Relevant gene selection using ANOVA-ant colony optimisation approach for malaria vector data classification

Micheal Olaolu Arowolo

Department of Computer Science,
College of Pure and Applied Sciences,
Landmark University Omu-Aran,
Kwara State, Nigeria
Email: arowolo.olaolu@lmu.edu.ng

Joseph Bamidele Awotunde*

Department of Computer Science,
Faculty of Information and Communication Sciences,
University of Ilorin,
Ilorin, Kwara State, Nigeria
Email: awotunde.jb@unilorin.edu.ng
Email: jabonnetbylinks@gmail.com
*Corresponding author

Peace Ayegba

Department of Computer Science,
College of Pure and Applied Sciences,
Landmark University Omu-Aran,
Kwara State, Nigeria
Email: ayegba.peace@lmu.edu.ng

Shakirat Oluwatosin Haroon-Sulyman

Department of Information and Communication Science,
Faculty of Information and Communication Sciences,
University of Ilorin,
Ilorin, Kwara State, Nigeria
Email: sulyman.sh@unlorin.edu.ng

Abstract: Recent progress in gene expression data research makes it possible to quantify and identify several thousand genes' expressions simultaneously. For malaria infection and transmission, gene expression data classification using dimensionality reduction is a standard approach in gene expression data analysis and proposed for this study. A major problem occurs in the reduction of high dimensional data, it plays a significant role in improving the precision of classification, allowing biologists and clinicians to correctly predict infections in humans by choosing a limited subclass of appropriate genes and deleting redundant, and noisy genes. The combination of a novel analysis of variance (ANOVA) with ant colony optimisation (ACO) approach as a hybrid feature selection to select relevant genes is suggested in this study to minimise the redundancy between genes, and SVM is used for classification. The proposed method's efficacy was shown by the experimental outcomes based on the high-dimensional of gene expression data.

Keywords: malaria vector; gene expression; analysis of variance; ANOVA; ant colony optimisation; data classification; support vector machine; machine learning; high-dimensional data; data analysis; vector-borne disease; multi-layer perception.

Reference to this paper should be made as follows: Arowolo, M.O., Awotunde, J.B., Ayegba, P. and Haroon-Sulyman, S.O. (2022) 'Relevant gene selection using ANOVA-ant colony optimisation approach for malaria vector data classification', *Int. J. Modelling, Identification and Control*, Vol. 41, Nos. 1/2, pp.12–21.

Biographical notes: Micheal Olaolu Arowolo is a faculty of the Department of Computer Science at the Landmark University, Omu-Aran, Nigeria. He holds a Bachelor's degree from the Al-Hikmah University, Ilorin, Nigeria and a Master's degree from the Kwara State University, Malete, Nigeria. He also completed his PhD from the Landmark University, Omu-Aran, Nigeria. 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 IEEE, IAENG, APISE, SDIWC, and an Oracle Certified Expert.

Joseph Bamidele Awotunde was a Lecturer II with the Computer Science Department, University, of Ilorin, Ilorin, Nigeria. He is the author of more than 40 articles, 15 edited books, and more than 15 conference proceedings in reputable outlets majorly indexed in SCOPUS and Web of Science. His research interests include information security, cybersecurity, bioinformatics artificial intelligence, internet of medical things, wireless body sensor networks, wireless networks, telemedicine, m-health/e-health, and medical imaging. He is a member of Computer Professional Registration Council of Nigeria (MCPN), and Nigeria Computer Society (MNCS).

Peace Ayegba is an Assistant Lecturer in the Department of Computer Science, Landmark University Omu-Aran. Her research areas are in algorithms, artificial intelligence and digital health.

Shakirat Oluwatosin Haroon-Sulyman is a Lecturer II in the Department of Information and Communication Science, Faculty of Communication and Information Sciences, University of Ilorin, Kwara State, Nigeria. She holds a Bachelor of Science in Computer Science and a Master's in Information Technology at the Al-Hikmah University, Ilorin, Kwara State and Universiti Utara Malaysia, Kedah, Malaysia respectively. She has published articles in both local and international journals, conference proceedings as well as chapters in books. Her research interests include: e-learning, social media usage, multimedia, system design, machine learning and text-mining.

1 Introduction

Malaria is a worldwide long plagued, vector-borne disease. Due to rampant changes in environmental and food patterns, malaria infects a lot of individuals. Mostly in West African areas, this disease is widely spread and seriously infecting people in those regions. The malaria vector-borne disease dataset is characteristically distorted. The dynamic nature of such data in machine learning technology is a challenge. Accurate prediction of smaller samples is very important for medical diagnosis and prognosis of the infection, particularly in the medical field (Sajana and Narasingarao, 2018).

Gene expression technologies such as RNA-Seq are a prevailing tool for describing gene expression designs in different cells, enabling various investigations like identifying innovative cell categories (Townes et al., 2019). The RNA-Seq experiments generate a huge amount of gene expression data, resulting in curse-of-dimensionality problems, imposing limitations on the technology's efficiency. Data pre-processing is not always possible, as data subsets can lead to loss and overfitting of information, directly impacting the accuracy and reliability of machine learning models (Zahoor and Zafar, 2020).

Much research has been carried out to overcome the curse of dimensionality by defining appropriate and impactful models from high-dimensional information to attain relevant characteristics. For the analysis of gene expression data, conventional models are not adequate and are a challenge for further study (Almugren and Alshamlan, 2019). Dimensionality reduction techniques for selecting relevant features from high dimensional data and reduce noises, such as infiltration tactics optimisation, particle swarm optimisation, genetic algorithm, information gain, swarm intelligence, SMOTE, among other mining approaches, have been utilised in the literature to achieve high-performance accuracies (Zahoor and Zafar, 2020; Tahir et al., 2019; Mahani and Ali, 2020; Saw and Myint, 2019). Several advances have been made using several hybridised techniques, which are capable of generating high accuracies. Yet, this research is opened to more advances due to variations in existing methods. Developing a comprehensive method that can improve models' accuracy reliability remains a challenging problem (Sato and Nakagawa, 2014).

In this study, a modified and improved ANOVA-ACO is suggested to select relevant genes as a subset for malaria vector data classification and build an overall model that helps fetch relevant and reliable features. They are refined and classified using support vector machine (SVM) to achieve reliable accuracy. Analysis of variance (ANOVA) provides the apparent analytical multiple factors in the experiment by expressing the intensity values (Arowolo et al., 2016). The ant colony optimisation (ACO) denoted an efficient sample base gene approach to unravel the swarm intelligence optimisation problem to enhance classification performance (Bir-Jmel et al., 2019). The ANOVA-ACO promises an optimal feature subset from high dimensional data. This study is beneficial to researchers and health practitioners for decision making, as this study have shown an insightful contribution compared with other investigators, it proposes to better improve knowledge adoptions.

The remainder of this study is organised as follows; Section 2 offers the study's literature review, Section 3 introduces the techniques and the algorithm proposed, Section 4 discusses the experimental findings and interpretation are identified. Section 5 discourses the conclusions and future directions.

2 Literature review

The ITO procedure uses parameter-free and based classifiers in creating high-precision and reliability classifications. The procedure generated two phases of outcomes. The lightweight infantry group (LIG) joins rapidly in finding non-local utmost. It generated similar outcomes of about 88% accuracy with the follow-up team using advanced tuning of about 99% to boost zero efficiencies. Every ITO is a base approach with the self-reliantly selected model. Subset technique of collection, pre-processing, and classifier and validation approaches. The active soldiers are collective for optimum outcomes by heterogeneous ensembles. Their approach discourses data insufficiency, versatility for optimal varied base classifiers, and capable of producing the HAHR approach similar to the results of the MAQC-II developed (Zahoor and Zafar, 2020). In previous research works, AI has been used in medical and in the biomedical field (Ayo et al., 2020; Satapathy et al., 2021), for the involvement of heart disease and diabetes (Oladipo et al., 2021; Berglund, 2015; Bhoi, 2017; Awotunde et al., 2021a) investigated diabetes proteins. Academics have used ANN, SVM, fuzzy logic systems, K-means classifier, and many other AI methods (Awotunde et al., 2021b; Mishra et al., 2020a).

A novel distributed method of selection of features was suggested to remedy the curse-of-dimensionality of information about microarrays. Their methodology is inspired by an educational process for creating project groups for the final year. To construct multiple balanced feature clusters, they utilised information gain, symmetrical uncertainty, and entropy. Each cluster trains the multi-layer perceptron (MLP), and the prevailing cluster is selected for more processing. Training and tuning the MLP is a very time-consuming task in itself. Educating many such clusters makes it resourceful and starving. Using MLP makes it probable to prematurely halt the procedure and pick clusters for further processing with maximum precision (Potharaju and Sreedevi, 2019; Mishra et al., 2020b).

Together with kernel ridge regression (KRR) classifier, an improved cat swarm optimisation algorithm for feature selection has been proposed. They presented a universal dataset, with KRR outperforming the wavelet kernel ridge regression (WKRR) and radial basis kernel ridge regression (RKRR). In comparison to multi-class datasets, their methodology works moderately improved on two-class datasets (Mohapatra et al., 2016).

Pattern classification based on binary gene expression data, with exciting learning machine (ELM) variations, such as online sequential ELM (OSELM) and extreme learning machine based on the kernel (KELM), was suggested. For the classification of microarray medical datasets, two variants are used in the KELM group. A modified cat swarm optimisation method is implemented to diminish the gene expression dataset's high dimensionality, resulting in a high amount and limited sample sizes. With a series of output metrics for the datasets, the efficacy of the proposed algorithm is checked (Chakravarty et al., 2020).

Awotunde et al. (2021c) proposed a method for predicting the incidence of malaria in Nigeria. Using Kwara State as a case study, both environmental and clinical data were used to forecast the malaria-endemic well. For the proposed scheme, the researchers used a deep learning algorithm as a classifier. Three locations were chosen with a 34-month periodic pattern from the Irepodun local government areas of Kwara State. Environmental factors affected how each location responded. Both factors are critical in malaria transmission and prediction, according to the findings. The LSTM algorithm is an effective tool for detecting cases of widespread malaria.

An enhanced artificial bee colony (ABC) algorithm was proposed to pick a small number of critical cancer genes, with predictive accuracy improvements. It is assumed that ABC's quest equation is good at discovery but bad at exploitation. They improved the ABC algorithm to overcome this limitation by adding the pheromone principle, which is one main mechanism of the ACO algorithm. A novel process of sequential bees interacts to segment their data. A publicly available dataset uses the proposed algorithm is tested after the parameters are logically calibrated to one of the datasets. The findings gotten are contrasted with other datasets. It has been shown that the efficiency of the proposed approach is superior (Moosa et al., 2016).

To pick the pertinent features, ANOVA, an arithmetical test grounded on MapReduce, was suggested. MapReduce-based K-nearest after function selection. The classification of the microarray data is also indicated by the neighbour (K-NN) classifier. On Hadoop, these algorithms are successfully implemented, and different datasets are used to perform comparative analysis (Kumar et al., 2015).

Data mining was used to construct the association rules, and the algorithms for mathematical, soft computing and development were utilised to pick the optimum classifier grounded on previous data. Three techniques were utilised to evaluate the disease category based on the dataset's attributes, and new tested data were generated using ten-fold, and 98% training, and 2% tested data (Singh et al., 2020).

The use of swarm intelligence algorithms in high-dimensional data for feature selection processes focusing on the subject matter is medical data classification. The findings show that intelligence algorithms for swarming have a good capability for the selection of features. They presented comparisons and different swarm substitute algorithms to be introduced for high dimensional classification in function selections (Saw and Myint, 2019).

An efficient relief-F-ACO approach, for tumour classification, was proposed. The investigational outcomes

of numerous gene expression datasets demonstrated that the proposed approach is operative and meaningfully reduces the dimensionality of gene expression datasets. It picks the maximum important genes with relevant precision for classification (Sun et al., 2019).

A hybrid method was proposed using MWIS-ACO-LS for the problem of feature selection, based on the grouping of the original graph-based feature selection method. They try reducing redundancy among genes considering association among decisive and maximising gene status (Fisher), and an improved ACO with local search procedure utilising the classifier. We evaluated MWIS-ACO-LS on well-replicated gene expression datasets to evaluate the proposed process. Our proposed approach's efficacy was shown by the investigational outcome classification issues (Bir-Jmel et al., 2019).

An efficient logistic approach based on F-logistic was suggested with complex alterations to classify the genes. To obtain usable data, using various approaches for levelling techniques, F-logistic was used for the profiling calculated and unevenly spread-out period points. In evaluation with additional efficient information methods, i.e., the useful ANOVA, they evaluated F-logistic efficiency through real and synthetic period sequence datasets. The actual dataset contains a period of sequence gene expression outlines for lasting effects of interferon β with multiple sclerosis ailment developments. In instance/control research focused on period-sequence expression outlines, F-logistic identifies complex alterations that are unfound by viable methods. F-logistic is successful for a period reliant on biomarker discovery, analysis, and treatment (Kayano et al., 2016).

The procedure described in the ANOVA is examined to pick the approach that mimics human disease greatest precisely in terms of genome-wide gene expression. For the same dataset, two widely used data fitting algorithms are tested and evaluated in machine learning. The implementation of individual algorithms is addressed, computational costs, advantages, and drawbacks of the individual procedure are analysed to evaluate cancer (Shamsaei and Gao, 2016).

For improving the accuracy of the artificial neural network, a hybrid classification improvement strategy such as genetic algorithm (GA), particle swarm optimisation (PSO), and fireworks algorithm (FWA) has been presented. Two trials are used to test the enhancing process. To begin, the suggested models are evaluated on five benchmark medical databases from the University of California, Irvine's (UCI) archive. The best-performing system is then employed in the second study, which concentrates on fine-tuning the parameters of the chosen algorithm by varying the number of iterations in ANNs with varied numbers of hidden layers. On a biological genetic sequence huge dataset obtained from The Cancer Genome Atlas (TCGA) archive, improved ANN with the three optimisation algorithms is investigated. GA and FWA are statistically significant, whereas PSO is not, and GA outperforms PSO and FWA in terms of performance. The process works well and improves with each phase as significant results are attained (Salman et al., 2018).

A multi-chromosome GEP technique was proposed for gene expression programming using multiple chromosomes (MC-GEP). To begin, an individual is made up of numerous chromosomes, each of which contains one or more genes. Secondly, each chromosome's expression, or mixtures of numerous chromosomes, can be used to identify an individual. Subsequently, chromosome rejoining is altered and carried out in an organised manner, similar to meiosis. In comparison to GEP, experiments demonstrate that MC-GEP can minimise the operating rate and accuracy of iterations (Wang et al., 2011).

A new approach for detecting heart disorders in respective heart valve data was proposed: pessimistic multi-granulation rough set-based classification for heart valve disease diagnosis. The supervised rapid reduce feature selection technique is used to choose essential features from heart valve data in this study. Only relevant features picked using supervised fast reduction from heart valve data are used in the classification approach. For the detection of heart valve dysfunction, a new classification approach based on pessimistic multi-granulation rough sets (PMGRS) is used in this work. Set approximations in multi-granulation rough sets are well-defined by many equivalence relations on the universe, resulting in an effective classification model. This is supported by experimental results, which show that the suggested approach outperforms conventional benchmark classification algorithms such as Naive Bayes, MLP, J48, and decision table classifiers in terms of classification performance (Azar et al., 2016).

Skin disease classification using deep learning neural networks with MobileNet V2 and LSTM suggested a computerised technique for classifying skin illness using deep learning neural networks with MobileNet V2 and LSTM (LSTM). The MobileNet V2 model has proven to be more efficient and accurate, and it can be used on small computational devices. For exact predictions, the proposed model is effective in maintaining stateful information.

The progression of pathological growth is measured using a grey-level co-occurrence matrix. The results were compared to other state-of-the-art models like fine-tuned neural networks (FTNN), convolutional neural networks (CNN), visual geometry group's (VGG) very deep convolutional networks for large-scale image recognition, and CNN architecture that expanded with few changes. The proposed method outperformed existing methods with over 85% accuracy using the HAM10000 dataset. Its robustness in recognising the impacted region significantly faster and with roughly twice the number of computations as the traditional MobileNet model results in little computational effort. A mobile application is also built for quick and accurate action. It assists patients and dermatologists in determining the type of disease from a photograph of the affected region during the early stages of skin disease. These findings suggest that the proposed system could assist general practitioners in quickly and accurately diagnosing

skin conditions, reducing complications and morbidity (Srinivasu et al., 2021).

Maximum power point tracking (MPPT) is critical for obtaining the maximum power from a photovoltaic (PV) system by ensuring optimal production under sunlight and temperature variations. The adaptive neuro-fuzzy inference system (ANFIS) is an algorithm-based MPPT that is built using a combination of an artificial neural network (ANN) and a fuzzy logic controller (FLC). Under temperature and irradiance change, the ANFIS algorithm is tested in MATLAB/Simulink and compared to the fixed step traditional perturb and observe (P&O) and gradient descent techniques. The ANFIS-MPPT technique provided a significant improvement in PV system performance over other techniques under various operating conditions, including faster convergence, stability in steady state, fewer oscillations around the MPP, and higher efficiency in tracking the maximum power from the PV system (Amara et al., 2019).

A robust, accurate, and intelligent control for trajectory tracking of a three-degree-of-freedom quadrotor unmanned aerial vehicle is proposed in this paper (UAV). The controller is based on the ant colony optimisation (ACO) algorithm and an ANFIS. Controlling nonlinear systems with intelligent control, such as fuzzy logic, is a good option. The ANFIS controller is used to recreate the quadrotor's reference trajectory in the 2D vertical plane. The ACO algorithm adjusts ANFIS parameters automatically in order to eliminate learning errors and increase the controller's quality. The proposed ANFIS controller adjusted by ACO is compared to ANFIS and proportional-integral-derivative (PID) controllers to assess its performance. The hybrid ANFIS-ACO controller, as expected, produces better outcomes than the other approaches established in the same study (Selma et al., 2020).

3 Materials and methodology

In this section, the methods and materials used, such as the dataset, workflow, performance evaluation, and algorithms are discussed extensively. This study uses a mosquito anopheles gambiae gene dataset to analyse the performance of the proposed model. The proposed algorithm is a hybridised feature selection algorithm known as ANOVA and ACO (ANOVA-ACO) algorithm. In this study, a hybrid technique is suggested to select features in high dimensional datasets. The best subset of features were selected and classified using SVM.

3.1 Datasets

High-dimensional data investigations have been discussed extensively. An ANOVA-ACO and SVM classification algorithm is proposed using mosquito anopheles gambiae gene expression data consisting of 14,914 gene features with 37 attributes of Uganda, mosquito's gene data (Isaacs et al., 2018), with its profile transcript contents, genes, descriptions, control, resistance, susceptible for combating malaria transmission, through anopheles gambiae mosquitoes which is an openly accessible data, it is tabulated in Table 1. MATLAB experimental tool is used to experiment, ANOVA-ACO is proposed and used to fetch relevant features. The selected were classified using the SVM.

Table 1 Dataset structures

Dataset	*Attributes*	*Instances*
Mosquito anopheles gambiae	36	14,914

3.2 Analysis of variance

ANOVA is a filter-based feature selection method used to analyse gene expression results and fetch for important treatment effects by selecting interesting p-value genes. This approach is dynamic and adopts the normal distributed population sample size with equal variance with mutually independent results (Mallika and Saravanan, 2010). The one-way ANOVA is used in this analysis. It achieves an investigation on the comparison of two or more classes for individual genes and yields a single p-value distinct classes from others. It is important.

The smallest values are the most greatly varied genes. If the p-value for the F-ratio is fewer than the difficult value of all the information provided in the ANOVA table, then the effect is important. The p-value is set to 0.05 in this study, where values less than outcomes in important effects, although any value more than these value outcomes to non-significant effects. The very small p-value suggests that there are extremely important variations between the means of the column. The likelihood of the F-value emerging from two similar distributions provides us with a measure of the importance of the sample variation and between-sample variation. Small p-values suggest a low likelihood of interesting genes because of the sampling of the sample variations' internal sample distribution. The p-values are used in this study for individual gene ranking and wise generating for pairs (Adebiyi et al., 2021).

ANOVA test conducts a variance analysis for each function, where the function describes the class variable. It uses the statistic value as ranking. The complex the F-score, the difference in the mean values between the conforming function groups are. The filter score is specified for each characteristic as;

$$ANOVA\ test\ (X_m) = \frac{\sum_{i=1}^{l} n_i \left(x_i^{(m)} - x_i^{(m)}\right)^2 \Big/ (i-1)}{\sum_{i-1}^{l} \sum_{j=1}^{n_i} \left(x_{ij}^{(m)} - x_i^{(m)}\right)^2 \Big/ (n-l)} \quad (1)$$

where l is the Y classes and $x_{ij}^{(m)}$, $i \in \{1, \ldots, l\}, j = j \in \{1, \ldots, n_i\}$, are the observed values of the features X_m for class i instances.

The $\bar{x}_i^{(m)} = \frac{1}{n}\sum_{j=1}^{n_i} x_{ij}^{(m)}$ are the mean values of X_m in-class i and $\bar{x}_i^{(m)} = \frac{1}{n}\sum_{i=1}^{n_i}\sum_{j=1}^{n_i} x_{ij}^{(m)}$ is the mean of X_m for all instances in the dataset (Bommert et al., 2020), Algorithm 1 shows the ANOVA procedure.

Algorithm 1 ANOVA Algorithm for feature selection

Input: Sample size Y * X features matrix.
Output: Selected N features
for each selected feature set do
i = 1, 2, …, n.
Calculate the value of features with the degree between classes with the total number of classes –1
Evaluate the value of the degree of classes with mean square error within classes.
Calculate the F-statistics value (Fi) by mean square error classes
Evaluate the p-value for individual f-statistics obtained in the dispersal table
if pi < 0.5 then
    Fetch the features fi
    Add fi to a feature matrix Xn
Else
    Discard fi features
End if
    Sort feature set in the population size
If size X > 5,000 then
    Select top feature sets
Else
    Keep X feature matrix
End if
End for
Return X feature matrix

3.3 Ant colony optimisation

ACO portrays the action of the use of real ants as a useful metaheuristic technique to solve several intricate problems gotten in the discounted list. ACO uses the pheromone level as the algorithm's ending principle (Singh et al., 2020). The ACO algorithm a wrapper-based method for selecting features using a probabilistic procedure to solve computational difficulties to minimise search pathway by finding the optimal track through graphs, that can typically be utilised for finding optimal feature subsets (Sun et al., 2019).

Let $Y_{pq}(0) = K$, where K is a constant, and the n^{th} and defines each path's position according to the number of pheromones, where $n = 1, 2, \ldots, m$. The n^{th} ant's probability shifts at the p^{th} moment from the I position to the q^{th} position, which is defined as

$$S_{pq}^{n}(y) = \frac{Y_{pq}^{\infty}(y)\varphi_{pq}^{\beta}(y)}{\sum sample \in allowednY_{pq}^{\infty}(y)\varphi_{pq}^{\beta}(y)} \quad (2)$$

The important relative track and $\infty > 0$; β is the relative important visibility $\beta < 0$ and calculated using a metaheuristic algorithm. Algorithm 2 shows the procedure the ACO uses.

Algorithm 2 ACO algorithm for feature selection

Input: Sample data; nmax; m; α; β; λ; τi0, 1; min; max.
Output: Total optimum result S.
Begin
Select the subset features of gene
Use Algorithm 1 hypothesis for the gene-similarity graph.
Fetch the initial subset of genes.
Apply ACO to the selected genes subset
Step 1: ACO joint with the ANOVA
Set pheromone matrix by ones.
for T = 1; T < nmax;
    for i = 1; i < m;
    build the path for optimum result (S) for ant based on probabilistic decision rules.
    Calculate the fitness of the optimum result using LOOCV.
if i == 1 then
    Soptimum = S
    end if
if f(Soptimum) ≤ f(S) then
    Soptimum = S
    end if
Update pheromones based on S.
    end for
    Return

3.4 Proposed approach (ANOVA-ACO)

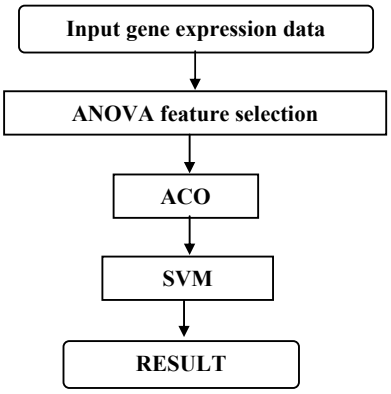
In this study, the ANOVA-ACO approach is proposed to fetch relevant features from high-dimensional data, which can help make relevant decisions for clinicians in the designs of drugs and approaches to eradicate malaria infections in humans. The ANOVA uses a 0.5 p-value in its decision make; it is a filter approach and not sufficient enough as noises have been observed. Adopting ACO to the selected features from the ANOVA is an interesting approach due to the important task of an ant search, which includes the rule generation, the prune rules, and updating of pheromone that helps enhance the algorithm (Yu et al., 2009; Arowolo et al., 2021a).

3.5 Support vector machine

The next step after pre-processing is the classification process, after reducing the dimensional complexity of data. The primary target in this experiment is classification. The

data is diagnosed (classified) at this point based on whether they are infected or not. It uses the SVM algorithm. The algorithm's outcomes are then compared and evaluated based on accuracy.

Figure 1 Flowchart for proposed ANOVA-ACO algorithm



The SVM classification algorithm is a useful tool to solve the problem of classification. It has a major advantage of high generalisation capability, absence of local minima, and suitability for a small-sample dataset, compared to conventional classification methods (Yu et al., 2009; Arowolo et al., 2021b). Given the dataset $S = \{(x_i, y_i) \mid x_i \in y_i \in \{-1, +1\}, i = 1, \ldots, N\}$, where x_i is a sample of the d-dimension, y_i is the corresponding class name, and N is the number of samples.

3.6 Performance evaluation

To validate the performance of the proposed ANOVA-ACO and SVM experiment, MATLAB was used to experiment. A standard performance metrics such as the classification accuracy, sensitivity, specificity, precision, recall, f-score, and confusion matrix is utilised, using four specific parameters known as the true positive (TP), true negative (TN), false negative (FN) and false positive (FP) (Basavegowda and Dagnew, 2020).

4 Result and discussion

In this study, an extensive experiment to validate the classification performance of a proposed ANVA-ACO-SVM algorithm is carried out. The experiment is performed using a malaria vector dataset.

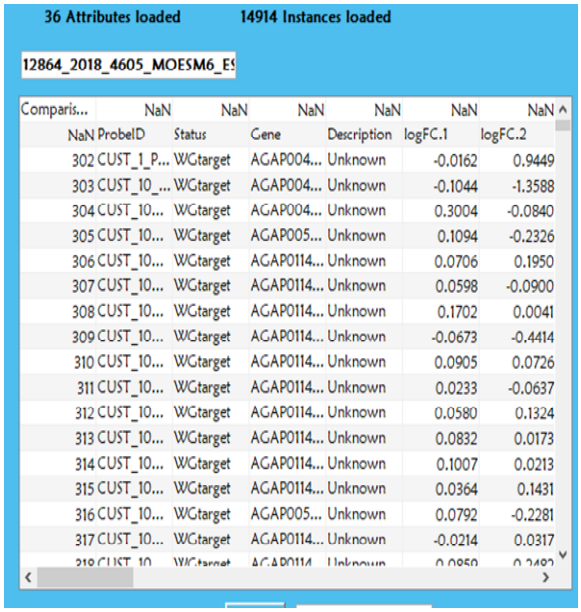
The descriptions of the gene expression datasets are shown in Table 1, comprising of the number of 36 samples and 14,914 numbers of gene samples.

These data are characteristically high-dimensional with small samples. ANOVA uses a p-value of 0.5 to fetch out relevant information from the high dimensional data using this experimental method. Five thousand nine hundred eighty-three features were selected, the reduced data are passed into the ACO to further fetch for the latent relevant subset of the features, and 1,641 features were selected. The amount of the sample population is $X = 5{,}000$ in Algorithm 1, the maximum number of iterations is set as 2,000 for the ACO, and since the amount Q of pheromone on the path from ants in iterations is related to the distance between notes p and q 1,560, one sets $Q = 3{,}000$.

The experimental operating system uses Windows 10, Intel Core i5, and 8 GB RAM with MATLAB 2015A.

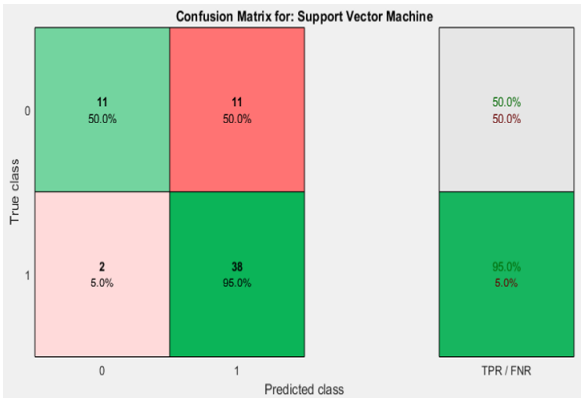
This proposed experiment evaluates the classification performance of our proposed algorithm in terms of classification accuracy. The classification accuracies of the ANOVA-ACO-SVM algorithm are compared with those of the state-of-the-art related. Figure 2 showed the loaded data.

Figure 2 Dataset for malaria vector (see online version for colours)



In this study, the selected data using ANOVA is passed into the SVM classifier, with ten folds cross-validation, 80% training, and 20% testing. Figure 3 shows the confusion matrix for the classification.

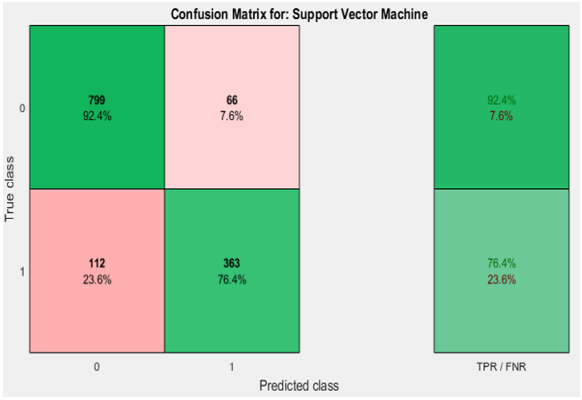
Figure 3 ANOVA-SVM confusion matrix (see online version for colours)



Notes: TP = 38; TN = 11; FP = 11; FN = 2.

To validate the gene expression data experiment, ANOVA-ACO is used, with ten folds cross-validation, with 80% training ad 20% testing. The confusion matrix of the experiment is shown in Figure 4.

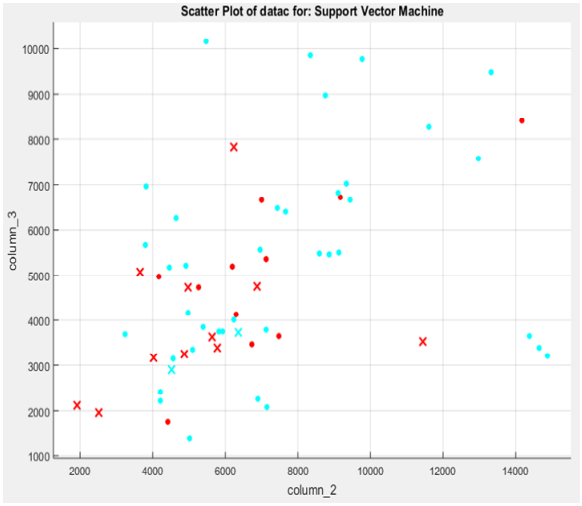
Figure 4 ANOVA-ACO-SVM confusion matrix (see online version for colours)



Notes: TP = 363; TN = 799; FP = 66; FN = 112.

To visualise the gene expression values, the scattered plot shows the selected features in Figure 5.

Figure 5 The scattered plot for the selected features, depicting the variables X and Y (see online version for colours)



In this study, dimensionality reduction approaches using ANOVA and ANOVA-ACO were carried out to classify malaria vector datasets. SVM was used as a classification algorithm to evaluate the performance of the experiment. Table 2 shows the performance evaluation and compares it with the state of the art in Table 3.

Table 2 Performance evaluation for the experiment

Performance metrics (%)	*ANOVA-SVM*	*ANOVA-ACO-SVM*
Accuracy	79.03	86.71
Sensitivity	95	76.42
Specificity	50	92.37
Precision	77.55	84.62
F1-score	85.39	70.53

In this study, a two-stage feature selection experiment was carried out using ANOVA-ACO and classified using SVM. The experiment achieved about 87% accuracy and compared with the state-of-the-art. The experiment shows a promising performance that clinicians can adopt in the decision-making of drugs and transmission resistance of medications and insecticides. In summary, the proposed method significantly reduced the dimensional gene expression dataset with relevant outcomes.

Table 3 Comparison of the proposed system with existing models

Authors	*Experiments*	*Results%*
Srinivasu et al. (2021)	MobileNet V2-LSTM	93
Zahoor and Zafar (2020)	ITO	88
Sun et al. (2019)	Fishers	83.8
Salman et al. (2018)	ANN-PSO	80.34
Sun et al. (2019)	Relief-F-ACO	83.2
Azar et al. (2016)	PMGRS	90

5 Conclusions

Identification and classification of relevant genes in high-dimensional clinical data have been explored widely. In this paper, a two-stage feature selection method, based on ANOVA-ACO, was proposed to reduce the dimensionality of dataset genes and improve classification performance. Firstly, ANOVA, as a filter-based feature selection approach, was used to select relevant information for a given malaria vector dataset. It presented relevant features with the p-value of 0.5 and 5,983 relevant features. To further improve the classification performance, the ACO algorithm as a wrapper-based method was used to select the optimal subtype of the selected features achieved from ANOVA. Finally, the SVM classification was used on the developed model. The experimental result achieved about 87% accuracy and shows that the proposed model is a promising one that clinicians can adopt for decision-making. However, the sensitivity was discovered to be relatively low and can be considered for improvement in future work by considering other classifiers such as the KNN or hybridisation of the SVM classifier. The results have proven and open gap for futuristic decisions, the ACO algorithm is focused to the optimisation of the classification specifications and signifies an accurate and consistent quality of interactions, as proved by this investigation, which was premised on the implementation and advancement of machine learning techniques for malaria infection. This study's main limitation is that the proposed method is its sufficiency for biological exploration of other ailments, as the promising approach cannot optimally balance sizes of gene features and classification accuracy. Hence further experiment is required to adopt the development model for other bioinformatics discoveries and further improve the classification using other optimisation strategies and classifiers like the PSO and the KNN.

References

Adebiyi, M.O., Arowolo, M.O. and Olugbara, O. (2021) 'A genetic algorithm for prediction of RNA-Seq malaria vector gene expression data classification using SVM kernels', *Bulletin of Electrical Engineering and Informatics*, Vol. 10, No. 2, pp.1071–1079 [online] https://doi.org/10.11591/eei.v10i2.2769

Almugren, N. and Alshamlan, H. (2019) 'A survey on hybrid feature selection methods in microarray gene expression data for cancer classification', *IEEE Access*, Vol. 7, pp.78533–78548 [online] https://doi.org/10.1109/ACCESS.2019.2922987.

Amara, K., Malek, A., Bakir, T., Fekik, A., Azar, A.T., Almustafa, K.M., Bourennane, E.B. and Hocine, D. (2019) 'Adaptive neuro-fuzzy inference system based maximum power point tracking for stand-alone photovoltaic system', *International Journal of Modelling, Identification and Control*, Vol. 33, No. 4, p.311 [online] https://doi.org/10.1504/IJMIC.2019.107480.

Arowolo, M.O., Abdulsalam, S.O., Saheed, Y.K. and Salawu, M.D. (2016) 'A feature selection based on one-way-ANOVA for microarray data classification', *Al-Hikmah Journal of Pure and Applied Sciences*, Vol. 3, pp.30–35.

Arowolo, M.O., Adebiyi, M.O., Adebiyi, A.A. and Olugbara, O. (2021a) 'Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier', *Journal of Big Data*, Vol. 8, No. 1, p.29 [online] https://doi.org/10.1186/s40537-021-00415-z.

Arowolo, M.O., Adebiyi, M.O., Aremu, C. and Adebiyi, A.A. (2021b) 'A survey of dimension reduction and classification methods for RNA-Seq data on malaria vector', *Journal of Big Data*, Vol. 8, No. 1, p.50 [online] https://doi.org/10.1186/s40537-021-00441-x.

Awotunde, J.B., Folorunso, S.O., Bhoi, A.K., Adebayo, P.O. and Ijaz, M.F. (2021a) 'Disease diagnosis system for IoT-based wearable body sensors with machine learning algorithm', *Intelligent Systems Reference Library*, Vol. 209, pp.201–222.

Awotunde, J.B., Ajagbe, S.A., Oladipupo, M.A., Awokola, J.A., Afolabi, O.S., Mathew, T.O. and Oguns, Y.J. (2021b) 'An improved machine learnings diagnosis technique for COVID-19 pandemic using chest X-ray images', *Communications in Computer and Information Science*, October, 1455 CCIS, pp.319–330.

Awotunde, J.B., Jimoh, R.G., Oladipo, I.D. and Abdulraheem, M. (2021c) 'Prediction of malaria fever using long-short-term memory and big data', *Communications in Computer and Information Science*, Vol. 1350, pp.41–53.

Ayo, F.E., Awotunde, J.B., Ogundokun, R.O., Folorunso, S.O. and Adekunle, A.O. (2020) 'A decision support system for multi-target disease diagnosis: a bioinformatics approach', *Heliyon*, Vol. 6, No. 3, p.e03657.

Azar, A.T., Kumar, S.S., Inbarani, H.H. and Hassanien, A.E. (2016) 'Pessimistic multi-granulation rough set-based classification for heart valve disease diagnosis', *International Journal of Modelling, Identification and Control*, Vol. 26, No. 1, p.42 [online] https://doi.org/10.1504/IJMIC.2016.077744.

Basavegowda, H.S. and Dagnew, G. (2020) 'Deep learning approach for microarray cancer data classification', *CAAI Transactions on Intelligence Technology*, Vol. 5, No. 1, pp.22–33 [online] https://doi.org/10.1049/trit.2019.0028.

Berglund, B. (2015) 'Environmental dissemination of antibiotic resistance genes and correlation to anthropogenic contamination with antibiotics', *Infection Ecology & Epidemiology*, Vol. 5, No. 1, p.28564.

Bhoi, A.K. (2017) 'Classification and clustering of Parkinson's and healthy control gait dynamics using LDA and K-means', *International Journal Bioautomation*, Vol. 21, No. 1, p.19.

Bir-Jmel, A., Douiri, S.M. and Elbernoussi, S. (2019) 'Gene selection via a new hybrid ant colony optimization algorithm for cancer classification in high-dimensional data', *Computational and Mathematical Methods in Medicine*, pp.1–20 [online] https://doi.org/10.1155/2019/7828590.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J. and Lang, M. (2020) 'Benchmark for filter methods for feature selection in high-dimensional classification data', *Computational Statistics & Data Analysis*, Vol. 143, p.106839 [online] https://doi.org/10.1016/j.csda.2019.106839.

Chakravarty, S., Bisoi, R. and Dash, P.K. (2020) 'A hybrid kernel extreme learning machine and improved cat swarm optimization for microarray medical data classification', in *Data Analytics in Medicine*, pp.779–814, IGI Global [online] https://doi.org/10.4018/978-1-7998-1204-3.ch042.

Isaacs, A.T., Mawejje, H.D., Tomlinson, S., Rigden, D.J. and Donnelly, M.J. (2018) 'Genome-wide transcriptional analyses in anopheles mosquitoes reveal an unexpected association between salivary gland gene expression and insecticide resistance', *BMC Genomics*, Vol. 19, No. 1, p.225 [online] https://doi.org/10.1186/s12864-018-4605-1.

Kayano, M., Matsui, H., Yamaguchi, R., Imoto, S. and Miyano, S. (2016) 'Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection', *Biostatistics*, Vol. 17, No. 2, pp.235–248 [online] https://doi.org/10.1093/biostatistics/kxv037.

Kumar, M., Rath, N.K., Swain, A. and Rath, S.K. (2015) 'Feature selection and classification of microarray data using MapReduce based ANOVA and K-nearest neighbor', *Procedia Computer Science*, Vol. 54, pp.301–310 [online] https://doi.org/10.1016/j.procs.2015.06.035.

Mahani, A. and Ali, A.R.B. (2020) 'Classification problem in imbalanced datasets', in *Recent Trends in Computational Intelligence*, IntechOpen [online] https://doi.org/10.5772/intechopen.89603.

Mallika, R. and Saravanan, V. (2010) 'An SVM based classification method for cancer data using minimum microarray gene expressions', *World Academy of Science, Engineering and Technology*, Vol. 38, pp.543–547.

Mishra, S., Mallick, P.K., Tripathy, H.K., Bhoi, A.K. and González-Briones, A. (2020a) 'Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier', *Applied Sciences*, Vol. 10, No. 22, p.8137.

Mishra, S., Tripathy, H.K., Mallick, P.K., Bhoi, A.K. and Barsocchi, P. (2020b) 'EAGA-MLP – an enhanced and adaptive hybrid classification model for diabetes diagnosis', *Sensors*, Vol. 20, No. 14, p.4036.

Mohapatra, P., Chakravarty, S. and Dash, P.K. (2016) 'Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system', *Swarm and Evolutionary Computation*, Vol. 28, pp.144–160 [online] https://doi.org/10.1016/j.swevo.2016.02.002.

Moosa, J.M., Shakur, R., Kaykobad, M. and Rahman, M.S. (2016) 'Gene selection for cancer classification with the help of bees', *BMC Medical Genomics*, Vol. 9, No. S2, p.47 [online] https://doi.org/10.1186/s12920-016-0204-7.

Oladipo, I.D., Babatunde, A.O., Awotunde, J.B. and Abdulraheem, M. (2021) 'An improved hybridization in the diagnosis of diabetes mellitus using selected computational intelligence', *Communications in Computer and Information Science*, Vol. 1350, pp.272–285.

Potharaju, S.P. and Sreedevi, M. (2019) 'Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance', *Clinical Epidemiology and Global Health*, Vol. 7, No. 2, pp.171–176 [online] https://doi.org/10.1016/j.cegh.2018.04.001.

Sajana, T. and Narasingarao, M.R. (2018) 'A comparative study on imbalanced malaria disease diagnosis using machine learning techniques', *Journal of Advance Research in Dynamical and Control Systems*, Vol. 10, No. 2, pp.552–561.

Salman, I., Ucan, O., Bayat, O. and Shaker, K. (2018) 'Impact of metaheuristic iteration on artificial neural network structure in medical data', *Processes*, Vol. 6, No. 5, p.57 [online] https://doi.org/10.3390/pr6050057.

Satapathy, S.K., Bhoi, A.K., Loganathan, D., Khandelwal, B. and Barsocchi, P. (2021) 'Machine learning with ensemble stacking model for automated sleep staging using dual-channel EEG signal', *Biomedical Signal Processing and Control*, Vol. 69, p.102898.

Sato, I. and Nakagawa, H. (2014) 'Approximation analysis of stochastic gradient Langevin dynamics by using Fokker-Planck equation and ITO process', in *International Conference on Machine Learning*, PMLR, Beijing, China, pp.982–990.

Saw, T. and Myint, P.H. (2019) 'Swarm intelligence based feature selection for high dimensional classification: a literature survey', *International Journal of Computer*, Vol. 33, No. 1, pp.69–83.

Selma, B., Chouraqui, S. and Abouaïssa, H. (2020) 'Hybrid ANFIS-ant colony based optimisation for quadrotor trajectory tracking control', *International Journal of Modelling, Identification and Control*, Vol. 34, No. 1, p.13 [online] https://doi.org/10.1504/IJMIC.2020.108913.

Shamsaei, B. and Gao, C. (2016) 'Comparison of some machine learning and statistical algorithms for classification and prediction of human cancer type', *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp.296–299 [online] https://doi.org/10.1109/BHI.2016.7455893.

Singh, D., Paul Choudhury, J. and De, M. (2020) 'A comparative study of meta heuristic model to assess the type of breast cancer disease', *IETE Journal of Research*, pp.1–12 [online] https://doi.org/10.1080/03772063.2020.1775139.

Srinivasu, P.N., SivaSai, J.G., Ijaz, M.F., Bhoi, A.K., Kim, W. and Kang, J.J. (2021) 'Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM', *Sensors*, Vol. 21, No. 8, p.2852 [online] https://doi.org/10.3390/s21082852.

Sun, L., Kong, X., Xu, J., Xue, Z., Zhai, R. and Zhang, S. (2019) 'A hybrid gene selection method based on relief-F and ant colony optimization algorithm for tumor classification', *Scientific Reports*, Vol. 9, No. 1, p.8978 [online] https://doi.org/10.1038/s41598-019-45223-x.

Tahir, M.A.U.H., Asghar, S., Manzoor, A. and Noor, M.A. (2019) 'A classification model for class imbalance dataset using genetic programming', *IEEE Access*, Vol. 7, pp.71013–71037 [online] https://doi.org/10.1109/ACCESS.2019.2915611.

Townes, F.W., Hicks, S.C., Aryee, M.J. and Irizarry, R.A. (2019) 'Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model', *Genome Biology*, Vol. 20, No. 1, p.295 [online] https://doi.org/10.1186/s13059-019-1861-6.

Wang, B., Yao, M. and Zhu, R. (2011) 'Gene expression programming with multiple chromosomes', *International Journal of Modelling, Identification and Control*, Vol. 14, No. 4, p.235 [online] https://doi.org/10.1504/IJMIC.2011.043145.

Yu, H., Gu, G., Liu, H., Shen, J. and Zhao, J. (2009) 'A modified ant colony optimization algorithm for tumor marker gene selection', *Genomics, Proteomics & Bioinformatics*, Vol. 7, No. 4, pp.200–208 [online] https://doi.org/10.1016/S1672-0229(08)60050-9.

Zahoor, J. and Zafar, K. (2020) 'Classification of microarray gene expression data using an infiltration tactics optimization (ITO) algorithm', *Genes*, Vol. 11, No. 7, p.819 [online] https://doi.org/10.3390/genes11070819.