Journal of Engineering Science and Technology
Vol. 17, No. 4 (2022) 2891 - 2903

ICA LEARNING APPROACH FOR PREDICTING OF RNA-SEQ MALARIA VECTOR DATA CLASSIFICATION USING SVM KERNEL ALGORITHMS

MICHEAL OLAOLU AROWOLO

Department of Computer Science, Landmark University, Omu-Aran, Nigeria
E-mail: arowolo.micheal@lmu.edu.ng

Abstract

Parasites of malaria follow vague difference in sections in life as they expand across various stratospheres of mosquito vectors. There are transcriptomes of several thousand human parasites. Ribonucleic acid sequencing (RNA-Seq) is a prevalent gene expression technique haven led to a better understanding of genetic requests. RNA-Seq measures gene expression transcripts. Data from the RNA-Seq require methodological developments in machine learning techniques. Scientists have suggested many addressed learning for the study of biological evidence. Independent Component Analysis (ICA) algorithm is utilized in this analysis to collect latent components from a high-dimensional RNA-Seq malaria vector dataset and analyse its classification output utilizing classification algorithms for Support Vector Machine (SVM) Kernel. The effectiveness of this assay is tested on an RNA-Seq sample of mosquito Anopheles gambiae. The findings of the analysis hit important performance thresholds with a classification accuracy of 92% and 87% individually.

Keywords: Mosquito Anopheles, ICA, RNA-Seq, SVM,.

1. Introduction

Several large data sets have been generated through high-throughput next-generation sequencing technologies, helping biologists examine and discover troubling gene transcripts such as RNA associations and Ailments such as, among others, cancer, malaria, biology, family, physiology [1].

Anopheles gambiae are classes of blood-sucking mosquitoes in Africa which have large malaria carrier Plasmodium falciparum. Anopheles Mosquito is a deadly type of parasite malaria that is responsible for thousands of deaths. As novel antimalarials increase competition with antimalarial drug banquets, looking for innovative drugs needs improved biological knowledge of these species. How mosquito anopheles parasites assist complex regulation of gene expression has been a major problem requiring the development of an improved robust malaria vector transcription prognostic model [2].

By defining a provisional biological functional strategy by enhanced sequencing analysis, RNA-seq study generates sensitive biological insightful tests. RNA-Seq data include high-dimensional curse elimination, such as; diseases, noises, repetition, repetition, invalid, inter alia, wrong data [3]. Innovations have enhanced strategies for developing creative healthcare structures, for example, updated procedures, specialized patient fitness screening devices and other cancer and illness diagnostics. [4].

Various machine learning methods are advanced over the past decades with practical advancements to determine the tremendous volume of RNA-Seq and data expression of next-generation gene sequencing by analysing biologically applicable frameworks [5]. Several researchers used high-performance machine learning methods for evidence on the expression of the RNA-Seq genes [6-8].

This investigation proposes an ICA extraction attribute reduction protocol to assess the high dimensionality of data processing on gene expression, Classification methods for SVM are used to diagnose discrete biological structures which have better classification accuracies than can be suggested as an effective method for predicting which detecting new genes.

2. Related Works

A broad genetic archive of individuals with or without diseases relies on statistical approaches, and genes accountable for the development of diseases can be identified. Expressed Differential Genes are characterized using several methods. Machine learning (ML) procedures are essential to recognize variations between human genome-derived genes. In the study and diagnosis of gene expression profiles in different disorders, several machine learning methods are emulated and used. There is talk of the necessity for profiling gene expression and its approaches using specialized machine learning. It reviews a ton of research studies conducted in this field by scholars. There are still recognised work holes in the study of gene expressions [4].

Oh et al. [9] worked with blood-based genetic factor signature expression and machine learning to anticipate autism spectrum ailment and classify transcripts and could be used in the classification. Employing R-language platform for machine learning procedures for RNA information from Gene omnibus expression database.

Autism spectrum disorder reported by groups in the ranking cluster review relatively well-discriminated. Help vector machine and neighbour classifiers K-nearest are utilized to verify the data, resulting in an all-encompassing class estimation of 93.8% precision and 100% and 87.5% sensitivity and specificity, respectively.

Qi et al. [10] concentrated on RNA-Seq data clustering and classifying, undertaking an interactive analysis, reviewing the pros and cons of methods that have recently arisen as prevalent changes using clustering and classification methods, utilizing non-linear and linear methods with reducing sc-RNA-seq information dimensioning techniques, mixing and capturing.

By rating broad collections of RNA-Seq-measured genes [11], focused on supervised learning methods for RNA-Seq gene selection. They used variable range measurements created by the algorithm of random forest classification, defined by the channel of extreme pseudo-samples, utilizing variational auto-encoders and regressors to derive ranks from 323 to 1,210 samples of 12 RNA-Seq cancer datasets. Results revealed the latent of guided gene selection strategies centred on experience in RNA-Seq research and indicated the necessity to practise gene selection strategies in the study of gene expression.

lquicira-Hernandez et al. [12] used a supervised approach to data classification analysis on the RNA-Seq. They implemented a generalizable approach of extremely detailed incorporating impartial feature collection from a simplified dimensional space and machine learning inference approach, single-cell classification. Sc-Pred was applied to a mononuclear cell, colorectal tumour biopsies, pancreatic tissue, and circulating dendritic cells in the RNA-seq dataset. Sc-Pred is particularly efficient in classifying different cells.

Cui et al. [13] experimented on machine learning with an emphasis on RNA-DNA research, indicating low-expressed genomes theoretically impacted by PAH disease. They suggested a ground-breaking collection of features and advanced methods for classifying a trivial range of incredibly useful genes in machine learning algorithms. Studies revealed that clusters of genes with limited expression reveal modified types of PAH when forecasting and discriminating.

Shon et al. [14] reported on the characterization of data on gene expression using CNN for stomach cancer. They developed a deep learning classification methodology and demonstrated its claim to the information expressions collected from patients with stomach cancer. 60,483 data genes from 334 patients with stomach cancer were analysed using PCA, heat maps and CNN algorithms. They merged clinical knowledge of RNA-seq with data on gene expression, tested genes and analysed using a CNN algorithm, with a precision of 96 and 50%.

Reid et al. [15] operated on RNA-Seq discovery of concealed malaria parasite transcripts, illustrating the distinction of an RNA-seq technique deconvoluting transcriptional discrepancy for 500 separate rodents and human malaria parasites; discrete transcriptional signatures tucked within were found.

Tan and Gilbert [16] also operated on an ensemble algorithm to classify data on the cancer gene expression. On seven publicly available cancer microarray findings, and linked to the classification presentation of these techniques, C4.5 decision tree, bagged and enhanced ensemble decision trees classifiers were used. They detected ensemble learning using bagged and boosted benefited more than the sole decision trees in classification.

Song and Wang [17] collaborated with the analytical ensemble to develop a classification method for the gene expression of cancer results. With the Adaboost algorithm, a Recursive Feature Elimination was performed to pick suitable classification characteristics. Reporting there suggested a transition.

Tarek et al. [18] focused on classifying cancer for evidence on gene expression. We suggested an approach to the classification of the operational ensemble that improves the introduction of the description and the poise of the performance. The ensemble classifiers results are less contingent on a unique set of instructional individualities.

Lee and Lee [19] collaborated to build the tree model to classify the ensemble's selected characteristics. This analysis employed an ensemble-based set of functions with random trees and a wrapper method to develop the classification. Through the bagging procedure, wrapper system, and random trees, the probable knowledge grouping method of the ensemble produces a subset. The future approach excludes redundant attributes and uses a system of chance weighting to pick the right classification features. Using RF, SVM, and NB checks, the possible feature selection framework is evaluated, and its performance correlates with GA-SVM, GA-NB, FS-NB, FS-SVM, and GA-RF techniques. A rating precision of 92% is obtained by the process.

Tan et al. [20] worked on the analysis of multiple gene expression's function extractions, such as the PCA, ICA, PLS, and LLE. A description and program bases were given in the last section for each process of extraction of functionality.

Zahoor and Zafar [21] worked on classifying gene expression data with an infiltrated tactical optimizer algorithm to address scarce data and achieved an accuracy of 88%.

Arowolo et al. [22] using an optimized approach for classifying malaria infections by fetching relevant genes using a genetic algorithm with PCA and ICA, the classified their result using KNN and obtained about 90% accuracy. Hameed et al. [23] proposed a novel application to select and classify high dimensioned genes, and their proposed model obtained an accuracy of 90%.

3. Materials and Methods

The literature suggested various approaches for processing high-dimensional data. For a better result, Independent Component Analysis (ICA) and Ensemble classifier are being considered in this analysis to reduce the dimensionality of high-dimensional RNA-Seq results.

Included are 2457 instances of 7 gene attributes, data obtained from western Kenya containing mosquito genes between 2010 and 2012. The text profile file includes AGAP-012984, AGAP-002724, AGAP-003714, AGAP-004779, AGAP-009472, CPLCG3-AGAP-008446, CYP6M2-AGAP-008212 and CYP6P3-AGAP-002865, RNA-Seq genes, distinctions of deltamethrin-resistant and susceptible Anopheles gambiae mosquito transcriptome, openly accessible from figshare.com. A brief overview of the dataset is given in Table 1.

Table 1. Dataset features.

Dataset	Attributes	Instances
Mosquito Anopheles gambiae	**7**	**2457**

3.1. Methods

MATLAB was used to evaluate the data obtained from [24] as an experimental tool, to remove functionality, ICA was used. Using the ensemble algorithm method, the derived functions are evaluated [25].

Independent component analysis (ICA)

ICA is a valued branch of PCA, which has remained conservative because the individual bases were isolated by the visors from their linear grouping [26]. The original ICA reality is that the general PCA has uncorrelated properties. Based n x p on information matrix X, the rows in which *ri* (*j*=1 ..., *n*) are centred on analytical variables, and the columns *cj* (*j*=1 ..., *p*) are the entities of the equivalent variables.:

$$X = AS \tag{1}$$

A is an *n x n* fusion matrix with a full description, where *S* is an *n x p* is a base matrix with the need to be statistically unbiased as practicable. The original variables found in the *S* tables are independent components; to wit, the variables observed are linearly written individual components. By observing the exact linear combinations of the empirical variables, the individual components are obtained as mixing can be reversed as:

$$U = S = A\text{-}1X = WX \tag{2}.$$

The U are uncorrelated with zero means, *S* are the linear combined to construct *X*, *A* is a square and orthonormal matrix with Weighted *W* with an unmixing matrix.

3.2. Support vector machine (SVM)

SVM is an algorithm for thinking machines, proposed by Vapnik in 1992 [27]. The procedure works with the point of finding the best hyperplane which isolates in the input space between classes. SVM is a linear classifier; it is generated by combining the kernel ideas into high-dimensional workspaces to deal with nonlinear problems. For non-linear problems, SVM uses a kernel to train the data to narrowly spread the dimension. When tweaking the proportions, SVM should search for the ideal hyperplane which can distinguish a class from other classes [28]. The method for finding the best hyperplane using SVM, as shown by the adoption of Aydadenta and Adiwijaya (2018), is as follows:

i. Let $y_i \in \{y_1, y_2, \ldots, y_n\}$, where yi is the p-attributes and target class $z_i \in \{+1, -1\}$
ii. Assuming the classes +1 and -1 can be separated completely by hyperplane, as defined in Eq. (2):

$$v.y + c = 0 \tag{3}$$

From Eqs. (3), (4) and (5):

$$v.y + c \geq +1, \textit{ for class } +1 \tag{4}$$

$$v.b + c \leq -1, \textit{ for class } -1 \tag{5}$$

where the input data is y, v is the ordinary plane, and c is the positive relative to the coordinates in the centre field.

SVM strives to find hyperplanes that optimize the differences between two groups. In programming, increasing restrictions is a quadratic problem that is resolved by achieving the minimum value. The benefit of SVM is their ability to achieve an extensive range of high-dimensional data classification difficulties [29].

SVM is outstanding in comparison with other classification methods, with its exceptional classification adequacy [30]. SVM is classified into separable linear and nonlinear. SVM's has kernel functions that transform data into a higher dimensional space to make separations conceivable. Kernel functions are a family of algorithms for the interpretation or identification of patterns. Training Vectors *xi* is translated into higher dimensional planetary by ability. SVM considers a linear, separating hyperplane with the limit in this higher dimension space. The Error Form cost parameter is $C>0$.

Many SVM kernels exist, such as; the polynomial kernel, Radial base function (RBF), linear kernel, Sigmoid, Gaussian kernel, String Kernels, and others. A Kernel's decision is based on the current problem at hand, as it depends on which models to be evaluated, a few kernel functions have been found to perform admirably in for a wide variety of applications [31]. SVM-Polynomial Kernel and Gaussian Kernel are the prescribed kernel functions for this study.

3.2.1. SVM-Gaussian kernel

Gaussian kernel [32] is related to a general supposition of smoothness in all subordinates of the k-th order. Kernels that manage a certain prior data recurrence material can be built to represent earlier learning problems. Every input vector x is translated to an infinite-dimensional vector with all the polynomial extensions of the x components.

3.2.2. SVM Polynomial kernel

For example, a model of a polynomial kernel features conjunction up to the polynomial order. Radial base functions require circles in comparison to the linear kernel, which allows only the collection of lines (or hyperplanes).

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

3.2.3. SVM-Linear kernel function

The polynomial kernel, for example, is the least complex function for the kernel. In addition to a discretionary constant K, it is given by the inner product (a, b).

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

3.2.4. SVM-RBF kernel function

In SVM kernel functions, π, a, and b are kernel parameters, RBF is the basic kernel function because of nonlinear maps measurements in higher dimensional space than the linear kernel, which has fewer hyperparameters than the polynomial part [33].

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

In this study, ICA is suggested to best describe the gene expression data, to help enhance the classification accuracy. The ICA deletes redundant genes as a pre-processor and provides a latent component.

3.3. Performance evaluation

Machine-learning procedure success assessment involves a validation system of measurement. The uncertainty matrix is often used for the study of four characteristics of classification models; True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). This determines the examples categorized appropriately and erroneously from the illustration of the dataset provided for the model check [4]. The formulation defines quality metrics below [34].

A model's precision is measured utilizing four variables called TP, FP, TN, and FN. TP's substance seeks state as it appears. FP's output identifies the environment when it doesn't exist. TN's commodity does not find the state because it doesn't exist. FN's product does not consider the condition because it already exists. Accuracy: (TP + TN) / (TP+TN+FP+FN)

Sensitivity calculates positive for the quantity of suitably known instances. Focus: TP/ (TP+FN)

Specificity describes the number of cases that are fittingly remembered with individual negatives.

Specificity: TN / (FP+TN)

Precision: TP / (TP + FP)

Recall: TP / TP+FN

F-Score: 2 x (recall x precision) / (recall + precision)

3.4. Applications

Gene expression investigation provides an enhanced pathway for RNA-Seq identification. The necessity to investigate particular genes is obliging in evolving diverse applications such as improved therapy, cancer diagnosis, creation of genes and medications, tumour identification, diseases such as malaria, etc. Machine learning strategies to recognize the project and the gap in the results. It retains excellent algorithms as instruments specific to different fields.

The experiment is performed via MATLAB (Matrix Laboratory). MATLAB is an arithmetical modelling setting, and selected programming language created by MathWorks, enabling system panels, mapping of roles and data, implementation of procedures, development of operator boundaries, written in diverse languages, such as; C, Java, and Python [16]. Forecasting using the Malaria database with RNA-Seq technology using the MATLAB approach is the main point of this study. For this analysis, the device conformation uses the iCore2 processor, 4 GB RAM size, 64-bit Framework, and MATLAB 2015a implementing software.

4. Results

This research explores RNA-Seq innovation of vulnerable and tolerant genes, carrying 2457 instances of Mosquitoes Anopheles gambiae. To that the burden of dimensionality, the ICA algorithm has been applied to the results. ICA function dimensionality reduction extraction identifies and excludes uncorrelated attributes (variables) to determine maximal variation with fewer key components.

For this analysis, ICA is extended to evidence from the Mosquito Anopheles, which offers important gene knowledge that is valuable for further studies. Using MATLAB tool, classification algorithms apply SVM kernels to implement the model.

Using ICA as a method of reducing the dimensionality of extraction features, 25 gene features of the latent components were important.

SVM classification methodology with 10-fold cross-validation was applied to evaluate the classification with 0.05 parameter data holdout for planning and 5% for testing and validating the classifier accuracy.

Classification utilizes a learning evaluation process to eliminate the sampling biases, the planning and review methods are evaluated. The recorded measurement results are grounded on time of computation and proficiency parameters (Accuracy, Specificity, Sensitivity, Precision, F-score and Recall) [25].

This analysis relates the classification efficiency of models, with SVM classifiers to their kernels. The output result and matrix of the confusion are shown in the figures below. Figure 1 displays the data for evaluating the charged Mosquito Anopheles.

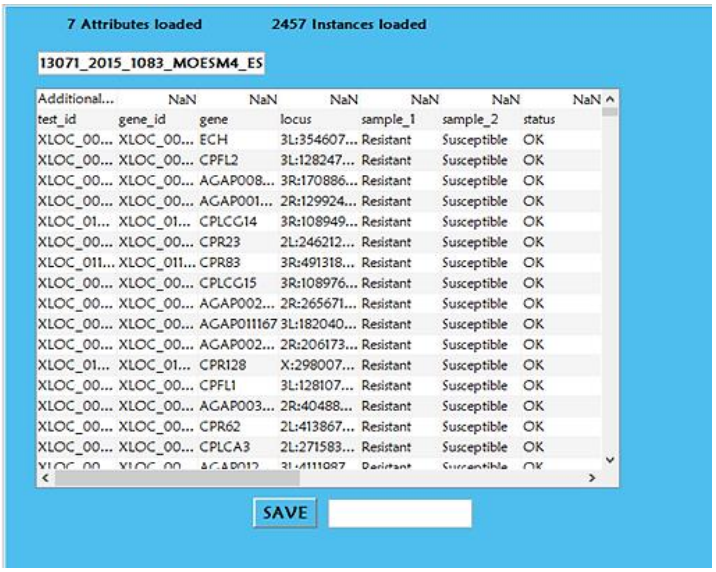

Fig. 1. Mosquito Anopheles gambiae loaded dataset.

ICA fetches the latent components from the loaded dataset shown in Fig. 1. The derived characteristics are translated into the classification of SVM, and the results are seen in the following Fig. 2. Figure 3 illustrates the fractured classification of plots x and y; the ambiguity matrix provides a solution for the performance measures as described in Figs. 4 and 5.

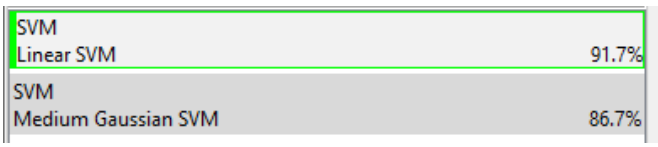

Fig. 2. Classification results using SVM-Kernels.

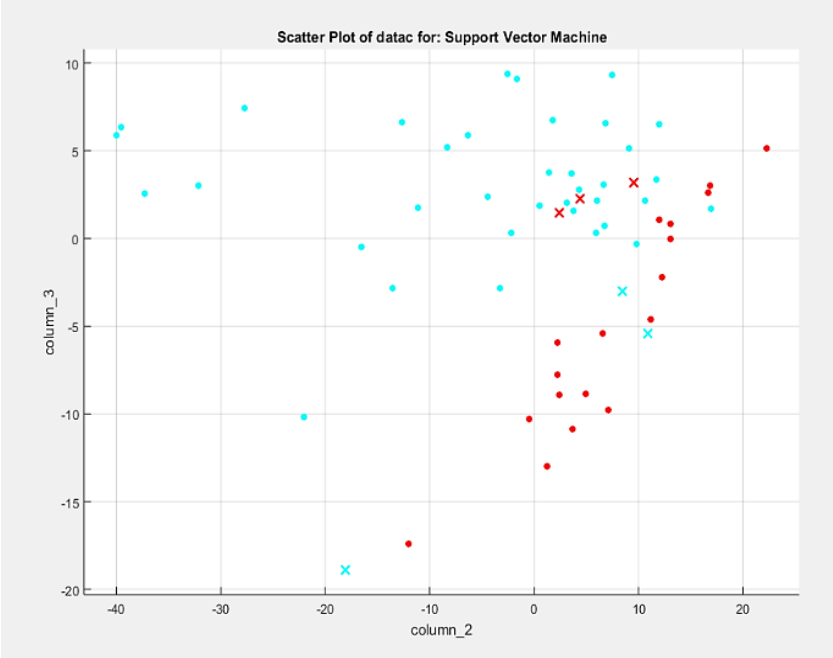

Fig. 3. Dataset scattered plot for SVM classification.

In this study, the evaluation of the selected optimal latent components has been classified using an SVM classifier, to evaluate the experiment, a confusion matrix is generated, which comprises of true positive, true negative, false positive and false negative values of classes, these values are used in generating the performance metrics evaluation, Figs. 4 and 5 show the labels and classes in the confusion matrix.

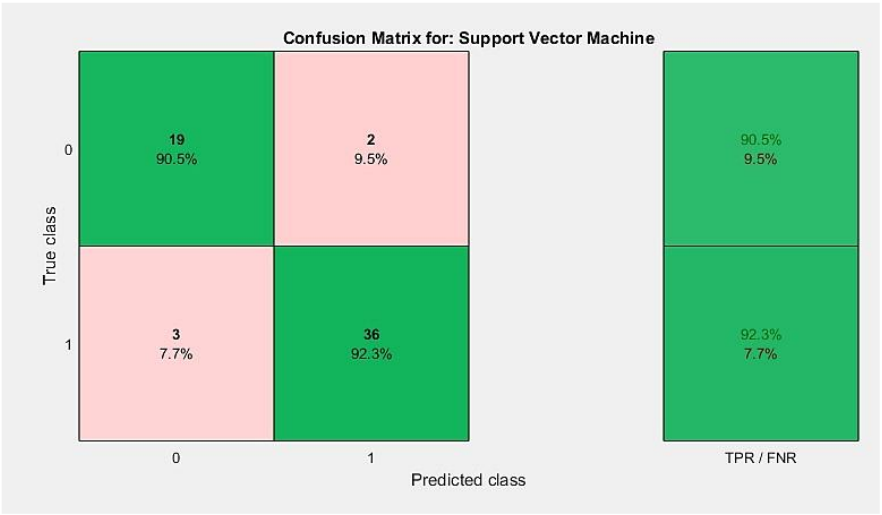

Fig. 4. Confusion matrix for the classification using linear-SVM (L-SVM) TP=36; TN=19; FP=2; FN=3.

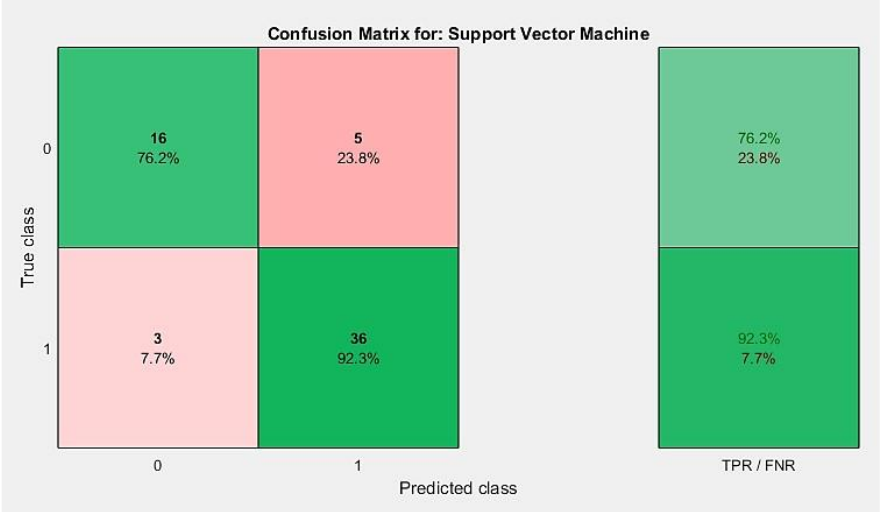

Fig. 5. Confusion matrix for the classification using SVM-RBF TP=36; TN=16; FP=5; FN=3.

RNA-Seq data was gotten for Mosquito Anopheles Gambiae to assess the output from:
https://figshare.com/articles/Additional_file_4_of_RNAseq_analyses_of_changes_in_the_Anopheles_gambiae_transcriptome_associated_with_resistance_to_pyrethroids_in_Kenya_identification_of_candidateresistance_genes_and_candidate resistance_ SNPs/4346279/1

2457 gene features were obtained, ICA was utilized as a basis for a lessening in dimensionality and 25 latent factor features were removed. Instead, these components are categorized using SVM classification algorithms to estimate their performance. The outcome demonstrates the machine-learning technology's success in embryos. The success findings are shown and compared in Table 2 for confirmation of the method. The finding reveals that L-SVM is outperforming RBF-SVM Tree in terms of less time of training and output precision.

This analysis evaluated and enhanced the malaria vector data classification, many studies were suggested in evaluations by investigators with the performance metrics seen in Table 2, Fig. 6 shows the classification and performance metrics diagram where Linear-SVM proves to outperform RBF-SVM in terms of precision, the results showed that, dimensionality reduction model using ICA function.

Table 2. Performance metrics table for the confusion matrix.

Performance Metrics	L-SVM Classification	RBF-SVM Classification
Accuracy (%)	92	87
Sensitivity (%)	92.3	92.3
Specificity (%)	90.5	76.2
Precision (%)	94.7	87.8
Recall (%)	92.3	92.3
F-Score (%)	93.5	90.0

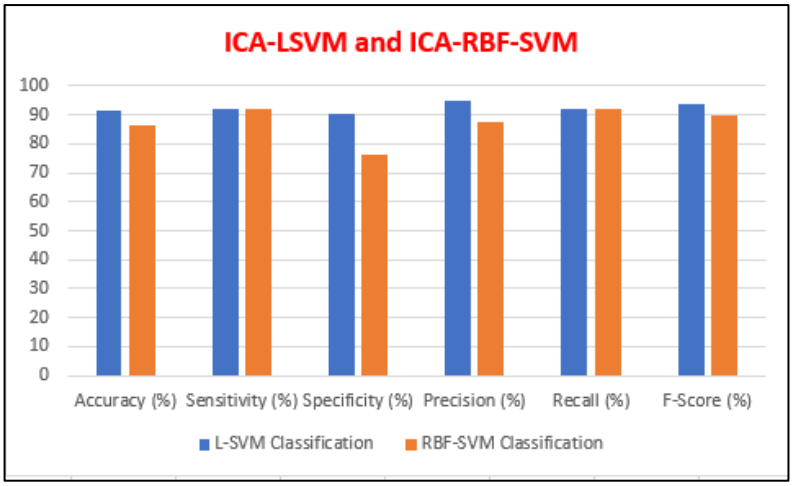

Fig. 6. Result classification chart.

This research strengthens and can be useful in human malaria ailment prognosis and diagnosis. The theoretical solution with machine learning methods such as model and classification procedures for the reduction of dimensionality.

System Dimensionality Reduction follows the ICA function extraction technique which employs SVM classifiers.

This study carried out performance analysis and assessment and the findings gotten were revealed, L-SVM outstrips the classification procedure SVM-RBF.

5. Conclusions

This study analysed and enhanced the malaria vector data classification, numerous works were suggested in researchers' analyses, the results showed that the model of dimensionality reduction with feature extraction approaches such as ICA can support advanced classification productivity such as SVMs.

It will be important to explore whether the function extraction models and algorithms can be enhanced by the recently proposed research. In future research, it is important to implement feature selection and other feature extraction approaches for proportional assessment and to demonstrate that other approaches that can be utilized to enhance classification efficiency related to the state of the art.

References

1. Sun, S.; Wang, C.; Ding, H.; and Zou, Q. (2019). Machine learning and its applications in plant molecular studies. *Briefings in Functional Genomics*, 19(1), 40-48.
2. Read, D.F.; Cook, K.; Lu, Y.Y.; Le Roch, K.G.; and Noble, W.S. (2019). Predicting gene expression in the human malaria parasite plasmodium falciparum using histone modification, nucleosome positioning, and 3D localization features. *PLOS Computational Biology*, 15(9), e1007329.
3. Arowolo, M.O.; Adebiyi, M.O.; Adebiyi, A.A. (2019). A dimensional reduced model for the classification of rna-seq anopheles gambiae data. *Journal of Theoretical and Applied Information Technology*, 97(23), 3487-96.

4. Sekaran, K.; and Sudha, M. (2019). Diagnostic gene biomarker selection for Alzheimer's classification using machine learning. *International Journal of Innovative Technology and Exploring Engineering*, 8(12), 2348-2352.
5. Johnson, N.T.; Dhroso, A.; Hughes, K.J.; and Korkin, D. (2018). Biological classification with RNA-SEQ data: Can alternatively spliced transcript expression enhance machine learning classifiers? *RNA*, 24(9), 1119-1132.
6. Libbrecht, M.W.; and Noble, W.S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), 321-332.
7. Jagga, Z.; and Gupta, D. (2014). Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. *BMC Proceedings, Proceedings of the Great Lakes Bioinformatics Conference* 2014, 8(S6), 1-7.
8. Anopheles gambiae 1000 Genomes Consortium; Data analysis group; Partner working group; Genetic diversity of the African malaria vector Anopheles gambiae. *Nature*, 552, 96-100.
9. Oh, D.H.; Kim, I.B.; Kim, S.H.; and Ahn, D.H. (2017). Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning. *Clinical Psychopharmacology and Neuroscience*, 15(1), 47-52.
10. Qi, R.; Ma, A.; Ma, Q.; and Zou, Q. (2019). Clustering and classification methods for single-cell RNA-sequencing data. *Briefings in Bioinformatics*, 21(4), 1196-1208.
11. Wenric, S.; and Shemirani, R. (2018). Using supervised learning methods for gene selection in RNA-SEQ case-control studies. *Frontiers in Genetics*, 9, Article 279, 1-9.
12. Alquicira-Hernandez, J.; Sathe, A.; Ji, H.P.; Nguyen, Q.; and Powell, J.E. (2019). *scPred*: Accurate supervised method for cell-type classification from single-cell RNA-SEQ data. *Genome Biology*, 20, 264.
13. Cui, S.; Wu, Q.; West, J.; and Bai, J. (2019). Machine learning-based microarray analyses indicate low-expression genes might collectively influence PAH disease. P*LOS Computational Biology*, 15(8): e1007264.
14. Shon, H.S.; Yi, Y.; Kim, K.O.; Cha, E.-J.; and Kim, K.-A. (2019). Classification of stomach cancer gene expression data using CNN algorithm of deep learning. *Journal of Biomedical Translational Research*, 20(1), 15-20.
15. Reid, A.J.; Talman, A.M.; Bennett, H.M.; Gomes, A.R.; Sanders, M.J.; Illingworth, C.J.; Billker, O.; Berriman, M.; and Lawniczak, M.K. (2018). Single-cell RNA-SEQ reveals hidden transcriptional variation in malaria parasites. *eLife* 7: e33105.
16. Tan, A.C.; and Gilbert D. (2003). Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics*, 2(3);S75-83.
17. Song, N.; and Wang, K. (2016). Design and analysis of ensemble classifier for gene expression data of cancer. *Advancements in Genetic Engineering*, 01(S1).
18. Tarek, S.; Abd Elwahab, R.A.; and Shoman, M. (2017). Gene expression based cancer classification. *Egyptian Informatics Journal*, 18(3), 151-159.
19. Lee, Y.; and Lee, C.-K. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, 19(9), 1132-1139.

20. Tan, C.S.; Ting, W.S.; Mohamad, M.S.; Chan, W.H.; Deris, S.; and Shah, Z.A. (2014). A review of feature extraction software for Microarray gene expression data. *BioMed Research International*, Volume 2014 |Article ID 213656, 1-15.
21. Zahoor, J.; and Zafar, K. (2020). Classification of Microarray gene expression data using an infiltration tactics optimization (ITO) algorithm. *Genes*, 11(7), 819.
22. Arowolo, M.O.; Adebiyi, M.O.; Adebiyi, A.A.; and Olugbara, O. (2020). Optimized hybrid heuristic based dimensionality reduction methods for malaria vector using KNN classifier. *Research Square*, 1-16.
23. Hameed, S.S.; Hassan, R.; Hassan, W.H.; Muhammadsharif, F.F.; and Latiff, L.A. (2021). HDG-select: A novel GUI based application for gene selection and classification in high dimensional datasets. *PLOS ONE*, 16(1): e0246039.
24. Bonizzoni, M.; Ochomo, E.; Dunn, W.A.; Britton, M.; Afrane, Y.; Zhou, G.; Hartsel, J.; Lee, M.; Xu, J.; Githeko, A.; Fass, J.; and Yan, G. (2015). RNA-SEQ analyses of changes in the anopheles gambiae transcriptome associated with resistance to pyrethroids in Kenya: Identification of candidate-resistance genes and candidate-resistance SNPs. *Parasites & Vectors*, 8, 474.
25. James, G.; Witten, D.; Hastie, T.; and Tibshirani, R. (2013). Statistical learning. Springer Texts in Statistics, 15-57.
26. Aydadenta, H.; and Adiwijaya (2018). On the classification techniques in data mining for microarray data classification. *Journal of Physics: Conference Series*, 971, 012004.
27. Polaka, I.; Tom, I.E.; and Borisov, A. (2010). Decision tree classifiers in bioinformatics. *Scientific Journal of Riga Technical University*. *Computer Sciences*, 42(1), 118-123.
28. Chang, C.-C.; and Lin, C.-J. (2011). LIBSVM. A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1-27.
29. Soofi, A.A.; and Awan, A. (2017). Classification techniques in machine learning: Applications and issues. *Journal of Basic & Applied Sciences*, 13, 459-465.
30. Khan, A.; Baharudin, B.; Lee, L.H.; and Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, 1(1), 4-20.
31. Suksut, K.; Kaoungku, N.; Kerdprasop, K.; and Kerdprasop, N. (2019). Improvement the imbalanced data classification with restarting genetic algorithm for support vector machine algorithm. *International Journal of Future Computer and Communication*, 8(2), 63-67.
32. Vanitha, C.D.A.; Devaraj, D.; and Venkatesulu, M. (2015). Gene expression data classification using support vector machine and mutual information-based gene selection. *Procedia Computer Science*, 47, 13-21.
33. Arowolo, M.O.; Abdulsalam, S.O.; Isiaka, R.M.; and Gbolagade, K.A. (2017) A Hybrid Dimensionality Reduction Model for Classification of Microarray Dataset. *International Journal of Information Technology and Computer Science* (*IJITCS*), 9(11), 57-63
34. rowolo, M.O.; Abdulsalam, S.O.; Isiaka, R.M.; and Gbolagade, K.A. (2018). A comparative analysis of feature selection and feature extraction models for classifying microarray dataset. *Computing and Information System*, 22(2), 29-38.