PCA Model For RNA-Seq Malaria Vector Data Classification Using KNN And Decision Tree Algorithm

Micheal Olaolu Arowolo
Department of Computer Science
Landmark University
Omu-Aran, Nigeria
arowolo.micheal@lmu.edu.ng

Marion Adebiyi
Department of Computer Science
Landmark University
Omu-Aran, Nigeria
marion.adebiyi@lmu.edu.ng

Ayodele Adebiyi
Department of Computer Science
Landmark University
Omu-Aran, Nigeria
ayo.adebiyi@lmu.edu.ng

Olatunji Okesola
Department of Computer Science
First Technical University
Ibadan, Nigeria
olatunjiokesola@tech-u.edu.ng

2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS) 978-1-7281-3126-9/20/$31.00 ©2020 IEEE 10.1109/ICMCECS47690.2020.240881

***Abstract*— Malaria parasites adopt unresolved discrepancy of life segments as they grow through various mosquito vector stratospheres. 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 analyzing 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 KNN and Decision Tree classification algorithms. The effectiveness of this experiment is validated on a mosquito anopheles gambiae RNA-Seq dataset. The experiment result achieved a relevant performance metrics with a classification accuracy of 86.7% and 83.3% respectively.**

Keywords—RNA-Seq, PCA, KNN, Decision Tree, Mosquito Anopheles

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 analyzing 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, KNN and Decision Tree classification methods are used to discover distinct biological frameworks and provide higher classification accuracies which can be recommended as an efficient technique for the prediction and detection tests of new genes for malaria.

This paper is structured as follows: Introduction, Related works, Materials and Methods, Results, Conclusion and References.

II. Related WORKS

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 recognized 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 analyzing 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 is current research gaps recognized in analyzing gene expressions [4].

Oh et.al, [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 neighbor 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 et.al, [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 et.al, [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 tumor biopsies, and circulating dendritic cells. They showed scPred classifies discrete cells with high accuracy.

Cui et.al, [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 et.al, [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 analyzed them with CNN deep learning algorithm. They got an accuracy of 95.96% and 50.51%.

Adam et.al, [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 et.al, [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 et.al, [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 et.al, [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 METHODS

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 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 x̄ and data covariance matrix S are as follows.

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

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

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

$$Su_t = \lambda_t u_t u_t^T Su_t = \lambda_t \qquad (3)$$

III.II.II Kth Nearest Neighbours

The gene data are classified using KNN algorithm. K-nearest neighbor is a supervised learning algorithm, the outcome of novel instance query is classified based on common K-nearest neighbor group. KNN algorithm uses locality classification as the estimate value of the new query instance. The purpose of this algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and only based on memory. The selected features are given as an input to this module. The K (number of nearest neighbors) values are chosen that are closest to the query point. The distance between the query-instance and all the training samples are calculated. The distance are then sorted and nearest neighbors based on the Kth minimum distance is determined. The category Y of the nearest neighbors is gathered. The simple majority of the category of nearest neighbors as the prediction value of the query instance is used. Any ties can be broken at random [21].

III.II.III Decision Trees

Decision tree classifiers recursively partition the instance space using hyperplanes that are orthogonal to axes. The model is built from a root node which represents an attribute and the instance space split is based on function of attribute values (split values are chosen differently for different algorithms), most frequently using its values. Then each new sub-space of the data is split into new sub-spaces iteratively until an end criterion is met and the terminal nodes (leaf nodes) are each assigned a class label that represents the classification outcome (the class of all or majority of the instances contained in the sub-space). Setting the right end criterion is very important because trees that are too large can be overfitted and small trees can be underfitted and suffer a loss in the accuracy in both cases. Most of the algorithms have a mechanism built in that deals with overfitting; it is called pruning. Each new instance is classified by navigating them from the root of the tree down lo a leaf, according to the outcome of the tests along the path [22]. Although decision trees produce efficient models, they are unstable – if the training data sets differ only slightly, the resulting models can be completely different for those two sets. Due to that, decision trees are often used in classifier ensembles.

III.II.IV 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. RESULTS

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 algorithms applies KNN and Decision Tree by employing MATLAB tool, to implement the model.

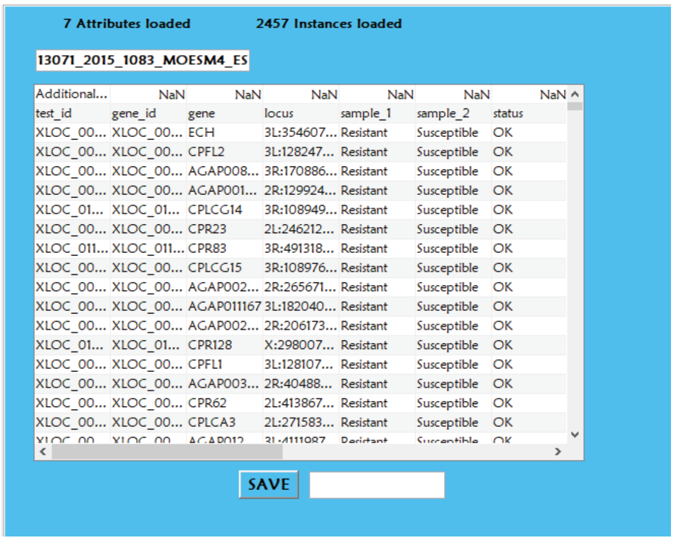
Using PCA as a feature extraction dimensionality reduction method, 1592 features of genes were significant and 45 latent components were achieved in 7.8486 Seconds.

A KNN and Decision Tree classification genomics, 10-folds cross validation were 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 KNN and Decision Tree classifiers, with 86.7 and 83.3% accuracy respectively. The result output and confusion matrix are shown below, in figure 2.



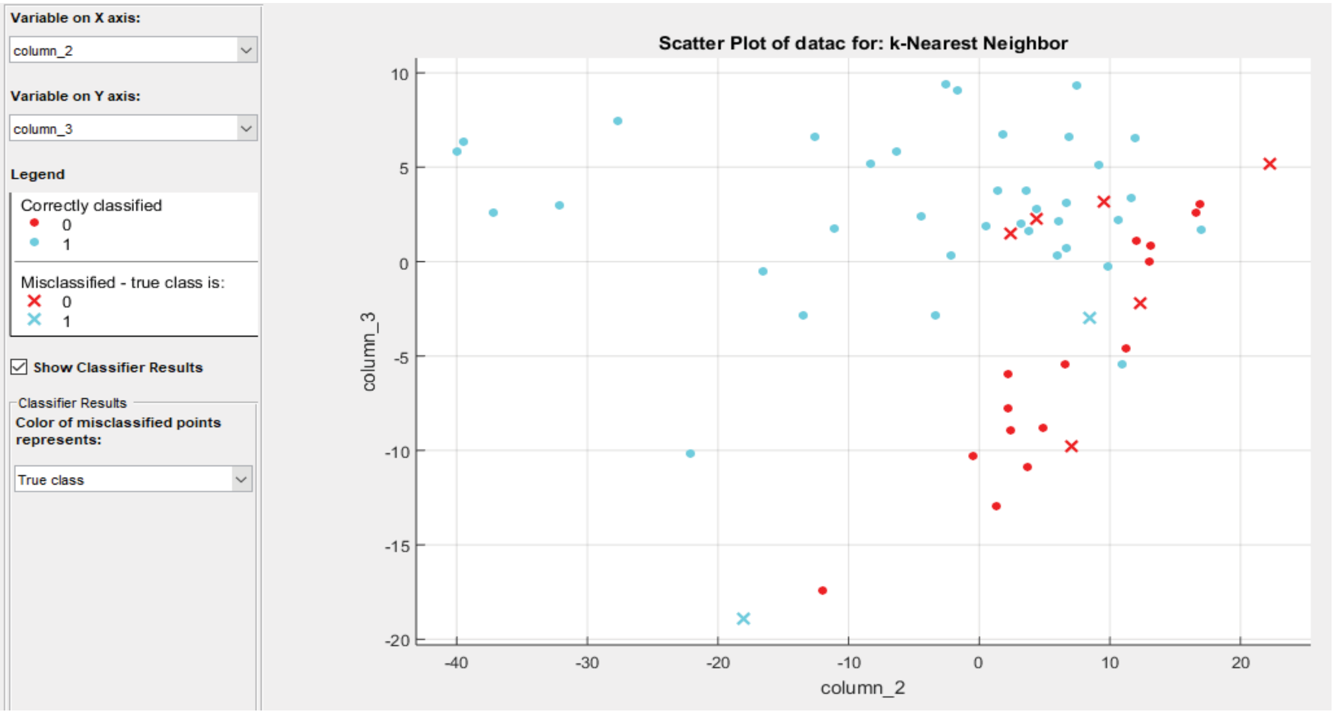
7 Attributes loaded 2457 Instances loaded

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

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

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.



Variable on X axis: column_2

Variable on Y axis: column_3

Legend

Correctly classified
- ● 0
- ● 1

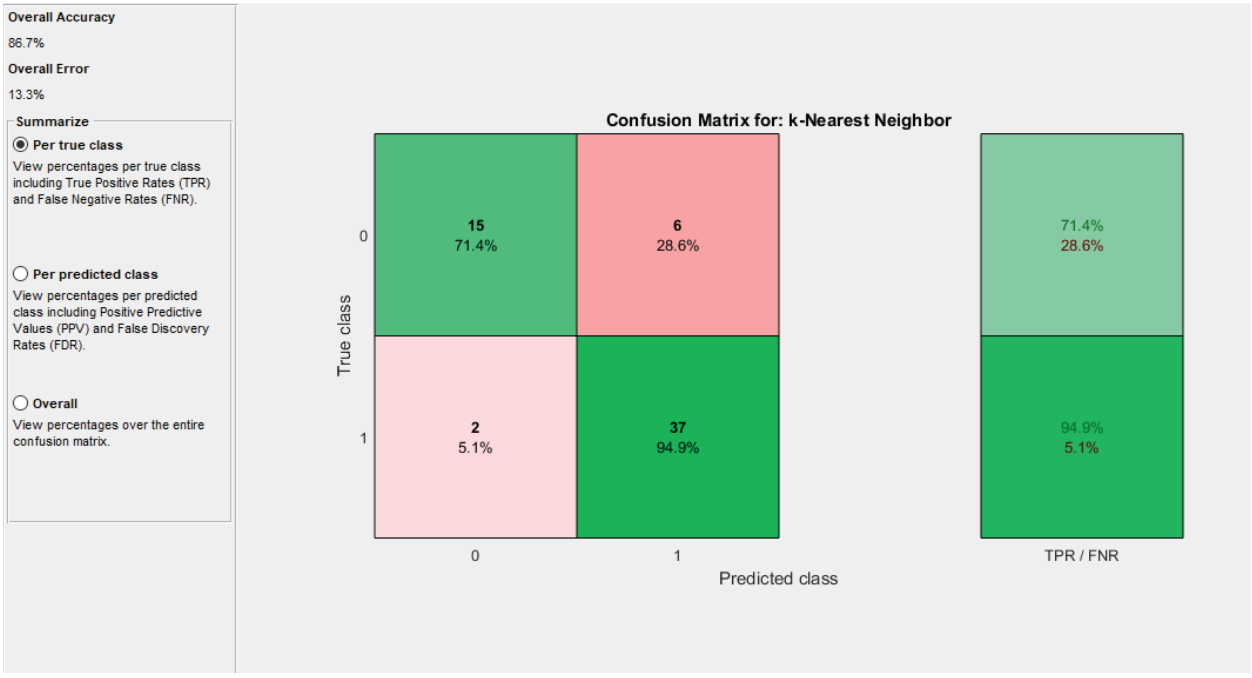
Misclassified - true class is:
- ✕ 0
- ✕ 1

☑ Show Classifier Results

Classifier Results

Color of misclassified points represents: True class

Fig. 2. Mosquito Anopheles Gambiae Data Scattered Plot for KNN Classification on MATLAB Environment.



Overall Accuracy

86.7%

Overall Error

13.3%

Summarize

● Per true class

View percentages per true class including True Positive Rates (TPR) and False Negative Rates (FNR).

○ Per predicted class

View percentages per predicted class including Positive Predictive Values (PPV) and False Discovery Rates (FDR).

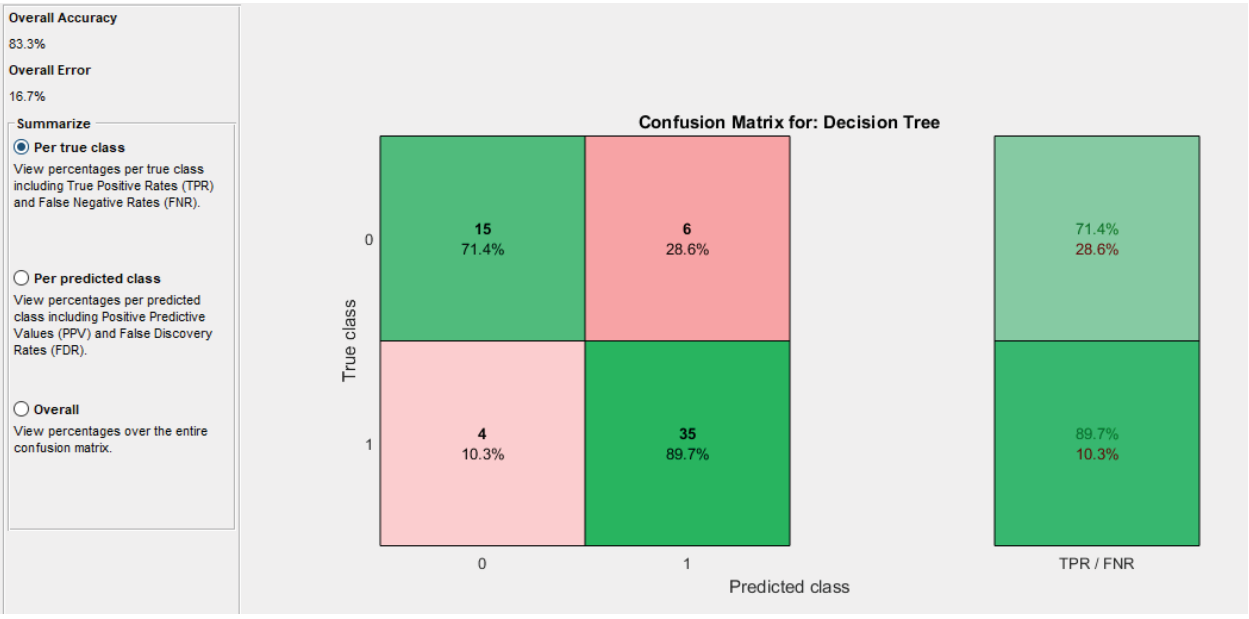
○ Overall

View percentages over the entire confusion matrix.

Confusion Matrix for: k-Nearest Neighbor

True class \ Predicted class	0	1	TPR / FNR
0	15 71.4%	6 28.6%	71.4% 28.6%
1	2 5.1%	37 94.9%	94.9% 5.1%

Fig 3. Overall Accuracy for and Confusion Matrix for the Classification of Mosquito RNA-Seq Data Using KNN TP=37; TN=15; FP=6; FN=2



Overall Accuracy

83.3%

Overall Error

16.7%

Summarize

● Per true class

View percentages per true class including True Positive Rates (TPR) and False Negative Rates (FNR).

○ Per predicted class

View percentages per predicted class including Positive Predictive Values (PPV) and False Discovery Rates (FDR).

○ Overall

View percentages over the entire confusion matrix.

Fig 3. Overall Accuracy for and Confusion Matrix for the Classification of Mosquito RNA-Seq Data Using Decision Tree Classifier
TP=35; TN=15;FP=6; FN=4

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 KNN outperforms Decision Tree in terms of less training time and accuracy performance.

TABLE 2. PERFORMANCE METRICS TABLE FOR THE CONFUSION MATRIX

Performance Metrics	KNN Classification	Decision Tree Classification
Accuracy (%)	86.7	83.3
Sensitivity (%)	94.9	89.7
Specificity (%)	71.4	71.4
Precision (%)	86.1	85.4
Recall (%)	94.9	89.7
F-Score (%)	90.3	87.5
Training Time (sec)	7.8486	7.8486

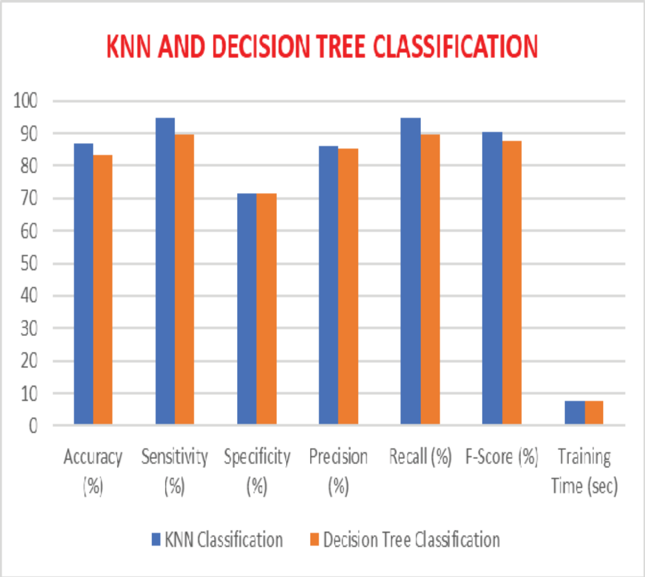


Fig 4. Comparative Chart for The Performance Metrics

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 classification output for KNN and Decision tree.

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 KNN and Decision Tree classifiers. This study performed the analysis and evaluation of the performance and the results obtained were shown, KNN outperforms the Decision Tree Classification algorithm.

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 KNN.

It would be interesting to investigate if recent proposed work can be improved feature extraction models and algorithms.

ACKNOWLEDGMENT

The author will like to thank the ICMECS 2020 for their reviews, the Co-authors and Landmark University Omu-Aran for their supports towards the advancement of this work.

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.

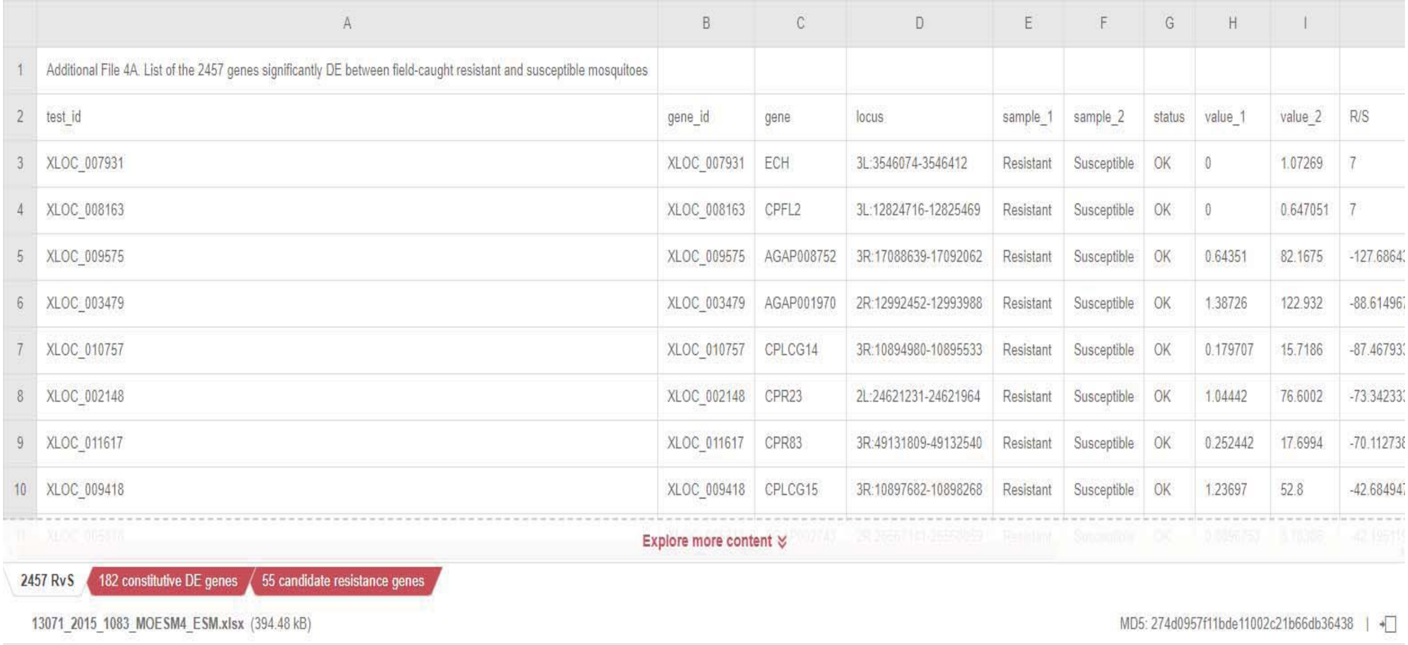
[28] S. Flennerhag. Machine Learning Ensemble, (2017). doi:10.5281/zenodo.1042144.

[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.



	A	B	C	D	E	F	G	H	I	
1	Additional File 4A. List of the 2457 genes significantly DE between field-caught resistant and susceptible mosquitoes									
2	test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	R/S
3	XLOC_007931	XLOC_007931	ECH	3L:3546074-3546412	Resistant	Susceptible	OK	0	1.07269	7
4	XLOC_008163	XLOC_008163	CPFL2	3L:12824716-12825469	Resistant	Susceptible	OK	0	0.647051	7
5	XLOC_009575	XLOC_009575	AGAP008752	3R:17088639-17092062	Resistant	Susceptible	OK	0.64351	82.1675	-127.6864
6	XLOC_003479	XLOC_003479	AGAP001970	2R:12992452-12993988	Resistant	Susceptible	OK	1.38726	122.932	-88.61496
7	XLOC_010757	XLOC_010757	CPLCG14	3R:10894980-10895533	Resistant	Susceptible	OK	0.179707	15.7186	-87.46793
8	XLOC_002148	XLOC_002148	CPR23	2L:24621231-24621964	Resistant	Susceptible	OK	1.04442	76.6002	-73.34233
9	XLOC_011617	XLOC_011617	CPR83	3R:49131809-49132540	Resistant	Susceptible	OK	0.252442	17.6994	-70.11273
10	XLOC_009418	XLOC_009418	CPLCG15	3R:10897682-10898268	Resistant	Susceptible	OK	1.23697	52.8	-42.68494

Explore more content

13071_2015_1083_MOESM4_ESM.xlsx (394.48 kB)

MD5: 274d0957f11bde11002c21b66db36438

Fig 5. Data Sample For Mosquito Anopheles Gambiae