mosquitoes with principal vectors of Plasmodium falciparum malaria in Africa. Mosquito Anopheles is one fatal kind of malaria parasite, liable for thousands of bereavements. As conflict to antimalarial medications banquets, innovative antimalarials rises, fetching for innovative medications necessitates enhanced biological understanding of this organisms. How mosquito anopheles parasite sustains specific regulation of gene expression has been a huge question that requires constructing an improved detailed prognostic model for malaria vector transcriptions [2] [8].

RNA-seq study generates responsive informative biological investigations by describing a tentative functional biological plan by improvement of sequencing study. RNA-Seq data requires the elimination of the curse of high-dimension, such as; disorders, noises, duplication, redundancy, irrelevant, inappropriate data, among others [3]. Recent technologies have improved methods in evolving innovative healthcare models such as modified treatments, smart human health monitoring systems, among other diagnoses of ailments and diseases [4].

Over the past decades, several machine learning tools have been developed with meaningfully innovations for analysing the huge amount of RNA-Seq and next generation sequencing gene data expression through learning the biologically relevant frameworks [5]. Several authors have exploited machine learning techniques for RNA-Seq gene expression data with variable success rates [6], [7].

This study suggests a PCA feature extraction dimensionality reduction procedure, to fetch out the high dimensionality in gene expression data analyzes, a Sub-space Ensemble classification method, to discover distinct biological frameworks is used to provide higher classification accuracy and it is recommended as an efficient technique for the prediction and detection of new genes for malaria.

II. RELATED WORK

Computational methods are applied on huge genetic dataset of persons with or without ailments, genes responsible for existence of ailments can be detected. Differentially Expressed Genes (DEG) are recognised by means of several procedures. Machine Learning (ML) procedures are important in recognizing the dissimilarity amongst genes gotten from human genome. Several methods on machine learning used in analysing and classifying gene expression profiles of numerous diseases are emulated. The need for gene expression profiling and its methods using various machine learning are discussed. Several research works done by researchers in this field are discussed. There are current research gaps recognised in analysing gene expressions [4].

Oh, Kim, Kim and Ahn, [9] worked on the estimate of Autism spectrum ailment with the aid of blood-based expression of gene signatures and machine learning, to identify transcripts that can be used in classification. They used RNA data from Gene expression omnibus database, using R language tool for machine learning algorithms. Ranked cluster analysis presented autism spectrum disorder remained comparatively well-discriminated from panels. Support vector machine and K-nearest neighbour classifiers are used to validate the data result

in an inclusive class estimate accuracy of 93.8% as well as a sensitivity and specificity of 100% and 87.5%, respectively.

Ren, Anjun, Qin and Quan, [10] worked on RNA-Seq data by clustering and classification by conducting an integrated assessment, they highlighted the pros and cons of approaches by using clustering and classification methods that have occurred lately as prevailing changes, with nonlinear and linear approaches with plunging dimension methods for scRNA-seq data, by integrating and providing a report of scRNA-seq data and download URLs.

Stephane and Ruhollah [11], worked on supervised learning approach for collection of RNA-Seq genes by ranking large ensembles of genes measured with RNA-Seq. they used variable rank measures produced by the random forests classification algorithm, they defined the EPS (extreme pseudo-samples) channel, using Variational Autoencoders and regressors to extract ranks of 12 cancer RNA-Seq datasets extending from 323 to 1,210 samples. There results proved the latent of supervised learning-based gene selection approaches in RNA-Seq trainings and highlight the necessity of using gene selection approaches on gene expression analysis.

Hernandez, Sathe, Ji, Nquyen and Powell [12], worked on RNA-Seq data classification using a supervised model. They presented a generalizable technique with vastly precise classification of single cells, by means of combining impartial feature selection from a condensed dimension space, and machine learning estimate technique. They applied scPred to RNA-seq dataset from mononuclear cells, pancreatic tissue, colorectal tumour biopsies, and circulating dendritic cells. They showed scPred classifies discrete cells with high accuracy.

Cui, Wu, West and Bai [13] worked on machine learning based on RNA-DNA analysis indicate low expressed genomes that might be collectively influenced PAH disease. They proposed an innovative feature selection and improved machine learning algorithm methods to classify an insignificant set of extremely useful genes. Outcomes showed that clusters of small-expression genes are revealing at predicting and distinguishing changed forms of PAH.

Shon, Yi, Kim, Cha and Kim [14] worked on classifying gene expression stomach cancer data using CNN. They developed a classification technique based on deep learning and proved its application to data expression gotten from stomach cancer patients. 60,483 genes of data from 334 stomach cancer patients in The Cancer Genome Atlas were assessed by principal component analysis (PCA), heatmaps, and the convolutional neural network (CNN) algorithm. They combined clinical data and RNA-seq gene expression data, examined genes, and analysed them with CNN deep learning algorithm. They got an accuracy of 95.96% and 50.51%.

Adam, Arthur, Hayley, Ana, Mandy, Christopher, Oliver, Matthew and Mara [15], worked on RNA-Seq revelation of hidden transcripts in malaria parasites by describing the variation of an RNA-seq procedure to deconvolute transcriptional disparity for about 500 distinct parasites of rodent and human malaria. They discovered concealed discrete transcriptional signatures.

Tan and Gilbert [16] worked on an ensemble machine learning algorithm for classification of cancer gene expression data. They focused on C4.5 decision tree, bagged and boosted ensemble decision trees, which are supervised machine learning procedures for cancer classification, on seven openly obtainable cancerous microarray data and related the classification presentation of these approaches. They detected that ensemble learning (bagged and boosted decision trees) does improved than single decision trees in classification.

Song, Wang, Xu, Xie, Chen and Wang [17] worked on designing an analytical ensemble classification approach for gene expression of data for cancer. A combinational Recursive Feature Elimination with Adaboost algorithm was carried out to select important features for classification. There results showed an enhancement.

Tarek, Elwahab and Shoman [18], worked on cancer classification for gene expression data. They proposed an operative ensemble classification approach that increases the presentation of the classification and the poise of the outcomes. Ensemble classifiers outcomes are less reliant on individualities of a sole training set.

Mohan, and Nagarajan [23], worked on improving tree model for classifying ensemble selected features. This learning used an ensemble-based feature selection with random trees and wrapper technique to advance the classification. The future ensemble knowledge classification technique originates a subset by means of the bagging, wrapper method, and random trees. The future technique eliminates the irrelevant features and chooses the optimal features for classification using a probability weighting principle. The future feature selection technique is evaluated using RF, SVM, and NB evaluations and compared their performance with the GASVMb, GANBb, FSNBb, FSSVMb, and GARFb methods. The technique attains a classification accuracy of 92.

Kamran, Kiana, Mojtaba, Sanjana, Laura and Donald [24], worked on classifying text algorithm survey. An outline of text classification algorithms is deliberated. The outline studied diverse text dimensionality reduction approaches, present algorithm methods, and assessments

III. MATERIALS AND METHOD

Numerous methods for analyzing high dimensional data have been proposed in literature. In this study, principal component analysis (PCA) and the ensemble classification algorithm is studied for dimensionality reduction of high dimensional RNA-Seq data to get a better performance.

III.I Material

2457 instances with 7 attributes of genes are used, the data was gotten from western Kenya, comprising of genes of mosquitos from 2010 to 2012. The transcription profiling file comprises of AGAP012984, AGAP0 02724, AGAP003714, AGAP004779, AGAP009472, CPLC G3 [AGAP008446], CYP6M2 [AGAP008212] and CYP6P3 [AGAP002865], RNA-Seq genes, variations in transcriptome of deltamethrin-resistant and vulnerable Anopheles gambiae mosquitoes in

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 1 (2020), pp. 163-169
 https://dx.doi.org/10.37624/IJERT/13.1.2020.163-169

western Kenya, are openly accessible dataset from figshare.com and financed by the National Institute of Health [19]. Table-1 demonstrates a concise description of the dataset.

Table 1. Dataset Features

Dataset	Attributes	Instances
Mosquito Anopheles Gambiae	**7**	**2457**

III.II Methods

MATLAB was used as an experimental tool to evaluate the data obtained from [19], PCA was used to extract features. The extracted features were used to performed classification using the ensemble algorithm approach [20].

III.II.I Principal Component Analysis (PCA)

PCA [10] is an unsupervised feature extraction dimensionality reduction procedure, it adopts normally distributed data, diagonalizes covariance matrix. The orthogonal alteration is used to transform a conventional latent linear correlation variable into linear independent variables. Problems with linear dimensionality reduction procedures is absorbing unrelated data facts in a lower dimensional section. PCA can visualize models and advance the clarification capability [14].

PCA is a broadly useful method for dimensionality reduction, feature extraction, compression of data, visualization of data, among others.

In this study, PCA was used to extract features of gene expression having alterations among samples. PCA determines the principal subspace dimensions, which exploits the variance of the predictable data. The illustration of the experimental value in this principal subspace develops a feature vector of detected values. Adopting [14], the sample mean x□ and data covariance matrix S are as follows.

$$\bar{\boldsymbol{X}} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{X_n} \qquad \textbf{(1)}$$

$$\boldsymbol{S} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \ (\boldsymbol{X_n} - \bar{\boldsymbol{X}})(\boldsymbol{X_n} - \bar{\boldsymbol{X}})^{\boldsymbol{T}} \qquad \textbf{(2)}$$

Adopting equations (**1**) and (**2**), the unit vector on the principal subspace that exploits the variance of a given data set as follows.

$$\boldsymbol{Su_i} = \boldsymbol{\lambda_i u_i u_i}^{\boldsymbol{T}}\boldsymbol{Su_i} = \boldsymbol{\lambda_i} \qquad \textbf{(3)}$$

The vector maximizing the variance of the predictable data develops an eigenvector, u_i, of matrix *S*, and the maximal variance size in the path of the eigenvector develops the eigenvalue λ_i. Principal subspace collected for the principal component resultant from PCA is composed of an eigenvector of M pieces of maximal eigenvalues for matrix *S*.

III.II.II Ensemble Classifier

Ensemble classifiers can be trained using on unrelated subsets of the training data, diverse parameters of the classifiers, or even with diverse subsets of features as in random subspace models [21].

Ensemble classifier comprises of integrating results of diverse classifiers to produce a concluding decision, it is frequently used for gaining highly accurate results. Ensemble classifiers are relatively common in machine learning complications, and can be employed in bioinformatics field. Classification decision is achieved by merging the decision of each classifier [22].

Ensemble approaches is machine learning techniques combines decisions to advance the performance of the general classification. Several terms have been discovered in the literature to signify comparable connotations such as; multi-strategy learning, aggregation, integration multiple classifiers, classifier fusion, combination, committee, and so on.

Ensemble classifier can have complete improved performance than the discrete base classifiers. The efficiency of ensemble approaches is extremely dependent on the unconventionality of error devoted by discrete learner. Ensemble approaches performance depends on the accuracy and the variety of the base learners, ensemble classification has common techniques; bagging and boosting.

Bagging (**b**oostrap aggregating) employs the training data by randomly changing the unique *T* training data by *N* items. The replacement training sets are called bootstrap duplicates with some instances not appearing while others appear more than once. The final classifier *C**(*x*) is built by combining *Ci*(*x*) where every *Ci*(*x*) has an equal vote.

AdaBoost (Adaptive Boosting) technique effects the training data. Originally, the algorithm allocates all instance *xi* with an equal weight. In each iteration *i*, the knowledge algorithm attempts to diminish the weighted error on the training set and yields a classifier *Ci*(*x*). The weighted error of *Ci*(*x*) is calculated and useful to inform the weights on the training instances *xi*. The weight of *xi* increases giving to its effects on the classifier's performance that allots a high weight for a misclassified *xi* and a small weight for an acceptably classified *xi*. The final classifier *C**(*x*) is constructed by a weighted vote of the discrete *Ci*(*x*) rendering to its accuracy built on the weighted training set [23].

Adopting Kamran et.al [24], they showed how a boosting algorithm works for datasets, then trained by multi-model designs (ensemble learning). These advances resulted in the AdaBoost (Adaptive Boosting). Presume constructing D_t such that $D_1(i) = \frac{1}{m}$ given D_t and h_t:

$$D_{i+1}\{i\} = \frac{D_t(i)}{Z_t} X \begin{cases} e^{-\alpha_t} \ if y_i = \ h_t(x_i) \\ e^{\alpha_t} \ if y_i \neq \ h_t(x_i) \end{cases} \qquad (4)$$

$$= \frac{D_t(i)}{Z_t} \exp(-\alpha y_i h_t(x_i)) \qquad (5)$$

Where Z_t states to the normalization factor and α_t is as follows;

$$\alpha_t = \frac{1}{2} in(\frac{1-\epsilon_t}{\epsilon_t}) \qquad (6)$$

Basic ensemble classification techniques namely: The Weighted Averaging (WA); Max Voting (MV) and Averaging. Max Voting (MV) exists [26-28] Ensemble learning have three

advanced combination techniques; Stacking (STK); Blending (BLD); Bagging (BAG); and Boosting (BOT) [29-32].

III.II.III Performance Evaluation

Evaluating the performance of machine learning model requires some validation metrics. Confusion matrix is mostly used in classification models to analyze four features; True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). It discovers the correctly and incorrectly classified illustrations from the dataset sample given to test the model [4]. Performance metrics with its formula are presented below [25].

Accuracy of a model is calculated using four measures called TP, FP, TN, and FN.

The product of TP finds the state when it is existing.

The product of FP finds the state when it is not existing.

The product of TN does not find the state when it is not existing. The product of FN does not find the state when it is existing.

Accuracy: (TP + TN) / (TP+TN+FP+FN)

Sensitivity computes the amount of fittingly recognized instances with positive positives.

Sensitivity: TP/ (TP+FN)

Specificity finds the amount of fittingly recognized instances with actual negatives.

Specificity: TN/ (FP+TN)

Precision: TP/ (TP+FP)

Recall: TP/TP+FN

F-Score: 2 x (Recall x Precision) / (Recall + Precision)

III.II.IV Applications

Gene Expression Analysis offers an improved path to identify RNA-Seq data. The need to discovering relevant genes are helpful in developing numerous applications like modified treatment, diagnosis of diseases, discovering genes and drugs, tumor classification, ailments such as typhoid, malaria, among others. Machine learning technology in finding the designs and the discrepancy between data. It owns great algorithms as tools that is applied on various fields.

MATLAB (Matrix Laboratory) is utilized to perform the experiment, due to its ease and beneficial programming environment for engineers, architects, scientists, researchers, among others. MATLAB is a multi-worldview arithmetical

processing environment and exclusive programming language established by MathWorks. It permits framework controls, plotting of functions and information, execution of algorithms, production of User Interfaces, written in different languages, such as; C, C++, C#, Java, Fortran and Python [16]. The principle point of this study is the prediction of the RNA-Seq technology utilizing the MATLAB tool by utilizing the Malaria database. The computer conformation for the purpose of this study uses iCore2 processor, 4GB RAM size, 64-bit System and MATLAB 2015a as the executing tools.

IV. RESULT

This study discovers RNA-Seq novelty holding 2457 instances of Mosquitoes Anopheles Gambiae data, with susceptible and resistant genes. PCA algorithm was implemented on the data to diminish the curse of dimensionality.

PCA feature extraction dimensionality reduction detects and eliminate uncorrelated Attributes (Variables), to decide maximum variance with a smaller number of Principal Components.

In this study, PCA is applied on the Mosquito Anopheles data, and gives significant gene information that is useful for further investigation.

Classification algorithm applies Ensemble Adaboost by employing MATLAB tool, to implement the model.

Using PCA as a feature extraction dimensionality reduction method, 1592 features of genes were significant and 45 latent components were achieved in 7.8486 Seconds.

An Ensemble Adaboost classification genomics, 10-folds cross validation was employed to assess the implementation of the performance of the classification models, using 0.05 parameter holdout of data for training and 5% for testing to check the accuracy of the classifiers.

The classifier uses a learning assessment protocol, the training and testing phases are evaluated as a 10-fold cross validation to eradicate the sampling biases. This protocol is implemented using MATLAB. The reported result of valuation is based on the computational time and performance metrics (Accuracy, Specificity, Sensitivity, Precision, F-score and Recall) [25].

This study compares the classification performance of the models, using Adaboost Ensemble classifier, with 93.3% accuracy. The result output and confusion matrix are shown below, in figure 2.

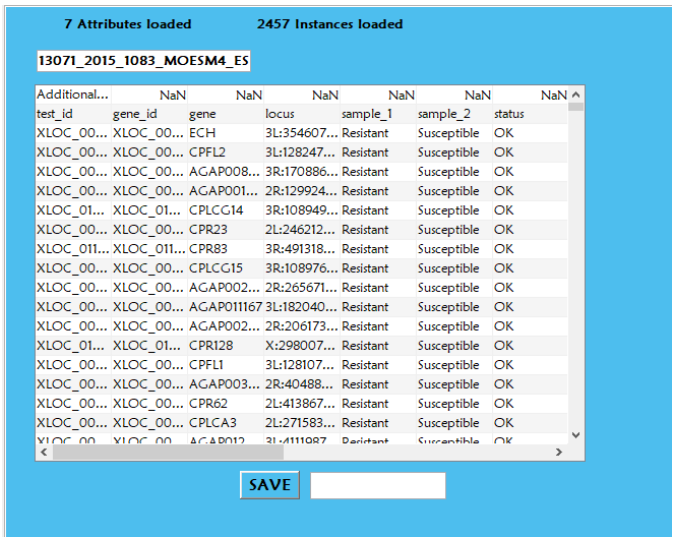

Fig. 1. Mosquito Anopheles Gambiae loaded data on MATLAB Environment.

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 1 (2020), pp. 163-169
 https://dx.doi.org/10.37624/IJERT/13.1.2020.163-169

This study used PCA in fetching latent components from the loaded data in figure one above. The extracted features are passed into ensemble classification and the result is shown in figure 3 below. The confusion matrix gives a solution to the performance metrics.

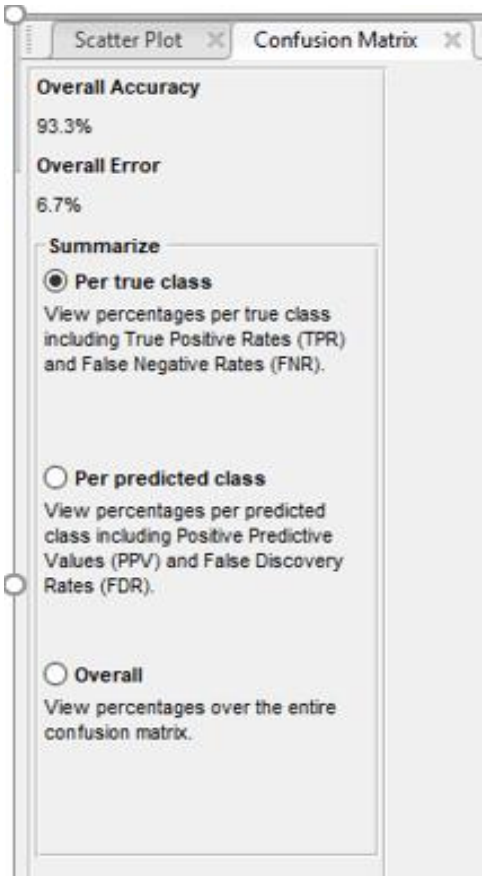

Fig. 2. Overall Accuracy

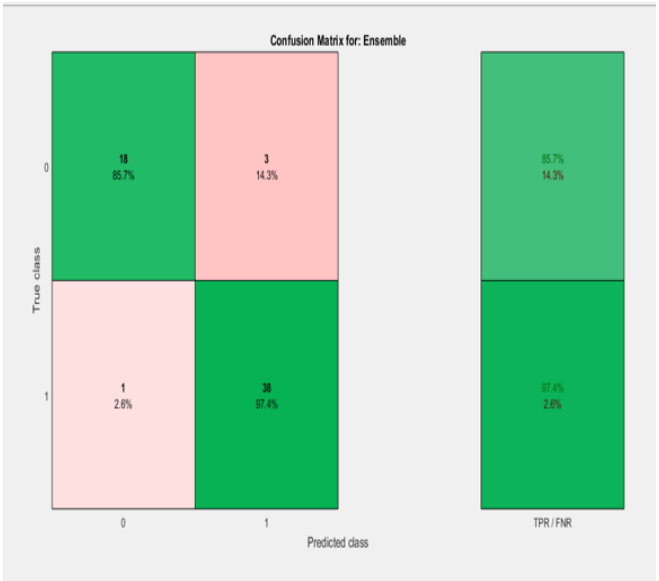

Fig. 3. Confusion Matrix for Ensemble Classification

To test the performance of datamining learning method, RNA-Seq data was downloaded for Mosquito Anopheles Gambiae 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

2457 gene feature were collected, PCA was used as a dimensionality reduction model, 1572 features were extracted with 45 latent components. These components are then classified using Ensemble classification to predict their performance. The result shows the effectiveness of machine learning technology in genes. To validate the approach, the performance results are shown and compared in the table 2 below. The result shows that SVM-polynomial kernel outperforms Gaussian kernel in terms of less training time and accuracy performance.

Table 2. Performance Metrics Table for the Confusion Matrix

Performance Metrics	Ensemble Classification
Accuracy (%)	93.3
Sensitivity (%)	97.4
Specificity (%)	85.7
Precision (%)	92.7
Recall (%)	97.4
F-Score (%)	93.7
ROC Curve (%)	99.6
Training Time (sec)	7.8486

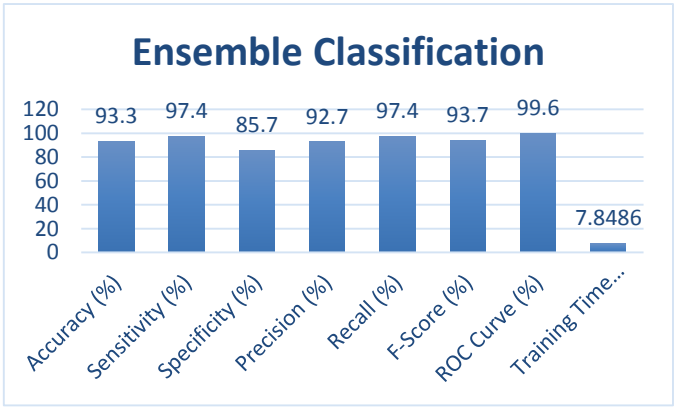

Fig. 4. Performance Metrics Graph

This study analyzed and improved the classification of malaria vector data, several works have been proposed in reviews by researchers using the performance metrics shown in figure 3 above, the results have proven that, dimensionality reduction model using PCA feature extraction methods can improve Ensemble classification output.

V. DISCUSSION

This study improves and can be efficient for the prognosis and diagnosis of malaria ailment in human. The proposed approach uses machine learning techniques such as dimensionality reduction model and classification algorithms.

Dimensionality reduction model uses the feature extraction model PCA and uses the Adaboost Ensemble classifier.

This study performed the analysis and evaluation of the performance and the results obtained were shown.

In future works, feature selection algorithms and other feature extraction methods can be introduced for comparative evaluation and to show if there are other methods that can be used to better the classification performance compared to the-state-of-art.

VI. CONCLUSION

This study analyzed and improved the classification of malaria vector data, several works have been proposed in reviews by researchers using the performance metrics shown in figure 3 above, the results have proven that, dimensionality reduction model using feature extraction methods such as PCA can help improve classification output such as Adaboost Ensemble.

It would be interesting to investigate if recent proposed work can be improved using other ensemble classifiers.

REFERENCES

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

[2] David FR, Kate C, Yank YL, Karine G, Roch L. 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] Arowolo MO, Adebiyi M, Adebiyi A. A Dimensional Reduced Model for the Classification of RNA-Seq Anopheles Gambiae Data. Journal of Theoretical and Applied Information Technology. 2019;97(23):3487-96.

[4] Karthik S, Sudha M. 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)P:182-191

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

[6] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genetics. 2015;16:321–332.

[7] Jagga Z, Gupta D. 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):96–100. doi:10.1038/nature24995

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

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

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

[12] Alquicira-Hernandez, J, Sathe, A, Ji, H.P, Nquyen Q, Powell JE. 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] Cui S, Wu Q, West J, Bai J. 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] Shon HS, Yi YG, Kim KO, Cha EJ, Kim KA. Classification of Stomach Canacer Gene Expression Data Using CNN Algorithm of Deep Learning. Journal of Biomedical Translation Research. 2019:20(1);15-20. doi.org/10.12729/jbtr.2019.20.1.015

[15] Adam JR, Arthur MT, Hayley MB, Ana RG, Mandy JS, Christopher JRI, 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] Tan AC, Gilbert D. Ensemble Machine Learning on Gene Expression Data for Cancer Classification. 2003:2(3);75-83.

[17] Song N, 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] Tarek S, 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] Mariangela B, 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] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with application in R. New York (NY): Springer; 2013.

[21] Nagi S, Bhattacharyya DK. Classification of Microarray Cancer Data Using Ensemble Approach. Network Modeling Analysis in Health Informatics and Bioinformatics. 2013:2;159-173.

[22] Sarah M, Ahmed IS, Labib ML. Classification Techniques in Gene Expression Microarray Data. International journal of Computer Science Mobile Computing. 2018:7(11);52-56

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

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

[25] Arowolo MO, 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] Guzman E, 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] Ren Y, 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] Flennerhag S. Machine Learning Ensemble, (2017). doi:10.5281/zenodo.1042144.

[29] Tsai CF, Hsu YF, Yen DC. A comparative study of classifier ensembles for bankruptcy prediction, Application Soft Computing Journal. 2014:24;977–984. doi:10.1016/j.asoc.2014.08.047

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

[31] Nisioti A, Mylonas A, Yoo PD, Member S, Katos V. From Intrusion Detection to Attacker Attribution : A Comprehensive Survey of Unsupervised Methods. IEEE Commun Surv Tutorials. 2018;PP(c):1.

[32] Hafizah S, Ariffin S, Muazzah N, Latiff A, Khairi MHH, Ariffin SHS, et al. 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):2724–30.