Sains Malaysiana 50(9)(2021): 2579-2589
http://doi.org/10.17576/jsm-2021-5009-07

Enhanced Dimensionality Reduction Methods for Classifying Malaria Vector Dataset using Decision Tree

(Peningkatan Kaedah Pengurangan Kedimensian untuk Mengelaskan Set Data Vektor Malaria menggunakan Pokok Keputusan)

MICHEAL OLAOLU AROWOLO*, MARION OLUBUNMI ADEBIYI & AYODELE ARIYO ADEBIYI

ABSTRACT

RNA-Seq data are utilized for biological applications and decision making for classification of genes. Lots of work in recent time are focused on reducing the dimension of RNA-Seq data. Dimensionality reduction approaches have been proposed in fetching relevant information in a given data. In this study, a novel optimized dimensionality reduction algorithm is proposed, by combining 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 Decision tree 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 from the high-dimensional input feature space. A feature ranking and earlier experience are used. 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 feature selection and classification in gene expression data analysis and specify that the approach is a capable accumulation to prevailing data mining techniques.

Keywords: Decision tree; independent component analysis; malaria vector; optimized genetic algorithm; principal component analysis

ABSTRAK

Data RNA-Seq digunakan untuk aplikasi biologi dan membuat keputusan untuk pengelasan gen. Banyak kajian kebelakangan ini memfokus untuk mengurangkan dimensi data RNA-Seq. Pendekatan pengurangan dimensi telah diusulkan dalam pengambilan maklumat yang relevan dalam data yang diberikan. Dalam kajian ini, algoritma pengurangan dimensi optimum baharu dicadangkan dengan menggabungkan algoritma genetik yang dioptimumkan dengan Analisis Komponen Utama dan Analisis Komponen Bebas (GA-O-PCA dan GAO-ICA), yang digunakan untuk mengenal pasti ciri subset optimum dan korelasi laten. Pengelas menggunakan Pokok keputusan pada kumpulan data terturun nyamuk anopheles gambiae untuk meningkatkan ketepatan dan kebolehan pengukuran dalam analisis ekspresi gen. Algoritma yang dicadangkan digunakan untuk mengambil ciri yang relevan berdasarkan ruang ciri input dimensi tinggi. Ciri pemeringkatan dan pengalaman sebelumnya digunakan. Prestasi model dinilai dan disahkan menggunakan ketepatan pengelasan untuk membandingkan pendekatan sedia ada dalam kepustakaan. Hasil uji kaji yang dicapai terbukti menjanjikan ciri pemilihan dan pengelasan dalam analisis data ekspresi gen dan menentukan bahawa pendekatan tersebut merupakan pengumpulan yang mampu dilakukan terhadap teknik perlombongan data yang berlaku.

Kata kunci: Algoritma genetik yang dioptimumkan; analisis komponen bebas; analisis komponen utama; Pokok keputusan; vektor malaria

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 in fetching 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 the gene expression data analysis, overfitting and curse of dimensionality have been known to deteriorate the classification capabilities of high dimensional input space called the curse of dimensionality. To overcome these challenges, several dimensionality reduction procedures have been suggested in literature, to determine an optimal subset genetic factors that can help show hidden features of genes and enhance their interpretability. The dimensionality reduction aim is to discover 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 researchers have addressed the problems of curse of dimensionality, using different dimensionality reduction approaches in analyzing the gene expression data classification, by selecting relevant genes. Metaheuristics and hybrid dimensionality reduction methods have also been proposed for gene selections and classification, yet they suffer from correlation, they are cumbersome and suffer from high throughputs, with increased computational time for fetching gene subsets (Cai et al. 2018; Marfaja & Mirjalili 2018). Systematic approach for fetching an optimal subset gene is a crucial issue. Feature selection (filter, wrapper, and embedded) (Chen et al. 2020; Liu et al. 2020; Tadist et al. 2019) and feature extraction (Aziz et al. 2017; Bajaj et al. 2020; Wenric & Shemirani 2018) (linear and non-linear) are dimensionality reduction approaches that have been established, these approaches have overcome several problems such as performance enhancement, yet there is need for improvements enhanced model and optimization for get better results (Panshaei et al. 2019). Finding an optimal subset of genes proficient to handle high dimension optimization difficulties with reasonable solutions is required.

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

Feature extraction techniques such as principal component analysis (PCA) (non-linear) and independent component analysis (ICA) (linear) are extensively used method (Aziz et al. 2017; Kong et al. 2018) are common capable methods for fetching subset of gene samples for classification and have received growing attentions in recent time (Mohan et al. 2014). Integrated approach has proven to be significant, due to their good performances and advantages for solving dimensionality problems that halt classification, it is of 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 (Arowolo et al. 2017; Chen et al. 2020; Hira & Gillies 2015; Lin & Zhang 2019; Liu et al. 2020; Pashaei et al. 2019; Pragadeesh et al. 2019; Shukla et al. 2019; Tadist et al. 2019; Wang et al. 2017). However, these experiments necessitate enhancements that can help in making decisions on how to eradicate transmission of malaria in west Africa, as it is a scourge in Africa (Hodgson et al. 2019).

The objective of this study was to carry out an enhanced 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 Decision tree 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

Transcriptional RNA-Seq data analysis uses the larvae of anopheles gambiae, obtained from a western part of Bungoma of Kenya. Comprising of deltamethrin-resistant and vulnerable mosquitoes profiling with tolerant response systems; the data comprises seven characteristics similar to the Test ID, Gene ID, Genomes, Locus, Susceptible, Vulnerable and predictive rank and 2457 instances (Arowolo et al. 2020b; Table 1).

TABLE 1. Features of the dataset

Dataset	Attributes	Instances
Mosquito Anopheles Gambiae	7	2457

METHODS

RNA-Seq is a frequently applied genomic data technology. It recognizes several facets of proteomics, an operational excellence genomic DNA innovation used with analysis of large transcription factors. Giving a better understanding of cells changes in gene expression, experimental approaches and improved performance (Zhao et al. 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 (Huynh et al. 2019). Gene expression Classification of RNA-Seq data has provided valuable evidence to classify and assess German treatments for sicknesses. The expression of genes is a genomic feature in the predominant method of RNA-Seq that measures and gains better understanding of various biological issues. The problem of diagnostic challenges is a major challenge for RNA-Seq, and as a result of high dimensions of protein sequence expression data, it gives unfitting results.

In this study, the dataset uses a mosquito anopheles gambiae. The samples of the genes are normalized using the MATLAB 2015 bioinformatics tool package. The samples are passed into the optimized genetic algorithm. A reduced sample are then achieved and passed into the PCA and ICA separately. The further reduced data are split into training and testing sets. The training set are the vigorous samples and testing set are the outstanding samples. Classification is the conducted using KNN (Figure 1).

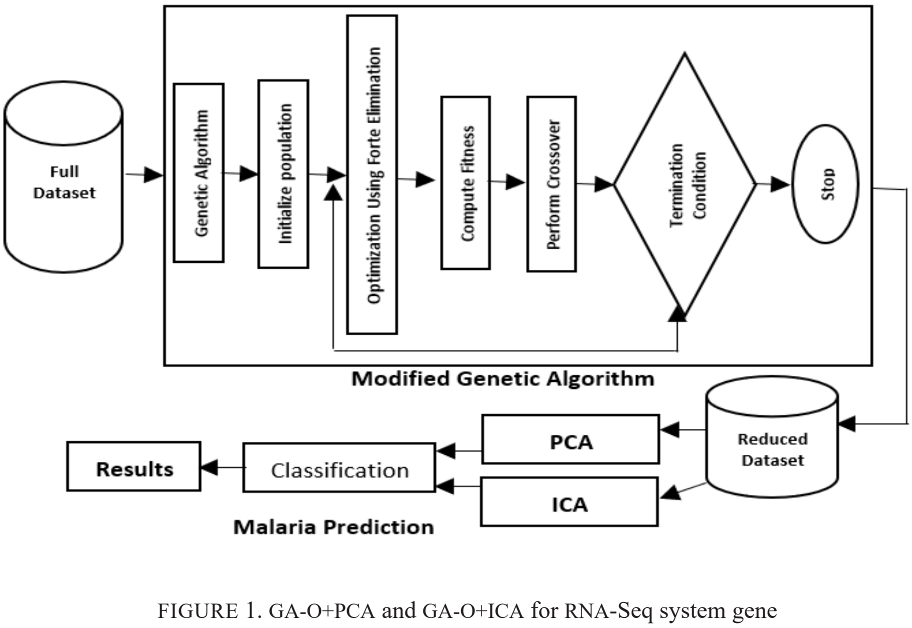


FIGURE 1. GA-O+PCA and GA-O+ICA for RNA-Seq system gene classification

DIMENSIONALITY REDUCTION

Reduction of dimensionality is a popular technique for removing noises and undesirable features. Gene expression dataset comprises computational complexity that leads to high computation intensity and deficiency of cross validation algorithm performance. Dimensionality reduction procedures are required to eliminate complexity, identifying relevant features and fetch features that interrupts the efficiency and function by increasing the attribute proportions of the samples. This method helps to reduce the risk of overfitting. The reduction of dimensionality is an important method known as the collection of features and extraction of features (Sahu et al. 2018; Shen et al. 2017).

FEATURE SELECTION

Feature selection is an indispensable phase in the evolution of a machine learning classification in innovations including RNA transcript for developing valuable exclusive identification features for transcription samples for training/testing modeling techniques. The selection of features allows the collection of partial knowledge to be implemented in classifiers and the removal of uninformative features to reduce the computational burden. 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 features is an efficient technique identified as the Filter, Wrapper and Embedded methods (Jabeen et al. 2018).

GENETIC ALGORITHM

The genetic algorithm is an evolutionary algorithm used to evaluate engine optimization problems to select wrapper-based features. GA may be based on true activities relating to social genes in the survival of the fittest foundation. GA consists of initial population growth, fitness evaluation, choice of family, mutation and crossover (Motieghader et al. 2017; Uma & Kirubakaran 2016).

In its uncluttered nature, GA is a combinatorial investigation process, with a collection of randomly chosen findings (phenotypic traits or entities) giving an optimum model for analytical purposes of evaluating the options that are desirable. Generally, each chromosomes or genotype has a set of properties characterized as binary 0s and 1s strings (Wang et al. 2017). While very sensitive to the initial population, GA has a weakness of optimality, with its answer value declining as problematic magnitude rises, it has been shown to produce fair quality solutions to boost it for gene sampling.

FEATURE EXTRACTION

Feature extraction is the process being used to identify significant factors, attributes or features which are found in information. Inside a collection of studies, instances of procedure of extracting features include the identification of patterns and the detection of common precedents. The use of noise removal to achieve a higher summary of its classification contains information with dimension arrays. Feature extraction enables revolutionary parameters of feature selection to reduce the prevailing computational burden. There have been double broad groups of algorithms for extracting the features, which included: linear (recognizes information on a lesser feature space, including Principal component analysis); non-linear (operates defined on a high-dimensional parameter with a lesser feature space, like ICA) for a non-linear connection among features (Hira & Gillies 2015).

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a feature extraction process for linear features, extensively utilized predominantly in biomedical research. PCA models feature spaces of high to lower dimensional space while recreating the k-dimensional unconnected features from field's specific n-dimension feature. PCA has identified that this is an efficient technique to identify high-dimensional genomic knowledge. It is typically used by information from RNA-seq. PCA explores orthogonal modification by converting a group of correlated variables into a group of uncorrelated variables. PCA for the review of exploratory outcomes. To evaluate the relationships between a group of variables and to minimize dimensionality, PCA can be used (Jain & Singh 2018).

INDEPENDENT COMPONENT ANALYSIS (ICA)

ICA helps in gathering hidden features from multi-dimensional information by removing multi-variate representations into independent non-gaussian components for the features to be linearly separable. ICA explores a correlation among information by decorrelating the knowledge through enhancing or minimizing the necessary information. ICA implements Decision X as a dimensional variation of the individual components S. If A indicates the independent vector of the balanced matrix W, input X's source extracted features are defined by columns A.

$$S = W\, to\, X. where\, X = A\, to\, S \qquad (1)$$

For biological information and recognition of relevant genes, ICA has been used extensively (Feng et al. 2020; Hashemi et al. 2018).

PCA is a linear transformation technique used to eliminate the complexity and number of characteristics. It is a 'non-linear' technique, although ICA is 'linear', while ICA is shown to perform better if an information is trained and tested (Hira & Gillies 2015).

CLASSIFICATION

Classification is a supervised learning algorithm in machine learning. It is a typical, beneficial process that attributes training data allocated to existing information from either a predetermined class label and determines it. The classification of buildings is carried out in two steps (Arowolo et al. 2017): First, the learning process, in which the classification model was developed with a class label giving a collection of training data; second, the model is used to predict class labels for hidden data and to calculate the accuracy of the KNN classifier.

K-NEAREST NEIGHBOR (KNN)

A supervised learning K-nearest neighbor classification technique for gene datasets performs the benefit of creative application event assessment of neighborhood 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 neighbors 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 neighbors. The unassuming prevalence of group of nearest neighbors is used as approximate number of instances of the question. Bonds can fragment randomly (Bose 2016).

For efficient, systematic study design, expanding the dimension of biological information is an important challenge. Conventional means for training complex models on multiple layers enhanced by texture features amenable to processing are relevant to use. In certain conventional methods use in dealing with large information, like RNA-Seq data, several difficulties are present. In general, the application of multiple dimensional reduction procedures can allow advantage of unique benefits where genomic selection acquired through one procedure is accepted as an input to the alternatives. In particular, procedures of extracting features can be used to effectively support feature selection, by using feature selection to choose the original gene subset, or by taking advantage of redundant gene elimination. A mixture of different feature extraction techniques can be applied to remove the initial feature subsets (Arowolo et al. 2020a; Motieghader et al. 2017; Sun et al. 2018).

In this report, an improved dimensionality reduction technique has been proposed for classifying malaria transmission Genomic DNA profiles.

RNA-Seq has tremendous potential for finding, defining, and tracing cell lines, but reduction of dimensionality helps to perceive the structures, but data remains difficult, current algorithms need the correct development to show suitable characteristics, enhanced dimensionality method has verified to be sufficient but needs effective algorithms to model. The classification technique proposed consists of three phases, namely: Selection of features, extraction of features, and group of classes.

Figure 2 illustrates the suggested enhanced system for classifying gene expression data for malaria. The framework consists of three subsystems, a subsystem for feature collection, a class-based subsystem for feature extraction, and a subsystem for classification. By adapting one algorithm listed herewith to pick an optimum subclass by assessing the fitness for chromosomes, the function selection subsystem uses an optimized GA. The PCA and ICA are suggested as subsystem because of its data projection of efficiency invariance with orthogonal orders and classified using Decision tree.

One of several implications of optimizing the genetic algorithm is, its evolved development of features of the algorithm, that in turn allows the various search level by exploring the optimal solution for generating high-quality solution at the same time and independently, this study uses an optimization of the set of genetic algorithm features to reduce the number of features and preserve the data. The extraction of characteristics is perfect for converting the reduced data into latent components, the richness of which is to reduce prosperity and all methods of dimensionality reduction that can be used for malaria classification are used.

Algorithm 1

```
Phase 1: formulate the c and d parameters (the row and column) then establish initial sample arbitrarily.
Phase 2: For i<population size
Phase 3: add tangent rate tan(xin/xin+1) of the angle between two vectors in adjacent dimensions for each pop(i)
Phase 4: if Phase 3=0, then
The value of the nth dimension of the discrete $i^{th}$ is modified to 0;
Do not change the value and move to step 6.
phase 4.1: if x>1, Phase 2:
Phase 4.2: measure the compatibility between $x_i$ and $x_j$ of persistence and Euclidean distance D
Phase 4.3: Inside the gap, measure the comparison principle $L=|X_i-X_j|<D$
Phase 4.4: if no comparison exists,
i. With the parallel of biallelic loci, eliminate individual fitness (SD (Xi, Xj)) and regular parallel $MSD_i$
ii. Compute the subpopulations M(t+1);
else
Merge N memory pool entities with a subpopulation grouped by descending order capability.
Phase 4.5: calculate the subpopulation * threshold
Phase 5: judge the convergence state
Phase 6: Calculate the number of 0 basics for the restructured pop in each dimension; remove this dimension if it is above the critical value Q
Phase 7: Get the population updated
Phase 8: The calculation of the fitness value F(i) of each person in the population
Phase 9: Setting up a new population
Phase 10: According to fitness, pick two entities from the population with the relational selection algorithm
Phase 11: If arbitrary (0, 1) < Pc, therefore proceed to Phase 12; otherwise, add Phase 13
Phase 12: For the two entities, add the crossover procedural mapping to the crossover probability Pc
Phase 13: If arbitrary (0, 1) < Pm, proceed to Stage 14
Phase 14: Using the mutation on the individuals or groups according to the mutation probability Pm.
Phase 15: Incorporating the two new people into the new population
Phase 16: Continue this procedure until generated by the N-th generation; otherwise, move to Phase 4
Phase 17: Transforming the sample size with a new population
Phase 18: Repeat this procedure until another group attains G, or switch to Procedure 8 else.
Phase 19: end
```

The genetic algorithm retrieves the information's essential factors. Using PCA and ICA algorithms independently to carry for the fundamental ideas, the identified features are supplied to the feature extraction steps. Decision tree classifiers is used for the analysis of learning process to yield the performance metrics.

```
Phase 1: Validate the information uploaded
Phase 2: Using the Optimized Genetic algorithm to implement a feature selection algorithm to the information
Phase 3: Implement the PCA Algorithm to the identified Phase 2 result features
Phase 4: Implement the Decision Tree Classification to the Phase 3 results
Phase 5: Assess the results
Phase 6: Repeat Step 3 using features extracted of the ICA feature
Phase 7: Implement Decision tree classification to level 6 performance.
Phase 8: Evaluate the outcomes
Phase 9: Evaluate the Step 5 and Step 6 results
```

All tests are carried out using an Intel Core 5 computer system, 16 GB RAM, 64-bit operating system. In C++ and MATLAB 2015, all algorithms are coded.

As a classification investigation, confusion matrix is generated to ensure better training performance and test sets evaluation metrics in terms of accuracy, specificity, sensitivity, precision, and recall (Arowolo et al. 2020b).

RESULTS AND DISCUSSION

An experiment on a malaria vector dataset collected from an accessible to the public source with 2457 instances and seven attributes was performed in this research (Arowolo et al. 2020b), Using MATLAB tool, an optimized genetic algorithm was used to select features from the dataset experimented, relevant features were selected with a threshold of 0.5, 708 gene subset features being important. For relevant correlations, the classifier proficiency is related to the state-of-the-art.

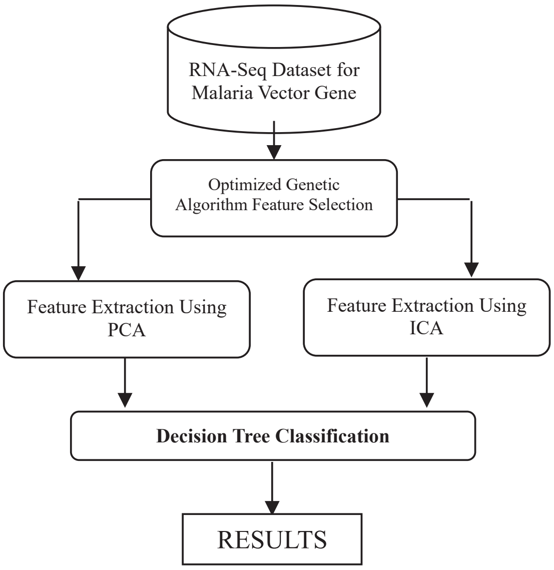

RNA-Seq Dataset for Malaria Vector Gene
Optimized Genetic Algorithm Feature Selection
Feature Extraction Using PCA
Feature Extraction Using ICA
Decision Tree Classification
RESULTS

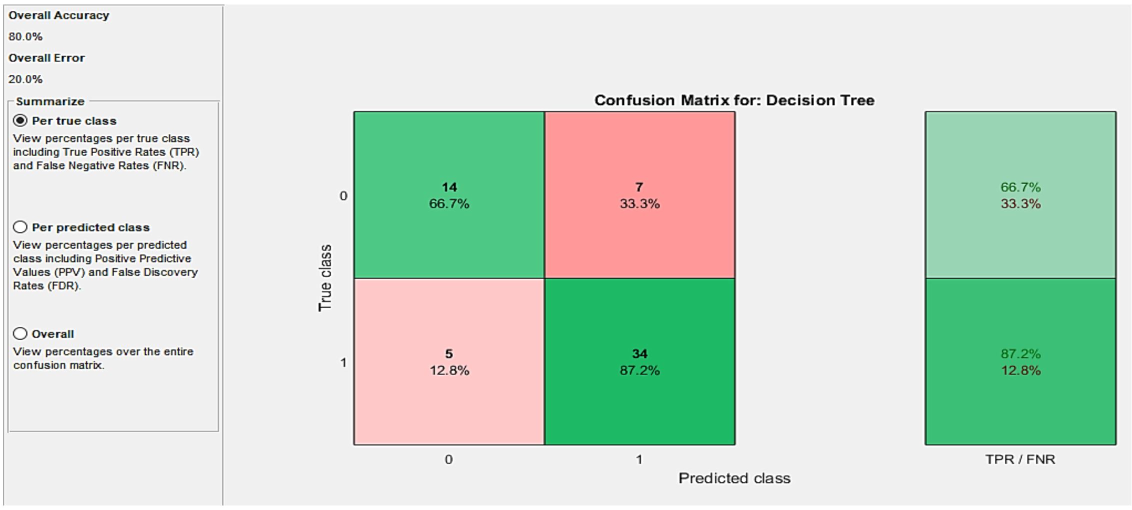
FIGURE 2. Suggested RNA-Seq Gene Classification Model

Firstly, seven hundred and eight selected features are passed into the PCA algorithm using the Optimized Genetic algorithm and ten latent variables are extracted in 1.4623 seconds. Using 10-fold cross-validation, the extracted features are then entered into the Decision tree classification algorithm and the Decision tree classification algorithm uncertainty matrix is evaluated. Secondly, the selected 708 features were passed into the extraction algorithm of the ICA feature; 25 latent variables were extracted in 0.42794 seconds. The extracted features are then transmitted using 10-fold cross-validation into the Decision tree classification algorithm and the confusion matrix of decision tree classification algorithms is evaluated.

Malaria vector reduced dimensionality classification, RNA-Seq data performed using GA-O+PCA+ Decision tree and GA-O+ICA+ Decision tree algorithms. Their success assessments are shown in Figures 3 and 4.

This research has many major consequences for the study of gene expression. The possible application of this research is to provide insight into biological and technological considerations that can clarify discovered gene structures and interpretations related to predictions, malaria diagnosis, and drug designs.

THE GA-O WITH PCA WITH DECISION TREE RESULTS


Overall Accuracy
80.0%
Overall Error
20.0%

Summarize

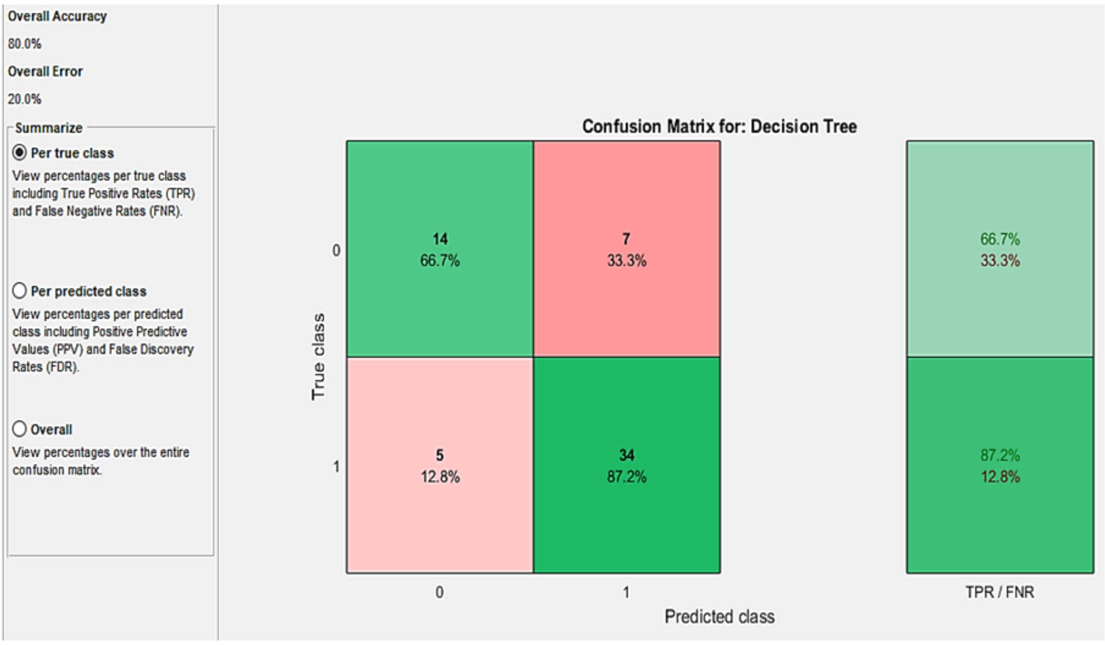
● Per true class
View percentages per true class including True Positive Rates (TPR) and False Negative Rates (FNR).

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

○ Overall
View percentages over the entire confusion matrix.

FIGURE 3. Confusion matrix for GA-O+PCA + Decision tree TP = 34; TN = 14; FP = 7; FN = 5

THE GA-O WITH ICA WITH DECISION TREE RESULTS



Overall Accuracy

80.0%

Overall Error

20.0%

Summarize

● Per true class

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

○ Per predicted class

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

○ Overall

View percentages over the entire confusion matrix.

FIGURE 4. Confusion matrix for GA-O+ICA + Decision Tree TP = 34; TN = 14; FP = 7; FN = 5

TABLE 2. Table of performance measures for the GA-O+PCA+ Decision Tree and GA-O+ICA+ Decision Tree Classification

Performance metrics (%)	GA-O+PCA+ Decision Tree	GA-O+ICA + Decision Tree
Accuracy	80	80
Sensitivity	87.2	87
Specificity	66.7	67
Precision	82.9	83
Matthews Correlation Coefficient	55.2	55
F1-score	85	85

As recorded in Table 2, the experiment achieved comparatively consistent performance using the algorithms applied. In this analysis, an optimized Genetic Algorithm feature selection strategy was used to minimize curse of dimensionality. In the second step, PCA and ICA algorithms were used as feature extraction to further fetch latent components from the reduced data. A Decision Tree Classification Algorithm with a 10-fold cross-validation parameter was used in the third stage. The outcome was an improved result, as shown in Table 3. Compared to the state-of-the-art, the accuracy showed an increase.

Several researchers have studied the problem of malaria classification using machine learning algorithms in order to provide a reliable detection and prediction approach for malaria transmission, the findings obtained in this study can be used for training needed predominance of malaria infection by clinicians by using this method to compile a pathologist curated dataset for training. The study of characterizing thousands of genes provides a deep insight into the problems of malaria classification with abundant data examined, drug development, malaria treatment prediction and diagnosis, and understanding gene functions with gene interaction under normal and abnormal conditions. This study's proposal improves the effects of classification success and demonstrates a lower reliance on the training set.

TABLE 3. Attributing the measures of results with other approaches

Performance metrics (%)	Generalized linear models + PCA (Feng et al. 2018)	GA+PCA+NN (Susmi et al. 2018)	GA+ Canonical correlation Analysis + Neural network (Susmi et al. 2018)	GA+PCA&CCA-NN (Susmi et al. 2018)
Accuracy	70	85.0	85	88

Conclusion

Data analysis of RNA-Seq offers useful and important benefits to the technology 's success, with tremendous helps to evolving the problems of gene expression descriptions. RNA-Seq's related applications include the reduction of dimensionality and classification approaches. Due to the great 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, integrated methods have also been used to classify gene expression results. The GA+PCA+ Decision tree outperformed the GA+PCA+ Decision tree-based approach by using GA feature selection with ICA and PCA feature extraction algorithms separately to perform a dimensionality reduction approach and test their assessment performance on Decision tree classification kernels.

This research therefore aims to apply integrated dimensionality reduction algorithms in future work on classifiers such as the ensemble, NN, to classify appropriate genes for classification of gene expression.

References

Arowolo, M.O., Adebiyi, M.O., Adebiyi, A.A. & Okesola, J.O. 2020a. PCA Model for RNA-Seq malaria vector data classification using KNN and decision tree algorithm. *International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*. pp. 1-8.

Arowolo, M.O., Adebiyi, M.O. & Adebiyi, A.A. 2020b. 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.

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.

Aziz, R., Verma, C.K. & Srivastava, N. 2017. Dimension reduction methods for microarray data: A review. *AIMS Bioengineering* 4(1): 179-197.

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: 107224.

Bose, J. 2016. Hybrid GA/KNN/SVM algorithm for classification of data. *BioHouse Journal of Computer Science* 2(2): 5-11.

Cai, J., Luo, J., Wang, S. & Yang, S. 2018. Feature selection in machine learning: A new perspective. *Neurocomputing* 300: 70-79.

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, Special Issue on Advances in Visual Analytics and Mining Visual Data* 37(5): e12553.

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(1): 54.

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.

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.

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.

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. *Biotechnology, and Biotechnological Equipment* http://dx.doi.org/10.1080/13102818.2017.1364977.

Hira, Z.M. & Gillies, D.F. 2015. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*. 2015: Article ID. 198363.

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(1): 418.

Hyunh, P-C., Nguyen, V-H. & Do, T.N. 2019. Novel hybrid DCNN-SVM model for classifying RNA-Sequencing gene expression data. *Journal of Information and Telecommunication* 3(4): 533-547.

Jabeen, A., Ahmad, N. & Raza, K. 2018. Machine learning-based state-of-the-art methods for the classification of RNA-Seq data. In *Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics*, vol 26, edited by Dey, N., Ashour, A. & Borra, S. New York: Springer, Cham. pp. 133-172.

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.

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* 45(5): 501-520.

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.

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* 2020: Article ID. 6076272.

Mafarja, M. & Mirjalili, S. 2018. Whale optimization for wrapper feature selection. *Applied Soft Computing* 62: 441-453.

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.

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.

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.

Pragadeesh, C., Jeyaraj, R., Siranjeevi, K., Abishek, R. & Jeyakumar, G. 2019. Hybrid feature selection using micro genetic algorithm on microarray gene expression data. *Journal of Intelligent and Fuzzy Systems* 36(3): 2241-2246.

Sahu, B., Dehuri, S. & Jagadev, A. 2018. A study on relevance of feature selection methods in microarray data. *The Open Bioinformatics Journal* 11: 117-139.

Shen, L., Jiang, H., He, M. & Liu, G. 2017. Collaborative representation-based classification of microarray gene expression data. *PLoS ONE* 12(12): e0189533.

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.

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.

Susmi, S.J., Nehimiah, H.K. & Kannan, A. 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.

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.

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.

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.

Wang, L., Wang, Y. & Chang, Q. 2017. Feature selection methods for big data bioinformatics: A Survey from the search perspective. *Methods* 111: 21-31.

Wenric, S. & Shemirani, R. 2018. Using supervised learning methods for gene selection in RNA-Seq case-control studies. *Frontiers in Genetics* 9: 297.

Zhao, S., Fung-Leung, W-P., Bottner, A., Ngo, K. & Liu, X. 2014. Comparison of RNA-Seq and microarray in transcriptome profiling of activated t-cells. *PLoS ONE* 9(1): e78644.

Department of Computer Science
Landmark University
Omu-Aran
Nigeria

*Corresponding author; email: arowolo.olaolu@lmu.edu.ng

Received: 6 October 2020
Accepted: 21 January 2021