Indonesian Journal of Electrical Engineering and Computer Science
Vol. 21, No. 2, February 2021, pp. 1073~1081
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v21.i2.pp1073-1081

Predicting RNA-Seq data using genetic algorithm and ensemble classification algorithms

Micheal Olaolu Arowolo[1], Marion O. Adebiyi[2], Ayodele A. Adebiyi[3] and Olatunji J. Okesola[4]
[1,2,3]Department of Computer Science, Landmark University, Omu-Aran, Kwara State, Nigeria
[4]Department of Computer Science, First Technical University, Ibadan, Nigeria

Article Info

Article history:

Received Feb 12, 2020
Revised Apr 13, 2020
Accepted May 3, 2020

Keywords:

Ada boost ensemble
Bagging ensemble
Genetic algorithm
Malaria vector
RNA-Seq

ABSTRACT

Malaria parasites accept uncertain, inconsistent life span breeding through vectors of mosquitoes stratospheres. Thousands of different transcriptome parasites exist. A prevalent Ribonucleic acid sequencing (RNA-seq) technique for gene expression has brought about enhanced identifications of genetical queries. Computation of RNA-seq gene expression data transcripts requires enhancements using analytical machine learning procedures. Numerous learning approaches have been adopted for analyzing and enhancing the performance of biological data and machines. In this study, a Genetic algorithm dimensionality reduction technique is proposed to fetch relevant information from a huge dimensional RNA-seq dataset, and classification uses Ensemble classification algorithms. The experiment is performed using a mosquito Anopheles gambiae dataset with a classification accuracy of 81.7% and 88.3%.

Corresponding Author:

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

1. INTRODUCTION

Next-generation sequencing technology high-throughput has produced large wide-ranging datasets, this gigantic expanse of data authorizes biologists in examining and realizing challenging transcriptions of genes, for example, relatives in diseases and RNA for example contagions (malaria), cancer, transmissible, genetics, biological, and others [1]. Blood-sucking mosquitoes such as mosquito anopheles with key vectors of malaria Plasmodium falciparum originates from Africa. Anopheles mosquitoes are a deadly malaria parasite, responsible for demises of thousands. Antimalarial suppositories spread, state-of-the-art antimalarials treatment upsurges, fetching for ground-breaking medications requires improved consideration of these living organisms. Mosquito parasite tolerates precise parameter of expression of genes takes a massive query, making an enhanced systematic predictive model for malaria vector transcripts [2-3]. Approachable revealing genetic inquiries have been made in RNA-Seq study by unfolding a cautious purposeful biological strategy by enhancement of sequencing study. RNA-Seq data necessitates the removal of the high-dimensionality curse, such as; disorders, sounds, recurrence, unconnected, severance, unsuitable data, and others [4]. Current skills consist of enhanced methods in developing ground-breaking medical care models, for example, keen human well-being treatment systems, enhanced treatments, among other detects of ailments and complaints [5].

Some machine learning approaches have been conventional in the recent era with persuasive novelties for studying the enormous sum of the RNA next-generation sequencing gene expression data over

studying the biologically applicable outlines [6]. Researchers have extensively worked on machine learning methods for RNA-Seq data expression having rates of success variable [7, 8]. Computational approaches have been useful on an enormous genomic dataset of publics diseases, genes in charge of the presence of conditions can be distinguished. Several measures have been observed by Differentially Expressed Genes (DEG). In identifying the difference between genes contracted from the human genome, the machine learning process is vital. Quite a lot of machine learning approaches proposed in analyzing and classifying gene expression profiling of many emulated diseases. Profiling of gene expression data and its approaches utilizing some machine learning are of significant. A lot of investigations have been carried out, with existing investigational predictable openings [5]. Blood-based gene expression disease signs and machine learning, to detect transcriptions for classification [9], with RNA data from omnibus gene expression data with machine learning tools and algorithms are projected. RNA-Seq data dimensionality reduction, clustering and classification have been proposed, directing mutual evaluation, the importance of procedures, occurring recently as predominant variations, with direct and indirect approaches with RNA-seq data dimensionality reduction methods, by the statement of RNA-seq data dimensionality reduction procedures [10].

A Genetic algorithm dimensionality reduction feature selection procedure is carried out, to draw and analyze high dimensional gene expression data, Ensemble classification algorithm approaches are carried out to regulate discrete genetic backgrounds that distributes classification accuracies which suggestable for effective prediction and detection approaches of innovative genes for malaria contagions in human.

2. METHOD PROPOSED

The planned framework for this study is tabulated in Figure 1, vital knowledge in predicting machine learning burden on high-dimensional gene expression RNA-Seq data, into subordinate dimensional dataset is proposed. This study fetches out imperative data in a specified dataset by employing Genetic Algorithm feature selection process as a phase, Ensemble classification algorithms are linked to estimate the performance of the RNA-seq malaria vector dataset.

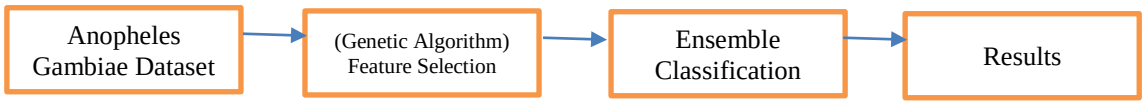


Figure 1. Proposed framework

A supervised learning procedure for RNA-Seq gene ranking huge ensembles group of measured RNA-Seq genes, carried out a mutable rank procedure made from random forests classification algorithm, using Autoencoder variations and regressors to abstract levels of 12 RNA-Seq cancerous datasets holding about a thousand samples. A concealed supervised learning-based feature selection procedure in RNA-Seq training was demonstrated and conferred using feature selection approaches on the gene expression dataset investigation [11]. A supervised classification RNA-Seq data model was presented using a simplified procedure with an infinite accurate classification of single cells, merging independent feature selection dimensional reduced model and machine learning procedure. Sc-Pred RNA-seq dataset from pancreatic muscle, mixing dendritic cells, colorectal tumour material elimination, and mononuclear cells were applied and presented a high-performance accuracy [12]. RNA-DNA machine learning investigation showing low genome expressions influencing PAH ailment was proposed, using an advanced feature selection and enhanced machine learning procedure for classifying irrelevant but very beneficial genes, the results displayed clusters of unrelated expression genes that reveal predicting and distinctive transformed PAH [13]. Classification of gene expression gastrointestinal tumor dataset using deep learning approach was proposed, using about 60,000 genes from 334 gastrointestinal tumor patient's data, PCA, heatmaps, and the CNN algorithm were proposed using scientific, and RNA-seq gene expression data investigation and classification accuracy of 95.96% and 50.51% were achieved [14]. An RNA-Seq disclosure of concealed transcriptions in malaria parasites was proposed by unfolding the dissimilarity of an RNA-seq process to free difference of transcripts for different mosquitos and revealed hidden distinct transcriptional signs [15].

Classification with ensemble machine learning procedure for cancerous data expression was proposed using C4.5, bagging and boosting ensemble supervised machine learning measures cancer data classification on seven open-sourced malicious microarray data, the bagging and boosting ensemble learning classification approach showed a better performance accuracy [16]. An investigative ensemble classification

method for gene expression for cancerous data was proposed using a Recursive Feature Elimination association feature selection approach to fetch important features, an Adaboost ensemble classification algorithm was used for classification, and the outcome displayed a relevant improvement. Cancerous gene expression data classification was done using an ensemble classification method; the performance and outcomes of the result showed a reduced amount of dependent on originalities of a single training dataset [17]. A metaheuristics technique for fetching genes and RNA/DNA data classification by briefing existing advances of metaheuristic-based methods in the embedded technique of feature selection approach was proposed, emphasizing helpful and integrating problem-specific data relevance into the examination operatives of developments. A ranking coefficient of linear SVM classifier was used in the local operative investigation for feature selection and classification [18]. A fault investigation for training engines using GA and classification learners, the approach lessens the computational complication and advances the accuracy to about 97% [19]. Tree model enhancement for classifying certain ensembled features was proposed using an ensemble-based feature selection, random trees and wrapper-based feature selection system in developing a classification model, and the ensemble data classification procedure initiates a subclass using the bagging, wrapper dimensionality reduction method, and random trees. This procedure removes the unconnected features and picks the best features for classification with a probability weighting value. The study was evaluated and compared with a classification accuracy of 92% [20]. An ensemble-feature selection implementation procedure using R-package tool was proposed, several feature selection techniques were combined with regularized outputs to a quantifiable ensemble ranking, feature selection procedures were combined, and used [21].

3. RESEARCH METHOD

High dimensional data investigations have been discussed extensively, a Genetic Algorithm and Ensemble classification algorithm is proposed using an RNA-Seq data consisting of 2457 instances with seven attributes of western Kenya, mosquito's gene data [22] with its profile transcript contents, RNA-Seq genes, transcript variations of deltamethrin- resistant and vulnerable Anopheles gambiae Kenyan mosquitoes which is an openly accessible data on figshare.com [23-24], it is tabulated in Table 1. MATLAB experimental tool is used to carry out the experiment, GA is proposed and used to fetch relevant features. The selected were classified using the Ensemble algorithm [25].

Table 1. Dataset structures

Dataset	Attributes	Instances
Mosquito Anopheles Gambiae	7	2457

3.1. Genetic algorithm

GA is a proficient method for investigating suitable features from high dimensional datasets, and predominant GA are wrapper-based feature selection methods. Quite a lot of limitation procedures for genetic algorithm exists, where alteration and crossover operatives persist and commonly connected to binary constraint values. Appropriate features are recognized using a genetic algorithm [26]. The RNA has *N* number of features representing features with values 0 and 1 as selected and unselected, correspondingly. Addressing the importance of features, GA is used in finding the ideal feature subset by means of the nominated figure of features for complex classification presentation. The general construction of the GA is defined in Algorithm 1 below by adopting [27]:

Algorithm 1. Genetic algorithm

```
Require: Initialize the parameters nPop = m, t_max, t = 0;
Ensure: Optimal feature subset with the highest fitness value.
       1: while (t<=t_max) do
       2:     Create pop m, t_max;
       3:     For k = 1 to m do
       4:           Parents [m_1, m_2] = system selection (m, nPop)
       5:           Child = Xor[m_1, m_2]
       6:           M u = mutation [Child}
       7:     End for
       8:     Replace m with Child_1, Child_2, …, Child_m
       9:     t = t+ 1;
       10:    End while
       11:    Store the Highest fitness value;
```

m is the population size, *r* is a random number lying flanked by 0 to 1, signifies the nominated chrome or unselected feature with a threshold δ set value to be 0.5, and α is the threshold number of features nominated. The significant problems of the precise method are selecting the maximum fitting features from the predictable datasets.

3.2. Ensemble classifier

Ensemble classifiers are trained using distinct sectors of the training data, diverse constraints of the classifiers, or varied sectors of features as in a model of random subspace [28]. Ensemble classifier includes integrating outcomes of numerous classifiers to yield a final result; it is regularly used for the acquisition of extremely accurate results. Ensemble classifiers are quite mutual in machine learning problems and can be active in the bioinformatics field. The classification result is attained by the inclusion of a choice of individual classifier [29]. Ensemble approaches are 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, multiple integration classifiers, classifier synthesis, grouping, committee, and so on. Ensemble classifier takes complete improved presentation than discrete based classifiers. The efficiency of ensemble approaches is extremely dependent on the unconventionality of fault devoted by the discrete learner. Ensemble approaches performance hinge on the accuracy and variety of the base learners, and ensemble classification has common techniques; bagging and boosting.

Bagging (bootstrap aggregating) employs the training data by arbitrarily changing the unique *T* training data by *N* items. The additional training sets are called bootstrap duplicates with some occurrences unappealing even though appearing consecutively. The classifier *C**(*x*) is built by combining *Ci*(*x*) where each *Ci*(*x*) has an equivalent vote.

AdaBoost (Adaptive Boosting) technique affects the training data. Originally, the procedure allows all instance *xi* with equal weight. In separate iteration *i*, the knowledge procedure reduces the weighted error on the training set and yields a classifier *Ci*(*x*). The weighted error of *Ci*(*x*) is calculated with use to inform the weights on the training instances *xi*. The weight of *xi* rises, giving to its effects on the classifier's outcome that allows a high weight for a misclassified *xi* and a small weight for an acceptably classified *xi*. The concluding classifier *C**(*x*) is built by a weighted vote of the discrete *Ci*(*x*) rendering to its accuracy built on the weighted training set [30-33].

Implementing 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 to construct D_t such that $D_1(i) = \frac{1}{m}$ given D_t and h_t:

$$D_{i+1}\{i\} = \frac{D_t(i)}{Z_t} X \begin{cases} e^{-\alpha_t} if y_i = h_t(x_i) \\ e^{\alpha_t} if y_i \neq h_t(x_i) \end{cases} \quad (3)$$

$$= \frac{D_t(i)}{Z_t} \exp(-\alpha y_i h_t(x_i)) \quad (4)$$

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

$$\alpha_t = \frac{1}{2} in(\frac{1-\epsilon_t}{\epsilon_t}) \quad (5)$$

Basic ensemble classification techniques:
Weighted Averaging (WA); Averaging and Max Voting (MV).
Max Voting (MV) exists [31] Ensemble learning has three combination advanced techniques; Stacking (STK); Blending (BLD); Bagging (BAG), and Boosting (BOT) [32-37].

3.3. Performance evaluation

Performance evaluation of machine learning technique entails validation metrics such as a confusion matrix, used for analyzing classification models features, discovering the classified illustrations from the given model of tested dataset model samples [5] using the performance metrics formulas [22, 27].

3.4 Applications

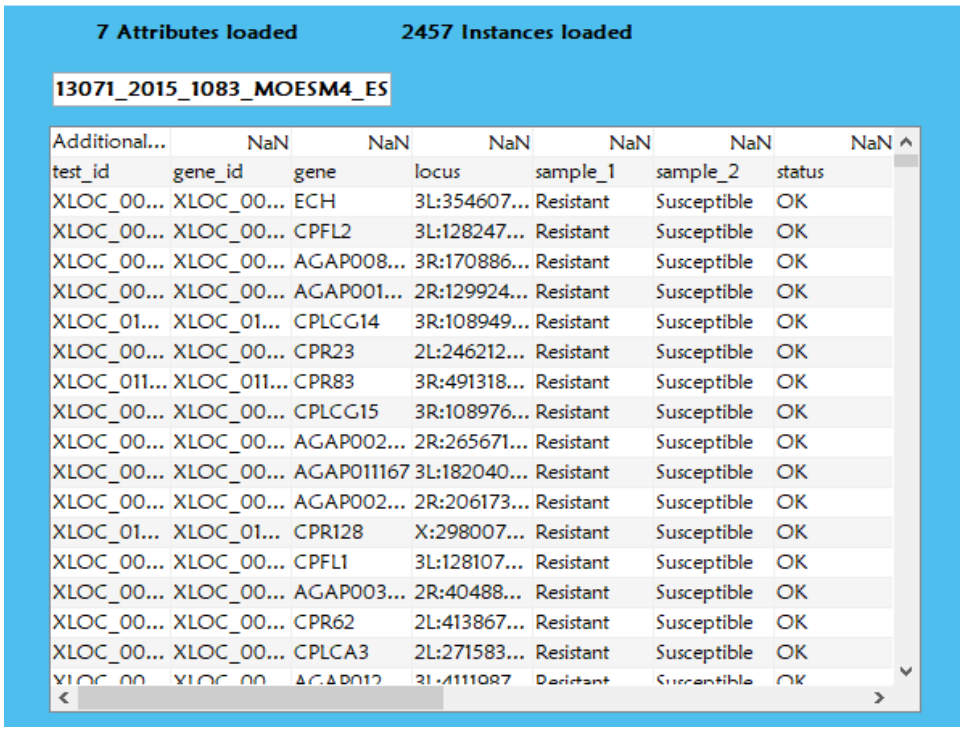
Gene analysis expression projects an improved approach in identifying RNA-Seq data, fetching for relevant essential genes for developing applications like treatments, genes and drugs discoveries, diagnosis, classification of cancerous diseases, malaria, fever, and so on. Finding the machine learning data designs

requires a great algorithm and tools used by several experiments. MATLAB tool is used to carry out the experiment [38-39]. Predicting RNA-Seq technology using MATLAB tool, malaria vector data, and computer resolution conformation uses iCore2 processor, 8GB RAM size, 64-bit System and MATLAB 2015 tool.

4. RESULTS AND DISCUSSION

RNA-Seq innovation with Mosquitoes Anopheles Gambiae data having 2457 susceptible and resistant genes as shown in Figure 2 below is implemented by using Genetic algorithm on the data to reduce the curse of dimensionality and fetch the optimal subset of data, remove uncorrelated attributes, and choose determined variance with a reduced number of subset features in the variable. The GA gives important gene data for a suitable study. The ensemble classification algorithm is used. Using GA as a feature selection method, with a threshold of 0.5, 708 optimal subset features of genes were significant.

The classifier uses an ensemble classification learning evaluation procedure, the training and testing segments use 10-fold cross-validation for eliminating selection partialities using MATLAB. Evaluation outcome is constructed using the computational time and performance metrics [27]—classification performance with Ada-Boost and Bagging Ensemble classification models, with 93.3% and 95% accuracy respectively. The result procedures are shown in Figures



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

Figure 2. Loaded data on MATLAB environment

GA is employed to fetch related components from the dataset, as shown in figure two, selected features are classified using an ensemble algorithm. The result of the confusion matrix is shown in the figures beneath. The confusion matrix shows a resolution to the performance metrics Ada-Boost and bagged ensemble classification algorithms are used and achieves an accuracy of 81.7% and 88.3% respectively. Figure 3 and Figure 4 shows the confusion matrices used in evaluating the performance of the experiment; it comprises of the True and False Positives and Negatives.

Testing the performance of RNA-Seq data [39], with 2457 gene features, GA was employed to eliminate irrelevant features in the data, 708 features were carefully chosen as a subset in the data. The selected features are passed into the ensemble classification model envisage their performance. The outcome proves the efficiency of machine learning ability on genes, validating the method, the results revealed in Table 2 with Bagged Ensemble outperforming the Adaboost ensemble algorithm in terms of accuracy with 88.3% to 81.7%.

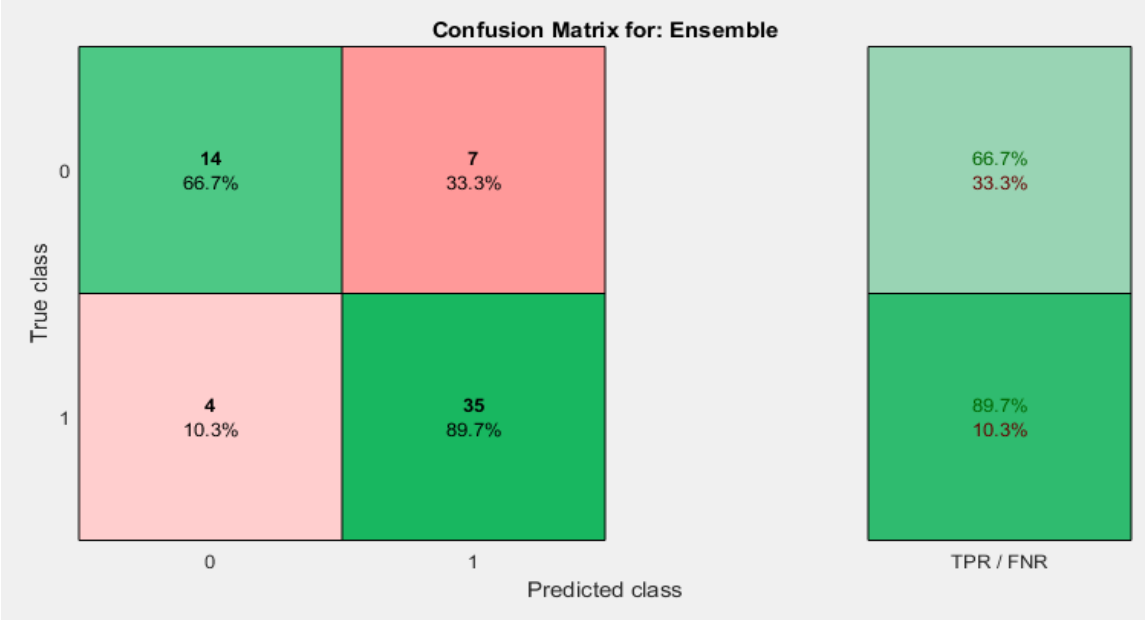


Figure 3. Confusion matrix for classifying mosquito RNA-Seq data with ada-boost ensemble classifier
TP=35; TN=14; FP=7; FN=4

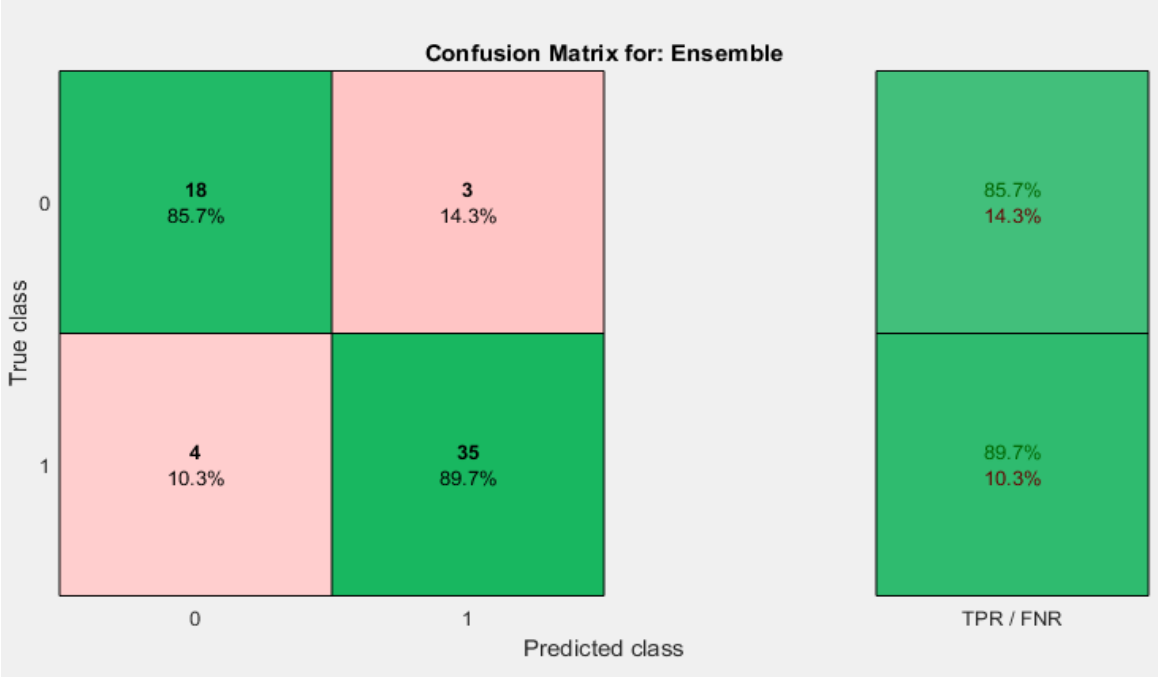


Figure 4. Confusion matrix for the classification of mosquito RNA-Seq data using bagged ensemble classifier
TP=35; TN=18; FP=3; FN=4

Table 2. Performance metrics table for the confusion matrix

Performance Metrics	Ada-Boost Ensemble Classification	Bagged Ensemble Classification
Accuracy (%)	81.7	88.3
Sensitivity (%)	89.7	89.7
Specificity (%)	90.6	85.7
Precision (%)	83.3	92.1
Recall (%)	89.7	92.1
F-Score (%)	86.4	92.1

5. CONCLUSION

In this study, an improved and efficient prediction and analysis of malaria ailment in human is carried out using machine learning procedures such as genetic algorithm and Ensemble algorithms. This study analyzed and evaluated the performance, and the showed the obtained results of the employed Classification algorithms, the bagged ensemble classifier outperforms the Ada-boost. The improved classification of malaria vector data is carried out with other numerous works, and the results show dimensionality reduction model with Genetic Algorithm feature selection approach, is helpful and can advance the classification results such as ensemble. Investigating further on feature selection models and algorithms will be of great value to get a suitable model for enhancing RNA-Seq technology.

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*, vol. 19, no. 1, pp. 40-48, 2019. doi: 10.1093/bfgp/elz036.

[2] FR David, C Kate, Y L Yank, G Karine, and L Roch. "Predicting Gene Expression in the Human Malaria Parasite Plasmodium Falciparum Using Histone Modification, Nucleosome Positioning, and 3D Localization Features," *PLOS Computational Biology*, vol. 15, no. 9, 2019, doi: 10.1371/journal.pcbi.1007329.

[3] Anopheles gambiae 1000 Genomes Consortium; Data analysis group; Partner working group; "Genetic diversity of the African malaria vector Anopheles gambiae," *Nature*, 2017. doi: 10.1038/nature24995

[4] 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*, vol. 97, no. 23, pp. 3487-96, 2019.

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

[6] NT Johnson, A Dhroso, KJ Hughes, D Korkin, "Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers?" *RNA*, vol. 24, no. 9, pp. 1119–1132, 2018. doi: 10.1261/rna.062802.117.

[7] MW Libbrecht, and WS Noble. "Machine learning applications in genetics and genomics," *Nat Rev Genetics*, vol. 16, no. 6, pp. 321–332, 2015.

[8] Z Jagga, and D Gupta, "Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms," *BMC Proceedings*, vol. 8, no. 2, pp. 1-9, 2014.

[9] Q Ren, M. Anjun, M. Qin, Z. Quan. "Clustering and classification methods for Single-cell RNA-Seq data," *Briefings in Bioinformatics*, pp.1-13, 2019.

[10] W. Stephen, S. Ruhollah. "Using supervised learning methods for gene selection in RNA-Seq case-control studies. frontiers in genetic," *Bioinformatics and Computational Biology*, vo. 9, pp. 1-6, pp. 1-6, 2018. doi: 10.3389/fgene.2018.00297

[11] 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*, vol. 20, no. 1, pp. 1-17, 2019, doi: 10.1186/s13059-019-1862-5

[12] S. Cui, Q. Wu, J. West, and J. Bai. "Machine learning-based microarray analyses indicate low-expression genes might collectively influence PAH disease," *PLOS Computational Biology*, vol. 15, no. 8, 2019. doi: 10.1371/journal.pcbi.1007264

[13] 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*, vol. 20, no. 1, pp.15-20, 2019. doi: 10.12729/jbtr.2019.20.1.015

[14] 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*, vol. 7, 2018. doi: 10.7554/eLife.33105

[15] A.C. Tan, Gilbert D., "Ensemble machine learning on gene expression data for cancer classification," *Ensemble Machine Learning on Gene Expression Data for Cancer Classification*, vol. 2, no. 3, pp. 75-83, 2003.

[16] 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*, vol. 5, no. 1, pp. 1-7, 2016. doi:10.4172/2169-0111.1000152

[17] S. Tarek, Elwahab RA, Shoman M. "Gene expression based cancer classification," *Egyptian Informatics Journal*, vol. 18, pp. 3, pp. 151-159, 2017, doi: 10.1016/j.eij.2016.12.001.

[18] RN Toma, AE Prosvirin, and JM Kim, "Bearing fault diagnosis of induction motors using a genetic algorithm and machine learning classifiers," *Sensors Basel*, vol. 20, no. 7, 2020.

[19] A.C. Tan, Gilbert D., "Ensemble machine learning on gene expression data for cancer classification," *Applied Bioinformatics*, vol. 3, pp. 1-10, 2003.

[20] U Neumann, N Genze and D Heider. "EFS: An ensemble feature selection tool implemented as r-package and web application," *BioData Min*, vol. 10, no. 1, pp. 1-9, 2017.

[21] 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*, vol. 22, no. 2, pp. 29-38, 2018.

[22] MO Arowolo, MO Adebiyi, AA Adebiyi and O Okesola. "PCA model for RNA-Seq malaria vector data classification using KNN and decision tree algorithm*," 2020 International Conference in Mathematics, Computer Engineering and Computer Science*, pp. 1-8, 2020. doi: 10.1109/ICMCECS47690.2020.240881.
[23] 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*, vol. 8, no. 1, pp. 1-13., 2015. doi: 10.1186/s13071-015-1083-z.
[24] Mariangela Bonizzoni, Eric Ochomo, William Dunn, Monica Britton, Yaw Afrane, *et al.*, "Additional file 4: of RNA-seq analyses of changes in the Anopheles gambiae transcriptome associated with resistance to pyrethroids in Kenya: identification of candidate-resistance genes and candidate-resistance SNPs," 2015. [Online]. Available: 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
[25] G. James, Witten D, Hastie T, Tibshirani R., "An introduction to statistical learning with application in R," New York NY: Springer, 2013.
[26] B Duval and J-K Hao, "Advances in metaheuristics for gene selectio and classification of microarray data," *Briefings in Bioinformatics*, vol. 11, no. 1, pp. 127-141, 2010.
[27] M.O. Arowolo, S.O., Abdulsalam, R.M. Isiaka and K.A. Gbolagade. "A hybrid dimensionality reduction model for classification of microarray dataset," *International Journal of Information Technology and Computer Science*, vol. 9, no. 11 pp. 57-63, 2013.
[28] Nagi S, Bhattacharyya DK. "Classification of microarray cancer data using ensemble approach," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 2, pp. 159-173, 2013.
[29] Sarah M, Ahmed IS, Labib ML., "Classification techniques in gene expression microarray data," *International journal of Computer Science Mobile Computing*, vol. 7, no. 11, pp. 52-56, 2018.
[30] Tan AC, Gilbert D. "Ensemble machine learning on gene expression data for cancer classification," *Applied Bioinformatics*, vol. 3, pp. 1-10, 2003.
[31] 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*, pp. 771–776, 2015. doi:10.1109/ASE.2015.88.
[32] Ren Y, Suganthan PN, Srikanth N. "Ensemble methods for wind and solar power forecasting: A state-ofthe-art review," *Reneweable Sustainable Energy Revolution*, vol. 50, no. 4, pp. 82-91, 2015. doi: 10.1016/j.rser.2015.04.081.
[33] Flennerhag S. Machine Learning Ensemble, 2017. doi:10.5281/zenodo.1042144.
[34] Kamran K, Kiana JM, Mojtaba H, Sanjana M, Laura B, Donald B. "Text Classification Algorithms: A Survey," *Information MDPI*, vol. 10, no. 4, pp. 62-68, 2019.
[35] Tsai CF, Hsu YF, Yen DC. "A comparative study of classifier ensembles for bankruptcy prediction*," Application Soft Computing Journal*, vol. 24, pp. 977–984, 2014, doi: 10.1016/j.asoc.2014.08.047
[36] Mayr A, Binder A, Gefeller O, Schmid M. "The evolution of boosting algorithms from machine learning to statistical modelling," *Methods Informatics and Medicine*, vol. 53, no. 6, pp. 419-427, 2014.
[37] 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*, vol. 20, no. 4, pp. 3369-3388. 2018.
[38] 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.*, vol. 8, no. 2, pp. 2724-30, 2018.
[39] DH Oh, IB Kim, SH Kim, and DH Ahn, "Predicting Autism Spectrum Disorder Using Blood-based Gene Expression Signatures and Machine Learning," *Clin Psychopharmacology Neuroscience*, vol. 15, no. 1, pp. 47-52, 2017. doi:10.9758/cpn.2017.15.1.47

BIOGRAPHIES OF AUTHORS

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

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

Professor Adebiyi, Ayodele Ariyo, 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 an MBA degree from University of Ilorin, Ilorin Nigeria. He had his M.Sc and PhD degree in Management Information System (MIS) from Covenant University, Nigeria, respectively. His research interests include the application of soft computing techniques in solving real-life problems, software engineering and information system research. He has successfully mentored and supervised several postgraduate students at Masters and PhD level. He has published widely in local and international reputable journals. He is a member of Nigerian Computer Society (NCS), the Computer Registration Council of Nigeria (CPN) and IEEE member.

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 PhD in Computer Sciences. He is a member of the 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 the University of South Africa. His research interests include Cybersecurity, biometrics, and Software engineering. He has several publications in scholarly journals and conference proceedings both local and international.