# Optimized Hybrid Heuristic Based Dimensionality Reduction Methods for Malaria Vector Using KNN Classifier

**Micheal Olaolu Arowolo** ( arowolo.olaolu@gmail.com )
Landmark University https://orcid.org/0000-0002-9418-5346
**Marion Olubunmi Adebiyi**
Landmark University
**Ayodele Ariyo Adebiyi**
Landmark University
**Oludayo Olugbara**
Durban University of Technology

**Research**

**Keywords:** Classification, Dimensionality Reduction, Hybrid, Malaria Vector.

**Posted Date:** November 18th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-107396/v1



**Version of Record:** A version of this preprint was published on February 4th, 2021. See the published version at https://doi.org/10.1186/s40537-021-00415-z.

# Abstract

RNA-Seq data are utilized for biological applications and decision making for the classification of genes. A lot of works in recent time are focused on reducing the dimension of RNA-Seq data. Dimensionality reduction approaches have been proposed in the transformation of these data. In this study, a novel optimized hybrid investigative approach is proposed. It combines an optimized genetic algorithm with Principal Component Analysis and Independent Component Analysis (GA-O-PCA and GAO-ICA), which are used to identify an optimum subset and latent correlated features, respectively. The classifier uses KNN on the reduced mosquito Anopheles gambiae dataset, to enhance the accuracy and scalability in the gene expression analysis. The proposed algorithm is used to fetch relevant features based on the high-dimensional input feature space. A fast algorithm for feature ranking is used to select relevant features. The performances of the model are evaluated and validated using the classification accuracy to compare existing approaches in the literature. The achieved experimental results prove to be promising for selecting relevant genes and classifying pertinent gene expression data analysis by indicating that the approach is a capable addition to prevailing machine learning methods.


# Introduction

A major problem in the bioinformatics field is the collection of genes from high-throughput biological data. The gene expression data are known for having small samples with large irrelevant and redundant noisy genes. Gene expression data analysis comprises of small and large samples with irrelevant and redundant gene sequences. These gene sequences depreciate classification learning model performances. Dimensionality reduction techniques have been used severally. It has been used to fetch relevant discriminative subsets from the gene expression data; it also assists in saving computational burdens and improving classification prediction accuracy (Pashaei et al. 2019).

In gene expression data analysis, overfitting and curse of dimensionality have been known to deteriorate the classification capabilities. It comprises of high dimensional input space called the curse of dimensionality. Overcoming the curse of dimensionality challenges, several dimensionality reduction techniques have been exploited in literature. They are determining optimal subset genes helpful for revealing hidden features of genes and enhance their interpretability is a major. The dimensionality reduction aim is to discover the trivial subset of genes that can help improve prediction performances, which will be helpful to clinicians in decision making and treatments (Shukla et al. 2019).

Several authors have addressed the problems of the curse of dimensionality. Metaheuristics have also been proposed, yet approaches suffer from correlations, high throughputs, and increase in computational time for fetching gene subsets (Cai et al. 2018; Marfaja & Mirjalili, 2018). A systematic approach to fetching an optimal subset gene is a crucial issue.

Feature selection (filter, wrapper and embedded) (Tadist et al. 2019; Liu et al. 2020; Chen et al. 2020;) and feature extraction (Aziz et al. 2017; Wenric & Shemirani, 2018; Bajaj et al. 2020) (supervised and

unsupervised) are dimensionality reduction approaches that have been established, these approaches have overcome several problems such as performance enhancement, yet there is need for improvements hybrid model and optimization for getting better results (Panshaei et al. 2019).

Finding an optimal subset of genes proficient at handling high dimension optimization difficulties with reasonable solutions is required.

Genetic algorithm (GA), is a feature selection technique; it is a wrapper-based which is represented by an optimization technique. GA is said to be adaptive heuristic search approach that finds an optimal subset of features in complex problems such as high dimensionality (Chiesa et al. 2020). GA is proficient for finding optimal subsets on high dimensional data and have been used extensively, yet they are computationally expensive and prone to overfitting. Overcoming this limitation, optimization strategies have been used to ensure better performances for finding optimum feature subsets and classification accuracy.

Principal component analysis (PCA) (non-linear) and Independent component analysis (ICA) (linear) are appropriate feature extraction methods that have been extensively used (Aziz et al. 2016; Kong et al. 2018) are common capable methods for fetching subset of gene samples for classification and have received growing attention in recent time (Mohan et al. 2014). The hybrid approach has proven to be significant, due to their good performances and advantages for solving dimensionality problems that halt classification, it is of the essence to come up with efficient models that are computationally fast and easy to implement for classification of gene expression data analysis (Chuang et al. 2012).

Several experiments have been carried out in literature (Hira & Gillies, 2015; Wang et al. 2017; Arowolo et al. 2017; Pragadeesh et al. 2019; Lin & Zhang, 2019; Tadist et al. 2019; Pashaei et al. 2019; Shukla et al. 2019; Chen et al. 2020; Chen et al. 2020; Liu et al. 2020). However, these experiments necessitate enhancements that can help in making decisions on how to eradicate the transmission of malaria in West Africa, as it is a scourge in Africa (Hodgson et al. 2019).

This study proposes a hybrid dimensionality reduction model for the classification of malaria vector data. Based on the approaches, an optimized genetic algorithm (GA-O) is used to fetch out subset relevant genes. The PCA and ICA are used on the subset data, to fetch latent components in the data. Combining GA-O with PCA and GA-O with ICA, are classified using KNN on a Mosquito Anopheles gambiae dataset. This study proposes to improve the classification complexities such as the computational cost, fetching relevant subset genes and relationship among genes that can be used by clinicians for decision making.

# Materials And Methods

### Datasets

RNA-Seq for gene expression data analysis uses the mosquito Anopheles gambiae (*Ag.*) larvae, from Kenya western region. It comprises of deltamethrin susceptible and resistant mosquitoes' profile with

considerate resistance devices; with 7 attributes relating to the Tests, Genes, Genes identities, Locus, Susceptible, Resistant, Status and a predictor from 2457 instances (Arowolo et al. 2020).

**Table 1. Features of the dataset**

| Dataset | Features | Samples |
|---|---|---|
| Mosquito (*Ag.*) Anopheles Gambiae | 7 | 2457 |

## Methodology

RNA-Seq gene expression data is a widely utilized technology for diagnostic of several diseases, such as cancer, malaria, among others. It recognizes several aspects of transcriptomes, which is a principal existing technology for high-throughput genetic factors. Providing enhanced insight for transcriptome cells, alternative therapies and improved determinations (Zhao & Leung, 2014), it identifies early secret variations occurring in conditions of disease by reacting to different environments and other training therapeutics, producing sufficient quantities of sequencing data (Hyung et al. 2019). Gene expression Classification of RNA-Seq data has provided valuable evidence to classify and assess German medications for diseases. The expression of genes is genomic factors in the predominant method of RNA-Seq quantifying and gaining a better understanding of various biological tissues. The problem of diagnostic challenges is a major challenge for RNA-Seq, and owing to the high dimensional gene data expression, it gives unfitting results.

In this study, the dataset uses a mosquito Anopheles gambiae. The samples of the genes are normalized using the MATLAB tool package. The samples are passed into the optimized genetic algorithm. A reduced sample is then achieved and passed into the PCA and ICA separately. The further reduced data are split into training and testing sets.  The training set is the vigorous samples, and testing set is the outstanding samples. Classification is conducted using KNN.

## Dimension Reduction

A recognized technique for eliminating unwanted noise and unnecessary features is dimensionality reduction. Gene expression data comprises of high dimensional features that amount to computational weightiness, depriving the performance of classification models. To eliminate redundancy. Obtaining irrelevant features that interrupt efficiency with activity by reducing the samples of feature ratios, dimensionality reduction procedures are essential. This method helps in reducing risks of overfitting. Reducing the dimensionality is an important method known as the collection of features and extraction of features (Shen et al. 2017; Sahu et al. 2018).

## Feature Selection

Technologies such as RNA-seq transcriptomes, constructing relevant particular feature identifiers for sequences transcript is essential, to train and test models. Feature selection is important to create a better

classification performance. Selection of features allows choosing of suitable elements for in classification model performances by removing irrelevant and redundant features which minimize the curse of dimensionality. It helps to make the classification phase learning procedure successful and increases the success model. For example, extensive information feature selection process; RNA-Seq data involves supervised and unattended decision-making learning. For classification problems, rank characteristics conferring significance are important, and selecting the best will advance the prediction model's performance. The collection of feature selection is an efficient technique identified as a filter, wrapper and embedded types (Jabeen et al. 2017).

### Genetic Algorithm (GA)

Genetic algorithm is a wrapper based evolutionary algorithm for selecting relevant features, used in investigating engine optimization problems. In the survival of the fittest base, GA is based on actual activities linked to human genetic factors. GA is made up of initial population development, fitness assessment, parent selection, crossover and mutation (Uma & Kirubakaran, 2016; Motieghader et al. 2017).

GA is an investigative discovery method, in a simple procedure, with a sample of randomly generated outcomes (phenotypes or entities) offering an acceptable value for the main purpose of computing the beneficial results. Respective chromosomes or genotypes typically comprises of sets of properties categorized as binary strings of 0’s and 1’s (Wang et al. 2017). While very sensitive to the initial population, GA has a weakness of optimality. Its result quality declines as problem dimensions rise, it has been shown to produce reasonable quality solutions to boost it for gene sampling.

### Feature Extraction

Extraction of features is a technique used in the identification of important features, characteristics or features existing in data. Feature extraction technique examples are the identification of patterns and the detection of public instances in a set of identifications. Data with dimensional loads include the use of feature extraction, for producing a clearer explanation of characterizations. Feature extraction allows revolutionary selected feature variables to decrease the presence of the curse of dimensionality. There are two broad collections of feature extraction procedures, explicitly: linear (assumes data on low-dimensional subspace, such as PCA) and non-linear (assumes a low-dimensional subspace, characterized with a high-dimensional feature, such as ICA) for a non-linear relation between features (Hira & Gillies, 2015).

### Principal Component Analysis (PCA)

PCA is a method of linear feature extraction; it is commonly used primarily in genetic studies. Through reformatting the k-dimensional discrete features from exclusive n-dimension feature field, PCA projects feature spaces from high to lower dimensions. PCA has acknowledged that it is an important method for the exploration of high-dimensional knowledge on gene expression. It is widely used for RNA-seq data. By

transforming a set of correlated variables into a set of uncorrelated variables, investigating orthogonal alteration. PCA for the study of experimental results. PCA may be used to analyze the relationships between a set of variables and to minimize dimensionality (Jain & Singh, 2018).

### Independent Component Analysis (ICA)

Disintegrating multivariate signs into independent non-gaussian for statistically independent components, ICA supports finding hidden features from multidimensional details. By decorrelating the data, ICA seeks a connection between information by manipulating or lessening the relevant data. As a linear combination of the independent components I, ICA adopts Opinion X. If B means columns of B define the separate weighted matrix R, the basis feature vectors of observation X.

$$I = R \text{ to } X,\ X = B \text{ to } S \qquad (1)$$

For biological information, recognition and other reasons, ICA have been used extensively (Hashemi et al. 2018; Feng et al. 2020).

PCA is a linear alteration technique, used to minimize the dimension and number of features. It is a "non-linear" algorithm, while ICA is "linear," if a data is preprocessed, ICA has been shown to perform better (Hira & Gillies, 2015).

### Classification

In data mining techniques, classification is a supervised learning method. It is a common, supportive task that gives and predicts class labels specified from the predefined class label to current data. The building of classification is comprised of two steps (Arowolo et al. 2017):

- The learning process, in which the classification model was developed with a class label giving a collection of training data.
- The model predicts the class labels for concealed data and to calculates the accuracy of the KNN classifier.

### K-Nearest Neighbor (KNN)

A supervised learning $K^{th}$ nearest neighbour classification technique for gene datasets performs the benefit of creative application event assessment of neighbourhood classification. The KNN algorithm classifies creative entities based on examples, characteristics and training models. KNN classifiers do not train models to suit but are retention-based. The selected features are assumed to be inputs for segments. The K value of the closest neighbours is selected nearest to the spot of the question. Based on the minimum determined distance of Kth, detachment between query-instance and training models is taken into account and sorted. Group Y is taken from the closest neighbours. The unassuming prevalence of groups of nearest neighbours is used as the approximate number of instances of question. Bonds can fragment randomly (Bose, 2016).

Increasing the dimensionality of biological data is a major problem for simple, predictable research methods. It is important to use traditional approaches for learning complex strategies on several layers moved by morphological processes interested in processing. Several complexities are involved in most typical procedures used to deal with high-dimensional data, such as the RNA-Seq data. The combination of different methods for reducing dimensionality will, in essence, take advantage of unique advantages where subset genes obtained from a procedure is supported as an input to the other. In general, feature extraction techniques support feature selection proficiently, by using feature selection to pick the original subset of genes, or by taking advantage of redundant gene elimination. Extracting primary subset features, combining various feature extraction methods can be useful (Motieghader et al. 2017; Sun et al. 2018; Arowolo et al. 2020).

An effective dimension reduction method to classify malaria vector data was suggested in this report.

RNA-Seq has tremendous potential for finding, defining and tracing cell lines. Still, the reduction of dimensionality helps to perceive the structures. Still, data remains difficult, and current algorithms need the correct development to reveal suitable characteristics, fusion approach proves to be strong but necessitates effective procedures to model.

The classification technique proposed consists of three Phases, namely:

- Selection of features
- Extraction of features
- Category of category

Figure 1 illustrates the projected hybrid system for classifying malaria gene expression dataset. The framework consists of three subsystems, a subsystem for feature collection, a class-based subsystem for feature extraction, and a subsystem for classification.

By adopting one algorithm below to pick an optimum subset by assessing the chromosome fitness, the function selection sub-system uses an optimized GA. The function extraction subsystem uses PCA and ICA because of its data projection of efficiency invariance along with impertinent orders. The standard of the researches is categorized using KNN.

Significances of genetic algorithm optimization are its evolutionary dispensation of the algorithm's features; it helps numerous search point which simultaneously and independently explores the optimal result to produce a good result. In this study, an optimization of the collection of genetic algorithm features to minimize numbers of features and maintain discriminant features. The extraction of features is ideal; it transforms reduced data to latent elements, the productivity is to lessen prosperity and suffer from both methods of reduction of dimensionality used for classification of malaria.

### Algorithm 1:

Phase 1: formulate the a and b parameters then establish the initial population arbitrarily.

Phase 2: For i<population size

Phase 3: calculate the tangent rate tan(xin/xin+1) of the intricate angle among two vectors in end-to-end dimensions for individual pop(i)

Phase 4: if Phase 3=0, update the value of the nth dimension of ith discrete to 0; do not update value, continue phase6

phase 4.1: if x>1, Phase 2:

Phase 4.2: measure compatibility between xi and xj of persistence and Euclidean distance D

Phase 4.3: compute evaluation attitude in the distance L=|Xi-Xj|<D

Phase 4.4: if no comparison,

1. Eliminate discrete fitness with the equivalent of biallelic loci (SD (Xi, Xj)) and consistent parallel MSDi
2. compute subpopulations M(t+1);

else

Combine N entities in memory pool with subpopulation prepared by fitness in descending order

Phase 4.5: calculate subpopulation * threshold

Phase 5: judge convergence condition

Phase 6: calculate the number of 0 fundamentals in each dimension for the restructured population; if above critical value Q, delete the dimension

Phase 7: get an updated population

Phase 8: calculate fitness value F(i) of respective discrete in the population

Phase 9: set new population

Phase 10: pick two entities from population rendering fitness with relational selection algorithm

Phase 11: if random (0, 1) < Pc, then move on to Phase 12; else, implement Phase 13

Phase 12: apply the crossover operative rendering to the crossover probability Pc on the two entities

Phase 13: if random (0, 1) < Pm, move to Phase 14

Phase 14: apply mutation operator to mutation probability Pm on two individuals

Phase 15: add two new individuals into a new population

Phase 16: Repeat process till N-th generation is generated; else, return to Phase 4

Phase 17: change population with the new population

Phase 18: Reiterate procedure till number of groups exceeds G; else, return to Phase 8

Phase 19: end

Genetic algorithm realizes pertinent features in the data using the optimized genetic algorithm in algorithm 1. The carefully chosen features are used by the PCA and ICA feature extraction phases distinctly to fetch for fundamental components. KNN classifiers are used to analyze the performance metrics for the learning procedure.

Phase 1: Preprocess imported data

Phase 2: Apply Optimized Genetic algorithm

Phase 3: Apply PCA algorithm on selected features from Phase 2

Phase 4: Apply KNN Classification on Phase 3 outcome

Phase 5: Evaluate performance

Phase 6: Repeat Phase 3 using ICA algorithm

Phase 7: Apply KNN algorithms on level 6 output.

Phase 8: Evaluate performance

Phase 9: Compare results of Phase 5 and Phase 6

Experiments in this study are performed using Intel Core 5 with a 16GB RAM, and 64-bit Operating system. All algorithms were coded in C++ on MATLAB 2015 environment platform.

The confusion matrices were used as the classification evaluation to certify comparable training and testing performances of the experiments in terms of accuracy, sensitivity, among other metrics (Arowolo et al. 2020).

# Results And Discussion

This study proposes a malaria vector dataset classification, using a public dataset, with 2457 samples and 7 features (Arowolo et al. 2020), on a MATLAB tool. The dataset was investigated using an optimized genetic algorithm to pick pertinent features in the data, using 0.5 thresholds, 708 significant subset

features were selected. Classifier ability associated with the state-of-the-art was used for required evaluations.

The selected 708 features by the Optimized Genetic algorithm is first conceded into PCA algorithm with an extracted output of 10 latent variables in 1.4623 seconds. The results of the extracted features are classified using the KNN classification algorithm with 10-fold cross-validation. The KNN Confusion matrix was then evaluated using the performance metrics analysis.

The 708 selected features furthermore were conceded into the ICA algorithm and extracted 25 latent variables in 0.42794 seconds. The latent features were classified on KNN with 10-folds cross-validation, and the confusion matrix is evaluated.

Dimensional reduced malaria vector data was carried out, using GA-O + PCA + KNN and GA-O + ICA + KNN algorithms and the performance evaluations of the experiments are tabulated below.

This study shows numerous significant suggestions for analyzing data gene expressions. The potential application of this experiment is to give relevant understanding into genetic and technical deliberations that can clarify revealed structures and elucidations for genes appropriate for predictions, analysis, detections of malaria infections, transmissions and drug designs.

# The Ga-o With Pca With K-nn Results

# The Ga-o With Ica With K-nn Results

Table 2
Performance Metrics Table for the GA-O + PCA + K-NN and GA-O + ICA + KNN Classification

| Performance Metrics (%) | GA-O + PCA + K-NN | GA-O + ICA + K-NN |
|---|---|---|
| Accuracy | 88.3 | 90 |
| Sensitivity | 100 | 100 |
| Specificity | 52.4 | 52.4 |
| Precision | 79.6 | 86.7 |
| Recall | 100 | 100 |
| F-score | 88.6 | 92.88 |

As stated in Table 2, this study attained reliable performances with useful algorithms comparatively.

This study proposed a hybrid dimension reduction approach using an optimized Genetic algorithm. PCA and ICA algorithms were used as on the selected features. KNN algorithm, using 10-fold cross-validation

parameter, was used to classify the experiment. The result showed an enhanced result, as revealed in Table 2. Compared to the state-of-the-art, the accuracies presented an improvement.

Providing a dependable discovery and prediction method for malaria infection and transmission, numerous investigators have studied underlying classification problems using machine learning methods. Results achieved can be proposed to train prevalent malaria infections by clinicians, through the use of this procedure to compile curated diagnostic dataset to train classifiers and increase approaches for datasets to increase the dataset size significantly, concerning the overfitting difficulties related the training of datasets. The study of illustrating thousands of genes suggests unfathomable understanding into malaria classification complications with ample of data discoveries, for drug finding, prediction and diagnosis of malaria treatments as well as understanding roles of genes with the communication between the genes in common and irregular situations. This study grew the classification performance results and demonstrated a less dependence training set.

Table 3
Comparative of the Performance Metrics with the State-of-the-art

| Performance Metrics (%) | GLM + PCA (Feng et al. 2017) | GA + PCA + NN (Sumsi et al. 2018) | GA + CCA + NN (Sumsi et al. 2018) | GA + PCA&CCANN (Sumsi et al. 2018) |
|---|---|---|---|---|
| Accuracy | 70 | 85.0 | 85 | 88 |

# Conclusion

Data analysis of RNA-Seq offers valuable and important benefits to the technology ‘s success, with tremendous helps to evolve the problems of gene expression profiling. RNA-Seq’s related applications include the reduction of dimensionality and classification approaches. Due to the curse of dimensionality bound in the data of gene expression, it is a critical problem. Several strategies have been proposed to develop the technology, predict and detect diseases extracted from samples, and the reduction of dimensionality has proved to overcome these challenges. Yet, there is a need to undertake further inquiries. Recently, hybrid methods have also been used to classify gene expression results. GA + ICA + KNN outperformed the GA + PCA + KNN based method by performing a dimensionality reduction method using GA with ICA and GA with PCA algorithms discretely and evaluating their performance on KNN classification kernels.

For futuristic works, this study proposed the application of hybrid dimension reduction procedures on other classifiers such as the ensemble, neural networks, decision trees, to identify the relevant classification of the gene expression data.

# Abbreviations

RNA-Seq

Ribonucleic Acid Sequencing; GA:Genetic Algorithm; GA-O:Optimized Genetic Algorithm; PCA:Principal Component Analysis; ICA:Independent Component Analysis; KNN:Kth Nearest Neighbor; NN:Neural Network; DNA:Deoxyribonucleic Acid; MATLAB:Mathematical Laboratory; ID:Identity; CCA:Canonical Component Analysis; GLM:Generalized Linear Model.

# Declarations


### Acknowledgements

The author would like to thank Landmark University for supporting this work with all the needful experiments in this research.

### Funding

There is no funding presently for this work.


### Authors' contributions

MO Arowolo contributed by carrying out the research as a PhD student under the Supervision and mentoring of Prof. O Olugbara, Prof. AA Adebiyi and Dr MO Adebiyi, who took the role for technical issues. They also advised all process for this work. MO Arowolo wrote the manuscript, while O Olugbara, MO Adebiyi and AA Adebiyi revised the manuscript. All authors read and approved the final manuscript.

### Author information


Micheal Olaolu Arowolo is faculty member, of the Department of Computer Science, Landmark University, Omu-Aran. He is a PhD student in computer science. His research interests are machine learning, data mining, gene expression analysis and computer arithmetic. arowolo.olaolu@lmu.edu.ng

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. marion.adebiyi@lmu.edu.ng

Professor Ayodele Ariyo Adebiyi is a Professor of Computer Science. 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 BSc degree in Computer Science and an MBA degree from University of Ilorin, Ilorin Nigeria. He had his MSc and the 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. ayo.adebiyi@lmu.edu.ng

Prof Oludayo Olugbara graduated with a first-class Bachelor of Science (Hons) in Mathematics from the University of Ilorin in 1991, he was a junior research fellow in at the University of Ilorin, after completing the national youth service corps. In 1993 he commenced his Master's Degree in Mathematics with specialization in Computer Science at the University of Ilorin and completed the degree in 1995. He holds a PhD degree in Computer Science from the University of Zululand in South Africa. He is a Professor of Information Technology at the Durban University of Technology in South Africa. He is a holder of academic awards and scholarships, including the International Federation of Information Processing (IFIP) TC2 sponsored by Microsoft Research Cambridge in 2007 and respected research paper award at International Conference on Machine Learning and Data Analysis, organized by the IAENG International Association of Engineers, San Francisco, the USA in 2012. He is a University Scholar at the University of Ilorin, Member of Marquis Whos' Who in the World (USA), Member of the Association for Computing Machinery (ACM, USA), Member of Computer Society of South Africa (CSSA) and other academic associations. He was awarded honorary referee of the Maejo International Journal of Science and Technology, Thailand in 2007-2010 and 2011. In December 2015, He was awarded an outstanding scientist by the Center for Advanced Research and Design of Venus International Foundation in India. He became an established researcher courtesy of the National Research Foundation (NRF) of South Africa rating in 2017. He has examined several postgraduate theses, dissertations and assessed research publications for professorial appointments both nationally and internationally. He has published widely, and he is a reviewer for many reputable journals. oludayoo@dut.ac.za

### Availability of data and materials

The datasets for this study are available on request to the corresponding author.

### Competing interests

The authors declare that they have no competing interests.

### Author Details

1,2,3 Computer Science Department, Landmark University, Omu-Aran, Nigeria.

4 Department of Computer Science and Information Technology, Durban University of Technology, Durban 4001, South Africa.

## References

1. Pashaei, E., Pashaei, E., & Aydin, N. 2019. Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization. *Genomics* 111(4): 669-686.
2. Shukla, A.K., Singh, P., & Vardhan, M. 2019. A new hybrid wrapper TLBO and SA with SVM approach for gene expression data. *Information Sciences* 503: 238-254.
3. Cai, J., Luo, J., Wang, S., Yang, S. 2018. Feature selection in machine learning: a new perspective, Neurocomputing.
4. Mafarja, M., & Mirjalili, S. 2018. Whale Optimization for wrapper feature selection. *Applied Soft Computing.* 62: 441-453.
5. Tadist, K., Najah, S., Nikolov, N.S., Mrabti, F., & Zahi, A. 2019. Feature selection methods and genomic big data: a systematic review. *Journal of Big Data.* 6(79).
6. Chen, C-W., Tsai, Y-H., Chang, F-R., & Lin, W-C. 2020. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems.* 37(5).
7. Liu, Y., Ju, S., Wang, J., & Su, C. 2020. A New Feature Selection Method for Text Classification Based on Independent Feature Space Search. *Mathematical Problems in Engineering.*
8. Aziz, R., Verma, C.K., & Srivastava, N. 2017. Dimension reduction methods for microarray data: a review. *AIMS Bioengineering,* 4(1): 179-197.
9. Wenric, S., & Shemirani, R. 2018. Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies. *Frontiers in Genetics.*
10. Bajaj, V., Taran, S., Khare, S.K., Sengur, A. 2020. Feature extraction method for classification of alertness and drowsiness states EEG signals. *Applied Acoustics.* 163.
11. Chiesa, M., Maioli, G., Colombo, G.J., & Piacentini, L. 2020. GARS: Genetic Algorithm for the identification of a Robust Subset of features in high-dimensional datasets. BMC Bioinformatics. 21(54).
12. Mohan, A., Rao, M.D., Sunderrajan, S., & Pennathur, G. 2014. Automatic classification of protein structures using physicochemical parameters. Interdiscip. Sci.: Comput. Life Sci., 6: 176-186.
13. Kong, W., Vanderburg, C.R., Gunshin, H., Rogers, J.T., & Huang, X. 2018. A review of independent component analysis application to microarray gene expression data. Biotechniques Future Science. 45(5).
14. Chuang, L., Chu, Y., Li, J.C., Yang, C. 2012. A hybrid BPSO-CGA approach for gene selection and classification of microarray data. Journal of Computational Biology. 19: 68-82.
15. Hodgson, S.H., Muller, J., Lockstone, H.E., Hill, A.V.S., Marsh, K., Draper, S.J., & Knight, J.C. 2019. Use of gene expression studies to investigate the human immunological response to malaria infection. Malaria Journal. 18(418).
16. Pragadeesh, C., Jeyaraj, R., Siranjeevi, K., Abishek, R., & JJeyakumar, G. 2019. Hybrid feature selection using micro genetic algorithm on microarray gene expression data. Journal of Intelligent and Fuzzy Systems. 36(3): 2241-2246.

17. Lin, Z., & Zhang, G. 2019. Genetic algorithm-based parameter optimization for EO-1 Hyperion remote sensing image classification. European Journal of Remote Sensing. 50(1): 124-131.
18. Wang, J., Du, P., Niu, T., & Yang, W. 2017. A novel hybrid system based on a new proposed algorithm—Multi-Objective Whale Optimization Algorithm for wind speed forecasting. Applied Energy. 208: 344-360.
19. Hira, Z.M., & Gillies, D.F. 2015. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Advances in Bioinformatics.
20. Arowolo, M.O., Abdulsalam, S.O., Isisaka, R.M., & Gbolagade, K.A. 2017. A hybrid dimensionality reduction model for classification of microarray dataset. International Journal of Information Technology and Computer Science. 9(11): 57-63.
21. Arowolo, M.O., Adebiyi, M.O., Adebiyi, A.A., & Okesola J.O. 2020. 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 (ICMCECS). 1-8.
22. Zhao, S., Leung, W-P. F-., Bottner, A., Ngo, K., and Liu, X. 2014. Comparison of RNA-Seq and microarray in transcriptome profiling of activated t-cells, PLoS One, 9(1).
23. Huyng, P-C., Nguyen, V-H., & Do, T. 2019. Novel hybrid DCNN-SVM model for classifying RNA-Sequencing gene expression data. 533-547.
24. Shen, L., Jiang, H., He, M. & Liu, G. (2017). Collaborative representation-based classification of microarray gene expression data. Plos One, 12(2).
25. Sahu, B., Dehuri, S., & Jagadev, A. 2018. A study on relevance of feature selection methods in microarray data. The Open Bioinformatics Journal; Bentham, 11: 117-139.
26. Jabeen, A., Ahmad, N., & Raza, K. 2017. Machine Learning-based State-of-the-art Methods for the Classification of RNA-Seq Data.
27. Uma, S.M., & Kirubakaran, E. 2016. A hybrid heuristic dimensionality reduction technique for microarray gene expression data classification: a blending of GA, PSO and ACO. International Journal of Data Mining, Modelling and Management, 8(2): 160-179.
28. Motieghader, H., Najafi, A., Sadeghi, B., & M-Nejad, A. 2017. A Hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. Informatics in Medicine Unlocked, 9: 246-254.
29. Wang, L., Wang, Y., & Chang, Q. 2017. Feature selection methods for big data bioinformatics: A Survey from the search perspective. Methods, 111: 21-31.
30. Jain, D., & Singh, V. 2018. An efficient hybrid feature selection model for dimensionality reduction," International Conference on Computational Intelligence and Data Science, Procedia Computer Science. 123: 333-341.
31. Feng, C., Liu, C., Zhang, H., Guan, R., Li, D., Zhou, F., Liang, Y., & Feng, X. 2020. Dimension reduction and clustering models for single-cell RNA-Seq data: A comparative study. International Journal of Molecular Sciences. 21(2181): 1-21.

32. Hashemi, F. S. G., Ismail, M. R., Yusop, M. R., Hashemi, M. S. G., Shahraki, M. H. N., Rastegari, H., Miah, G., & Aslani, F. (2018). Intelligent mining of large-scale bio-data: Bioinformatics applications. Reviews; Bioinformatic, Biotechnology, and Biotechnological Equipment. 28(1).
33. Bose, J. 2016. Hybrid GA/KNN/SVM Algorithm for Classification of Data. BioHouse Journal of Computer science. 2(2). 5-11.
34. Sun, L., Kong, X., Xu, J., Xue, Z., Zhai, R., & Zhang, S. 2019. A hybrid gene selection method based on Refief-F and Ant colony optimization algorithm for tumor classification," Nature Research Academics. 9(8978).
35. Arowolo, M. O., Adebiyi, M.O., Adebiyi, A.A. An efficient PCA Ensemble learning approach for prediction of RNA-Seq malaria vector gene expression data classification. International Journal of Engineering Research and Technology. 13(1): 163-169.
36. Susmi, S.J., & Nehimiah, H.K. 2018. Hybrid Dimensionality Reduction Techniques with Genetic Algorithm and Neural Network for Classifying Leukemia Gene Expression Data". Indian Journal of Science and Technology, 9(1): 1-8.
37. Feng, C., Lu, S., Zhang, H., & Feng, X. 2018. Dimension Reduction and Clustering Models for Sc-RNA Sequencing Data. International Journal of Molecular Sciences. 21: 1-21.

# Figures

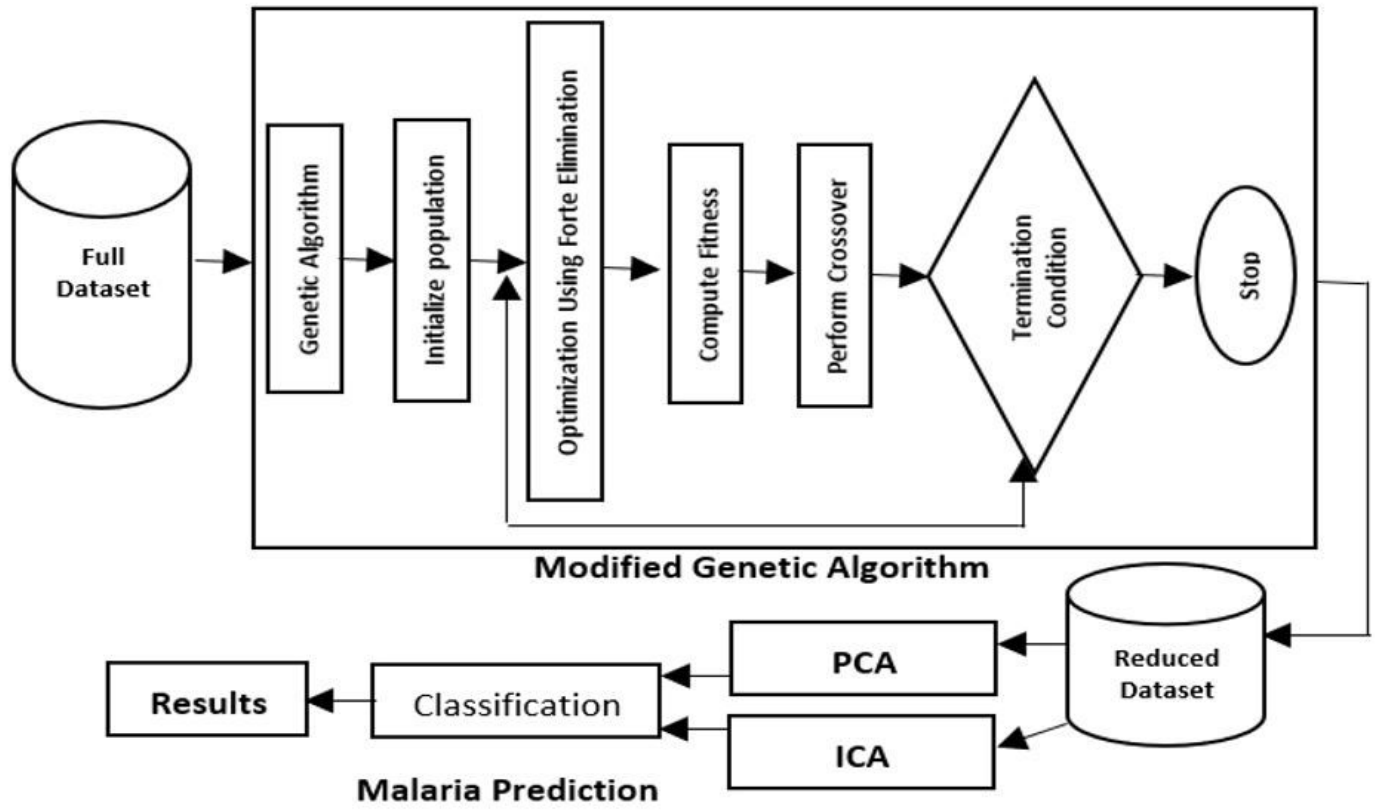



**Figure 1**

proposed framework for RNA-Seq gene classification

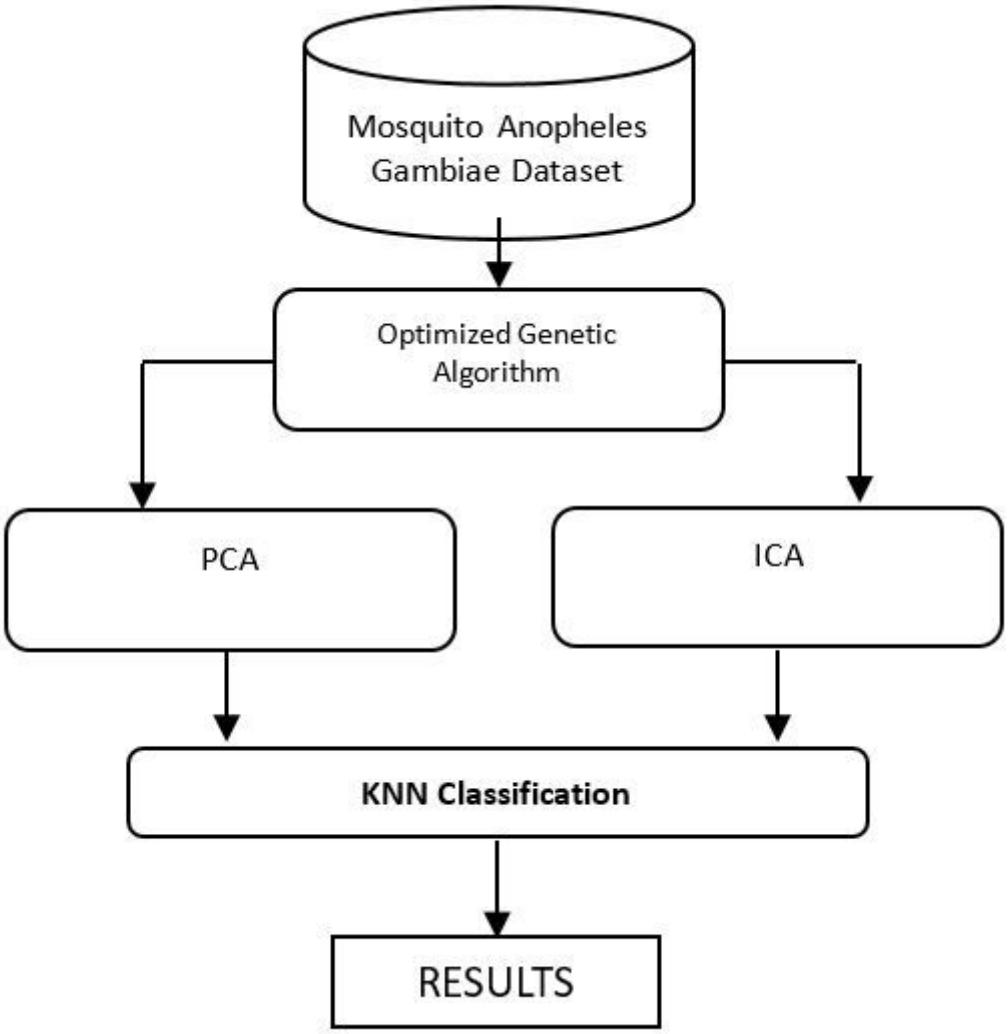
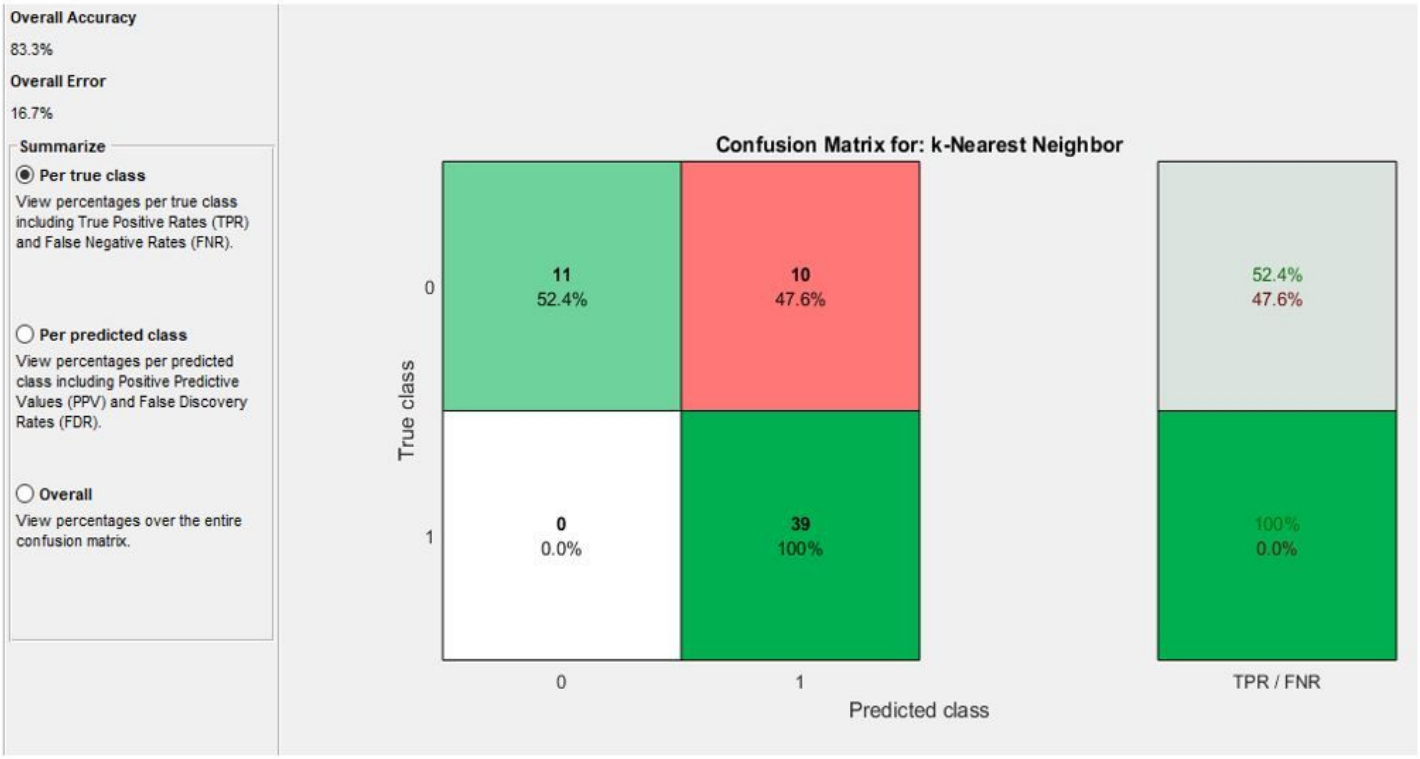


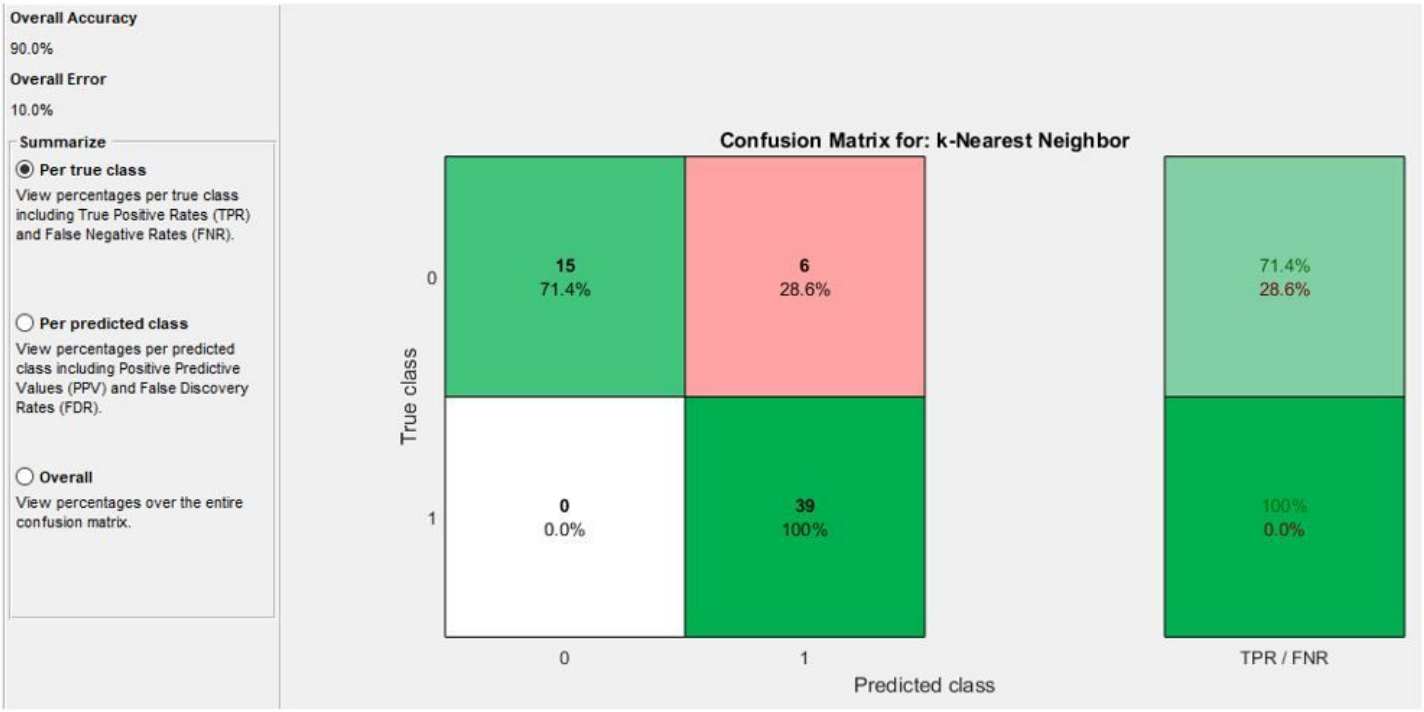

**Figure 2**

Proposed Classification Framework for the Gene Expression Data Analysis




## Figure 3

Confusion matrix for GA-O+PCA+K-NN TP=39; TN=11; FP=10; FN=0.




## Figure 4

Confusion matrix for GA-O+ICA+K-NN TP=39; TN=15; FP=6; FN=0.