International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 12 (2020), pp. 5012-5027
© International Research Publication House. http://www.irphouse.com

A Development of a Systematic in-Silico Analysis for RNA-Seq Anopheles Gambiae Data

[1]Marion Adebiyi, [2]Samuel Oladayo Olawepo, [3]Micheal Olaolu Arowolo and [4]Aderonke Anthonia Kayode

[1,2,3,4]Department of Computer Science, Landmark University, Nigeria.

*ORCIDs: 0000-0001-7713-956X (MARION), 0000-0001-8454-9241(OLAWEPO),
0000-0002-9418-5346 (AROWOLO), 0000-0001-5440-5546(ADERONKE)*

Abstract

In recent times, the Anopheles gambiae has developed resistant gene which have emanated as RNA-Seq data. The classification and evaluation of biological models is a relevant function of the RNA-Seq gene expression data. A significant aspect of data analytics is clustering. This research introduces an analysis on a reduced data using K-means clustering algorithm with Support Vector Machine classification method such as confusion matrix. Various computational performance metrics and accuracy such as F_1 score, Balanced accuracy, Matthew's correlation coefficient and Jaccard index. This study reveals the possible detection of insecticide resistance gene and usage in successful malaria illness diagnosis and treatment with an accuracy of 98%.

Keywords: RNA-Seq, K-Means, Clustering, SVM, Anopheles gambiae

I. INTRODUCTION

Diseases are medical conditions that negatively affect a living organism which are associated with signs and symptoms. The type of diseases varies from infectious, genetic and parasitic diseases such as malaria. Malaria is a severe communal medical worry in the sub-Saharan nations associated with high mortality and morbidity. The blood-borne pathogen called *Plasmodium falciparum* is referred to as the deadliest type of malaria caused by the female Anopheles specie mosquitoes which leads to the emergence of drug resistance [10]. The primary malaria vectors are the *Anopheles gambiae sensu strictu* (*Anopheles gambiae*) which emerges as distinct species [19]. *Anopheles gambiae* has being detected to have existence in both behavioural and physiological resistance forms. Bioinformatics develops with the concept of vast volumes of biological knowledge being processed, collected and evaluated. Bioinformatics is used for analytical and computational methods for In-silico study of biological applications with mathematical and statistical techniques such as transcriptomics. Transcriptomics is said to be the qualitative and quantitative study of Ribonucleic Acid (RNA) on a genome scale [13]. This can be said to be an approach to derive comprehensive wide dataset analyses of associated characteristics such as genes, proteins, DNA modifications among others. The emergence of transcriptomics has introduced the Ribonucleic Acid sequencing (RNA-Seq) gene analyses which possesses various tools to estimate the thousands of genes simultaneously by finding useful patterns. RNA-Seq provides an opportunity for discovery of novel transcripts of genes which is used to analyze the continuously changing cellular transcriptome [4]. RNA-Seq comprises of numerous single stranded short sequences which are individual elements immobilized on a solid support to represent the single gene. Gene expression shows thousands of expressions with different time points which displays correlation between two tissue samples like a diseased and normal sample. In-silico is more similar to natural modelling carried out through computer simulations which speed up the rate of gene expression discovery. The computer-based In-silico processes has been used to carry out analysis, interpretation and visualization of large dataset from various sources. Overtime, source of unknown variations in data is described by researchers and correlated with other established instances based on correlation of certain attributes within well-defined laws. The data mining area is used to collect valuable knowledge, find the concealed variations and similar parameters for a broad dataset. The types of data mining methods are decision trees, association, prediction, classification, clustering, sequential patterns among others. Data classification and data clustering are among the most important tools in data mining which enables researchers to identify specific information [37]. Classification of Data is a method of data mining used to identify predefined groups dependent on each object in a data collection. Datamining are supervised or unsupervised approach which deals with known and unknown domains. The known domains refer to supervised classification while the unknown domain refers to unsupervised classification. Clustering technique is a vital study area of datamining that is an unsupervised data mining method used to position individual resources of specific classes without previous information of specific class attributes to discover complexity in the results. Clustering is a primary and essential phase in data analysis aimed at transforming unlabeled dataset into discrete collection of data structures through studying the principle of classified data. Clusters identify hidden patterns which are automatically linked by learning the data pattern with clusters of similar patterns related to another cluster. In Bioinformatics, clustering is as important as it is in image processing, learning and pattern recognition, a bedrock of further research in this study thereby making use of the K-means clustering technique. Clustering algorithms remain a model based or distance based. An example of clustering is the K-means among other clustering which is an iterative refinement technique that is aimed at separating observations into clusters with the nearest mean used as the prototype of the cluster.

In this study, a clustering model using K-Means procedure is developed, SVM classifier is used to assess the quality of RNA-Sequencing data for malaria vector, Anopheles gambiae. This study develops an approach for deploying K-means algorithm to analyse the Anopheles gambiae RNA-Seq data with SVM classifier for evaluating and validating the quality metrics such as accuracy, specificity, sensitivity, precision, Recall, F-score and computational time of the genomic data.

II. RELATED WORK

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

Jelili, Ojeniyi and Obagbuwa in 2010, [16] worked on K-Means clustering analysis technique to examine data on academic performance of students. The K-Means clustering was involved in a deterministic statistical analysis with Euclidean distance to analyze student's performance. This paper reveals the predictive ability of clustering algorithms and statistical techniques. 79 student's results for nine courses offered by each student were used in a qualitative data analysis approach to test the differences of resemblance and the effects of numerical explanation for the success evaluation. Execution time focuses on the frequency and size of system. The proposed technique is not limited to model academic forecasts but also exist as an improved version for modelling. This study proposes fuzzy models by making use of two course dataset results to forecast students' academic patterns. An alternative method described is rough Set theory by using Rosetta toolkit to analyze student data. The purpose of using this toolkit is to assess data in relation to identifying association between the affecting factors and student grade.

Tajunisha and Saravanan in 2010, [32] used the Principal Component analysis (PCA) for dimensional dataset heuristics and reduction method which limit the volume of iterations of cluster distance estimation. The primary value of centroid relies on the effects of implementation of K-means algorithms. The application of PCA on microarray data before the clustering method application is to improve the precision of tests collected based on the general assumption that improvement entails the implementation of the centroid variables achieved similar to the ultimate solution by the suggested process. In this paper, the results obtained through k-means algorithm on a microarray data having centroids which are prepared arbitrarily and centroid generated with PCA. The time complexity of the technique was less than the average k-means execution time with random centroid initialization. Data reduction technique is advisable to achieve accuracy and efficiency in a large high dimensional dataset.

In this study Khalilian, Mustapha, Suliman and Mamat in 2010, [17] discusses a method that improve the K-Means algorithm performance, making use of divide and conquer strategy in relation to compatibility and equivalence concept. This study presents experimentation on a dataset which find suitable partner with similar personality for 8 pages of psychiatric personality assessment. The hidden patterns are extracted which reveals the property of K-Means algorithm as a qualified tool suitable for big sized data set and dimension yielding low quality clusters, whereas Hill Climbing algorithm is capable of constructing high quality attributes but possesses complexity issues making it unsuitable for huge datasets having high dimension data. Combination of both techniques allowed a subdivision based on certain criterion. The end results of the proposed application demonstrate remarkably improved efficiency and accuracy in cluster formation by creating structured clusters.

Velmurugan (2012), [35] worked on input data points based on arbitrary distribution to study the performance and quality of clustering procedures like K-Medoids and K-Means. Data type selected for processing is dependent on the clustering algorithm. The distribution of the arbitrary data points is clustered accordingly. Partition based algorithms have been observed to perform well to analyse small or standard datasets which identify cluster of spherical shape. The outcome of all the algorithms can be analysed according to an exact data point figure and the computing time needed for each algorithm. Analysis of time complexity involves the principle of computational complexity used to define in the best case the usage of computing resources by an algorithm and represented the worst-case execution period. Algorithm behaviour is analysed on the performance benchmark by evaluating the computation cycles an individual algorithm takes in processing the datasets. Quality evaluation focuses on the study with distance estimation among two points of data. In this study, it was discovered that the total k-Means execution period is comparatively lower than the k-Medoids algorithm. K-means is defined by all partition-based clustering methods as the optimal method for portioning dataset into set of groups as regards their various patterns. In K-means algorithm, the outcome is dependent on starting points referred as centroids. Dimensionality reduction is a major function for determining centroid. Several methods are introduced to improve k-means efficiency which shows near perfect accuracy.

In 2015, Pierson and Yau, [26] worked on a zero inflated single cell gene expression investigation, they built a dimensionality reduced technique, 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 regulation data providing strength compared to varied limitations. 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, [8] worked on an innovative fusion dimension reduction method used 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 with fusion levelled covariance estimators. It diminishes information of dimensionality thereby picking each quantity of probabilistic PCs (PPCs) being held, they presented and created observed consistent Akaike's information criterion (CAIC), Akaike's information criterion (AIC), and information theoretic measure of complexity (ICOMP) rule of Bozdogan. Openly accessible small 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 biomedical applications and bioinformatics.

In 2016, Wenyan, Xuewen and Jingjing, [36] worked on feature selection for cancer classification for disease utilizing microarray data expression. This paper used information on microarray genes which decides genes marker pertinent to a sort of malignancy. They researched a separation-based element choice strategy for two grouping issue. So as to choose the best marker genes, the Bhattacharyya separation can be actualized which quantify series of uniqueness in gene expression levels. The SVM classification was uses the best fetched marker genes. The execution of marker gene classification and selection can be represented in both recreation studies with double genuine information analyses by proposing an innovative gene selection technique used for SVM classification. The proposed scheme firstly ranked every gene according to importance of Bhattacharyya distances among indicated classes. The best subset genes can be chosen to accomplish the very least misclassification rate which occurs in developed SVMs preceded by a forward selection algorithm. 10-fold cross-validation is connected in locating various optimal parameters for SVM by concluding optimum subset gene. Subsequently, classification models are built and trained. Model classification is estimated by calculation for testing set. The execution of the projected B/SVM technique with SVM-RFE and SWKC/SVM gives normal misclassification amount (1.1%) and high normal retrieval rate (95.7%).

In 2017, Nancy and Vijay Kumar, [38] worked on Alzheimer's infection determination by utilizing dimensionality reduction based on KNN classification algorithm for analysing and classifying the Alzheimer malady and mild 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 investigator's information archive - Uniform Data Set. 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, [33] worked on dimensionality reduction approach using PCA and Factor Analysis for bioinformatics 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, [12] 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, overdispersion, 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 overscattering. 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, [6] used neural network algorithm which limit the dimensions of RNA-Seq single cell 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, a technique based on neural network (NN) was created and tested for recovery and investigation of a scRNA-Seq data. Different NN structures were displayed, some fuse prior biological learning which is used to originate a reduced dimension depiction of the data for a single cell expression. They demonstrate that NN technique enhances earlier strategies in the capacity to accurately cluster cells in analyses not used in the training and the capacity to effectively derive state or cell type 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 an approach for diminishing RNA-Seq dimensionality datasets, for increasing interpretability and preventive data bad luck, creating novel uncorrelated factors increasingly 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, [7] proposed a single cell RNA-

Sequential data by means of deep variational autoencoder using an unsupervised feature extraction model. The VASC models the failure and fetches the nonlinear ranked feature representation of high dimensional information. There result was tested on about 20 datasets, the VASC showed a better performance with broader compatibility features.

In 2018, Etienne, Leland and John, [9] worked on the dimensionality reduction model used for imagining single-cell data by means of UMAP. A uniform manifold approximation, nonlinear dimensionality-reduction method and prognosis (UMAP), was established for the investigation of any kind of high-dimensional data. UMAP is applied to biological data, which involves the use of three well-characterized mass cytometry and single-cell RNA sequencing datasets. Associating five other tools with the act of UMAP, UMAP was revealed to provide the fastest execution time with highest reproducibility and the most meaningful organization of cell clusters. In this work, the usage of UMAP was highlighted which improves interpretation and visualization of single-cell data.

In 2018, Jiarui, Anne and Sohrab, [15] worked on an interpretable dimensionality reduction model of single cell transcriptome data with broad 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.

In 2019, Geng, Baitang and Tieliu, [11] proposed an ScRNA-Seq technologies and its relating computational analysis, in this review, there was 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, [22] worked on the current practices in ScRNA-Seq analysis, by formulating present best- practice endorsements for steps based on self-determining comparison studies. The best-practice references were combined 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.

In 2019, Tamim, Lieke, Davy, Dylan, Hailiang, Marcel, and Ahmed, [31] worked on a comparative analysis of automatic cell identification methods for scRNA-Seq data, they benchmarked 22 classification approaches that repeatedly allot cell characteristics that includes general-purpose classifiers and single-cell-specific. The approaches were presented to have being assessed using 27 openly accessible single-cell RNA sequencing datasets of diverse models, species, knowledges, and levels of intricacy. Two experimental survey were employed to predict the performance of each approach for within dataset predictions (intra-dataset) and across datasets (inter-dataset) based on accuracy, percentage of unclassified cells, and time of computation. The method's sensitivity is evaluated with input characteristics, number of cells per population, and their accuracy through various specification rates and datasets. For large datasets with multiple groups or deep annotations, the major classifiers work better on a number of datasets which involves reduced accuracy. Overall, the widely used support vector machine classifier possess the highest output over the various derived experiments.

In 2019, Ren, Anjun, Qin and Quan, [27] worked on the clustering and classification models for ScRNA-Seq data. This paper reflects on a systematical reassessment of integrated tools and methods, which highlights the pros and cons of each approach. Close attention was ascribed to clustering and classification methods as well as discussing the various approach that have being discovered in recent times as powerful alternatives, which includes linear and nonlinear methods with descending dimension methods. In conclusion, the emphasis was on classification and clustering approach for scRNA-seq data, in particular, integrated methods, that provide a systematic illustration of scRNA-seq data and download URLs.

A substantial implementation of RNA-Seq data for gene expression prediction and classification of biological models data analysis was proposed [24], and reveals a comparative analysis on a concentrated data PCA dimensional method by analysing the presentation characteristics of classification using SVM kernels such as the SVM-Polynomial kernels and SVM-Gaussian kernels. This work uses PCA feature extraction algorithm to derive the latent elements that can help improve the classification of a mosquito anopheles gambiae data by making use of SVM polynomial kernel and Gaussian kernel on a reduced dimensional data that employs PCA algorithm.

III. MATERIALS AND METHODS

The proposed approach for this study analysed the dataset used and the framework which is discussed below.

III.I Dataset Used for Analysis

The RNA-Seq data was used to study the variations in deltamethrin-resistant as well as susceptible Anopheles gambiae mosquitoes of western Kenya, a freely accessible dataset from figshare.com and sponsored by the National Health Institute [23]. The dataset comprises of various list of genes such as; AGAP002724, AGAP009472, AGAP004779, AGAP012984, AGAP003714, CPLC G3 [AGAP008446], CYP6P3 [AGAP002865] and CYP6M2 [AGAP008212], mosquito genes from western Kenya in the year 2010 and 2012 with the dataset comprising of 2457 genes and 2 cells.

MATLAB is a mathematical rendering environment with multi-worldview and a free programming language implemented by MathWorks. It makes frame controls, feature and knowledge mapping, algorithm execution, user interface development, written in different languages, such as; C, C++, C#, Java, Fortran and Python [2]. MATLAB is a versatile numerical and visualization framework which can be built, developed and applied to solve fairly complex science and engineering problems. MATLAB R2019b is used exclusively for the implementation of this work.

III.II Experimental Methodology

This study summarizes the proposed framework in figure 1 below. The major interest is to predict a machine learning process with an effective clustering algorithm such as K-Means which data input is further analyzed with the training and then classified by the classification module using the Support Vector Machine (SVM) which reveals the performance and optimization of classification results.

The detailed methodology in this study is as follows:

i) Use K-Means to cluster the mosquito Anopheles gambiae data.

ii) Use Support Vector Machine as a classifier on the data.

iii) Analyse the results in terms of accuracy, specificity, sensitivity, precision and computational time performance metrics.

iv) Model with K-Means cluster, followed by classifying the clustered data using SVM classifier.

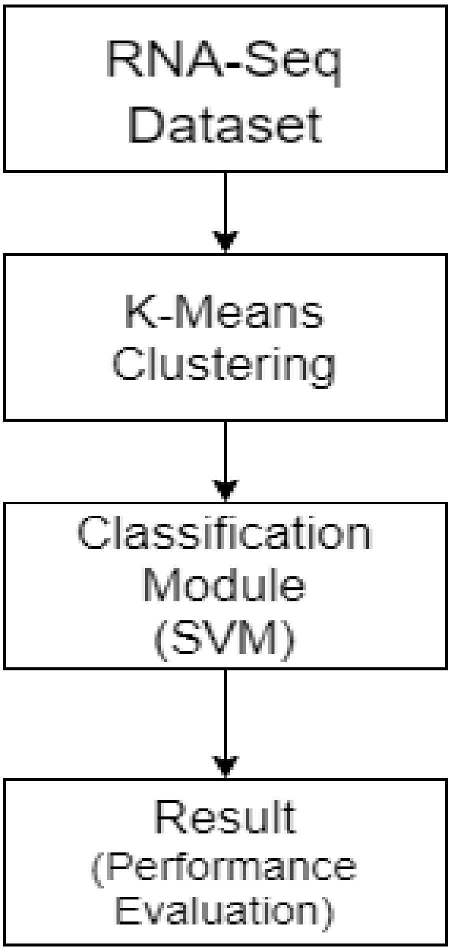


Figure 1: Proposed Framework.

This study reflects on analysis and interpretation of data using data mining techniques and MATLAB tools. The data derived was compiled in Microsoft Excel 2013 software. The MATLAB software was used for data clustering technique after classification technique was applied to study and illustrate how well the clusters in gene resistance and susceptible are revealed and analysed. Application and clustering method with classification is used as benchmark datasets i.e. the RNA-Seq Anopheles gambiae dataset is performed. The process reveals the important information in the dataset which highlights the instances of resistance and susceptible genes. The developed model for this system was implemented and developed on MATLAB R2019b which is a Fourth generational programming language with object oriented based procedures. The *Anopheles gambiae* dataset was experimented with 13362 instances obtained from malaria patients with 2457 gene expression levels.

III.III Clustering

Clustering organizes related objects into similarity based groups of different density with shapes. Clustering is referred to as a data mining method that creates similar characteristics automatically from suitable cluster of objects (Muhammad, 2015). Clustering methodology is used without previous awareness of the distinct category properties and composition in the data to position individual objects in specific categories. In data analysis, clustering is considered as the essential and vital phase. Clustering technique is unsupervised which identifies a cluster as members inside a similar object of cluster known as homogeneity and dissimilar object of cluster known as heterogeneity. Clustering is used in various entities like web mining, market research, Geo-informatics, Image processing, pattern recognition and Bioinformatics. In this work, the K-Means clustering algorithm which is used to analyse the Microarray dataset so as to eliminate the redundant attributes from the dataset. When the redundant attributes are removed, the dataset is converted to the classifier system which is used for training and testing data.

III.III.I K-Means

K-Means is an iterative technique used to partition objects within a cluster into a specified number of clusters, k. K-Means algorithm can be identified as a centroid-based algorithm which is based on central vector concept which is relative to the distance between objects. The results of K-Means algorithms depend on the initial points called centroids [18]. In optimization, the centroid-based clustering are computationally intensive which results to approximate solutions which are generated by local optimization. There has been motivation for design and implementation of wide variety of clustering techniques based on different fundamental algorithms to provide a systematic and comprehensive evaluation of performance [39]. MicroRNA, mRNA and large scale protein expression data are used to give precise analysis of gene and identification of regulatory mechanisms [29]. The time dependencies between elements makes the model very effective for solving problems related with time series data. In this work, K-means will be implemented upon to cluster the mRNA data set with an SVM classifier.

N: *number of data objects*

K: *number of clusters*

objects[N]: *array of data objects*

clusters[K]: *array of cluster centers*

membership[N]: *array of object memberships*

```
while $\delta$/N > threshold
    $\delta \leftarrow 0$
    for i ← 0 to N-1
        for j ← 0 to K-1
            distance ← | objects[i] - clusters[j] |
            if distance < $d_{min}$
                $d_{min}$ ← distance
                n ← j
        if membership[i] ≠ n
            $\delta \leftarrow \delta + 1$
            membership[i] ← n
        new_clusters[n] ← new_clusters[n] + objects[i]
        new_cluster_size[n] ← new_cluster_size[n] + 1
        for j ← 0 to K-1
            clusters[j][*] ← new_clusters[j][*] / new_cluster_size[j]
            new_clusters[j][*] ← 0
            new_cluster_size[j] ← 0
```

III.IV Classification

A final performance is dependent on the influence of the classification model being used. The different types of classifier also vary with performance on any given dataset. The classification module uses the concept of Support Vector Machine (SVM) for learning method that can effectively deal with large datasets. SVM has been proven in recent times to be the best and relatively accurate classifier.

III.IV. I Support Vector Machine (SVM)

The evolution of support vector machine which is simply known as a tool built for binary classification by Vapnik and colleagues at Bell laboratories [5], [39] with algorithm improvements by others. The evolution of Support Vector Machine (SVMs) can be said to be new techniques which have acquired immense popularity owing to satisfactory performance in a wide range of machine learning problems with strong theoretical foundations in statistical learning theory [20], [28]. The SVM is a supervised learning system, which produces input-output mapping functions through a collection of labelled training data. The SVM are used to solve separable linear as well as separable non-linear problems. SVM is more robust with different dataset which makes it more efficient for training with both the supervised and unsupervised learning [40]. In general, the primary function of SVM is to maximize the marginal distance to the hyperplane and to enhance the distinct feature between the two dataset categories.

$$W_0 + \mathbf{W}^T \mathbf{X}_{pos} = 1 \quad (1)$$

$$W_0 + \mathbf{W}^T \mathbf{X}_{neg} = -1 \quad (2)$$

If w0 + wTxtest > 1, the sample xtest is considered toward the right of the positive hyperplane.

If w0 + wTxtest < -1, the sample xtest is considered toward the left of the negative hyperplane.

For subtraction,

$$\mathbf{W}^T (\mathbf{X}_{pos} - \mathbf{X}_{neg}) = 2 \quad (3)$$

As stated by the adoption of [1], the process of locating the best hyperplane while utilizing the SVM is listed below:

$$\text{Let } y_t \in \{y_1, y_2, \ldots, y_n\}, \quad (4)$$

where y_t is the *p* attributes and target class $E_t \in \{+1, -1\}$

Assume the classes +1 and -1 can be separated completely by hyperplane.

$$v.y + c = 0 \quad (5)$$

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

$$v.y + c \geq +1, \quad \text{for class } +1 \quad (6)$$

$$v.b + c \leq -1, \quad \text{for class } -1 \quad (7)$$

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 [24]. Expansion of margins is a quadratic programming problem which is done by calculating the average point. SVM has the advantage of being able to manage a variety of classification issues in high dimensional data [30].

SVM is excellent in comparison with other classification methods, with its outstanding classification applicability [18]. SVM is grouped into linear and non-linear separable. SVM has kernel functions which change data into a higher dimensional space that makes it possible to accomplish separations. Kernel functions are a collection of algorithms used to identify or analyse data. Training vectors *xi* is mapped into higher dimensional space by the capacity Φ. In the higher dimension space, SVM reveals a linear separating hyperplane with the maximum. 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 Kernel's decision relies upon the current problem to be solved, since it relies upon what models are to be analysed, a couple of kernel functions in a large assortment of applications [3]. The prescribed kernel function for this study is the SVM-Polynomial Kernel and Gaussian Kernel.

III.V Performance Evaluation Metrics

The performance evaluation metrics of classifier is evaluated in terms of classification time, accuracy, specificity, sensitivity, precision, f score, balanced accuracy, Matthew's correlation coefficients and Jaccard index. The terms below are adopted from [37].

Accuracy = (TP + TN) / (TP + TN + FP + FN) % (8)

Specificity = TN / (TN + FN) % (9)

Sensitivity = TP / (TP + FN) % (10)

Precision = TP / (TP + FP) (11)

F_1 score = 2TP / (2TP + FP +FN) (12)

BAcc = 1 / 2 (TP / (TP + FN) + TN / (TN + FP)) (13)

MCC = (TP.TN – FP. FN) / $\sqrt{PP.PN.RP.RN}$ (14)

J = TP / (TP + FP +FN) = TP / n - TN' (15)

Where:

TP (True Positives) = correctly classified positive cases,

TN (True Negatives) = correctly classified negative cases,

FP (False Positives) = incorrectly classified negative cases,

FN (False Negatives) = incorrectly classified positive cases.

RP (Real Positives) = correctly classified positive cases,

RN (Real Negatives) = incorrectly classified negative cases,

PP (Predictive Positives) = correctly classified positive cases,

PN (Predictive Negatives) = incorrectly classified negative cases,

Accuracy is the probability that a diagnosing test is performed correctly.

Specificity (true negative fraction) is the probability of a diagnosing test being negative, provided that the individual has no disease.

Sensitivity (true positive fraction) is the probability of a diagnosing test being positive, provided that the individual has the disease.

F_1 score is referred as the harmonic mean of Positive Predictive Value and sensitivity.

Balanced Accuracy overcomes most bias issues within an imbalanced learning dataset, which is an average of sensitivity and specificity.

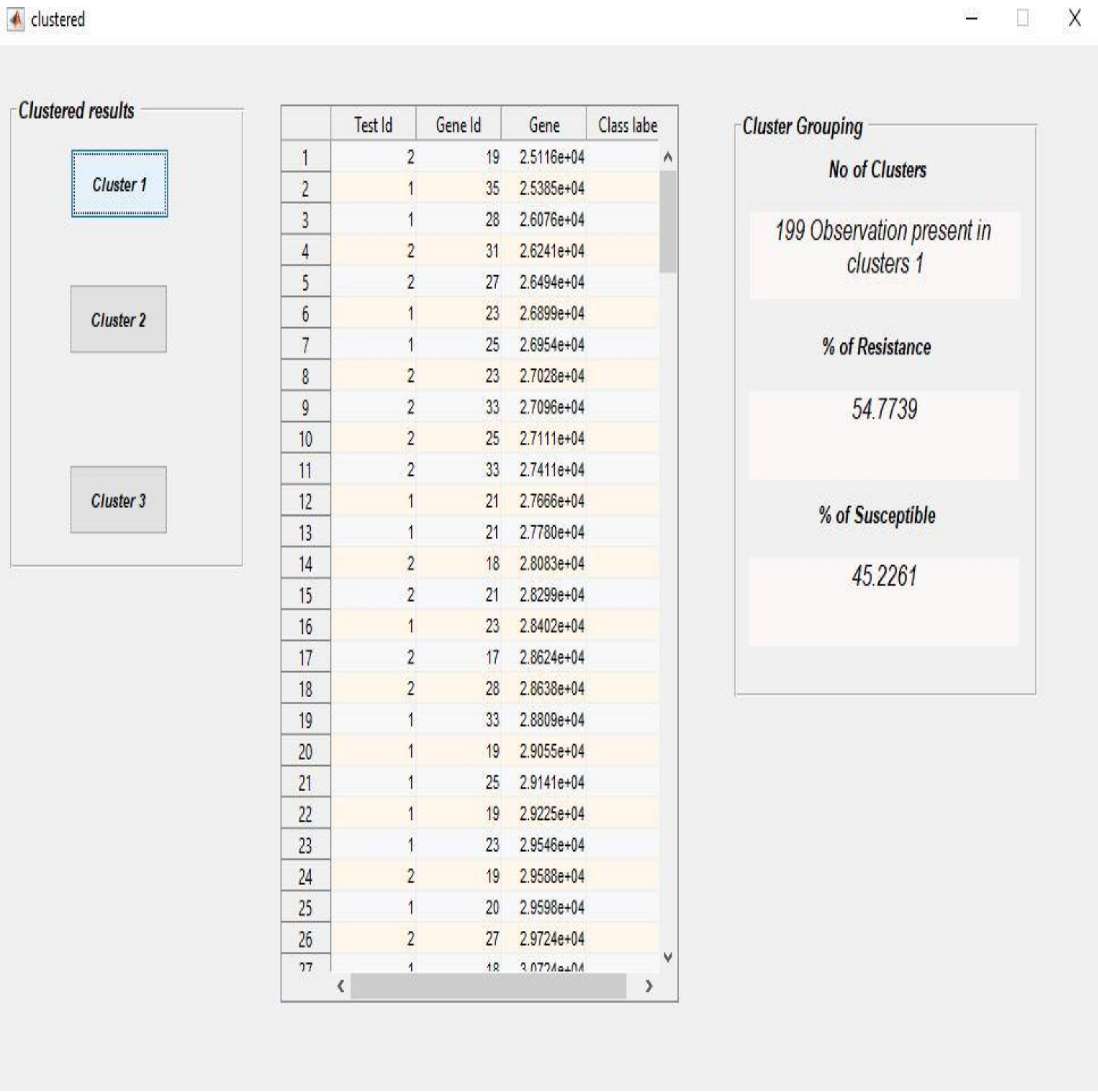
Matthew's Correlation Coefficients is applied in evaluating predictors with imbalanced dataset.

Jaccard index displays the rate of true positives between all the samples, either positive predictions or real positives.

IV. RESULT

This study shows RNA-Seq data which contains 2457 samples of *Anopheles Gambiae* mosquitoes which comprises of both resistant and susceptible genes. K-Means algorithm was executed on the data to increase the classification of biological useful classify clusters by using it for computational analysis. K-Means helps in time reduction and storage for relevant clusters to eliminate redundancy. This study reveals the application of K-Means on the generated data, to identify maximum variance and derive an enrich cluster which will code as a potential insecticidal target. Classification algorithm applies SVM Polynomial kernel and Gaussian kernel which is utilized by MATLAB to implement the model. Using K-Means as a clustering method, 300 out of 2457 genes were significant and achieved in 4.852 seconds. In bioinformatics, the SVM kernel classifier methods is generally recognized in machine learning approaches, 8-folds cross validation was used to assess performance of the execution, using 15 parameter holdout value for training and testing the accuracy of the classifiers. To every classifier, a supervised learning protocol is utilized using MATLAB. This report results from the computational time and performance evaluation such as Accuracy, Specificity, Sensitivity, Precision, F-score, Balanced Accuracy, Matthew's correlation coefficient and Jaccard index). This study assesses the performance of model classifications. The MATLAB environment is used for implementation of the RNA-Seq data. In recent times, MATLAB has been proven to be powerful and efficient with machine learning data analysis with good experimental metrics. The MATLAB R2019b was used for the analysis and executed on windows 10 Operating System. The default view of MATLAB has a blank screen with command examples and comments. The dataset used is selected from the left-hand-corner where the K-Means folder is picked. The dataset is in xlsx format which is loaded into the GUI. The GUI comprises of series of codes which the dataset will be analysed on. The main.m command is executed after the dataset is preloaded successfully.

Figure 2 shows the cluster 1 result analysis with 199 observation present, 54.7739% of resistance, 45.2261% of susceptible gene from the data loaded.



	Test Id	Gene Id	Gene	Class labe
1	2	19	2.5116e+04	
2	1	35	2.5385e+04	
3	1	28	2.6076e+04	
4	2	31	2.6241e+04	
5	2	27	2.6494e+04	
6	1	23	2.6899e+04	
7	1	25	2.6954e+04	
8	2	23	2.7028e+04	
9	2	33	2.7096e+04	
10	2	25	2.7111e+04	
11	2	33	2.7411e+04	
12	1	21	2.7666e+04	
13	1	21	2.7780e+04	
14	2	18	2.8083e+04	
15	2	21	2.8299e+04	
16	1	23	2.8402e+04	
17	2	17	2.8624e+04	
18	2	28	2.8638e+04	
19	1	33	2.8809e+04	
20	1	19	2.9055e+04	
21	1	25	2.9141e+04	
22	1	19	2.9225e+04	
23	1	23	2.9546e+04	
24	2	19	2.9588e+04	
25	1	20	2.9598e+04	
26	2	27	2.9724e+04	
27	1	18	3.0724e+04	

No of Clusters

199 Observation present in clusters 1

% of Resistance

54.7739

% of Susceptible

45.2261

Figure 2: Cluster 1 Loaded and Analysed Clustered Result

199 genes were clustered

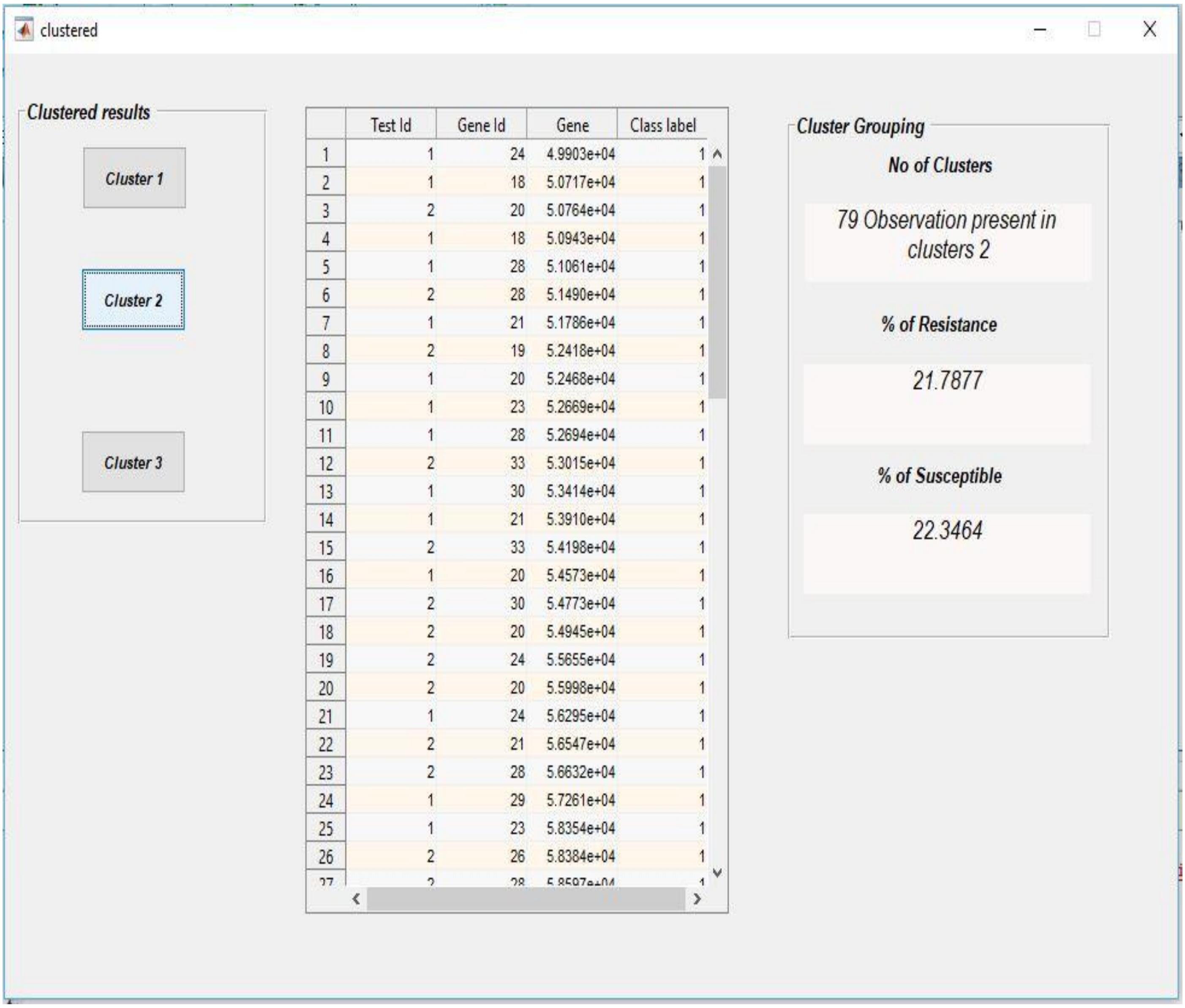
54.7 % = 107.46 resistance

45.2 % = 89.98 susceptible

Resistance + Susceptible = 108 + 90 = 198 genes

1 gene is ungrouped

Figure 3 shows the cluster 2 result analysis with 79 observation present, 21.7877% of resistance, 22.364% of susceptible genes from the data loaded.



	Test Id	Gene Id	Gene	Class label
1	1	24	4.9903e+04	1
2	1	18	5.0717e+04	1
3	2	20	5.0764e+04	1
4	1	18	5.0943e+04	1
5	1	28	5.1061e+04	1
6	2	28	5.1490e+04	1
7	1	21	5.1786e+04	1
8	2	19	5.2418e+04	1
9	1	20	5.2468e+04	1
10	1	23	5.2669e+04	1
11	1	28	5.2694e+04	1
12	2	33	5.3015e+04	1
13	1	30	5.3414e+04	1
14	1	21	5.3910e+04	1
15	2	33	5.4198e+04	1
16	1	20	5.4573e+04	1
17	2	30	5.4773e+04	1
18	2	20	5.4945e+04	1
19	2	24	5.5655e+04	1
20	2	20	5.5998e+04	1
21	1	24	5.6295e+04	1
22	2	21	5.6547e+04	1
23	2	28	5.6632e+04	1
24	1	29	5.7261e+04	1
25	1	23	5.8354e+04	1
26	2	26	5.8384e+04	1
27	2	28	5.8597e+04	1

Cluster Grouping

No of Clusters

79 Observation present in clusters 2

% of Resistance

21.7877

% of Susceptible

22.3464

Figure 3: Cluster 2 Loaded and Analysed Clustered Result

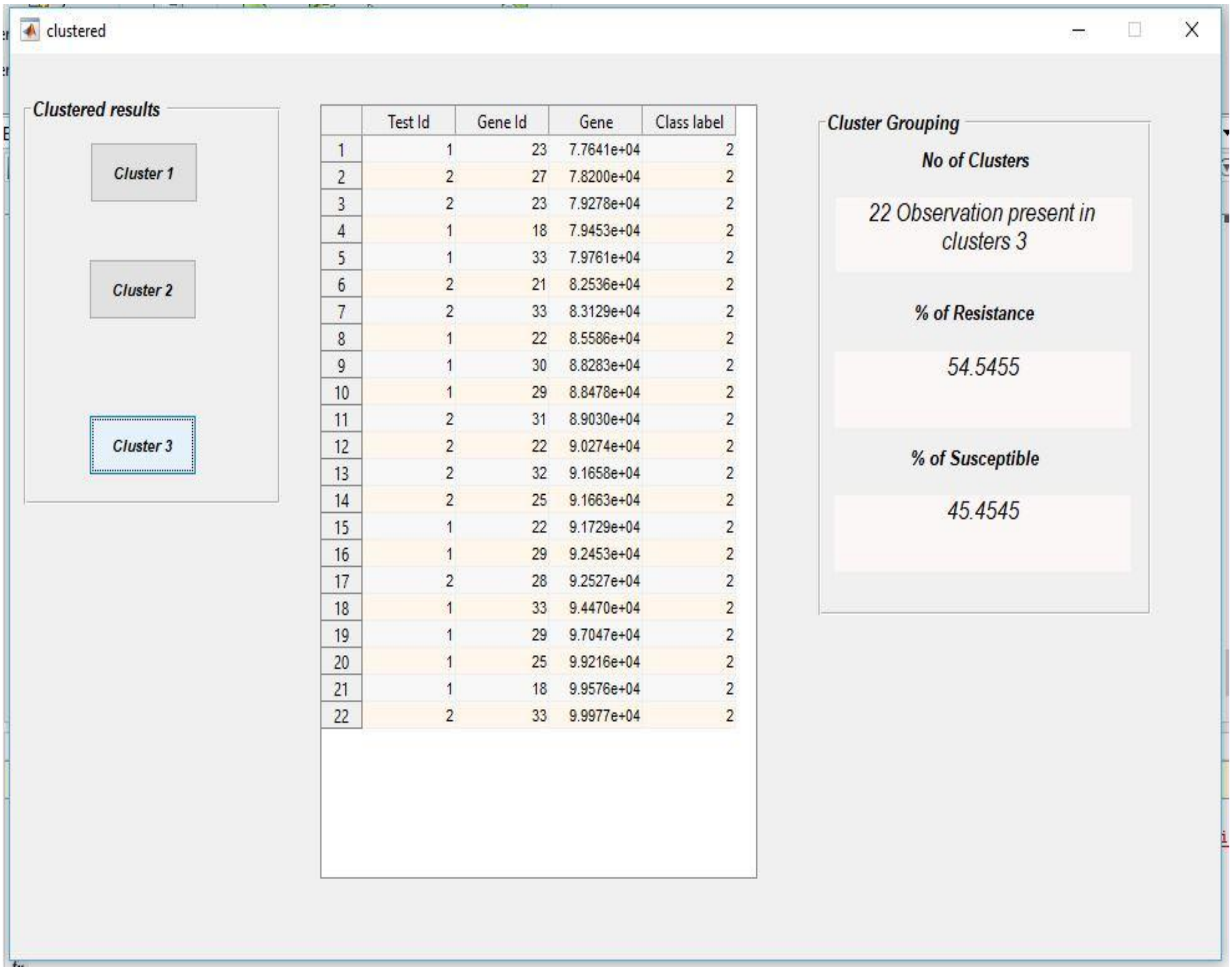
79 genes were clustered

21.78 %	=	17.2	resistance
22.34 %	=	17.6	susceptible

Resistance + Susceptible = 17 + 18 = 35 genes

44 genes ungrouped

Figure 4 shows the cluster 3 result analysis with 22 observation present, 54.5455% of resistance, 45.4545% of susceptible genes from the data loaded.



	Test Id	Gene Id	Gene	Class label
1	1	23	7.7641e+04	2
2	2	27	7.8200e+04	2
3	2	23	7.9278e+04	2
4	1	18	7.9453e+04	2
5	1	33	7.9761e+04	2
6	2	21	8.2536e+04	2
7	2	33	8.3129e+04	2
8	1	22	8.5586e+04	2
9	1	30	8.8283e+04	2
10	1	29	8.8478e+04	2
11	2	31	8.9030e+04	2
12	2	22	9.0274e+04	2
13	2	32	9.1658e+04	2
14	2	25	9.1663e+04	2
15	1	22	9.1729e+04	2
16	1	29	9.2453e+04	2
17	2	28	9.2527e+04	2
18	1	33	9.4470e+04	2
19	1	29	9.7047e+04	2
20	1	25	9.9216e+04	2
21	1	18	9.9576e+04	2
22	2	33	9.9977e+04	2

Cluster Grouping

No of Clusters

22 Observation present in clusters 3

% of Resistance

54.5455

% of Susceptible

45.4545

Figure 4: Cluster 3 Loaded and Analysed Clustered Result

22 genes were clustered

54.54 %	=	11.99	resistance
45.45 %	=	9.99	susceptible

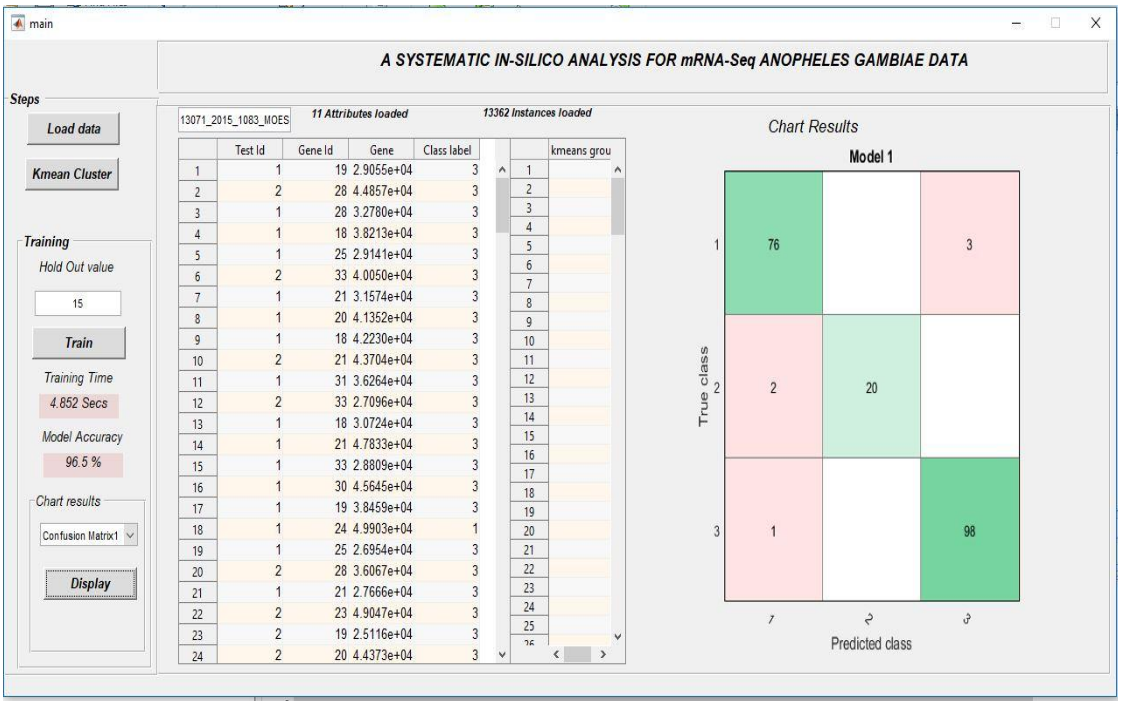
Resistance + Susceptible = 12 + 10 = 22 genes

0 gene is ungrouped

4.2 Classification

Confusion Matrix 1

The trained datasets are further classified with the Support Vector Machine (SVM) which obtained 96.5% classification accuracy. The figure 5 shows the chart result in confusion matrix 1. The confusion matrix obtained a training time of 4.852 secs and Model 1 with true class and predicted class.



A SYSTEMATIC IN-SILICO ANALYSIS FOR mRNA-Seq ANOPHELES GAMBIAE DATA

Steps

Training

Hold Out value: 15

Training Time: 4.852 Secs

Model Accuracy: 96.5 %

Chart results: Confusion Matrix1

13071_2015_1083_MOES

11 Attributes loaded

13362 Instances loaded

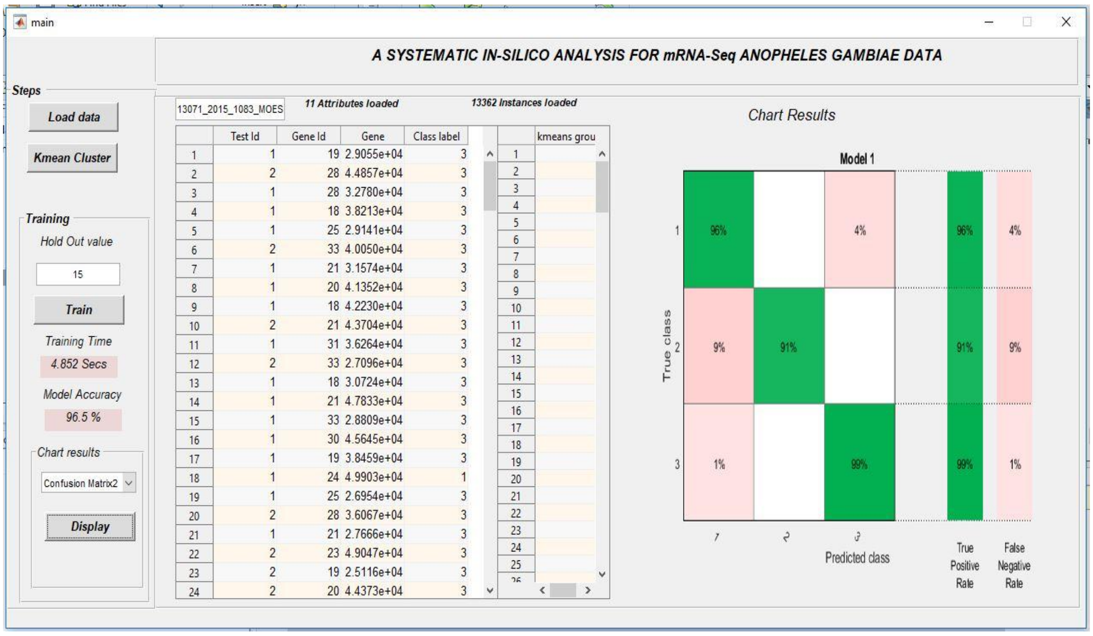
	Test Id	Gene Id	Gene	Class label
1	1	19	2.9055e+04	3
2	2	28	4.4857e+04	3
3	1	28	3.2780e+04	3
4	1	18	3.8213e+04	3
5	1	25	2.9141e+04	3
6	2	33	4.0050e+04	3
7	1	21	3.1574e+04	3
8	1	20	4.1352e+04	3
9	1	18	4.2230e+04	3
10	2	21	4.3704e+04	3
11	1	31	3.6264e+04	3
12	2	33	2.7096e+04	3
13	1	18	3.0724e+04	3
14	1	21	4.7833e+04	3
15	1	33	2.8809e+04	3
16	1	30	4.5645e+04	3
17	1	19	3.8459e+04	3
18	1	24	4.9903e+04	1
19	1	25	2.6954e+04	3
20	2	28	3.6067e+04	3
21	1	21	2.7666e+04	3
22	2	23	4.9047e+04	3
23	2	19	2.5116e+04	3
24	2	20	4.4373e+04	3

	kmeans grou
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	

Figure 5: Confusion Matrix 1 chart result

Confusion Matrix 2

Figure 6 shows the chart result in confusion matrix 2 which took a total time of 4.852 secs for Model 1 with True Positive Rate yields 96% and False Negative Rate yields 1%.



A SYSTEMATIC IN-SILICO ANALYSIS FOR mRNA-Seq ANOPHELES GAMBIAE DATA

Steps

Training

Hold Out value: 15

Training Time: 4.852 Secs

Model Accuracy: 96.5 %

Chart results: Confusion Matrix2

13071_2015_1083_MOES

11 Attributes loaded

13362 Instances loaded

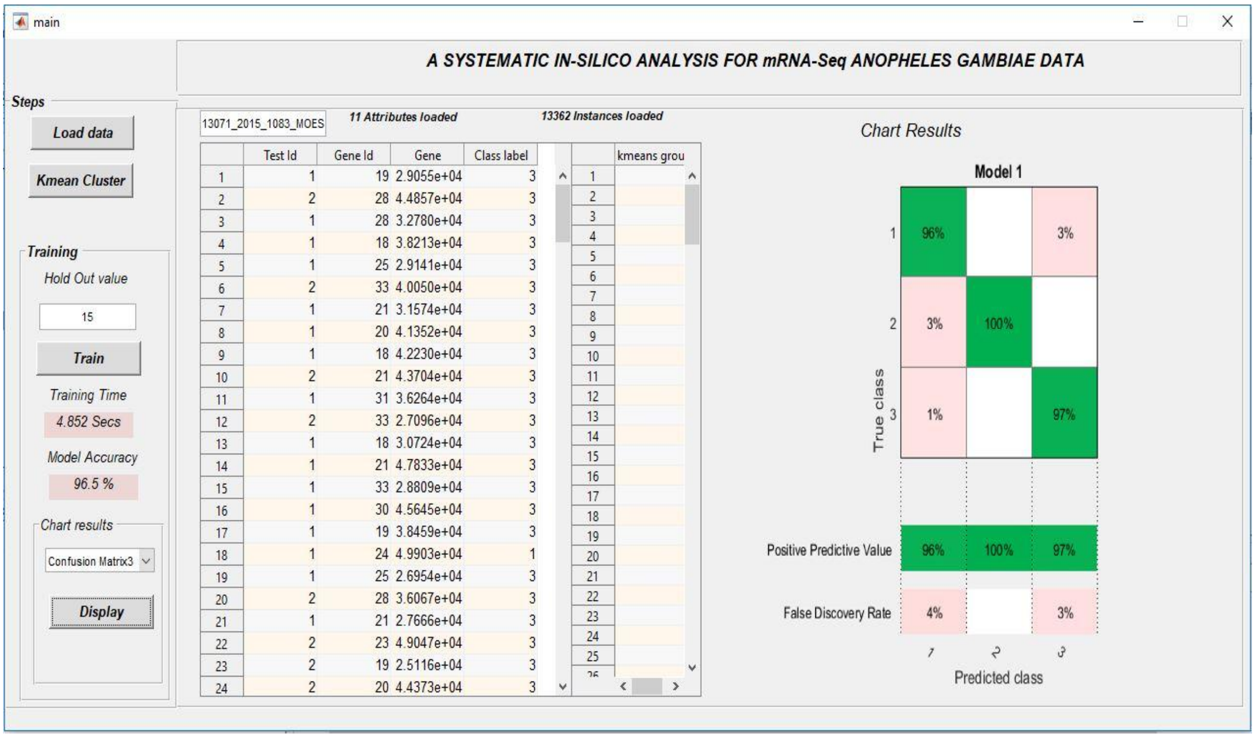
	Test Id	Gene Id	Gene	Class label
1	1	19	2.9055e+04	3
2	2	28	4.4857e+04	3
3	1	28	3.2780e+04	3
4	1	18	3.8213e+04	3
5	1	25	2.9141e+04	3
6	2	33	4.0050e+04	3
7	1	21	3.1574e+04	3
8	1	20	4.1352e+04	3
9	1	18	4.2230e+04	3
10	2	21	4.3704e+04	3
11	1	31	3.6264e+04	3
12	2	33	2.7096e+04	3
13	1	18	3.0724e+04	3
14	1	21	4.7833e+04	3
15	1	33	2.8809e+04	3
16	1	30	4.5645e+04	3
17	1	19	3.8459e+04	3
18	1	24	4.9903e+04	1
19	1	25	2.6954e+04	3
20	2	28	3.6067e+04	3
21	1	21	2.7666e+04	3
22	2	23	4.9047e+04	3
23	2	19	2.5116e+04	3
24	2	20	4.4373e+04	3

	kmeans grou
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	

Figure 6: Confusion Matrix 2 chart result

Confusion Matrix 3

Figure 7 shows the chart result in confusion matrix 3 which took a total time of 4.852 secs for Model 1 with Model Accuracy 96.5% with Positive Predictive Value of 96% and False Discovery Rate yields 3%.



13071_2015_1083_MOES

11 Attributes loaded

13362 Instances loaded

Hold Out value: 15

Training Time: 4.852 Secs

Model Accuracy: 96.5 %

Chart results: Confusion Matrix3

	Test Id	Gene Id	Gene	Class label
1	1	19	2.9055e+04	3
2	2	28	4.4857e+04	3
3	1	28	3.2780e+04	3
4	1	18	3.8213e+04	3
5	1	25	2.9141e+04	3
6	2	33	4.0050e+04	3
7	1	21	3.1574e+04	3
8	1	20	4.1352e+04	3
9	1	18	4.2230e+04	3
10	2	21	4.3704e+04	3
11	1	31	3.6264e+04	3
12	2	33	2.7096e+04	3
13	1	18	3.0724e+04	3
14	1	21	4.7833e+04	3
15	1	33	2.8809e+04	3
16	1	30	4.5645e+04	3
17	1	19	3.8459e+04	3
18	1	24	4.9903e+04	1
19	1	25	2.6954e+04	3
20	2	28	3.6067e+04	3
21	1	21	2.7666e+04	3
22	2	23	4.9047e+04	3
23	2	19	2.5116e+04	3
24	2	20	4.4373e+04	3

	kmeans grou
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	

Figure 7: Confusion Matrix 3 chart result

The performance analysis of classification using Support Vector Machine on RNA-Seq *anopheles gambiae* dataset derived 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_candidateresistance_SNPs/4346279/1

2457 gene features were obtained, the K-Means method was used which achieves higher value with the dataset on optimization parameters such as accuracy, sensitivity, specificity, f-score, balanced accuracy, matthew's correlation coefficient, jaccard index, timing and prediction. When the dataset is clustered, by application of K-Means algorithm, some valuable data are discovered and the algorithm accuracy is increased by removing redundant data. The clustering method using the K-Means algorithm for RNA-Seq data plays an essential role, it optimizes the performance of clustering which is further enhanced by the SVM classification performance algorithm in terms of accuracy, sensitivity, specificity, f-score and precision. The result displays the quality of machine-learning innovation in genes. For further validation, the results of the performance will be displayed and compared in the table 1 below.

Table 1: Execution Results Table

Performance Metrics	Cluster 1	Cluster 2	Cluster 3
Accuracy (%)	97.8	97	98
Specificity (%)	96.2	96	99
Sensitivity (%)	99	99	99
Precision (%)	97	96	97
F-Score (%)	98	98	98
Balanced Accuracy (%)	97	98	98
Matthew's correlation coefficient (%)	95	95	96
Jaccard Index (%)	96	95	96
Training Time (Secs)	4.852	4.852	4.852

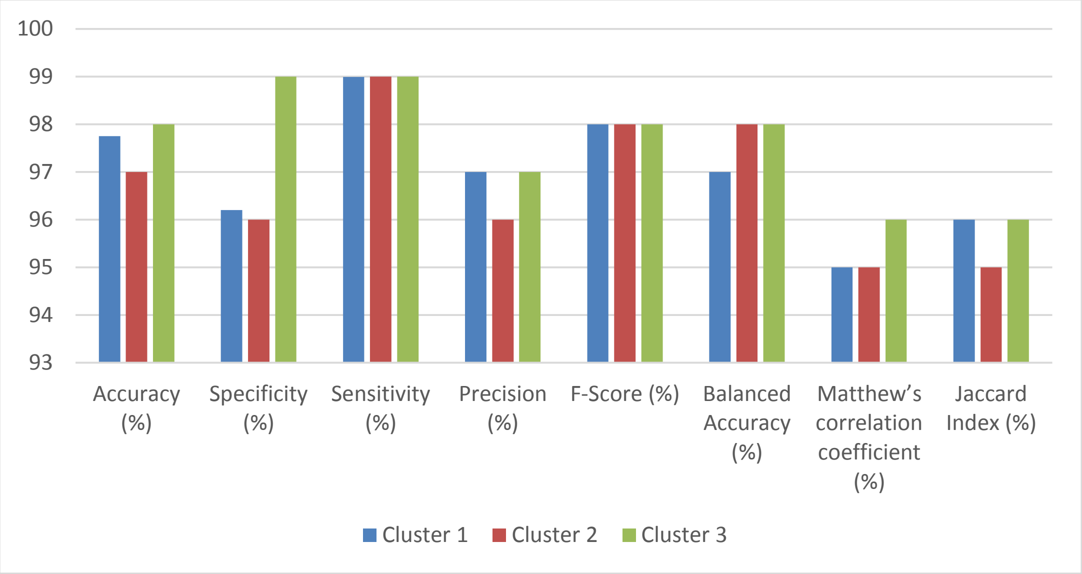


Figure 8: Performance Metrics Graph

The importance of the study is to evaluate and enhance the classification of malaria dataset for diagnosis and treatment plan of malaria ailments in individuals. There are various techniques that has been proposed over time by variety of researchers using the quality metrics shown in figure 8 above, the results reveal that clustering method such as the K-Means is very effective and efficient in classification output such as SVM. In Figure 9 below, it shows the scattered plot for clustered performance metrics which is extracted using the K-Means technique. A generated scattered plot of input data of a SVM is suitable for model classification which highlights the support vectors, maximum margin of hyperplanes within the three-classes in the dataset.

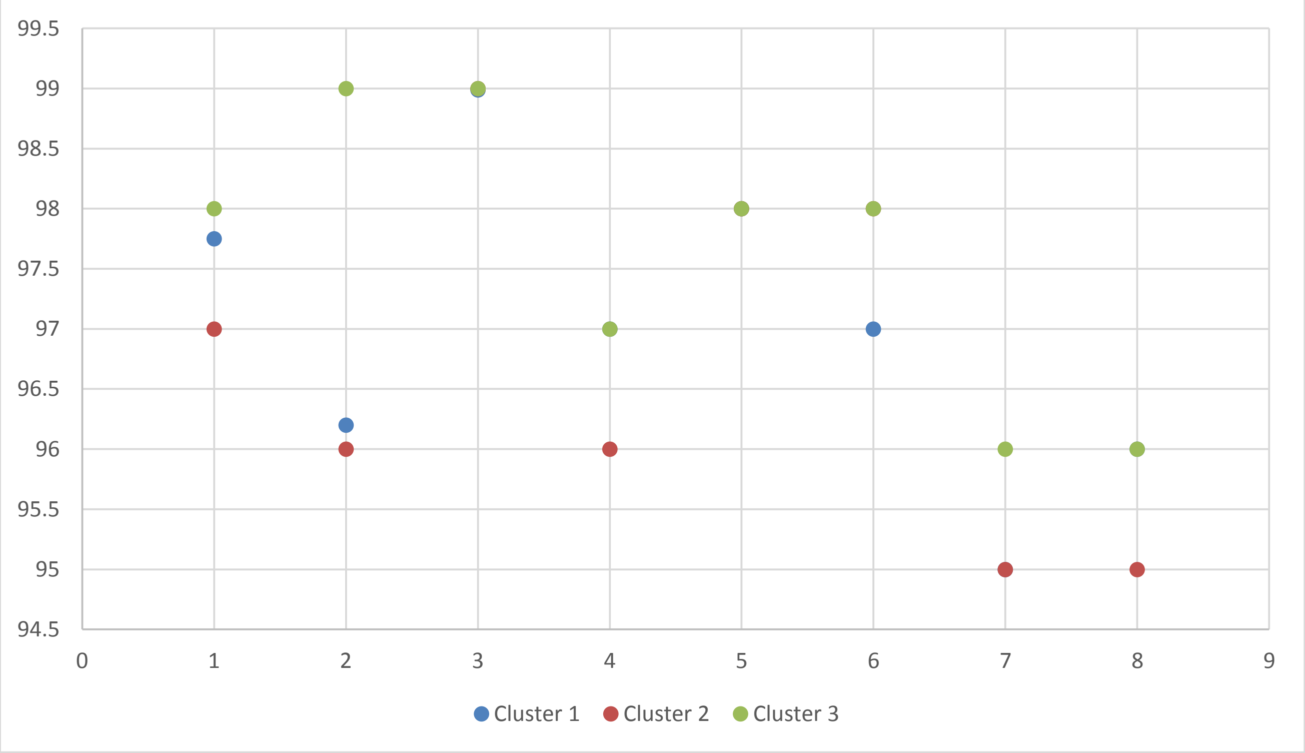


Figure 9: Scattered Plot for Performance Metrics Clusters

The scattered plot is to visualize large amount of data by identifying the clusters in the distribution of data. This plot displays the dimensions taken by the performance metric clusters. The classifier results are further represented by graph and plot which gives a better virtualization of the analysis derived and interprets it in clear data representation that displays the cluster with optimal result. In this scattered plot, it reveals that cluster 3 possess a better performance metrics.

V. DISCUSSION

This study enhances and can be successful in modern malaria illness diagnosis and treatment. The implemented method uses machine learning technique such as model and classification algorithms for clustering analysis and optimization. The redundancy elimination and increase in accuracy and effectiveness which uses the K-Means clustering method and the SVM classifier kernels which obtained a better accuracy performance. This study carried out the performance analysis and evaluation which displayed the results derived. The grouped data in cluster 1, 2 and 3 have confirmed the existence of resistance and susceptible genes with appropriate metrics per cluster. The SVM classifier further analyses the grouped data into charts with true class, predicted class among others. The summary of the analysis results reveals that cluster 3 is more efficient with the various performance metrics used. This study can be used successfully in modern malaria illness diagnosis and treatment. The implemented machine learning tools are efficient for clustering analysis and optimization. K-Means clustering with the SVM classifier obtained a better accuracy performance. For future recommendations, clustering algorithms and other optimization method can be implemented for quantitative evaluation to reveal if there exist more techniques which is useful to improve the efficiency of classification with an advanced approach.

VI. CONCLUSION

Tremendous progress has been made in recent years on RNA-seq progress, with improvement in enhancing the efficiency metrics and performance which is influenced by observational method and data analysis. Mosquito is considered as a lethal insect which has different species. Several studies have emanated from the urgency of identifying and treating the malaria vector into group of clusters. This study uses K-Means algorithm with SVM classification techniques for clustering analysis to improve accuracy and eliminate redundancy in the dataset thereby deriving an enriched cluster. The K-Means algorithm used in this system handles the correlation of the data more efficiently. SVM classification is used to classify the data and results which are displayed in a chart format. This proposed system reveals the performance optimization of RNA-seq data on gene expression. The system performs an Insilco clustering analysis in a simpler way which reduces the complexity and errors when compared with conventional model. Thus, the clustering analysis can be done more efficiently using proposed system. This experiment shows that all the 3 K-Means clusters are useful for classification, while it iterates between the number of susceptible genes, resistance genes with the total number of observations present thereby eliminating discriminative information. The classification is of high essentiality so as to analyze and remove clustering problems. This research work contributes to knowledge by, combining clustering algorithm with an enhanced classifier as a stage in mining RNA-Seq dataset, it also developed a malaria vector resistance prediction model using RNA dataset from Anopheles gambiae. For the contribution of knowledge as regards bioinformatics in the field of machine learning, more performance measures should be implemented and executed with more evaluation protocols for multi label predictors such as Coverage (Cov), Aiming (Aim), Absolute-True-Rate (ATR) and Absolute-False-Rate (AFR). The future studies can be done to develop a hybrid system which detects the unlabeled data with less time complexity and more classifiers to test the system with different datasets.

ACKNOWLEDGEMENTS

The authors appreciate the contribution of the entire embers of the Department of Computer Science and Centre for Research and Development, Landmark University, Omu-Aran, Nigeria for fully sponsoring the publication of this research.

Data Availability

All data are included in this paper.

Funding

The funding of the publication of this research is funded by the Centre for Research and Development, Landmark University, Omu-Aran.

Consent

All participants provided written informed consent to publish this study.

Conflicts of Interest

Authors declare that they have no conflicts of interest.

REFERENCES

[1] 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. 2018; pages 1-10. doi :10.1088/1742-6596/971/1/012004

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

[3] 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) 2012;1(2):185-189.

[4] Bonizzoni, M., Ochomo, E., Dunn, W. A., Britton, M., Afrane, Y., Zhou, G., … Yan, G. 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, 2015;8(1). doi:10.1186/s13071-015-1083-z

[5] Burges, C. A tutorial on support vector machine for pattern recognition, Data Min, 1998.

[6] 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, 2017;45(17): 1-11. Doi: 10.1093/nar/gkx681.

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

[8] 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. 2015;Volume 1, pages 1-14. http://dx.doi.org/10.1155/2015/370640

[9] Etienne, B., Leland, M., John, H. Dimensionality Reduction for Visualizing Single Cell Data Using UMAP. Nature Biotechnology. 2018;37(1), pp 1-13. Doi:10.1038/nbt.4314.

[10] Francis, E. G. History of the discovery of the malaria parasites and their vectors. Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London WC1E 7HT UK. *Parasites & Vectors*, 2010;**3**(5).

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

[12] 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, 2017;12(8): pp1-24.

[13] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18:83. https://doi.org/10.1186/s13059-017-1215-1.

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

[15] 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.

[16] Jelili, O. O., Ojeniyi, O. O., and Obagbuwa, I. C. Application of K-Means Clustering Algorithm for Prediction of Students' Academic Performance. International Journal of Computer Science and Information Security (IJCSIS), Vol. 7, No. 1, 2010.

[17] Khalilian, M., Mustapha, N., Suliman, M., N., and Mamat, M., A. A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets. International Multi Conference of Engineers and Computer Scientists (IMECS). Vol. I. March 17, 2010.

[18] Khan, A., Baharudin, B., Lee, L.H., and Khan, K. A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology, 1(1): pp1-17, 2010.

[19] Lanzaro GC, Lee Y. Speciation in Anopheles gambiae – The Distribution of Genetic polymorphism and patterns of reproductive isolation among natural populations. In S. Manguin, editor. Anopheles mosquitoes – new insights into malaria vectors. INTECH 2013. DOI: 10.5772/56232.

[20] Long, X., Cleveland, W., and Yao, Y. Methods and Systems for Identifying and Localizing Objects based on Features of the Objects that are Mapped to a Vector: Google Patents. 2011.

[21] Majhi, S.K., and Biswal, S. Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer, Karbala International Journal of Modern Science (2018), https://doi.org/10.1016/j.kijoms.2018.09.001. 2018;Pp 1-14.

[22] Malte, D, L., and Fabian, J, T. Current Best Practices in Single Cell RNA-Seq Analysis: Tutorial. Molecular System Biology. 2019;15(6). Doi:10.15252/msb.20188746.

[23] 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. 2015;8(474): pp1-13. https://doi.org/10.1186/s13071-015-1083-z

[24] Arowolo, M, O., Marion Adebiyi and Ayodele Adebiyi. "A Dimensional Reduced Model for the Classification of RNA-SEQ Anopheles Gambiae Data". Journal of Theoretical and Applied Information Technology, JATIT, 2019;Vol.97, No. 23, pp. 3487-3496.

[25] Muhammad Ali Masood., and Khan, M, N, A. "Clustering Techniques in Bioinformatics", IJMECS, 2015;vol.7, no.1, pp.38-46.DOI: 10.5815/ijmecs.2015.01.06.

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

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

[28] Salcedo-Campos, F., Diaz-Verdejo J., and Garcia-Teodoro., P. Segmental parameterization and statistical modelling of e-mail headers for spam detection. Information Science., 2012;Vol. 195, pp.

45-61.

[29] Seo J, Jin D, Choi C-H, and Lee H. Integration of MicroRNA, mRNA, and Protein Expression Data for the Identification of Cancer-Related MicroRNAs. PLoS ONE 2017;12(1): e0168412. doi:10.1371/journal.pone.0168412

[30] Soofi, A.A., and Awan, A. Classification Techniques in. Machine Learning: Applications and Issues. Journal of Basic and Applied Sciences, 2017;Volume 13, pp 459-465.

[31] 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. 2019;20(194). Pp.1-14.

[32] Tajunisha and Saravanan. Performance analysis of Kmeans with different initialization methods for high dimensional data. International Journal of Artificial Intelligence & Applications (IJAIA), 2010;Vol.1, No.4.

[33] 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, 2017;8(5): pp.515-426.

[34] Vapnik, V. The Nature of Statistical Learning Theory, second ed., Springer, New York. 1999.

[35] Velmurugan, T. Efficiency of K-Means & K-Medoids Algorithms for Clustering Arbitrary Data Points. International Journal of Computer Technology & Applications (IJCTA), 2012;Vol. 3 (5).

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

[37] Yasen Jiao and Pufeng Du. Performance Measures in Evaluating Machine Learning Based Bioinformatics Predictors for Classifications. Quantitative Biology, 2016;4(4): pp320-330. doi: 10.1007/s40484-016-0081-2.

[38] Balamurugan, M., Nancy, A., And Vijaykumar, S. Alzheimer’s Disease Diagnosis by Using Dimensionality Reduction Based on KNN Classifier. Biomedical & Pharmacology Journal 2017;10(4): pp1823-1830.

[39] Angelo Duò, Robinson M.D and Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 2; peer review: 2 approved] F1000Research 2018;**7**:1141. (https://doi.org/10.12688/f1000research.15666.2)

[40] Zhen W, Jin C, and Ming Q. Non-parallel planes support vector machine for multi-class classification. Int Conf Logistics Syst Intell Manag, 2010;1:581-585.