WALAILAK JOURNAL
Applied Sciences
http://wjst.wu.ac.th https://doi.org/10.48048/wjst.2021.9849

An Adaptive Genetic Algorithm with Recursive Feature Elimination Approach for Predicting Malaria Vector Gene Expression Data Classification using Support Vector Machine Kernels

Micheal Olaolu AROWOLO[1,*], Marion Olubunmi ADEBIYI[1], Chiebuka Timothy NNODIM[2], Sulaiman Olaniyi ABDULSALAM[3] and Ayodele Ariyo ADEBIYI[1]

[1]Department of Computer Science, Landmark University, Omu-Aran, Kwara State, Nigeria
[2]Department Department of Mechanical Engineering, Landmark University, Omu-Aran, Kwara State, Nigeria
[3]Department of Computer Science, Kwara State University, Malete, Nigeria

([*]Corresponding author's e-mail: arowolo.micheal@lmu.edu.ng)

Received: 13 April 2020, Revised: 20 July 2021, Accepted: 27 July 2021

Abstract

As mosquito parasites breed across many parts of the sub-Saharan Africa part of the world, infected cells embrace an unpredictable and erratic life period. Millions of individual parasites have gene expressions. Ribonucleic acid sequencing (RNA-seq) is a popular transcriptional technique that has improved the detection of major genetic probes. The RNA-seq analysis generally requires computational improvements of machine learning techniques since it computes interpretations of gene expressions. For this study, an adaptive genetic algorithm (A-GA) with recursive feature elimination (RFE) (A-GA-RFE) feature selection algorithms was utilized to detect important information from a high-dimensional gene expression malaria vector RNA-seq dataset. Support Vector Machine (SVM) kernels were used as the classification algorithms to evaluate its predictive performances. The feasibility of this study was confirmed by using an RNA-seq dataset from the mosquito Anopheles gambiae. The technique results in related performance had 98.3 and 96.7 % accuracy rates, respectively.

Keywords: RNA-seq, Adaptive genetic algorithm, Recursive feature elimination, Malaria vector, Support Vector Machine kernels

Introduction

The rising of next-generation sequencing systems has brought in a huge volume of data that helps scientists to analyze and discover challenging sequences of genomes, including relations in pathogens and Ribonucleic acid (RNA), including certain diseases e.g., malaria, tumors, biological hormones, neurological, among many others [1].

Mosquitoes that feed on blood, e.g., mosquito Anopheles, are the primary vectors of Plasmodium falciparum malaria. Anopheles mosquitoes carry the malaria parasite, which has killed thousands of people. As the fight against antimalarial suppositories intensifies and demand for state-of-the-art, antimalarials medication rises, the quest for ground-breaking drugs necessitates a good biological knowledge of these species. The question on how the mosquito anopheles parasite embraces unique gene expression parameters has become a huge one, necessitating the creation of a more detailed inductive model for malaria vector transcriptions [2,8].

By enhancing the sequencing sample, it is possible to make accessible unveiling genetic investigations in the RNA-seq analysis by unveiling a careful meaningful biological approach. The curse

of high-dimension data, such as sounds, disorders, duplication, meaningless, redundancy, and unfitting data, must be removed from RNA-seq data [3]. Current capabilities have developed strategies to designing ground-breaking healthcare models, involving sensitive public health clinical practice, innovative procedures, as well as other illness and ailment evaluations, among many other issues [4].

However, many artificial intelligence techniques with insightful innovations have recently been developed for analyzing the huge amount of the next-generation sequencing RNA gene expression sequencing data by training clinically relevant structures, such as Random Forest, J48, SMO, Naïve Bayes, among others [5]. In the clinical therapy of malaria patients, progress is also being made toward customized and predictive medicine. Machine learning techniques are widely used for prognosis and diagnosis to assess various and complex cancer-related data. Many researchers have tried to use data analysis techniques to investigate RNA-seq gene expression data, in varying levels of success [6,7].

This research attempts to establish an adaptive genetic algorithm with Recursive feature elimination feature selection procedures for obtaining high dimensionality in gene expression research analysis. SVM kernels classification methods are used to evaluate distinct genetic contexts and achieve classification precision, which can be proposed as an important method for the detection and development of innovative genetics for malaria infection.

Reviews of related works

Large genetic datasets, computational methods are useful for detecting genes responsible for the existence of diseases. Differentially Expressed Genes (DEG) can be identified using a variety of methods. Machine Learning (ML) methods are useful for recognizing variations between genes extracted from the human genome. There are several methods of machine learning that have been used to investigate and identify gene expression profiles of various diseases. The importance of gene expression unfolding and methods for doing so using a variety of machine learning techniques are developed. There are several studies in this area that are being discussed. Recent study flaws in gene expression analysis have been discovered [4].

Oh *et al.* [9] suggested using blood-based transcription of gene indications and machine learning to formative transcripts in classification to evaluate Autism variation disorder. With data mining algorithms, RNA details from the genomic omnibus repository were used in the R tool framework. Autism variation infection prevailed extremely very well tables, according to classified clusters review. When SVM and KNN classifiers were being used to accept results, the overall class evaluation accuracy was 93.8 %.

Qi *et al.* [10] recommended a gene cluster analysis and classification by demonstrating an integrated evaluation. They centered on the benefits and drawbacks of strategies by with clustering and classification frameworks lately as prevailing variations, with linear and non-linear methodologies with dimension reduction strategies for small conditional RNA-seq (scRNA-seq) data, by combining and delivering an RNA-seq clustering and classification.

By ranking large ensembles of genes determined with RNA-seq, Wenric and Shemirani [11] developed a supervised learning approach for collecting RNA-seq genes. They utilized variable rank measures derived from random forests classification, described the EPS (extreme pseudo-samples) vector, and extracted ranks from 12 RNA-seq cancer datasets 323 to 1210 observations with Autoencoder variations and regression coefficient. The findings demonstrated the utility of supervised learning-based feature sampling techniques in RNA-seq training and addressed the importance of gene selection procedures in gene selection.

Alquicira-Hernandez *et al.* [12] used a supervised model to classify RNA-seq results. They demonstrated a generally applicable procedure for high precision single-cell classification by integrating unbiased feature selection from a simplified dimension space with a machine learning prediction strategy. RNA-seq datasets from mononuclear cells, pancreatic tissue, colorectal tumor biopsies and systemic dendritic cells were used to evaluate scPred. They revealed that scPred could accurately identify discrete cells.

Cui *et al.* [13] used RNA and Deoxyribonucleic acid (DNA) research to discover lower evolved genomes that could be influencing Pulmonary Arterial Hypertension (PAH) disease systematically. To

identify an irrelevant group of immensely helpful genes, they proposed a revolutionary feature collection and improved machine learning algorithm methods. The findings revealed that clusters of small-expression genes helped predict and recognize different types of PAH.

Shon *et al*. [14] used a Convolutional Neural Network to identify gene expression data from stomach cancer patients (CNN). They created a deep learning-based classification technique and used a stomach cancer patient to illustrate its implementation to data expression. Principal Component Analysis (PCA), heatmaps, and the CNN algorithm were used to evaluate 60,483 genes from 334 stomach cancer patients in The Cancer Genome Atlas. They analyzed genes and tested them using CNN after combining clinical data and RNA-seq gene expression data analysis. They got an accuracy of 95.96 and 50.51 %.

By describing the variation of an RNA-seq process to deconvolute transcription factors difference for around 500 different pathogens of migrant and individual malaria, Reid *et al.* [15] focused on RNA-seq revealing of secret transcripts in malaria parasites. They found distinct gene expression variations that were previously unknown.

Tan and Gilbert [16] built an ensemble machine learning model for cancer transcriptional model evaluation. Several public information was utilized malignant genomic results, they based on a C4.5 decision tree, bagged, and boosted ensemble decision trees, that are also supervised machine learning methods for cancer classification, and compared the classification approaches. In a classification experiment, ensemble learning (bagged and boosted decision trees) outperformed single decision trees.

Song *et al.* [17] collaborated on the creation of an analytical ensemble classification method for cancer gene expression results. To pick important features for classification, researchers used a combination of Recursive Feature Elimination and the Adaboost algorithm. There was an improvement in the results.

Tarek *et al.* [18] collaborated on a cancer classification algorithm based on gene expression results. They introduced an efficient ensemble classification method that improves the description of the classification as well as the predictability of the results. The findings of ensemble classifiers are less dependent on the uniqueness of a single training set.

By highlighting transform-based of metaheuristic-based methods in embedded systems of feature selection, Duval and Hao [21] established an evolutionary computing framework for finding genes and classification of RNA/DNA data. They demonstrated the utility and value of combining real concern relevant information into discovery operations of such a procedure. The research looked at how linear classifier rating constants, such as SVM, can be beneficially used in successful local discovery for feature selection and classification.

Shukla *et al.* [22] developed a novel framework based on a genetic algorithm by developing an innovative modified feature selection algorithm using a filter-wrapper-based technique for problem classification and resolving shortcomings of conventional approaches. The analysis utilized 5 UCI genetic samples with a variety of instances and dimensionality. The experiments established that the proposed methodology better advanced the important reduction of features and outperformed the state-of-the-art with a minimum accuracy of 40.04 % and the highest accuracy of 99.32 % using k-NN and SVM.

The tree model for classifying ensemble selected features was improved by Tan and Gilbert [23]. To improve the classification, this learning employed an ensemble-based feature selection with random trees and a wrapper procedure. The bagging, wrapper process and random trees were used in the prospective ensemble knowledge classification method to generate a subset. Using a probability weighting principle, the potential approach excluded redundant features and selected the best features for classification. The output of the potential feature selection method was evaluated by comparing it to the GASVMb, GANBb, FSNBb, FSSVMb and GARFb methods using Random Forest, SVM and Naive Bayes analyses. The method achieved a classification accuracy of 92 %.

Kowsari *et al.* [24] examined a range of dimensionality reduction methods for gene expression analysis, including PCA, Independent Component Analysis, Partial Least Square and Locally Linear Embedding. The approach covered discussions as well as the software's objective.

Materials and methods

In this paper, several methods for investigating high-dimensional data ware proposed. For huge dimensionality reduction of RNA-seq results, an adaptive genetic algorithm with RFE and SVM classification algorithms was considered in the analysis. The data were obtained from western Kenya and contained genes of mosquitos from 2010 to 2012. There were 2457 instances with 7 attributes of genes. The expression sequencing data included AGAP003714, AGAP004779, AGAP012984, AGAP0 02724, AGAP009472, CPLC G3 [AGAP008446], CYP6M2 [AGAP008212] and CYP6P3 [AGAP002865], Genomic DNA sequences, mutations in the transcriptome of deltamethrin-resistant and susceptible Anopheles gambiae insects in Kenya. The datasets are factually described in **Table 1**.

Table 1 Dataset features.

Dataset	Attributes	Instances
Mosquito Anopheles gambiae	7	2457

Methods

To evaluate the data collected from [19], MATLAB was utilized as a system development phase, and an adaptive genetic algorithm with Recursive Feature Elimination (A-GA-RFE) was used to select relevant features. Using the SVM algorithm approach [20], the selected features were used to improve the classification results.

Adaptive genetic algorithm (A-GA)

GA is a reliable method for fetching relevant features from high-dimensional datasets. Wrapper-based strategies to feature selection are currently common. There are a variety of parameterizations for genetic algorithms, but mutation and crossover operations are still mostly based on binary parameter concepts. A genetic algorithm is used to identify appropriate features [21]. The RNA has N attributes, each of which has a value of 1 or 0 to indicate whether it is selected or unselected. To address the importance of features, GA can be used to find an optimal feature subset for complex classification presentation using the nominated figure of features. In Algorithm 1, the Adaptive GA's specific structure is defined by using [22]. A-GA (adaptive genetic algorithm) with adaptive parameters provides innovative crossover and mutation operators to address disadvantages including time complexity and falling out of optimum points. The level of solutions that genetic algorithms can achieve in terms of accuracy and computational efficiency is heavily influenced by the probabilities of crossover (Pc) and mutation (Pm) [32]. Rather than using standardized Pc and Pm metrics, AGA uses population data from each generation to adaptively change the Pc and Pm to preserve population diversity while tolerating convergence capability. m is the population size, r is a random number between 0 and 1, chrome represents the nominated or unselected function with the help of a threshold fixed value of 0.5, and is the number of features selected at the threshold. Selecting the most appropriate features from the recognized datasets is one of the method's main challenges.

Recursive feature elimination (A-RFE)

RFE eliminates features iteratively at each level and re-ranks the remaining features by retraining the remaining features. RFE can automatically delete a feature if it is deemed to be a poor feature. When combined with other features, a poor feature can still be useful. As a result, simply eliminating redundant or poor features can hurt classification results. We also tested the classification output after removing a possible poor feature in terms of the value calculated to determine the significance of that feature [32].

Algorithm 1 Adaptive genetic algorithm recursive feature elimination (AGA-RFE)

Require: Initialize the population parameters randomly nPop = m, tmax, t = 0;
Ensure: Evaluate the optimal feature subset with the highest fitness value.

```
while (t <= tmax) do
Create pop m, tmax;
Select fittest Chromosomes
For k = 1 to m do
Apply crossover on rand population < Pc
Parents [m1, m2] = system selection (m, nPop)
        Child = Xor[m1, m2]
        M u = mutation [Child}
End for
Replace m with Child1, Child2, …, Childm
t = t + 1;
End while
Obtain the gene subset and Input subset data with labels and step to the size of RFE.
Store the Highest fitness value;
F:  Set of ranked selected features
R:  Set the remaining selected features {1,2, 3…., n}
S:  Set the score of the criterion functions for the feature set.
```

Support Vector Machine (SVM)

Vapnik proposed SVM as a machine learning system in 1992 [26]. The goal of SVM is to find the best hyperplane in the input space that differentiates between classes. SVM is a linear classifier; it is developed to function with non-linear problems by joining the kernel ideas in high-dimensional workspaces. SVM uses a kernel to train the data in non-linear problems, to expand the dimension broadly. SVM can search for the ideal hyperplane that can distinguish a class from other classes as the dimensions are modified [26]. According to Aydadenta and Adiwijaya (2018), the protocol for determining the best hyperplane using SVM is as follows using Eqs. (1) - (2):

$$\text{Let } y_i \in \{y_1, y_2, \ldots, y_n\}, \quad (1)$$

where y_i are the p-attributes and target class:

$$z_i \in \{+1, -1\} \quad (2)$$

Assume that the groups +1 and −1 can be absolutely separated by hyperplane, as defined in Eqs. (3) - (5):

$$v.y + c = 0 \quad (3)$$

From Eqs. (3) - (5) are gotten:

$$v.y + c \geq +1, \text{ for class } +1 \quad (4)$$

$$v.b + c \leq -1, \text{ for class } -1 \quad (5)$$

where y is the input data, v is the ordinary plane and c is the positive relative to the center-field coordinates. SVM intends to discover hyperplanes that maximize margins between 2 classes. Finding the minimal point is the solution to enhancing margins, which is a combinatorial optimization problem. SVM

has the benefit of being able to handle a wide range of classification problems in high-dimensional data [27].

SVM stands out among other classification algorithms because of its exceptional classification suitability [28]. SVM is grouped into linear and non-linear separable. SVM has kernel functions that change data into a higher dimensional space making it conceivable to achieve separations. Kernel roles are a class of procedures for design investigation or recognition. Training vectors xi is plotted into higher dimensional space by the capacity Φ. SVM gets a linear separating hyperplane with the uppermost in this higher dimension space. C > 0 is the forfeit parameter of the fault period.

Several SVM kernels exist encompassing the polynomial kernel, Radial basis function (RBF), linear kernel, Sigmoid, Gaussian kernel and String Kernels. The decision of a kernel relies upon the current issue at hand. Since it relies upon what models are to be analyzed, a couple of kernel functions commendably in for an extensive assortment of applications [29]. The prescribed kernel function for this study is the SVM-polynomial kernel and gaussian kernel.

SVM-gaussian kernel

Gaussian kernel [30] makes a comparison to an overall responsiveness premise in all k-th order subordinates. To represent previous challenges in learning, kernels coordinating such a preceding correspondence component of the data can be evolved. Each input vector x is transformed into an infinite-dimensional vector that contains all-degree polynomial extensions of x's components.

SVM-polynomial kernel

For instance, a polynomial kernel model features conjunction up to the directive of the polynomial. Radial basis functions permit spheres in disparity with the linear kernel, which permits just selecting lines (or hyperplanes) using Eq. (6).

$$K(y_a, y_j) = (\gamma y_a^S y_b + q)^e, \gamma > 0 \quad (6)$$

SVM-linear kernel function

For instance, the polynomial kernel is the least complex kernel function. It assumes the inner product (a,b) in addition to a discretionary constant K using Eq. (7).

$$K(y_a, y_b) = y_a^S y_b \quad (7)$$

SVM-RBF kernel function

In SVM kernel functions, **γ**, a, and b are kernel constraints, Radial Basis Function (RBF) is the fundamental kernel function due to the nonlinearly maps tests in higher dimensional space different from the linear kernel, it has fewer hyperparameters than the polynomial portion [31] using Eq. (8).

$$K(y_a, y_b) = \exp(-\gamma||y_a, y_b||^{\ 2}), \gamma > 0 \quad (8)$$

Performance evaluation

Validation metrics are needed when measuring the performance of a classification model. True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) are the 4 components that a confusion matrix is often used to assess in classification models (FN). It identified the representations that were incorrectly and correctly recognized from the dataset sample used to evaluate the model [4]. The method for calculating performance metrics is described [25].

The 4 categories TP, FP, TN, and FN are used to determine a model's accuracy.

When the state is present, the product of TP finds it.

When a state does not exist, FP's product finds it.

The result of TN does not locate the state when it is not present.

When a state already exists, the FN product does not find it.

Applications

Gene expression analysis provides a better way to find RNA-seq results. The need to identify related genes aids in the development of a variety of applications, including improved treatment, cancer detection, drug discovery, tumor recognition and infections such as typhoid and malaria, among many others. Machine learning application is used to locate models and data discrepancies. It has excellent algorithms as methods that can be used in a variety of fields.

The experimental investigation was carried out by using MATLAB (Matrix Laboratory), which provides a simple and useful programming model for scientists, architects, scientists, and researchers, among others. MathWorks developed MATLAB, a multi-worldview arithmetical processing environment, and proprietary programming language. It allows for system controls, plotting of functions and details, algorithm execution, and the development of User Interfaces in languages like C, C++, C#, Fortran, Java, and Python [16]. The main goal of this research is to predict the results of RNA-seq technology using the MATLAB tool and the Malaria database. This study's device configuration includes an iCore2 processor, a 64-bit operating system, 4GB of RAM, and MATLAB 2015a as the implementing system.

Results and discussion

This research has uncovered an RNA-seq novel with 2457 instances of Mosquitoes Anopheles gambiae data, including susceptible and resistant genes. The loaded data is shown in **Figure 1**. The sample was transferred to a genetic algorithm to eliminate the dimensionality curse. To determine maximum variance with a smaller proportion of subset features in the variable, an adaptive GA with RFE feature selection dimensionality reduction fetches the optimum subset of data and eliminates uncorrelated attributes (variables).

AGA-RFE was used on mosquito anopheles data in this analysis, and it provides important gene information that can be used for further research. To implement the model, classification algorithms use SVM kernels and the MATLAB tool. With a threshold of 0.5, 589 optimal subset features of genes were important using AGA-RFE as a feature selection dimensionality reduction tool.

To assess the implementation of the classification models' results, 10-fold cross-validation was used with SVM classification algorithms, with 0.05 parameter holdout of data for training and 5 % for testing to verify the classifiers' accuracy. To eliminate sampling biases, the classifier employed a learning evaluation procedure in which the training and testing phases were assessed using a 10-fold cross-validation method. MATLAB was used to implement this protocol. The calculation time and performance metrics (Accuracy, Specificity, Sensitivity, Precision, F-score, and Recall) [25] were used to calculate the valuation result.

This study compared the models' classifier accuracy with 98.3 and 96.7 % precision, respectively, using L-SVM and RBF-SVM classifiers. **Figure 2** depicts the result production as well as the confusion matrix.

7 Attributes loaded **2457 Instances loaded**

13071_2015_1083_MOESM4_ES

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

Figure 1 Data loaded into the MATLAB environment by the mosquito Anopheles gambiae.

This analysis employed AGA-RFE to retrieve specific components from the loaded data shown in **Figure 1**. The SVM classification was applied to the selected features, and the result of the confusion matrix obtained from the evaluation is shown in **Figures 2** and **3**. The efficiency metrics were solved by using the uncertainty matrix. This study achieves a 98 % accuracy using the L-SVM classification kernel and a 97 % accuracy using the RBF-SVM kernel classification system. Other output metrics are shown in tabular form in **Table 2**.

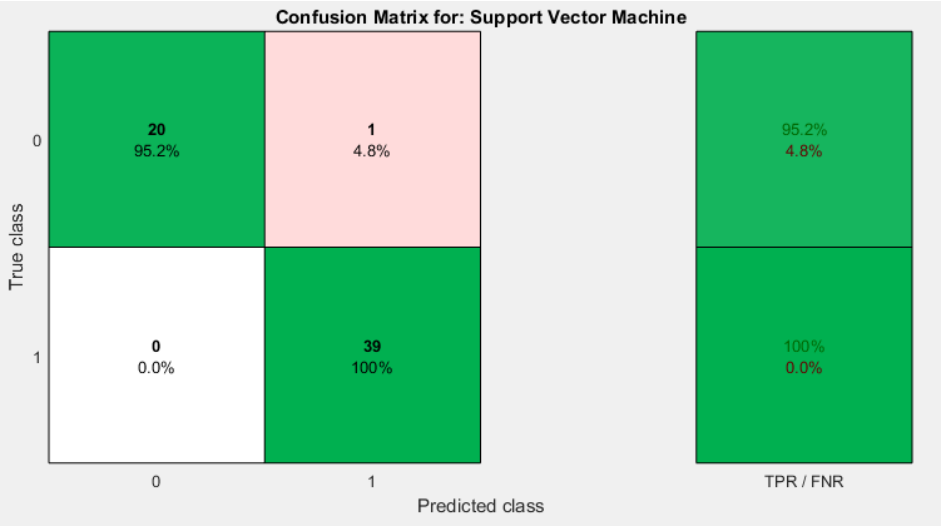

Figure 2 The confusion matrix for mosquito RNA-seq data classification using AGA-RFE-L-SVM, TP = 39; TN = 20; FP = 0; FN = 1.

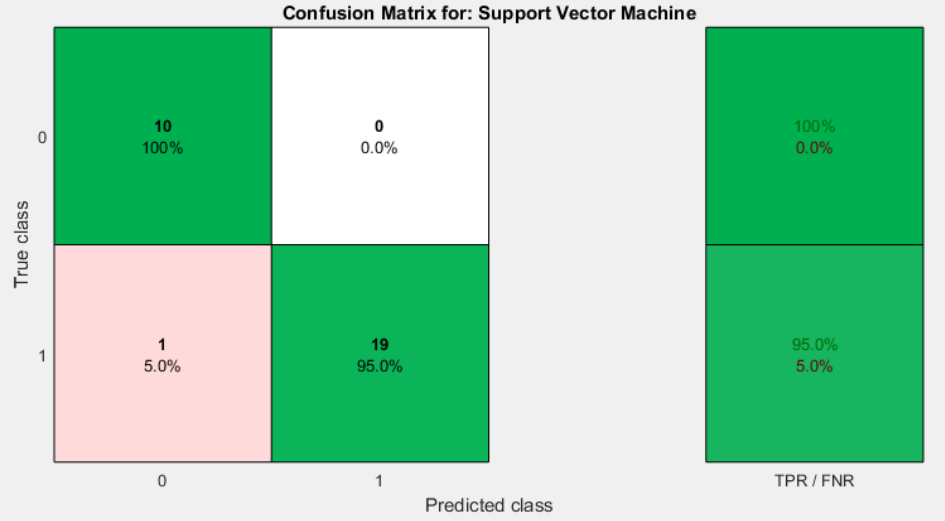

Figure 3 The confusion matrix for the classification of mosquito RNA-seq data classification using AGA-RFE-RBF-SVM, TP = 19; TN =10; FP = 0; FN = 1.

RNA-seq data for nosquito Anopheles Gambiae was downloaded to test the performance of the data mining learning method from 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.

AGA-RFE was used as a dimensionality reduction model, and 589 features were chosen as a subset in the data. To predict their results, these components were, then, classified using Ensemble classification. The outcome demonstrated the utility of machine learning technology in the field of genetics. The performance results are shown and compared in **Table 2** below to support the strategy. RBF-SVM outperforms L-SVM in terms of training time and accuracy rate, according to the results.

Table 2 Performance metrics table for the confusion matrix.

Performance metrics	L-SVM classification	RBF-SVM classification
Accuracy (%)	98.3	96.7
Sensitivity (%)	97.5	95.0
Specificity (%)	100	100
Precision (%)	100	100
Recall (%)	95.24	90.91
F-Score (%)	96.36	97.44

Conclusions

The findings of tis study are useful for the prognosis and diagnosis of human malaria. Machine learning methods, such as dimensionality reduction models and classification algorithms, were used in the proposed method. The AGA-RFE feature selection model and SVM classifiers were applied in the dimensionality reduction model. The results of this study's analysis and evaluation of output were shown, using the SVM classification algorithm.

Several studies have been proposed in the evaluations by the investigators by using the performance metrics obtained, and the results have proven that dimensionality reduction models using feature extraction methods, e.g., AGA-RFE, could help improve classification output such as SVM. It will be important to see if the feature selection models and algorithms proposed in recent work can be improved. The limitation of this study is based on the available dataset used. Should larger data are feasible, using other classifiers, e.g., Random forest, can be introduced in the future works.

References

[1] S Sun, C Wang, H Ding and Q Zou. Machine learning and its applications in plant molecular studies. *Briefings Funct. Genom.* 2019; **19**, 40-8.
[2] DF Read, K Cook, YY Lu, KGL Roch and WS Noble. Predicting gene expression in the human malaria parasite plasmodium falciparum using histone modification, nucleosome positioning, and 3D localization features. *PLoS Comput. Biol.* 2019; **15**, e1007329.
[3] MO Arowolo, M Adebiyi and A Adebiyi. A dimensional reduced model for the classification of RNA-seq Anopheles gambiae data. *J. Theor. Appl. Inform. Tech.* 2019; **97**, 3487-96.
[4] S Karthik, and M Sudha. A survey on machine learning approaches in gene expression classification in modelling computational diagnostic system for complex diseases. *Int. J. Eng. Adv. Tech.* 2018; **8**, 182-91.
[5] NT Johnson, A Dhroso, KJ Hughes and D Korkin. Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers? *RNA* 2018; **24**, 1119-32.
[6] MW Libbrecht and WS Noble. Machine learning applications in genetics and genomics. *Nat Rev Genet.* 2015; **16**, 321-32.
[7] 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 Proc.* 2014; **8**, S2.
[8] The Anopheles gambiae 1000 Genomes Consortium. Genetic diversity of the African genetic diversity of the African malaria vector Anopheles gambiae. *Nature* 2017; **552**, 96-100.
[9] DH Oh, IB Kim, SH Kim and DH Ahn. Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning. *Clin. Psychopharmacol. Neurosci.* 2017; **15**, 47-52.
[10] R Qi, A Ma, Q Ma and Q Zou. Clustering and classification methods for single-cell RNA-seq data. *Brief. Bioinform.* 2020; **21**, 1196-208.
[11] S Wenric and R Shemirani. Using supervised learning methods for gene selection in RNA-seq case-control studies. *Front. Genet.* 2018; **9**, 1-6.
[12] J Alquicira-Hernandez, A Sathe, HP Ji, Q Nquyen and JE Powell. scPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol*. 2019; **20**, 264.
[13] S Cui, Q Wu, J West and J Bai. Machine learning-based microarray analyses indicate low-expression genes might collectively influence PAH disease. *PLoS Comput. Biol.* 2019; **15**, e1007264.
[14] HS Shon, YG Yi, KO Kim, EJ Cha and KA Kim. Classification of stomach canacer gene expression data using CNN algorithm of deep learning. *J. Biomed. Transl. Res.* 2019; **20**, 15-20.
[15] AJ Reid, AM Talman, HM Bennett, AR Gomes, MJ Sanders, CJR Illingworth, O Billker, M Berriman and MKN Lawniczak. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *Elife* 2018; **7**, e33105.
[16] AC Tan and D Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Appl. Bioinformatics* 2003; **2**, S75-S83.
[17] N Song, K Wang, M Xu, X Xie, G Chen and Y Wang. Design and analysis of ensemble classifier for gene expression data of cancer. *Adv. Genet. Eng.* 2016; **5**, 1000152.
[18] S Tarek, RA Elwahab and M Shoman. Gene expression based cancer classification. *Egypt. Informat. J.* 2017; **18**, 151-9.

[19] M Bonizzoni, E Ochomo, WA Dunn, M Britton, Y Afrane, G Zhou, J Hartsel, MC Lee, J Xu, A Githeko, J Fass and G Yan. 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 Vector* 2015; **8**, 474.
[20] G James, D Witten, T Hastie and R Tibshirani. *An introduction to statistical learning: With application in R.* Springer, New York, 2013.
[21] B Duval and JK Hao. Advances in metaheuristics for gene selectio and classification of microarray data. *Brief. Bioinform.* 2010; **11**, 127-41.
[22] AK Shukla, P Singh and M Vardhan. A new hybrid feature subset selection framework based on binary genetic algorithm and information theory. *Int. J. Comput. Intell. Appl.* 2019; **18**, 1950020.
[23] AC Tan and D Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Appl. Bioinformatics* 2003; **3**, S57-83.
[24] K Kowsari, KJ Meimandi, M Heidarysafa, S Mendu, LE Barnes and DE Brown. Text classification algorithms: A survey. *Information* 2019; **10**, 150.
[25] AM Olaolu, SO Abdulsalam, IR Mope and GA Kazeem. A comparative analysis of feature selection and feature extraction models for classifying microarray dataset. *Comput. Inform. Syst.* 2018; **22**, 29-38.
[26] H Aydadenta and Adiwijaya. On the classification techniques in data mining for microarray data classification. *J. Phys. Conf. Series.* 2018; **971**, 012004.
[27] CC Chang and CJ Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Tech.* 2011; **2**, 27.
[28] A. Khan, B. Baharudin, L.H. Lee, K. khan, K. A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology. 2010: 1; pp. 1-17.
[29] HP Bhavsar and M Panchal. A review on support vector machine for data classification. *Int. J. Adv. Res. Comput. Eng. Tech.* 2012; **1**, 185-9.
[30] CDA Vanitha. D Devaraj and M Venkatesulu. Gene expression data classification using support vector machine and mutual information-based gene selection. *Proc. Comput. Sci.* 2015; **47**; 13-21.
[31] MO Arowolo, SO Abdulsalam, RM Isiaka and KA Gbolagade. A hybrid dimensionality reduction model for classification of microarray dataset. *Int. J. Inform. Tech. Comput. Sci.* 2017; **9**, 57-63.
[32] AK Shukla. Multi-population adaptive genetic algorithm for selection of microarray biomarkers. *Neural Comput. Appl.* 2020; **32**, 11897-918.
[33] XW Chen and JC Jeong. Enhanced recursive feature elimination. *In*: Proceedings of the 6th International Conference on Machine Learning and Applications, Cincinnati, OH, USA. 2007, p. 429-35.