Journal of Theoretical and Applied Information Technology
15th December 2019. Vol.97. No 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A DIMENSIONAL REDUCED MODEL FOR THE CLASSIFICATION OF RNA-SEQ ANOPHELES GAMBIAE DATA

[1]MICHEAL OLAOLU AROWOLO, MARION ADEBIYI, AYODELE ADEBIYI
[1]Department of Computer Science, Landmark University, Omu-Aran Nigeria
E-mail: [1]arowolo.micheal@lmu.edu.ng, marion.adebiyi@lmu.edu.ng, ayo.adebiyi@lmu.edu.ng

ABSTRACT

A significant application of gene expression RNA-Seq data is the classification and prediction of biological models. An essential component of data analysis is dimension reduction. This study presents a comparison study on a reduced data using Principal Component Analysis (PCA) feature extraction dimension reduction technique, and evaluates the relative performance of classification procedures of Support Vector Machine (SVM) kernel classification techniques, namely SVM-Polynomial kernels and SVM-Gaussian kernels. An accuracy and computational performance metrics of the processes were carried out. A malaria vector dataset for Ribonucleic Acid Sequencing (RNA-Seq) classification was used in the study, and 99.68% accuracy was achieved in the classification output result.
Keywords: *RNA-Seq, PCA, SVM-Gaussian Kernel, SVM-Polynomial Kernel, Malaria Vector*

1. INTRODUCTION

In biological learning, Next-generation sequencing (NGS) has been expansively utilized. General NGS information is the Ribonucleic Acid sequencing (RNA-seq); it is utilized to test the anomalies of mRNA expression in ailments. In difference with microarray advancements, RNA-Seq talks about significant data that presents explicit inventiveness of narrative protein isoforms with various compound scopes of uncovered qualities.

In recent era, RNA-Seq has become an expansively utilized genome-wide expression profile for figuring substance cells, because of their capacity of determining potential heterogeneities in cell populaces [1].

Since the advancement of RNA tasks as a notable intermediary among genome and proteome, finding and estimating gene expression have been the unmistakable conduct in biological science [2]. Sequencing RNA has a control where by finding and evaluation can be consolidated in a single high-throughput sequencing measure called RNA-sequencing (RNA-seq). Adoption of RNA-seq has expanded in the genomics society and has become a regular toolbox utilized by the research society with regards to biological sciences. A few RNA-seq conventions, investigations, challenges, among others have been looked into reviews to achieve RNA-seq learning appropriately.

No foremost prospective or good channel for the assorted variation of claims and investigation state where RNA-seq is utilized. Researches and adoption of systematic methodologies on living being and their objectives have advanced [3].

A thriving RNA-seq study must have a major prerequisite of creating information with the possibilities of responding to biological inquiries of concern. This is practiced by characterizing an investigational aim, series intensity and replicating reasonable biological plans under examination and by advancement of sequencing research, ensuring that information achievement does not end up being tainted with redundant views. One critical part of the RNA-Seq information is the expulsion of the scourge of high-dimension, for example, noises, commotions, repetition, redundancy, immaterial as well as irrelevant data, among others [4]. Because of high-measurement of biological information challenges, dimension reduction techniques are vital. RNA-Seq information has turned out to be a potential high-throughput procedure to simultaneous profile transcriptomes of substantial information [5]. RNA-Seq has key advantages, for example, the capacity to spot narrative transcripts, precision, and dynamic range [6]. Thousands of quality genes are instantaneously communicated and stated in RNA-Seq, expression levels of genes are habitually difficult, discovering an active low-dimensional illustration of RNA-Seq information is important. A few dimension reduction procedures utilized for RNA-Seq data investigation and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

information study to expel noises related to explicit information exist [7]. A notable dimension reduction procedure is based on principal component analysis (PCA). An unsupervised dimension reduction method that does not differentiate amid precise cell types directly, it has been extensively utilized, yet it proclaims interesting features is known as PCA [8].

This study proposes a computational PCA dimensionality reduction technique, to deal with the issue of curse of high dimensionality in gene expression space and analyzes SVM kernel classification methods. This study exhibits the robustness of this techniques, regarding to noises and sampling on RNA-Seq Anopheles Gambiae dataset, which affects the prognosis and diagnosis of malaria ailments.

2. RELATED WORKS

The related work shows the extensive researches that has been carried out to model this study.

In 2015, Pierson and Yau. [4] worked on a dimensionality reduction model for zero inflated single cell gene expression investigation, they built a dimensionality reduced technique, zero inflated factor analysis (ZIFA), the dropout attributes were modelled expressly, and demonstrate that it advances modelling precision on biological and simulated datasets. They improved the PPCA and FA context to represent dropout and deliver a benign technique for the dimensionality reduction of single-cell gene expression data giving robustness contrary to such vulnerabilities.

Without dropouts, the method is basically equal to PPCA or FA. Hence, users can utilize ZIFA as an immediate additional with the advantage that it consequently represents dropouts while remedial endeavors might be required with standard PCA. There procedure varies from methodologies, for example, the numerous variations of strong PCA, which mean to show corrupted perceptions. ZIFA regards dropouts as genuine perceptions, not exceptions, whose event properties have been described utilizing an observationally educated factual model.

In 2015, Esra, Hamparsum, and Sinan [9] worked on an innovative fusion dimension reduction method for small high dimensional gene expression datasets with information intricacy principle for cancer classification. Their study addressed the restrictions inside the setting of Probabilistic PCA (PPCA) by presenting, building up innovative methodology utilizing most extreme entropy covariance matrix and its fusion levelled covariance estimators. To diminish the dimensionality of the information and pick the quantity of probabilistic PCs (PPCs) to be held, they presented and created observed consistent Akaike's information criterion (CAIC), Akaike's information criterion (AIC), and the information theoretic measure of complexity (ICOMP) rule of Bozdogan. Openly accessible undersized benchmark informational collections of six were breaking down to demonstrate the effectiveness, adaptability, and flexibility of their methodology with fusion smoothed covariance matrix estimators, that does not decline to play out the PPCA to diminish the measurement and to do regulated characterization of malignancy bunches in high measurements. Their proposed technique can be utilized to take care of new issues and difficulties existing in the investigation of NGS information in bioinformatics and biomedical applications.

In 2016, Wenyan, Xuewen and Jingjing [10] worked on feature selection for cancer classification for disease utilizing microarray data expression. This paper used information on microarray gene expression level to decide marker genes that are pertinent to a sort of malignancy. They researched a separation-based element choice strategy for two-gather grouping issue. So as to choose marker genes, the Bhattacharyya separation is actualized to quantify the uniqueness in gene expression levels. They used SVM for classification with utilization of the selected marker genes. The execution of marker gene selection and classification are represented in both recreation studies and two genuine information analyses by proposing an innovative gene selection technique for classification based on SVMs. In the proposed scheme, they firstly ranked every gene according to the importance of their Bhattacharyya distances among the indicated classes. The optimal gene subset is chosen to accomplish the least misclassification rate in the developed SVMs following a forward selection algorithm. 10-fold cross-validation is connected to locate the optimal parameters for SVM with the concluding optimal gene subset. Subsequently, the classification model is trained and built. The classification model is estimated by its prediction for testing set. The execution of the proposed B/SVM technique with that of SVM-RFE and SWKC/SVM gives normal misclassification rate (1.1%) and high normal recovery rate (95.7%).

In 2017, Nancy and VijayKumar, [11] worked on Alzheimer's infection determination by utilizing dimensionality reduction based on KNN classification algorithm for analyzing and classifying the Alzheimer malady and mild

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

cognitive mutilation are available in the datasets. Their study gave more precision rate, accuracy rate and sensitivity rate to give a better output. This paper proposed a narrative dimensionality reduction based KNN classification Algorithm dissected the Alzheimer's illness present in the datasets. With the algorithm, the dataset was separated into 3 classes; first class having the Alzheimer's disease (AD), second class was having the normal outcome, third class having the mild cognitive impairment. The information's were taken from the researcher's data dictionary - Uniform Data Set (RDD-UDS).

The relative investigations between the current PNN classification procedures with the proposed KNN classification demonstrated that high measure of normal accuracy, sensitivity, specificity precision, recall, jaccard and dice coefficients furthermore diminish the information dimensionality and computational multifaceted nature. Their future work, stated that the feature extraction and classification algorithm will improve the classification performance.

In 2017, Usman Shazad, and Javed [12] worked on PCA and Factor Analysis for dimensionality reduction of bio-informatics data, they utilized the dimensionality reduction model of bioinformatics information. These systems used Leukemia dataset and its attributes was decreased. An investigation was exhibited on reducing the number of attributes using PCA and Factor Analysis. Leukemia data was used for the analyses. PCA was carried out on the dataset and 9 components were chosen out of the 500 components. The Factor Analysis was implored to extract the critical features.

In 2017, Gökmen, Dincer, Selcuk, Vahap, Gozde, Izzet, and Ahmet, [13] worked on a simulation study for the RNA-Seq data classification, they contrasted a few classifiers including PLDA renovation, NBLDA, single SVM, bagging SVM, CART, and random forest (RF). They analyzed the impact of a few parameters, for example, over-dispersion, number of genes, sample size, and classes, differential expression rate, and the transform technique on model performances. A broad modeled study was conducted and the outcomes were contrasted using the consequences of two miRNA and two mRNA exploratory datasets. The outcomes uncovered that expanding the differential expression rate, sample size, and transformation method on model presentation. RNA-Seq data classification requires cautious consideration when taking care of data over-scattering. They ended up that count-based classifier, the power changed PLDA and as classifiers, vst or rlog changed RF and SVM classifiers might be decent decision for classification.

In 2017, Chieh, Siddhartha, Hannah, and Ziv [8] used neural network algorithm to reduce the dimensions of single cell RNA-Seq data containing a few new computational complexities. These incorporate inquiries concerning the top strategies for clustering scRNA-Seq data, recognizing unique cells, and deciding the capacity of explicit cells dependent on their expression profile. Addressing these problems, they created and tested a technique based on neural network (NN) for the analysis and recovery of single cell RNA-Seq data. They showed different NN structures, some fuse prior biological learning, and utilized these to acquire a reduced dimension representation of the single cell expression data. They demonstrate that the NN technique enhances earlier strategies in the capacity to accurately group cells in analyses not utilized in the training and the capacity to effectively derive cell type or state by questioning a database of a huge number of single cell profiles. Database queries (utilizing a webserver) will empower investigators to characterize cells better while investigating heterogeneous scRNA-Seq tests.

In 2017, Ian and Jorge [14] reviewed recent ongoing advancements in PCA as a strategy for diminishing the dimensionality of RNA-Seq datasets, for expanding interpretability and yet limiting data misfortune by making new uncorrelated factors that progressively maximize variance. This study presented the essential thoughts of PCA, talking about what it can, can't do and after that depict a few variations of PCA and their application.

In 2018, Dongfang and Jin, [5] proposed a single-cell RNA-Sequential data using a deep variational autoencoder using an unsupervised feature extraction model. The VASC models the dropout and fetches the nonlinear hierarchical feature representation of the high dimensional data. There result was tested on about 20 datasets, the VASC showed a better performance with broader compatibility features.

In 2018, Etienne, Leland and John [34] worked on the dimensionality reduction model for visualizing single-cell data using UMAP. A nonlinear dimensionality-reduction method, uniform manifold approximation and projection (UMAP), was established for the investigation of any kind of high-dimensional data. They applied UMAP to biological data, using three well-characterized mass cytometry and single-cell RNA sequencing datasets. Associating the act of UMAP with five other tools, we find that UMAP provides the fastest

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

run times, highest reproducibility and the most meaningful organization of cell clusters. The work highlights the use of UMAP for improved visualization and interpretation of single-cell data.

2018, Jiarui, Anne and Sohrab [31] worked on an interpretable dimensionality reduction model of single cell transcriptome data with deep generative models by working on a robust model called the SCVIS, this captured and showed the lower dimensional structure in single cell gene expression of the data. A simulated demonstration of the lower dimensional data was presented, which preserved the local and global structures in the data. They used scvis to analyze four single-cell RNA-sequencing datasets, demonstrating interpretable two-dimensional representations of the high-dimensional single-cell RNA-sequencing data.

2019 Geng, Baitang and Tieliu [30] proposed an ScRNA-Seq technologies and its relating computational analysis, in their review, they provided an overview of currently available single-cell protocols and discussed several techniques for several RNA-Seq Data analysis such as their gene expressions, mapping, cell clustering, imputation, normalization, feature selection, feature extraction, among others.

Malte and Fabian, 2019 [32] worked on the current practices in ScRNA-Seq analysis, by formulating present best-practice endorsements for steps based on self-determining comparison studies. They combined these best-practice references into a workflow, applied to a public dataset to demonstrate its training. This review serves as a workflow tutorial for innovative participants into the field, and help established users update their analysis pipelines.

2019, Tamim, Lieke, Davy, Dylan, Hailiang, Marcel, and Ahmed [33] worked on a comparative analysis of automatic cell identification methods fro scRNA-Seq data, they benchmarked 22 classification approaches that repeatedly allot cell characteristics including single-cell-specific and general-purpose classifiers. The presentation of the approaches is assessed using 27 openly accessible single-cell RNA sequencing datasets of diverse sizes, knowledges, species, and levels of intricacy. They used 2 experimental setups to estimate the performance of each method for within dataset predictions (intra-dataset) and across datasets (inter-dataset) based on accuracy, percentage of unclassified cells, and computation time. We further evaluate the methods' sensitivity to the input features, number of cells per population, and their performance across different annotation levels and datasets. We find that most classifiers perform well on a variety of datasets with decreased accuracy for complex datasets with overlapping classes or deep annotations. The general-purpose support vector machine classifier has overall the best performance across the different experiments.

2019, Ren, Anjun, Qin and Quan [35] worked on the clustering and classification models for ScRNA-Seq data. In their paper, they systematically review the integrated methods and tools, highlighting the pros and cons of each approach. They paid attention to clustering and classification methods as well as discussing methods that have emerged recently as powerful alternatives, including nonlinear and linear methods and descending dimension methods. Finally, we focus on clustering and classification methods for scRNA-seq data, in particular, integrated methods, and provide a comprehensive description of scRNA-seq data and download URLs.

3. METHODOLOGY

The proposed approach for this study analyzed and discussed the dataset used and framework, it is discussed in details below.

3.1 DATASET USED FOR ANALYSIS

The RNA-seq information was employed to analyze the variations in transcriptome of deltamethrin- resistant and vulnerable Anopheles gambiae mosquitoes in western Kenya, an openly accessible dataset from figshare.com and financed by the National Institute of Health [15]. The dataset comprises of list of genes such as; AGAP012984, AGAP0 02724, AGAP003714, AGAP004779, AGAP009472, CPLC G3 [AGAP008446], CYP6M2 [AGAP008212] and CYP6P3 [AGAP002865], genes in mosquitoes from western Kenya in the year 2010 and 2012 with the dataset comprising of 2457 genes and 2 cells.

MATLAB (Matrix Laboratory) is utilized to perform the experiment, due to its ease and beneficial programming environment for engineers, architects, scientists, researchers, among others. MATLAB is a multi-worldview arithmetical processing environment and exclusive programming language established by MathWorks. It permits framework controls, plotting of functions and information, execution of algorithms, production of User Interfaces, written in different

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

languages, such as; C, C++, C#, Java, Fortran and Python [16]. The principle point of this study is the prediction of the RNA-Seq technology utilizing the MATLAB tool by utilizing the Malaria database. Table-1 demonstrates a concise description of the dataset.

Table 1: Dataset Features

Dataset	Attributes	Instances
Anopheles Gambiae Mosquitoes	2	2457

3.2 Experimental Methodology

This study summarizes the proposed framework in figure-1 below. The fundamental idea is to predict machine learning task on high dimensional RNA-Seq data, for cells and genes into lower dimensional dataset. The plan is adjusted to fetch out important data in a given dataset by utilizing PCA feature extraction method as a stage. To evaluate the performance of RNA-seq dataset, two Support Vector Machine (SVM) classification algorithms are compared.

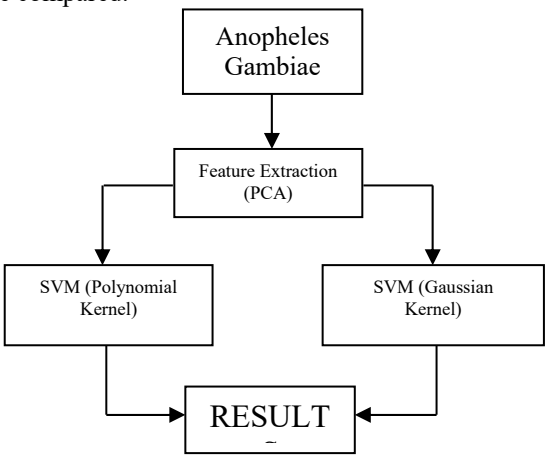

Figure 1: Proposed Framework.

RNA sequencing (RNA-seq) is the next generation sequencing technology to think about in transcriptome. It is utilized as an option to microarrays for gene expression analysis, without the need to earlier realize the RNA succession. RNA-seq offers progressively precise information and applications including identification of gene fusion, variations, alternative joining, post-transcriptional changes as well as analysis of small RNAs, such as; tRNA or miRNA profiles. A total image of the RNA substance can be gotten from low quantity biological samples. A few expository advances are basic for an effective portrayal and evaluation of the transcriptome. Bioinformatics tools are proposed for quality control, information handling, annotation, quantification and representation for translation and biological science investigation for understanding RNA-seq information.

3.3 Principal Component Analysis (PCA)

PCA: a broadly utilized unsupervised feature extraction dimensionality decrease technique, because of its simplicity, [17]. It utilizes a direct procedure to implant information into a linear subspace of lower dimensionality. PCA maps each occurrence of the given dataset present in a dimensional space to a j dimensional subspace such that $j < a$. The set of j new dimensions produced are known as the Principal Components (PC) and every principal component is coordinated towards maximum variance without the variance previously represented in the preceding components [18]. PCA is widely the most prevalent (unsupervised) linear technique; it builds a low-dimensional representation of the data that depicts much of the difference in the data as could be expected. This is done by finding a linear premise of decreased dimensionality for the information, in which the amount of difference in the data is maximal. PCA computation transformation matrix U adopted [19] and given as:

$$U = (\sum_{i=1}^{n}(B_i - l)(B_i - l)^S) \quad (1)$$

Where;

n is the instances

Bi is the i-th instance

l is the mean vector of the input data

The given high-dimensional input data are standardized as each attribute falls within same range, to ensure that all attributes with larger domains in the data does not overwhelm attributes with little domain. PCA compute the symmetrical vector which gives a premise to standardized data. The input data are linear combination of PC. It is arranged in diminishing order of their quality or criticalness, the size of data can be decreased by weaker component implying that PCs with lower variance.

3.4 Classification

A few classification algorithms exist, for example, Logistic Regression, SVM, K-Nearest Neighbor, among others [20]. After reducing the dimensional complexity of data, the subsequent stage is the classification procedure. Classification is the fundamental goal; the analyzed data is classified. Two SVM kernels techniques were utilized: Polynomial Kernel and Gaussian Kernel. The results of the algorithms are analyzed and compared

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

based on computational time, training time and performance metrics such as accuracy.

3.4.1 Support Vector Machine (SVM)

SVM is a learning machine algorithm presented by Vapnik in 1992 [21] Aydadenta and Adiwijaya, 2018). The algorithm works with the point of finding the best hyperplane that isolates between classes in the input space. SVM is a linear classifier; it is created to work with nonlinear problems by joining the kernel ideas in high-dimensional workspaces. In non-linear issues, SVM utilizes a kernel in training the data with the goal of spreading the dimension widely. When the dimensions are tweaked, SVM will look for the optimal hyperplane that can separate a class from different classes [26]. As indicated by the adoption of Aydadenta and Adiwijaya (2018) [21], the procedure to locate the best hyperplane utilizing SVM is as follows:

i. Let

$$y_t \in \{y_1, y_2, \dots, y_n\}, \text{where } y_t \text{ is the } p - \text{attributes and target class } z_t \in \{+1, -1\}$$

ii. Assuming the classes +1 and -1 can be separated completely by hyperplane, as defined in equation 2 below

$$v.y + c = 0 \quad (2)$$

From equation (2), Equations (3) and (4) are gotten:

$$v.y + c \geq +1, \text{ for class } +1 \quad (3)$$

$$v.b + c \leq -1, \text{ for class } -1 \quad (4)$$

Where, *y* is the input data, *v* is the ordinary plane and *c* is the positive relative to the center field coordinates.

SVM intends to discover hyperplanes that maximizes margins between two classes. Expanding margins is a quadratic programming issue that is solved by finding the minimal point. The advantage of SVM is its capacity to manage wide assortment of classification problems in high dimensional data [22].

Compared to other classification methods, SVM is outstanding, with its exceptional classification adequacy [23]. SVM is grouped into linear and non-linear separable. SVM's has kernel functions that change data into a higher dimensional space to make it conceivable to accomplish separations. Kernel functions are a class of algorithms for pattern investigation or recognition. Training vectors *xi* is mapped into higher dimensional space by the capacity Φ. SVM discovers a linear separating hyperplane with the maximal in this higher dimension space. $C > 0$ is the penalty parameter of the error term.

There are several SVM kernels that exist such as; the polynomial kernel, Radial basis function (RBF), linear kernel, Sigmoid, Gaussian kernel, String Kernels, among others. The decision of a Kernel relies upon the current issue at hand, since it relies upon what models are to be analyzed, a couple of kernel functions admirably in for a wide assortment of applications [24]. The prescribed kernel function for this study is the SVM-Polynomial Kernel and Gaussian Kernel.

3.4.2 SVM-Gaussian Kernel

Gaussian kernel [25] compare to a general smoothness supposition in all k-th order subordinates. Kernels coordinating a certain prior recurrence substance of the data can be developed to reflect earlier issues in learning. Each input vector $\underline{x}$ is mapped to an interminable dimensional vector including all degree polynomial extensions of *x's* components.

3.4.3 SVM Polynomial Kernel

For instance, a polynomial kernel model features conjunction to the order of the polynomial. Radial basis functions permit circles in disparity with the linear kernel, which permits just selecting lines (or hyperplanes).

$$K(y_a, y_i) = (\gamma y_a^S y_b + q)^\varepsilon, \gamma > 0 \quad (5)$$

3.4.3 SVM-Linear Kernel Function

For instance, polynomial kernel is the least complex kernel function. It is given by the inner product *(a,b)* in addition to a discretionary constant K.

$$K(y_a, y_b) = y_a^S y_b \quad (6)$$

3.4.4 SVM-RBF Kernel Function

In SVM kernel functions, **γ**, *a,* and *b* are kernel parameters, RBF is the fundamental kernel function due to the nonlinearly maps tests in developed dimensional space, compared to the linear kernel, it has reduced hyper parameters compared to the polynomial portion.

$$K(y_a, y_b) = \exp(-\gamma \|y_a, y_b\|^2), \gamma > 0 \quad (7)$$

4. RESULTS

This study explores RNA-Seq innovation containing 2457 examples of Anopheles Gambiae Mosquitoes data, having susceptible and resistant genes. PCA algorithm was executed on the data to reduce the curse of dimensionality. PCA identifies and remove uncorrelated Attributes (Variables), to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

decide maximum variance with a smaller number of Principal Components. In this study, PCA is applied on the given data, to lessen the dimensionality issue and give significant gene information that is useful for further investigation. Classification algorithm applies SVM-Gaussian kernel and Polynomial kernel by utilizing MATLAB to implement the model.

Using PCA as a dimensionality reduction method, 1592 out of 2457 genes were significant and 45 latent components were achieved in 7.8486 Seconds.

A supervised SVM kernel classifier methods, is among the most well-established and popular machine learning approaches in bioinformatics and genomics, 10-folds cross validation was employed to evaluate the execution of the performance of the classification models, using 0.05 parameter holdout of data for training and 5% for testing to check the accuracy of the classifiers.

To each of the classifiers, a basic supervised learning assessment protocol is carried out. In particular, the training and testing stages are assessed as a 10-fold cross validation to eliminate the sampling bias. This protocol is implemented using MATLAB. The reported result of assessment is based on the computational time and performance metrics (Accuracy, Specificity, Sensitivity, Precision, F-score and Recall) [27]. This study compares the classification performance of the models, using SVM-Polynomial kernel, with 99.68% accuracy, the experiment was trained in 63.53Secs. The result output and confusion matrix are shown below, in figure 2.

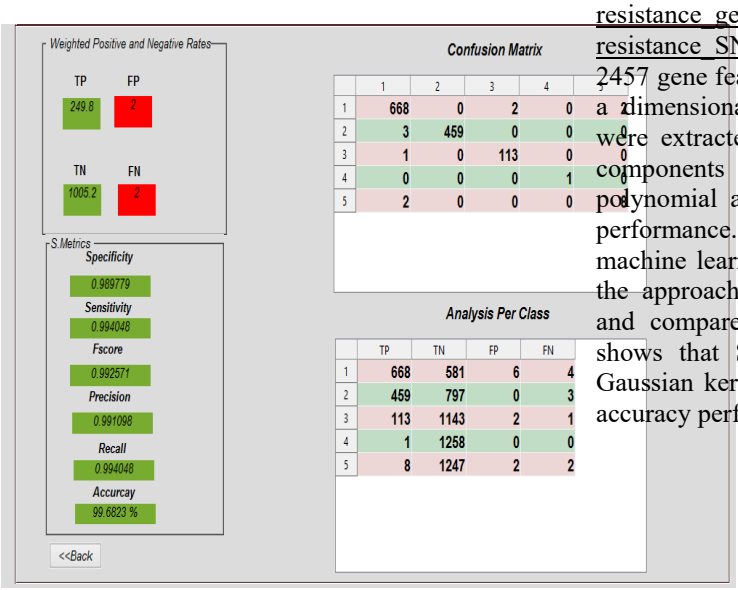

Figure 2: Confusion Matrix and Performance Metrics for SVM-Polynomial Kernel

Using SVM-Gaussian kernel, the extracted data was classified and accomplish 99.3% Accuracy, it trained the data in 337Sec. The outcome is shown in Figure 3 beneath

.

Figure 3: Confusion Matrix and Performance Metrics for SVM-Gaussian Kernel

To test the performance of datamining learning method, RNA-Seq data was downloaded for Mosquito Anopheles Gambiae [29] https://figshare.com/articles/Additional_file_4_of_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/4346279/1

2457 gene feature were collected, PCA was used as a dimensionality reduction model, 1572 features were extracted with 45 latent components. These components are then classified using SVM polynomial and Gaussian kernels to predict their performance. The result shows the effectiveness of machine learning technology in genes. To validate the approach, the performance results are shown and compared in the table 2 below. The result shows that SVM-polynomial kernel outperforms Gaussian kernel in terms of less training time and accuracy performance.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Execution Results Table

Performance Metrics	PCA-SVM-Polynomial Kernel	PCA-SVM-Gaussian Kernel
Accuracy (%)	99.68	99.39
Sensitivity (%)	99.40	99.71
Specificity (%)	98.97	97.10
F-Score (%)	99.25	98.60
Precision (%)	99.10	97.52
Recall (%)	99.40	99.70
Training Time (Secs)	63.53	337

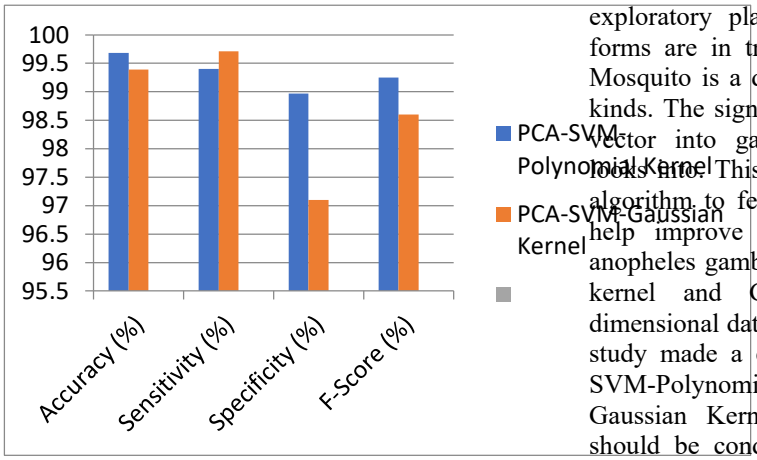

Figure 4: Performance Metrics Graph

The main finding is to analyze and improve the classification of malaria vector data, several works have been proposed in reviews by other researchers using the performance metrics shown in figure 4 above, this results have proven that, dimensionality reduction model using feature extraction methods such as PCA can help improve classification output such as SVM, yet SVM classifiers need the right kernels to help in projecting the capability of the output result.

5. DISCUSSION

The proposed approach improves and can be efficient for the prognosis and diagnosis of malaria ailment in human. The proposed approach uses machine learning techniques such as dimensionality reduction model and classification algorithms. Dimensionality reduction model uses the feature extraction model PCA and uses the SVM classifier kernels SVM-Polynomial and SVM Gaussian kernel. However, SVM-Polynomial kernel obtained a better accuracy performance. This study performed the analysis and evaluation of the performance and the results obtained were shown. In future works, feature selection algorithms and other feature extraction methods can be introduced for comparative evaluation and to show if there are other methods that can be used to better the classification performance compared to the-state-of-art.

6. CONCLUSION

In the past few years, remarkable works have been done on the innovation of RNA-seq, improvement as far as the execution measurements and productivity that are extraordinarily influenced by exploratory plan, activity and the data analysis forms are in trends to enhance the performance. Mosquito is a deadly insect comprising of various kinds. The significance of classification of malaria vector into gatherings has prompted numerous looks into. This study uses PCA feature extraction algorithm to fetch the latent components that can help improve the classification of a mosquito anopheles gambiae data by using SVM polynomial kernel and Gaussian kernel on a reduced dimensional data that employs PCA algorithm. This study made a comparative result that shows that SVM-Polynomial kernel outperforms SVM-Gaussian Kernel with 99.68%. Further studies should be conducted to improve performance of Machine Learning based methods by using more data and hybridized models.

7. ACKNOWLEDGEMENT

This work was funded by Landmark University, Omu-Aran Nigeria. I will like to thank my Co-authors and supervisors. We thank the School of Postgraduate Studies, Landmark University, the staff of the Department of Computer Science and College of Pure and Applied Sciences for their research support and enhancement program towards innovative contributions.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES:

[1] Aaron, T.L., Davis, J.M, and John, C.M. A Step-By-Step Workflow for Low-Level Analysis of Single-Cell RNA-Seq Data. F1000Research 1(5):1-62, 2016. https://doi.org/10.12688/f1000research.9501.2

[2] Ana, C., Pedro, M., Sonia, T., David, G., Alejandra, C., Andrew, M., Michał, W.S., Daniel, J.G., Laura, L.E., Xuegong, Z. and Ali M. A survey of best practices for RNA-seq data analysis. Genome Biology. 17(13): 1-19, 2016. DOI 10.1186/s13059-016-0881-8.

[3] Levin J.Z, Yassour M, Adiconis X, Nusbaum C, Thompson DA, and Friedman N. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. Volume 7, pages 709–715, 2010.

[4] Pierson, E., and Yau, C. ZIFA: Dimensionality Reduction for Zero-Inflated Single-Cell Gene Expression Analysis. Genome Biology. Volume 16. pages. 241-257, 2015.

[5] Dongfang, W., and Jin, G. VASC: Dimension Reduction and Visualization of Single-Cell RNA-Seq Data by Deep Variation Autoencoder. Genomics Proteomics Bioinformatics. 16(5): 320-331, 2018 Doi.org/10.1016/j.gpb.2018.08.03.

[6] Junhyong, K. Computational Analysis of RNA-Seq Data: From Quantification to High-Dimensional Analysis. University of Pennsylvania. pages 35-43, 2012.

[7] Bacher, R., and Kendziorski, C. Design and Computational Analysis of Single-Cell RNA-Seq Experiments. Genome Biology. 17(63) 2016.

[8] Chieh, L., Siddhartha, J., Hannah, K., and Ziv, B. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Research, 45(17): 1-11, 2017. doi: 10.1093/nar/gkx681.

[9] Esra, P., Hamparsum, B., and Sinan, Ç. A Novel Hybrid Dimension Reduction Technique for Undersized High Dimensional Gene Expression Data Sets Using Information Complexity Criterion for Cancer Classification. Computational and Mathematical Methods in Medicine. Volume 1, pages 1-14, 2015 http://dx.doi.org/10.1155/2015/370640

[10] Wenyan, Z., Xuewen, L., and Jingjing, Wu. Feature Selection for Cancer Classification Using Microarray Gene Expression Data. Biostatistics and Biometrics journals. 1(2):1-7, 2017.

[11] Balamurugan, M., Nancy, A., And Vijaykumar, S. Alzheimer's Disease Diagnosis by Using Dimensionality Reduction Based on KNN Classifier. *Biomedical & Pharmacology Journal* 10(4):1823-1830, 2017.

[12] Usman, A., Shazad, A., and Javed, F. Using PCA and Factor Analysis for Dimensionality Reduction of Bio-informatics Data. *(IJACSA) International Journal of Advanced Computer Science and Applications, 8(5): 515-426, 2017*

[13] Gökmen, Z., Dincer, G., Selcuk, K., Vahap, E., Gozde, E.Z., Izzet, P.D., Ahmet, O. A comprehensive simulation study on classification of RNASeq Data. PLoS One Journal, 12(8):1-24, 2017.

[14] Ian, T.J., and Jorge, C. Principal component analysis: a review and recent developments. Philosophical Transaction A Mathematical Physical Engineering Science. Volume 374, pages 1-21, 2017.

[15] Mariangela, B., Eric, O., William, A.D., Monica, B., Yaw, A., Guofa, Z., Joshua, H., Ming, L., Jiabao, X., Andrew, G., Joseph, F., and Guiyun, Y. 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 and Vector. 8(474):1-13, 2015. https://doi.org/10.1186/s13071-015-1083-z

[16] Bezanson, J., Karpinski, S., Shah, V., Edelman, A., (2012). A Fast-Dynamic Language for Technical Computing; Computational Engineering, Finance and Science. Computer Science Programming Language. 2012. arXiv:1209.5145

[17] Sofie, V. A comparative review of dimensionality reduction methods for high-throughput single-cell transcriptomics. Master's dissertation submitted to Ghent University to obtain the degree of Master of Science in Biochemistry and Biotechnology. Major Bioinformatics and Systems Biology. pages.1-88, 2017.

[18] Keerthi, K.V., Surendiran, B. Dimensionality reduction using Principal Component Analysis for network intrusion detection. Perspectives in Science. Volume 8, pages. 510—512, 2016.

[19] Elavarasan and Mani, K. A Survey on Feature Extraction Techniques. International Journal of Innovative Research in Computer and Communication Engineering. 3(1):1-4, 2015.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[20] Rimah, A., Dorra, B.A., and Noureddine, E. (2012). An Empirical Comparison of SVM and Some Supervised Learning Algorithms for Vowel recognition. International Journal of Intelligent Information Processing (IJIIP) 3(1):1-5, 2012.

[21] Aydadenta, H., and Adiwijaya. On the classification techniques in data mining for microarray data classification. International Conference on Data and Information Science, Journal of Physics: Conf. Series Volume 971. pages 1-10, 2018. doi :10.1088/1742-6596/971/1/012004

[22] Soofi, A.A., and Awan, A. (2017). Classification Techniques in. Machine Learning: Applications and Issues. Journal of Basic and Applied Sciences, Volume 13, pages 459-465, 2017

[23] Khan, A., Baharudin, B., Lee, L.H., khan, K. A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology, 1(1):1-17, 2010.

[24] Bhavsar, H., and Panchal, M.H. A Review on Support Vector Machine for Data Classification. International Journal of Advanced Research in Computer Engineering and Technology (IJARCET) 1(2):185-189, 2012.

[25] Devi, A.V, and Devaraj, D.V. Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection. Procedia Computer Science, Volume 47, pages 13 – 21, 2015.

[26] Chang C, Lin C. LIBSVM: A library for support vector machines. ACM TIST. 2(3):27, 2012.

[27] Nathan, T.J., Andi, D., Katelyn, J.H., and Dmitry, K. (2017). Biological classification with RNA-Seq data: Can alternative splicing enhance machine learning classifier?. bioRxiv
doi: http://dx.doi.org/10.1101/146340.

[28] Divya, J., and Vijendra, S. (2018). Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal. https://doi.org/10.1016/j.eij.2018.03.002

[29]https://figshare.com/articles/Additional_file_4_of_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/4346279/1

[30] Geng, G., Baitang, N., and Tieliu, S. Single Cell RNA-Seq Technologies and Related Computational Data Analysis. Frontiers in Genetics. 10(1); 317, 2019.. Doi.org/10.3389/fgen.2019.00317.

[31] Jiarui, D., Anne, C., and Sohrab, P.S. Interpretable Dimensionality Reduction of Single Cell Transcriptome Data with Deep Generative Models. Nature Research journal, Nature Communication. Volume 9, 2018. Doi:10.1038/s41467-018-04368-5.

[32] Malte, D.L., and Fabian, J.T. Current Best Practices in Single Cell RNA-Seq Analysis: Tutorial. Molecular System Biology. 15(6). Doi:10.15252/msb.20188746.

[33] Tamim, A., Lieke, M., Davy, C., Dylan, H., Hailaiang, M., Marcel, J.T.R., and Ahmed, M. A Comparison of Automated Cell Identification Methods for Single-Cell RNA Sequencing Data. Genome Biology. 20(194). Pages 1-14. 2019.

[34] Etienne, B., Leland, M., John, H. Dimensionality Reduction for Visualizing Single Cell Data Using UMAP. Nature Biotechnology. 37(1) 1-13. Doi: 10.1038/nbt.4314.

[35] Ren, Q., Anjun, M., Qin, M., and Quan, Z. (2019). Clustering and Classification Methods for Scingle Cell RNA-Seq Data. Briefings in Bioinformatics. Doi:10.1093/bib/bbz062.