https://afrjcict.net

Comparative Evaluation of Linear Support Vector Machine and K-Nearest Neighbour Algorithm Using Microarray Data on Leukemia Cancer Dataset

Ayomikun Kubrat Oladejo [1], Tinuke Omolewa Oladele [2], and Yakub Kayode Saheed [3]
[1,2]Department of Computer Science, University of Ilorin, Ilorin, Kwara State, Nigeria.
Email: [1]*oladejoayomikun@gmail.com*
[2]*tinuoladele@gmail.com*
&
[3]Department of Physical Sciences, Al-Hikmah University, Ilorin. Kwara State, Nigeria.
Email: yksaheed@alhikmah.edu.ng

ABSTRACT
High dimensionality affects the performance of classifiers, especially for microarray gene expression datasets. A lot of efficient dimensionality reduction techniques that transform these high dimensional data into a reduced form have been proposed for microarray data analysis and they perform well. However, these techniques need to be improved in systematic ways as regards to their performance metrics. This research work makes use of two dimensionality reduction strategies, feature selection and feature extraction, to address the problems of highly correlated data. In this study, analysis of micro array data was carried out on Leukemia cancer dataset, with the end goal of finding the smallest quality subsets for precise tumor arrangement. One-way ANOVA algorithm was used for selecting relevant variables and Principal component analysis (PCA) algorithm was used to remove the most relevant variables out of the ones that have been selected. The experimental analysis was carried out on matlabR2015a (8.5.0.197613) environment. The classification algorithms employed are support vector machine (SVM) and K Nearest Neighbour (KNN) as a classification method. Feature selection and feature extraction were combined into a generalized model to help to obtain a robust and efficient dimensional space. In this approach, redundant and irrelevant features are removed at each stage. The classification presents an efficient performance metrics in terms of accuracy, attaining 90% of SVM and 81.67% accuracy of KNN algorithm. The complexity of the proposed method is also significantly reduced.

Keywords: *Classifiers, K Nearest Neighbours (KNN), One-Way ANOVAs, Support Vector Machine (SVM), Microarray data.*

African Journal of Computing & ICT Reference Format:

Ayomikun Kubrat Oladejo, Tinuke Omolewa Oladele & Yakub Kayode Saheed (2018), Comparative Evaluation of Linear Support Vector Machine and K-Nearest Neighbour Algorithm using Microarray Data On Leukemia Cancer Dataset,
Afr. J. Comp. & ICT, Vol.11, No.2, pp. 1 - 10.

I. INTRODUCTION

Over the past two decades, the world has witnessed a true explosion of data, which has mainly been driven by innovative storage technology and the increasing popularity of the Internet. Today, an enormous measure of information is produced in the therapeutic area. A well-known source is microarray information. Microarray is a natural stage for get-together gene articulations [1].

Microarray experiment is a biological procedure to measure the activities of genes at a specific time frame applied to a subject, that is, pre-cancer screening, general health check and cancer remission check. It is designed for bioinformatics field to provide an insight for information on the gene interactions and cancer pathways with a potential for cancer diagnosis and prognosis, prediction of therapeutic responsiveness, discovery of new cancer groups and molecular marker identification [2-5]. Microarray experiment contains measurements for thousands of microscopic spot of DNA probes (that is DNA spots that have been complimentarily banded in the microarray experiment), however, only a small set of these probes are relevant to the subject of interest, for example, amongst 7129 probes in the leukaemia microarray data available from the Broad Institute, only about 1000 probes are relevant to the leukaemogenesis pathway [2]. Therefore, techniques for extracting the informative genes that underlies the pathogenesis of tumour cell proliferation, from high dimensional microarrays is necessary [4-7] and the need for computing algorithms to undertake such a complex task emerge naturally. This brings the theme of computational analysis in microarray studies to the forefront of research.

Microarray gene expression data is characterized by high feature dimensionality, sample scarcity and complex gene behavior (that is the interaction between genes within the data), which pose unique challenges in the development of computing algorithms in class prediction, cluster discovery and marker identification, with the aim of deriving a biological interpretation of the set of genes which underlies the cause of the disease. Additionally, microarray gene expression data may contain subgroup of cancer classes within a known class. This makes the analysis of microarray difficult. Thus, the first and foremost consideration for analysing microarray data is feature extraction. For class prediction, the extracted gene subset is used to avoid the over fitting problem on supervised classifiers and to achieve better predictive accuracy that generalizes well to unknown data [5]. Frequently, data preprocessing is required on microarray data to remove undesirable data characteristics with the idea of ensuring data integrity and improving classification performance. For instance, missing values in microarrays require some mathematical formulas to impute reasonable estimates to salvage the data. Feature reduction is the approach most commonly used to remove data redundancies.

In this study, analysis of micro array data was carried out on Leukemia cancer dataset, with the end goal of finding the smallest quality subsets for precise tumor arrangement. One-way ANOVA algorithm was used for selecting relevant variables and Principal component analysis (PCA) algorithm was used to eliminate the most relevant variables out of the ones that have been selected. This paper presents a comparative analysis of LSVM and KNN algorithms for micro array analysis on leukemia cancer. The rest of this paper is organized as follows. In section 1.2, literature review is presented, section 2 highlight the methodology, experimental analysis is showed in section 3. Section 4 is the results and discussion. In section 5, the paper is concluded.

II. LITERATURE REVIEW

A lot of research has addressed the topic of dimensionality reduction on classification of the microarray data by using different methods with different classifiers. A generic approach for classifying two types of acute leukemia was introduced by Golub et al. [2]. Several feature techniques have been proposed in the literature and survey of feature algorithms is included. Many researchers are involved in the study of goodness of a feature subset in determining an optimal one.

The work done by [8] assessed and thought about the effectiveness of various characterization techniques, including SVM, neural system, Bayesian grouping, decision tree (J48, ID3) and random forest strategies. Further, the productivity of the element choice techniques including bolster vector machine recursive element disposal (SVM-RFE), Chi-squared and relationship based component determination was looked at. Ten times cross approval was utilized to figure the exactness of the classifiers. To begin with, the characterization strategies were connected to all datasets without playing out any component choice. In many datasets, SVM and neural systems performed superior to other arrangement

techniques. In all cases SVM-RFE performed exceptionally well when it was connected with SVM arrangement techniques.

[9] displayed a gene selection scheme called ANOVA, which is utilized to locate the base number of qualities from microarray quality articulation that can be utilized as a part of arrangement of malignancy. The proposed positioning plan called two-way Analysis of Variance (ANOVA) was utilized for the determination of imperative qualities. The characterization can be found by the utilization of surely understood classifiers such as Support Vector Machines. The lymphoma dataset were utilized to show the viability of this approach. If the selected data contains missing values or exhaust cell sections, it must be preprocessed. This work incorporates three stages. Step 1 is an important gene selection stage using a scoring scheme called Analysis of Variance (ANOVA) method and afterward the best genes can be chosen with the most elevated scoring value from positioned data. The next step is the gene extraction using the principal component analysis (PCA) and the last step is classification capability of all gene combinations which can be performed with the utilization of the Support vector machine and K-Nearest Neighbor. The Selected genes are put into the classifier if exactness is not gotten to such an extent that classification is performed with a gene combination. The acquired outcomes utilizing ANOVA with SVM is then contrasted with the T - score strategy.

[10] offered a nonparallel plane proximal classifier (NPPC) troupe for malignancy order in light of microarray genes articulation profiles. A hybrid and computer supported analysis (computer aided design) structure is presented in light of channels and wrapper techniques. Least excess most extreme pertinence (MRMR) positioning technique is utilized for feature selection. The wrapper strategy is connected on those genes sets to diminish the computational weight and nonparallel plane proximal classifier (NPPC).

[11] featured the disclosure of differentially expressed genes (DEGs) in microarray data in their approach to construct an exact and savvy classifier. A T-Test highlight choice method and KNN classifier was connected on the Lymphoma data set to come to the DEGs and to breaking down the impact of these genes on the classifier accuracy, discretely.

The authors in [12] demonstrated how Support Vector Machine "SVM" has an excellent performance on classification and prediction; it is widely used on disease diagnosis or medical assistance. SVM only function well on two-group classification problems. Their study combined feature selection and SVM recursive feature elimination (SVM-RFE) to investigate the classification accuracy of multiclass problems for dermatology and zoo databases. The dermatology dataset contains 33 feature variables, 1 class variable and 366 testing instances; and the zoo dataset contains 16 feature variables, 1 class variable and 101 testing instances. The feature variable in the two datasets were sorted in descending order by explanatory power, and different feature sets were selected by the SVM-RFE to explore classification accuracy.

Taguchi method was also combined with SVM classifier in order to minimize parameters C and γ to increase classification accuracy for multiclass classification. Penalty parameter C represents the cost of the classification error of training data during the learning process, as determined by the user. When C is greater, the margin will be smaller, indicating that the fault tolerance rate will be smaller when a fault occurs. Otherwise, when C is smaller, the fault tolerance rate will be greater, where γ is the linear kernel function.

In [1] proposed different approaches to perform dimensionality reduction on high-dimensional microarray information. Distinctive component choice and highlight extraction strategies which plan to evacuate repetitive and unimportant highlights for new cases of classification can be precise were established. A well-known wellspring of its information is microarrays, a natural stage for social occasion quality articulations. Examining the microarrays can be troublesome because of the span of the information they create, and the entanglement connection among the diverse qualities makes investigation more troublesome and expelling overabundance highlights can enhance nature of the outcomes. A famous strategy for choosing huge highlights was introduced and a correlation between them was made.

III. METHODOLOGY

3.1 Feature Selection

A number of methods have been proposed for rule extraction from SVMs. Broadly speaking, these methods can be categorized into three main families which are: pedagogical, decomposition, and eclectic [13]. Some of these techniques till date still deliver moderately substantial administer sets, which constrains their clarification capacity [14]. Rule sets can just offer

clarification if the quantity of rules in the rule set is generally little and its order exactness is high. Simpler rules likewise offer better understanding and explanation [15]. To remove more understandable rules, irrelevant features which do not add to the grouping choice ought not to be in the rule predecessors. This highlights a requirement to consider feature selection as an integral part of rule extraction. In feature selection, one selects only those input dimensions that contain the relevant information for solving the particular problem. There are three categories of feature selection which are: filters, wrappers, and embedded techniques. This work focuses on filter-based approach.

3.2 Feature Selection Procedure

The feature selection procedure includes four important key steps; subset generation, subset evaluation, stopping criterion and result validation which are shown in Figure 1.

3.3 Proposed System

In this study, the researcher proposed to use feature selection method first, then feature extraction using Support Vector Machine algorithm in the first phase to reduce the dimensionality of the data by yielding the key attribute in the data. Thus, fewer numbers and smaller rules are obtained resulting in the improvement of the comprehensibility of the system. Lastly, classification Algorithm was employed using linear SVM Algorithm on Leukemia data set. The framework of the proposed system is shown in figure 2.

3.4 Experimental Analysis

3.4.1 Dataset Description

For implementing and testing the effectiveness of the algorithm, experiment will be performed on Leukemia dataset. The data set will be obtained from a genomic database. This dataset is going to be chosen because of its public accessibility and has previously been used for several Machine Learning studies [17]. The information required to be stored in the database is Leukemia cancer dataset [17]. It contains DNA microarray gene expression data. 7132 attributes and 35 instances are loaded from an excel spread sheet. The steps of the proposed study are as follows;

1. Load
2. Feature selection
3. Feature extraction
4. Classification

3.4.2 Feature Selection

In the feature selection mode, the feature is selected using the ANOVA t-test analysis at 95% confidence interval level which is at the 0.05 significance level. The obtained result is saved for future reference, so as to be passed into the feature extraction modules.

3.4.3 Principal Component Analysis (PCA) For Feature extraction

The result of the components extracted when PCA technique was used, a total of 20 components were extracted from the selected features. Analysis of Variance (ANOVA) is a hypothesis-testing technique used to test the equality of two or more population (or treatment) means by examining the variances of samples that are taken. ANOVA allows one to determine whether the differences between the samples are simply due to random error (sampling errors) or whether there are systematic treatment effects that cause the mean in one group to differ from the mean in another.

IV. RESULTS AND DISCUSSION

4.1 RESULTS

Figure 3 shows the confusion matrix result for the classified components which was extracted using SVM technique. The True Positive rate yields 85.7% and False Negative rate yields 92.3%. Confusion matrix gives the layout of the performance of a classification model (classifier) on a set of test data for which the true values are known.

Figure 3 shows the confusion matrix result for the classified components which was extracted using SVM technique.

TP=36 FP=3 FN=3 TN=18

Accuracy: This is the simplest scoring measure. It calculates the proportion of correctly classified instances.

Accuracy = (TP + TN) / (TP+TN+FP+FN)
(36 + 18) / (36 +18 +3 +3) = 54/60
Accuracy rate = 0.9
Percentage accuracy rate: 90.00%

Sensitivity: The sensitivity rate also known as the Recall or True Positive, tells us how likely the test will come back to positive on a sample that has the characteristic of Leukemia cancer:

Sensitivity = TP/ (TP+FN)
36/36+3 =36/39
Sensitivity Rate = 0.9231
Percentage Sensitivity Rate = 92.31%

Specificity*:* The specificity also known as the True Negative relates to the classifiers ability to identify negative results.

Specificity = TN/ (FP+TN)
18/ (3+18) = 18/21
Specificity Rate = 0.8571
Percentage Specificity Rate = 85.71%

Precision: This is a measure retrieved instances that are relevant.

Precision: TP/ (TP+FP)
36/ (36+3) = 36/39
Precision Rate = 0.9231
Percentage precision: 92.31%

4.2 Discussion

a. Comparative Evaluation of Linear SVM AND KNN

b.

Table 1: Comparative Evaluation of Linear SVM and KNN from the reduced dataset.

Performance Metrics	Classification using Linear-SVM	Classification Using KNN
Classifier Accuracy (%)	90	81.67
Sensitivity (%)	92.31	89.74
Specificity (%)	85.71	66.67
Precision (%)	92.31	83.33

Table 1, shows that the feature selection based on One-way-ANOVA method achieves necessary higher value in the datasets on the performance metrics such as the accuracy, specificity, and prediction when compared to the direct based method on the Leukemia cancer dataset.

Figure 5 shows the performance metrics of classification in terms of accuracy, sensitivity, specificity and precision using linear SVM and KNN.

V. CONCLUSION

In this study, analysis of micro array data was carried out on Leukemia cancer dataset, with the end goal of finding the littlest gene subsets for accurate cancer classification. The study employed SVM and KNN for classification and ANOVA for feature selection. ANOVA is an exceptionally powerful positioning plan while SVM is an adequately decent classifier contrasted with comparative mining approaches. As we have seen from the results in the Leukemia dataset, the gene combination that gives good separation may not be unique. In the Leukemia data set, 786 selected genes was classified using One-Way ANOVA feature selection. MATLAB R2015a (8.5.0.197613) was used to implement this procedure. The outcome demonstrates that the proposed decrease approach achieved promising results of the supplemented quantities of genes to the classifiers.

REFERENCES

[1] Zena, M. H., and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. London: Department of Computing, Imperial College.

[2] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H. Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286(5439):531-536.

[3] Dupuy, A., and Simon R. M. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. JNCI Journal of the National Cancer Institute, 99(2):147-157.

[4] Yu, J. J., Yu, A. A., Almal, S. M., Dhanasekaran, D., Ghosh, W. P., Worzel, and Chinnaiyan, A. M. (2007). Feature selection and molecular classification of cancer using genetic programming. Neoplasia, 9(4):292303, 2007.

[5] Wang, Y., Miller, D. J. and Clarke R. (2008). Approaches to working in high-dimensional data spaces: gene expression microarrays. British journal of cancer, 98(6):1023-1028.

[6] Osareh, A., and Shadgar, B. (2008). Classification and diagnostic prediction of cancers using gene microarray data analysis. Journal of Applied Sciences, 9(3):459-468.

[7] Zhang, J. T., Jiang, B., Liu, X., Jiang, and Zhao H. (2008).Systematic benchmarking of Microarray data feature extraction and classification. International Journal of Computer Mathematics, 85(5):803{811, 2008.

[8] **Pirooznia**, **M**., **Yang**, J. Y., **Yang**, M. Q. and **Deng**, Y. A. (**2008**). A Comparative Study of Different Machine Learning Methods on Microarray Gene Expression Data. BMC Genomics, 9.

[9] Bharathi, A. and Natarajan, A. M. (2010). Cancer Classification of Bioinformatics data

using ANOVA International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010 1793-8201.

[10] **Ghorai** S., Mukherjee A., Sengupta S., and **Dutta** P. K. (2011). Cancer classification from gene expression data by NPPC ensemble. IEEE/ACM Trans. Computational Biology and Bioinformatics; 8:659-671.

[11] Abeer, M., Basma, A. (2014). A Hybrid Reduction Approach for Enhancing Cancer Classification of Microarray Data. (IJARAI), Vol. 3, No.10. pp.1- 10. 2014.

[12] Mei-Ling Huang, Yung-Hsiang Hung, W. M. Lee, R. K. Li, and Bo-Ru Jiang (2014). SVM-RFE Based Feature Selection and Taguchi Parameters Optimization for Multiclass SVM Classifier. Hindawi Publishing Corporation, The Scientific World Journal Volume 2014, Article ID 795624, pp.1-10. http://dx.doi.org/10.1155/2014/795624.

[13] Andrews, R., Diederich, J., and Tickle, A. B. (1995). A Survey and critique of techniques for extracting rules from trained Artificial Neural Networks. Knowledge Based Systems. Vol.8, Issue 6, December 1995, pp.373-389.

[14] Barakat, N. H., and Bradley, A. P. (2007). Rule extraction from Support Vector Machines: A Sequential Covering Approach IEEE transactions on knowledge and data engineering.Vol. 19, no. 6, pp. 729-741, **2007.**

[15] Duch, W., Setiono, R., and Zurada, J. (2004). Computational intelligence methods for rule-based data understanding. Proceedings of the IEEE, Vol. 92, No. 5, pp. 771-805. May 2004.

[16] Solberg T. R., Sonesson A. K., Woolliams J. A., Odegard J, Meuwissen T. H. (2009). Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. 41:53. doi: 10.1186/1297-9686-41-53.

[17] Virtaneva, K. I., Wright, F. A., Tanner, S. M., Yuan, B., Lemon, W. J., Caligiuri, M. A., Bloomfield, C. D., Albert de la Chapelle and Krahe**,** R. (2001). Gene expression profiling reveals fundamental biological differences in AML with trisomy 8 and normal cytogenetics. Nature Genetics 27**, page** 65. ISSN 1546-1718 (online). doi:10.1038/87158.

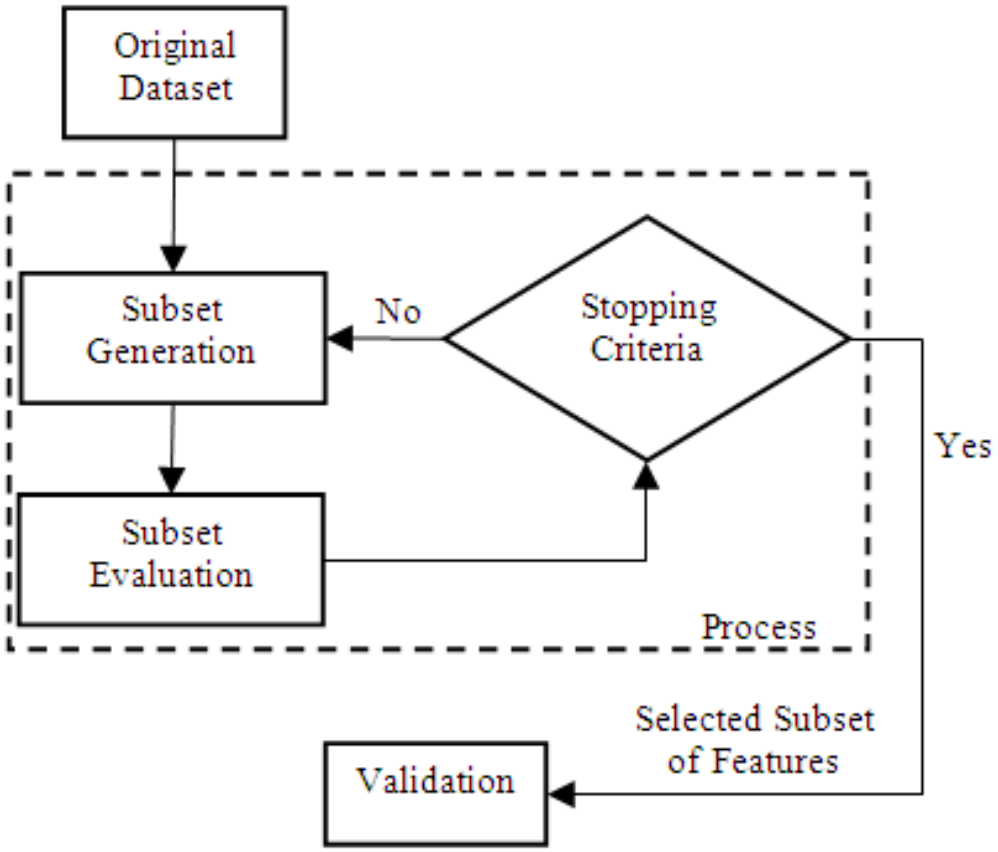

Figure 1: Feature Selection Method (adapted from [16])

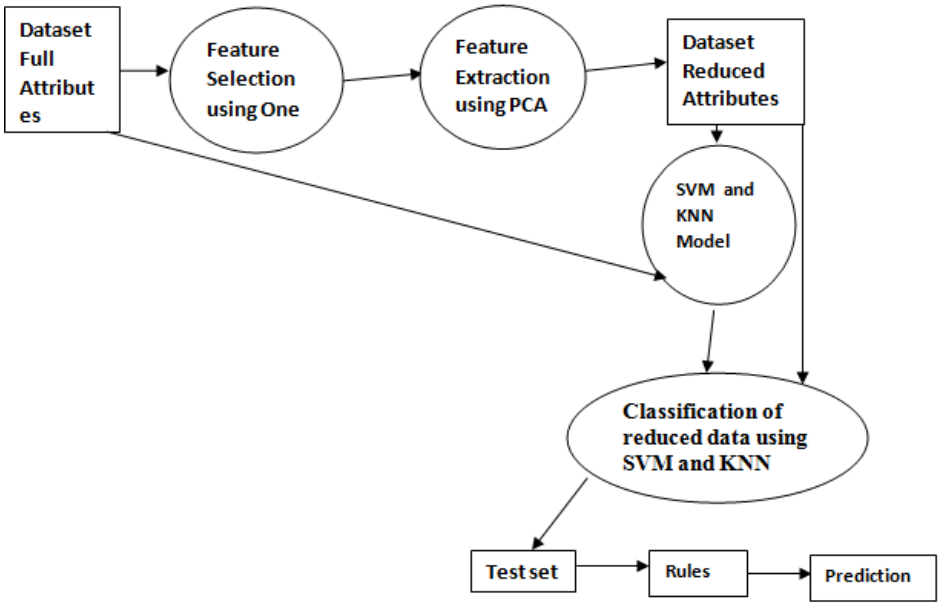

Figure 2: Framework of the Proposed System.

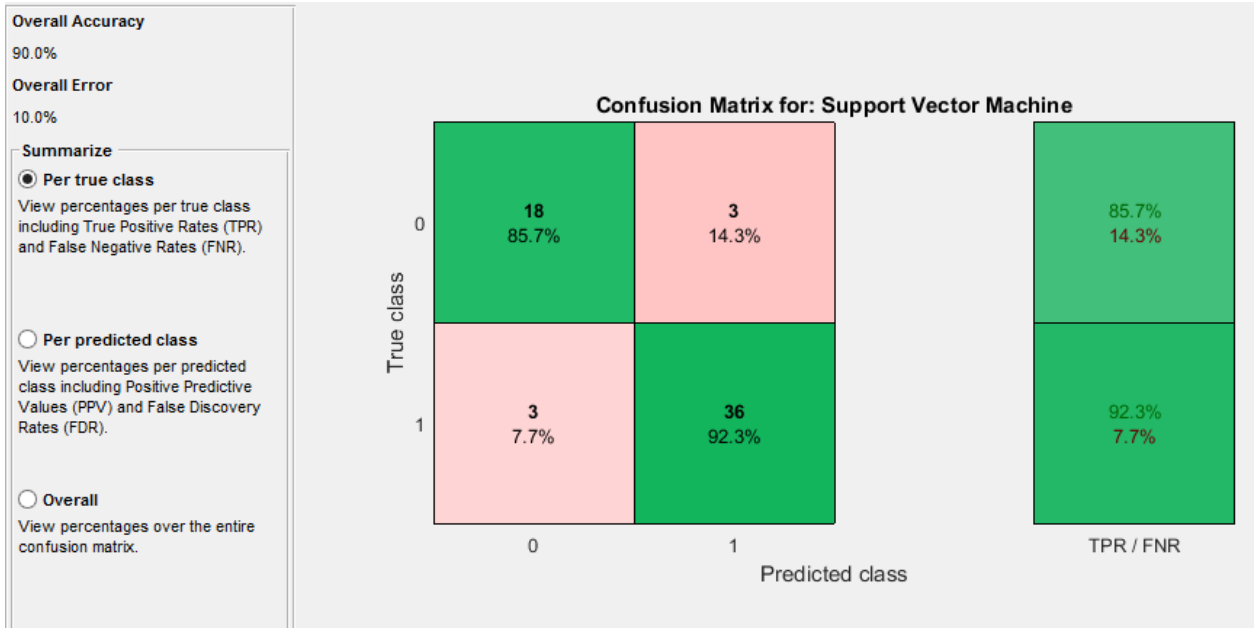

Figure 3: Confusion Matrix for the classification of selected features Using Linear-SVM.

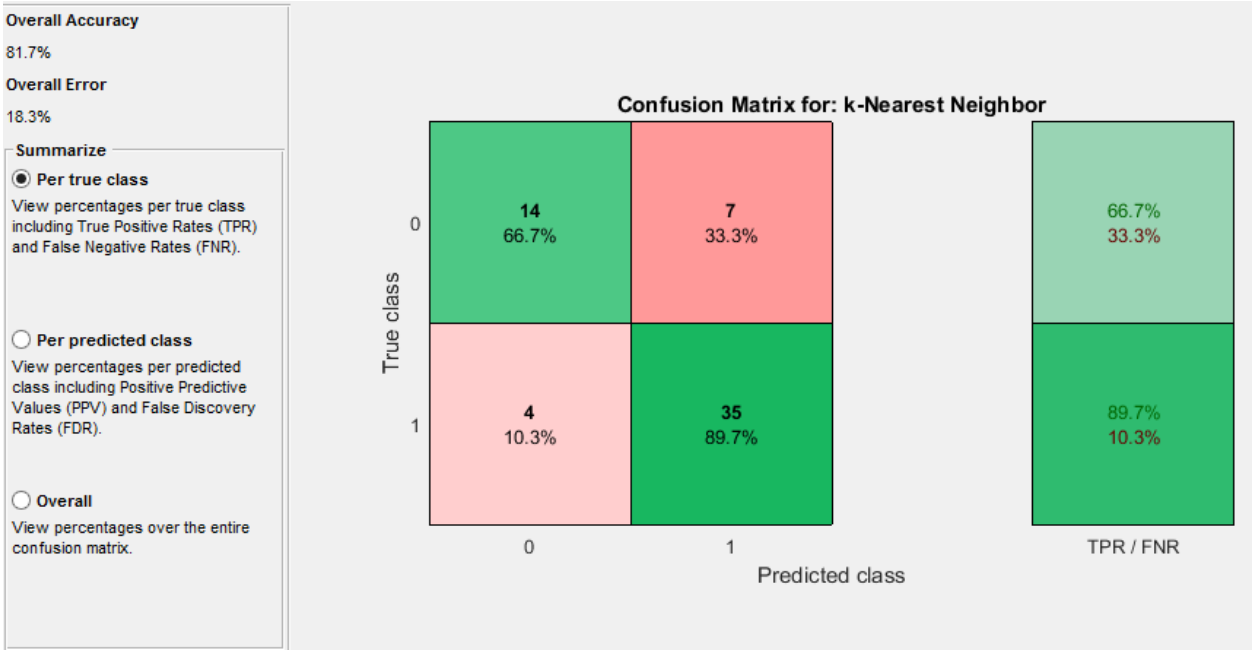

Figure 4: Confusion matrix for Leukemia dataset

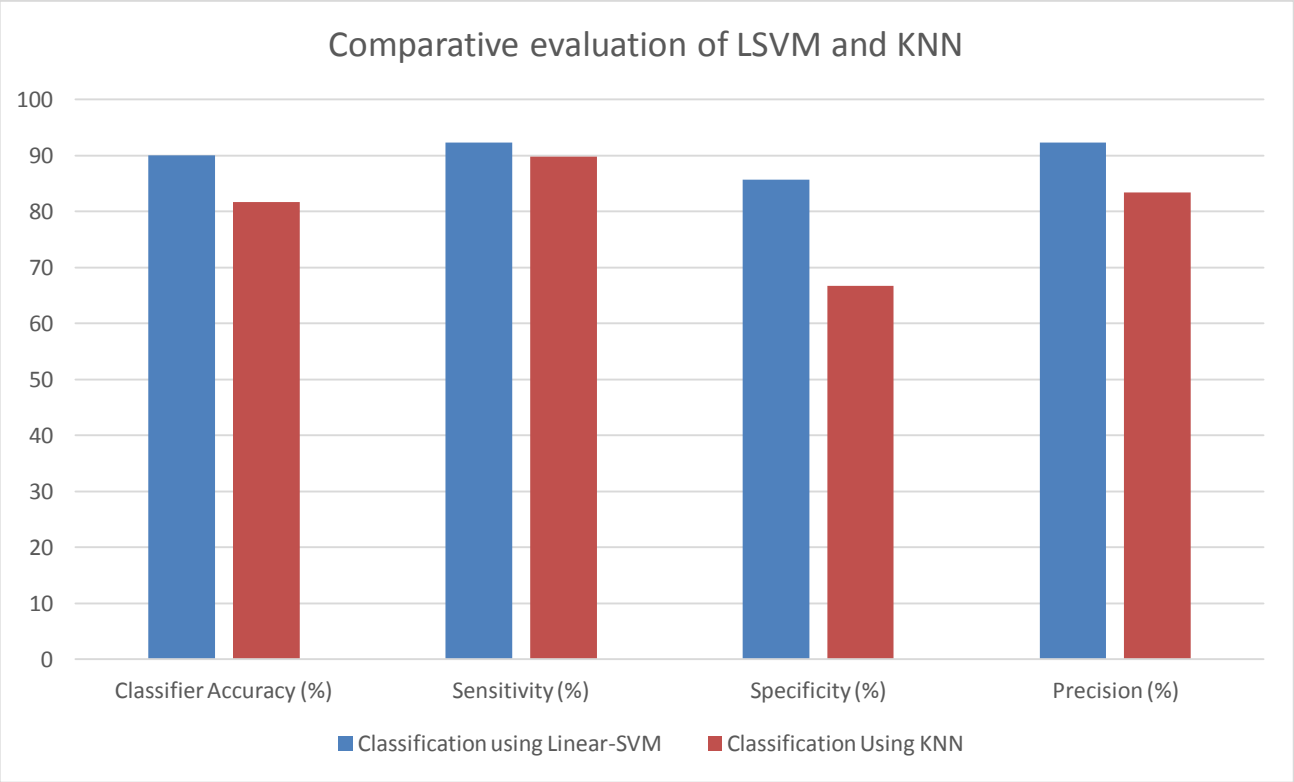

Figure 5: Performance metrics of L-SVM and KNN