International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
Article Received: *25 February 2023* ***Revised:*** *12 March 2023* ***Accepted:*** *30 April 2023*

Classification Algorithms for Cancer Detection Using Microarray Gene Expression Data and Multiple Attribute Sets

Simardeep Kaur[1], Dr. Maninder Singh[2]

[1]Department of Computer Science & Applications, DAV College, Abohar-152116, India

[2]Department of Computer Science, Punjabi University, Patiala-147002, India

Abstract

The ability to provide gene expression data in the form of expression profiles of thousands of genes across biological processes is one of the advantages of microarray technology. Based on the patterns shown by the gene expression data, the microarray assay's most intriguing artifact represents the differentiation of tumors. The current study uses the knowledge gathered from the microarray assay to classify leukemia into two categories: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Three computational intelligence approaches—Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and Feed Forward Neural Network (FFNN)—are used in this work to assess the efficacy of Principal Component Analysis (PCA), Canonical Correlation (CC), and Cosine Correlation (CosC) in addressing the difficulties of feature extraction. Accuracy, True Positive Rate (TPR), False Positive Rate (FPR), and Kappa-Coefficient (KC) are used to evaluate the performance of the combinations that have been put into practice. All nine combinations are subjected to a simulated examination using 500 samples that reflect leukemia gene expression data. With an average classification accuracy of 0.6231, experiments have demonstrated that PCA with FFNN performs better than all other combinations in terms of TPR, FPR, KC, and accuracy.

Keywords: Leukemia, cancer classification, and gene expression data.

1. Introduction

To prevent uncontrollable events and arrange successful treatment in advance, early cancer diagnosis is crucial [Round, 2017]. According to medical language, cancer is characterized by aberrant cell development that has the potential to multiply uncontrollably. One of the most deadly diseases that claims the lives of people each year is cancer. Global cancer statistics indicate that 4.3 million of the 18 million cancer cases recorded were leukemia cases, which accounts for 2.6% of all cancer cases reported by the World Wide Cancer Report [WCRF, 2018]. The first step in handling the matter is to know whether or not you have cancer and what kind of cancer you have. The connective tissue that connects all of the body's organs, tissues, and cells is blood. The main concern occurs when leukemia, or malignant development, is found in tissue that produces blood. Recent technological developments have demonstrated the great success of gene expression data in the early detection of cancer. Among the different types of cancer, leukemia is the focus of this work. Acute Myeloid Leukemia (AML) is a type of leukemia that primarily occurs when an acute disorder develops in bone marrow cells, platelets, or red blood cells (RBCs) [Khwaja et al., 2016], while Acute Lymphoblastic Leukemia (ALL) occurs when cancer develops in bone marrow lymphocytes [Pui et al., 2019]. Since these malignancies grow slowly and don't even exhibit recognizable signs, the condition could be lethal. In order to diagnose leukemia's stage [Seshi, 2006] and heterogeneity [Saadatpour et al., 2014], gene expression assays are quite helpful.

One kind of supervised learning technique used to determine the type of cancer based on expression array analysis is microarray data categorization. In reality, cancer is a genetic mutation manifested as unchecked cell division. In medical settings, the biological and phenotypic analysis of this gene expression aids in the diagnosis, prognosis, and treatment planning of asymptomatic cancer [Narrandes and Wayne, 2018]. In addition to histological assessment, leukemia can also be detected via images. Haematological assessment-based pathological investigations, however, are laborious, necessitate invasive sample, and are rarely repeated. Although image-based leukemia detection is somewhat less expensive, it does not identify the genetic foundation of cancer. Furthermore, interpreting microscopic pictures calls for a high level of skill

[Abdeldaim et al., 2018]. Genome profiling is made possible by technological advancements for the therapeutic treatment of cancer patients that emphasizes targeted therapy, where biomarkers are crucial. The changed regulation of biological pathways is measured using the retrieved gene profiles in terms of expression data for a single gene, a protein, or several genes. When making decisions about whether an expression is normal or malignant, this kind of information is actually helpful [Goossens et al., 2015]. In order to create a classifier model that could predict the diseases of the query sample data, a number of research in the literature used gene expression data as a general method for analyzing data samples of labelled gene expression. However, the association of massive gene expression profiles obtained in response to even a small experiment design makes this a computationally demanding task [Ganesh Kumar et al., 2012]. The feature selection phase is when the main issues with gene expression-based classification arise. As a step to improve classification models, researchers have employed a number of feature selection schemes in the literature, including statistical approaches like voting [Taylor and Kim, 2011], filtration [Alirezanejad et al., 2020], correlation method [Santhakumar and Santhakumar, 2020], nearest neighbor, least square, and logistic regression [Algamal et al., 2019].

When microarray technology is used, hundreds of genes' expression will be examined concurrently for each sample, resulting in a vast number of variables. For the purpose of classification, the gene expression offers uncontested and practical information. Nevertheless, it was discovered that the majority of classification methods fall short when gene expression deviates just a little from the predetermined expression profile. In order to effectively classify leukemia into ALL and AML using gene expression data, the given study aims to provide a framework to handle feature extraction issues and identify the optimal feature extraction and machine learning classifier combination.

Organization of the paper

In the introduction, the research addressed the difficulties of automatically classifying cancers based on gene expression data. Section 2 covers a survey of current classification methods with an emphasis on leukemia classification using data from microarray assays. The suggested methodology, which uses a variety of methods at the feature extraction stage to assess machine learning methods for leukemia classification, is described in Section 3. The findings are summed up in Section 4, and Section 5 concludes the paper.

2. Literature Review

This section reviews the most recent studies that have been published in the majority of reputable journals and are mostly concerned with the classification of leukemia using gene expression analysis.

In 2016, Sheikhpour and Aghaseram presented a method for detecting leukemia by analyzing the expression of 7129 genes that were taken from 72 leukemia patients. Information gain and correlation coefficient were used to make the selection, while SVM, linear discriminant, Naïve Bayes, and Multilayer Perceptron (MLP) neural networks were used in the experiment. The highest AML and ALL detection accuracies were obtained with SVM-based classification employing 87 genes that were chosen based on information gain. According to the study's findings, gene selection combined with computational intelligence methods may improve patient diagnosis and therapy [Sheikhpour and Aghaseram, 2016].

Using microarray gene expression data, including leukemia classes, Alshamlan et al. (2016) developed a swarm-based architecture to provide cancer categorization. The researcher utilized ABC to pick genes for a specific form of cancer from a bigger microarray sample. According to the simulation analysis, the maximum classification accuracy of 93.05% was attained by classifying cancer cases into ALL and AML groups using ABC with SVM, which included 14 genes. While taking into account a smaller number of genes that could include false positive results, the work had shown good classification accuracy. As a result, a sizable data set that represents malignant classes must be included [Alshamlan et al., 2016].

In order to differentiate between AML and ALL while examining the gene expression profiles of leukemia patients, Dwivedi (2018) suggested a supervised learning model based on Artificial Neural Network architecture. The neural network-based architecture performed better with the least amount of erroneous detection at the testing stage when compared to five other machine learning techniques: logical regression, SVM, k-nearest neighbor, and Naïve Bayes. Nevertheless, any feature reduction or extraction method that increases the cost of computing was not used [Dwivedi, 2018].

A leukemia classification method based on the Back Propagation Neural Network architecture was proposed by Sharma and Kumar in 2019. The researchers used PCA to extract the pertinent features that correspond to leukemia classes. The high dimensional data that ABC

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
Article Received: *25 February 2023* ***Revised:*** *12 March 2023* ***Accepted:*** *30 April 2023*

optimizes before feeding it to BPNN for classification is thereby reduced. A simulation investigation showed that the suggested ABC-BPNN architecture achieved an average accuracy of 98.72%, with room for improvement in terms of execution time [Sharma and Kumar, 2019].

The goal of Aghamaleki et al.'s 2019 study was to analyze gene expression data obtained from Gene Expression Omnibus datasets in order to identify a single leukemia class that reflected lymphoma type. Additionally, an ANN that was trained on the gene expression data collected from 53 patients was used to find biomarkers and detect leukemia. The simulation analysis showed a receiver operating characteristic (ROC) of 0.991, indicating great accuracy. Further authors suggested its scope in diagnosing other types of cancers with reliable results [**Aghamaleki et al., 2019**].

Arif and Shah (2020) illustrated the benefits of using feature-based analysis in the categorization of cancer. In order to solve the high dimensionality problems of microarray assays, the features were retrieved and reduced using statistical techniques that combined PCA with 6-feature selection approaches. Classification using SVM, LDA, and k-nearest neighbor came next. With the greatest accurate rate of 96%, the experiments revealed that the PCA-based feature extraction followed by SVM performed better than the other combinations. Binary classifiers were used in the work, and multiclass classifiers could be added [Arif and Shah, 2020].

In 2020, Sarder et al. conducted experiments on 7129 genes from 72 participants, of whom 47 were cancer patients and 25 served as controls. Multiple gene feature selection techniques were implemented to evaluate 6 machine learning techniques, namely, Adaboost, Artificial Neural Network (ANN), Linear Discriminant Analysis (LDA), Random Forest (RF), Naïve Bayes (NB) and Classification And Regression Tree (CART). Using the Least Absolute Shrinkage and Selection operator (LASSO) and NB combination for leukemia classes, the work showed a maximum accuracy of 99% [Sarder et al., 2020].

Using the Ant Colony Optimization methodology, Santhakumar and Logeswari have introduced a leukemia prediction method that concentrated on variable selection based on correlations found between the features. In order to improve the overall classification performance, Ant Lion Optimizations was used to select the best features. In order to maximize the feature data that the SVM classifier uses for training and classifying the gene expression data into leukemia classes with an accuracy of 90.91%, the work was based on approaches inspired by nature [Santhakumar and Logeswari, 2020].

3. Methodology

An online data source is used to retrieve the gene expression information of cancer patients. Predicting the type of cancer while analyzing gene expression data is the driving force behind machine learning evaluations, even in cases where a patient exhibits no symptoms at all. Three methods—principal component analysis (PCA), canonical correlation (CC), and cosine correlation (CoC)—are used to address the feature extraction problem with regard to gene expression data. To categorize the input data into two leukemia classes—ALL and AML—and normal, these dimensionally reduced expression data characteristics are given to three machine learning classifiers one after the other. LDA, SVM, and FFNN are the machine learning techniques used for classification. The performance of each combination of feature extraction technique and classifier is evaluated in terms of TPR, FPR, accuracy and KC. The flow chart of the proposed frame work is given in Figure 1.

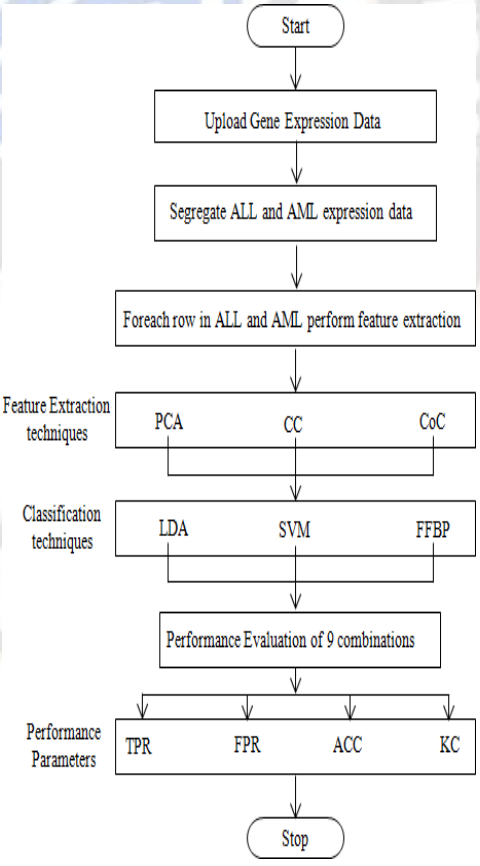

Figure 1: Flowchart representing proposed methodology

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
Article Received: 25 February 2023 *Revised:* 12 March 2023 *Accepted:* 30 April 2023

3.1. Data Source

The Gene Expression Dataset is used for experimentation in this study. It provides the findings of an examination of microarray data pertaining to the use of DNA microarrays for cancer detection. It is mostly used to classify patients with leukemia and offers data with tagged cancer classes to known tumour kinds. There are 38 training samples and 34 independent test samples in the original dataset. The expression results from bone marrow and peripheral blood samples match those from AML and ALL [Golub et al., 1999]. The Kaggle gene expression dataset [GED, kaggle] provides access to the dataset.

3.2. Feature Extraction Techniques

In the current work, the gene expression data is first divided into two categories of leukemia, and the vast amount of data is processed row by row. In order to select the most pertinent features that accurately reflect the kind of cancer, feature extraction is essential to the evaluation of big datasets. The fundamentals of feature-based analysis for cancer classification were illustrated by Arif and Shah [Arif and Shah, 2020]. Furthermore, when working with high dimensional microarray data, feature extraction is seen to be an essential component. Three feature extraction methods—PCA, CC, and CoC—are used in this work.

a. Principal Component Analysis (PCA)

It is the most widely used technique for high-dimensional data feature extraction. The goal is to preserve the dataset's variety while reducing the dimensionality of the collection of associated characteristics. Its fundamental idea reduces the linked traits to a considerably smaller set of characteristics known as primary components. The Principal Component (PC) is essentially a linear combination of one dimension's variables. The following is the expression for the Principal Component $'PC'$ for leukemia when taking into account gene expression datasets with $'d'$ dimensions:

$$PC = c1F1 + c2F2 + \ldots . + cdFd$$

where F stands for the original feature and $c1$ is the numerical coefficient of $F1$. There is some matrix transformation involved in the PCA process. The first step is data preparation and is done by subtracting the mean from the value and dividing it by the standard deviation $stand_{deviation}$ for each value of the considered variables.

$$Std_c = {}^{Fval - meanF}/_{\text{stand deviation}}$$

In this case, $Stdc$, $Fval$, and $meanF$ are the standard data, feature value, and mean of features, respectively. All of the variables are converted to the same scale after this standardization process. The sum of squares and cross-products calculated from the normalized data serve as a representation of the correlation matrix. When the expression data shows a greater difference, the correlation matrix turned out to be significant [Maruf et al., 2019]. The standardized correlation matrix that follows can be found as follows:

where Std' denotes the transposed matrix with $'n'$ number of features in the research, and $Stdc$ represents the center of each column on its mean.

The correlation matrix's $eignevecors$ and $eigenvalues$ are then computed.

$$(Correlationmatrix - evI)\, eV = 0$$

Where, ev is the $eignevalue$ and eV is the $eignevactor$

The new dataset is then generated after the principal components are selected by sorting the $eigenvactors$ from higher to lower $eigenvalue$. Adiwijaya et al. also used the essence of PCA-based feature reduction for different kinds of cancer detection protocols using microarray expression data [Adiwijaya et al., 2018], and Yan et al. used PCA for oncogene expression analysis in patients with leukemia [Yan et al., 2017].Inspired with these research works, PCA is incorporated in the present work to reduce the dimensionality of microarray data that represents a large data set of gene expression.

b. Canonical Correlation (CC)

It is the process of determining the link between the variables in a multidimensional dataset in statistical terms. When dealing with high-dimensional data, it has been widely utilized. Another name for the procedure is Canonical Correlation Analysis (CCA). When examining two sets of data, Knapp conducted a correlation analysis and explained that different statistical tests used to determine the significance and association can be regarded as a specific case of CCA [Knapp, 1978]. Stated differently, it is regarded as a method of drawing conclusions from cross-covariance matrices that involve two sets of random variables, X and Y,that show correlation.

$X = X_1, \ldots, X_n$ //variables in set X

$Y = Y_1, \ldots, Y_m$ // variables in set Y

Now the correlation matrix corresponding to all the variables reflects four combinations.

$CoR_{X,X} \rightarrow$ Correlations among variables in X set

$CoR_{Y,Y} \rightarrow$ Correlations among variables in Y set

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
Article Received: *25 February 2023* ***Revised:*** *12 March 2023* ***Accepted:*** *30 April 2023*

$CoR_{X,Y}$ → Correlations exhibited by variables from X set to Y set

$CoR_{Y,X}$ → Correlations exhibited by variables from Y set to X set

The CCA is further defined as the decomposition of the matrix that represents the combined correlation between two sets. Therefore, matrix representing the combined correlation between X and Y sets is represented as C_{mat} and singular decomposition value of C_{mat} is represented by the diagonal matrix (**Λ**) consisting of eigen values of C_{mat}.

$$C_{mat} = \begin{matrix} CoR^{-1} \\ X,X \end{matrix} \begin{matrix} CoR^{-1} \\ Y,Y \end{matrix} \begin{matrix} CoR_{X,Y} \end{matrix} \begin{matrix} CoR_{Y,X} \end{matrix}$$

Canonical correlation (CC) is calculated using the formula $CC = \sqrt{\Lambda}$, which is the square root of the eigenvalues. where the diagonal matrix of the combined correlation matrix showing two sets of variables is the eigenvalue, denoted by **Λ**. The proportion overlapped variance experienced by all variables is shown here by the eigen values. In other words, correlation $CoRX,Y$ and variance ($varX$, $varY$) between the two sets of random variables represent the total canonical correlation CCX,Y between two sets of random variables.The Canonical Correlation has demonstrated a broader range of uses. It was used in one work by Won et al. to determine statistical significance over aspects of multidimensional biological datasets [Won et al., 2020]. To estimate the growth of cancer, Fan et al., however, used feature aggregation and canonical correlation [Fan et al., 2019].

c. Cosine Correlation (CosC)

To ascertain how similar the features represented by vectors are, the cosine correlation is calculated. Additionally, Dubey et al. have successfully used it to find and choose features in high-dimensional datasets [Dubey et al., 2017]. Similarity analysis between the characteristics that reflect leukemia classes is carried out using Cosine similarity metrics, which were inspired by real-world implementations of Dubey's work. Using the following formula, the frequency vector for the cancer gene expression under consideration is calculated, and their size is then computed.

Where, F_{vect_i} represents the number of i^{th} terms in the feature F that consists of n number of variables or the attributes.

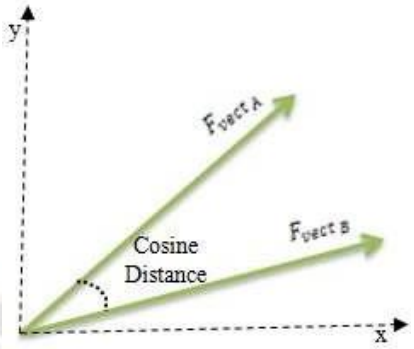

Figure 2 Cosine Similarity vectors The cosine similarity between two vectors is calculated as follows:

$$Cos(A,B) = \frac{F vect A \cdot F vect B}{|F_{vect_A}| * |F_{vect_B}|}$$

Where, the dot product of the two vectors, represented by F vect A. $Fvect\ B$, is equal to the sum of the ordered components. Zero value is given to the specific feature if no feature for the AML and ALL scenario can be located. This procedure aids in locating and choosing the most pertinent feature data needed for the classifiers covered in the following section to operate more effectively.

3.3. Classification Techniques

The three most widely used machine learning techniques for classifying leukemia gene expression data into ALL and AML categories are covered in this section. In their paper on the improved survival rates of cancer patients with earlier clarification of anomalies, Kumar and Greiner [Kumar and Greiner, 2019] have emphasized the significance of gene expression profiling. Additionally, Radakovich and colleagues demonstrated the use of computational intelligence for AML prediction based on leukemia patients' gene expression profiles [Radakovich et al., 2020]. The current study assesses three machine learning techniques one at a time in order to categorize AML and ALL classes, which are detailed below.

3.3.1. Linear Discriminant Analysis (LDA)

Fisher first presented this supervised learning classifier in 1936 [Mao et al., 2005]. It is a widely used modified version of Fisher linear discriminate analysis in statistical research. With each class denoting identical covariance matrices, the data is linearly classified into two or more classes. The method is predicated on the notion of providing the greatest amount of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
Article Received: *25 February 2023* ***Revised:*** *12 March 2023* ***Accepted:*** *30 April 2023*

differentiation between the classes while demonstrating the least amount of division within them. It is symbolized by the relationship below.

$$X' = \sum (X_c - X_n) - 1/2 \, (X_c + X_n)' \sum (X_c - X_n) > th_{val}$$

where X stands for the data matrix that contains feature data related to the expression of genes in malignant and non-cancerous tissues. The mean vector of the cancer group is represented by Xc, while the mean vector of the normal or control group is represented by Xn. In this case, $\sum$ stands for the sample co-variance matrix, and X' is the transposed data matrix of X, with th serving as the decision-making threshold. If $thval = 0$, it indicates that the two classes have similar data; if $thval > 0$, it indicates the cancer class; if not, it is assigned to the normal class. Although LDA is thought to be the best method, it does not work well with data that has more dimensions than the number of samples. This leads to a within-class matrix merging, which is also known as a small sample problem. It is discovered that LDA is unable to distinguish between classes if there are non-linearly separable entities. In order to solve this issue, kernel functions are employed [Tharwat et al., 2017].

1.1.1. Support Vector Machine (SVM)

SVM is a widely used binary classifier that creates a hyperplane to divide the two classes. SVM, among other classifiers, aims to differentiate classes by maximizing the distance between the two classes being studied by generating the biggest separation margin [Hsieh et al., 2010]. Classification of cancer is important not only for distinguishing between healthy and malignant people, but also for extending the time it takes to diagnose asymptomatic cases. Based on the training data set, which consists of gene expression features supplied at the input stage, the classifier makes predictions about the leukemia classes. Below is a description of the procedures used to classify and differentiate the two classes.

Algorithm 1: SVM based classification

1. $data_{test}$ → test feature data
2. Output: ALL_{class} and AML_{class} // leukaemia classes
3. Initialize SVM parameters
4. $kernel_{fun}$ → kernel function
5. for_{each} i in $data_{test}$
6. Analyse feature attributes
7. if $data_{test}(i)$ $belongs\ to$ $ALL_{gene-exp}$ // if matches with gene expression of ALL type
8. Assign $Cat_1 \rightarrow ALL_{class}$
9. $Else$
10. Assign $Cat_2 \rightarrow AML_{class}$
11. End_{if}
12. End_{for}
13. Returns: Leukemia classes → ALL_{class} and AML_{class}

The above algorithm summarizes the process involved in the separating two types of gene expression profiles. The test and training data are used at the input stage. The gene expression of unknown microarray tests is tested using the learned support vector in order to classify them into two forms of leukemia, which are represented by $ALicas\ and\ AMLclass$. It has been noted that the training samples, also known as support vectors (SVs), determine the final classification strength of SVM. In order to ascertain the overall classification performance, these SVs are examined as crucial points since they support the border of decision making. This indicates that the classifier's final judgment is influenced by training samples, particularly in the current instance where there were few samples available for ALL or AML patients. When used to bigger datasets, such gene expression profiles, kernel functions can be time-consuming and susceptible to overfitting [**Cawley et al., 2010**].

3.3.2. Feed Forward Neural Network (FFNN)

A neural network is a multilayered architecture and is modified according the classification requirements. In a Feed Forward Neural Network (FFNN), the neural network is trained by feeding the inputs in the forward direction at the input layer in order to compute the activation value of every neuron involved in the propagation. It is followed by the hidden or the middle layer that involves applications of weights for training and classification purpose. In the output layer the neuron activation values are computed again and aggregated to provide yield the results through the output layer. The difference between the observed output and the desired output reflects the error that incorporates in the classification process. In the present work, FFNN are implemented for gene expression data. A comprehensive discussion pertaining to various aspects of neural networks namely network modelling, simulation and experimentation were illustrated by Prieto et al., and co- researchers discussing practical

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
***Article Received:** 25 February 2023 **Revised:** 12 March 2023 **Accepted:** 30 April 2023*

applications to deal with real world challenges [**Prieto et al., 2016**]. The algorithmic structure of the neural network involved in the present study is discusses below.

Algorithm 2: FFNN for Leukemia Classification

1. Input: : $data_{train}$ → trained feature data
2. $data_{test}$ → test feature data
3. Output: Leukemia classes → ALL_{class} *and* AML_{class}
4. Initialize Neural Network parameters
5. E → Epochs
6. N → Neurons
7. MSE, $Gradient$, $Mutation\ and\ Validation\ Points$ → performance parameters
8. $Levenberg\ Marquardt$ → Training technique
9. $Random$ → Data Division
10. For_{each} *i in range of* $data_{train}$
11. $if\ data\ belongs\ to\ ALL_{class}\ generate\ Cat_1$
12. $if\ data\ belongs\ to\ AML_{class}\ generate\ Cat_2$
13. For_{end}
14. $Net = Newff(data_{train}, Cat_{data}, N)$ // *call nn*
15. $Class_{data} = Train(Net, data_{train}, Cat_{data})$
16. $Classification\ of\ test_{data}$
17. $class = SIM(Class_{data}, data_{test})$
18. $If\ class\ belongs\ to\ Cat_1,\ Assign\ ALL_{class}$
19. $If\ class\ belongs\ to\ Cat_2,\ Assign\ AML_{class}$
20. $Return$: NN structure *as* ALL_{class} *and* AML_{class}

The overall processing in of the neural network involves the feeding of feature data representing leukaemia classes along with the test data in the input layer. This is followed by training of neural network is trained on two type of classes and two categories are constructed as Cat_1 representing ALL class and Cat_2 representing AML class. Based on this information weights are assigned and simulation analysis classifies the test data into either ALL_{class} or AML_{class}. The performance of the classifiers is evaluated to using performance parameters as discussed in next section.

3.4. Performance Evaluation

The performance of all the classifiers implemented in the study is evaluated in terms of performance parameters namely, True Positive Rate (TPR), False Positive Rate (FPR), Accuracy (ACC) and kappa coefficient (KC). These parameters are computed by using confusion matrix corresponding to True Positive (T_p), True Negative (T_n), false positive (F_p) and False Negative (F_n) recognitions **Maniruzzaman** had described these statistical parameters in details in his study focussing on gene expression analysis using machine learning approaches [**Maniruzzaman et al., 2019]**. The performance parameters used in present study are as follows:

3.4.1. Accuracy (ACC): It is defined as the proportion of the true outcomes against the total number of population used in the experimentation. It is computed using following expression:

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

3.4.2. True Positive Rate (TPR): it defines the proportion of the true positive predictions made against the total number of positive outcomes. It is also termed as sensitivity. It is computed as follows:

TPR= TP/(TP+FN)

3.4.3. False Positive Rate (FPR): It corresponds to the probability of wrongly rejecting the expected outcomes. It is the ratio of number of false negative outcomes to the total number of negative outcomes. Mathematically it is represented as follows:

$F + T$

FPR= FP/(FP+TN)

3.4.4. Kappa Coefficient (KC): It is a statistical parameter that reflects the inter-reliability of classes during classification. In other words, it measures how closely the classification outcomes match the ground truth or the labelled expression data. It was introduced by Jacob Cohen in 1960 as a statistical meta-analysis and estimation [**Cohen, 1960**].

$KC = P0 - Pe / 1 - Pe$ Where, P_o corresponds to the probability of relative observed agreement within classes and P_e

Corresponds to the hypothetical probability of agreement obtained randomly by chance.

4. Results

The present section evaluates the implemented combination of feature extraction techniques and the machine learning classifiers with an aim to identify best feature selection and classification set to offer highest classification performance for leukaemia classes. The parameters used in the evaluation as discussed in the last section are analysed in the successive subsections.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
Article Received: *25 February 2023* ***Revised:*** *12 March 2023* ***Accepted:*** *30 April 2023*

4.1. TPR comparison

Table 1 provides a summary of the observed TPR values attained by combining PCA, CC, and CoC with LDA, SVM, and FFNN. Between 50 and 500 samples of the records corresponding to gene expression data are taken into consideration in each case. Figures 3, 4, and 5 show the graphical comparison of each situation, namely the TPR attained by utilizing LDA, SVM, and FFNN with PCA, CC, and CoC.

Table 1: TPR comparison of LDA, SVM and FFNN

Feature Extraction Technique	Number of Samples	TPR		
		LDA	SVM	FFNN
PCA	50	0.588	0.582	0.593
	100	0.746	0.684	0.684
	150	0.784	0.795	0.754
	200	0.829	0.846	0.858
	300	0.857	0.889	0.898
	400	0.886	0.902	0.934
	500	0.905	0.909	0.966
CC	50	0.564	0.571	0.573
	100	0.657	0.685	0.672
	150	0.715	0.798	0.754
	200	0.841	0.846	0.864
	300	0.879	0.894	0.902
	400	0.897	0.903	0.941
	500	0.913	0.918	0.974
CoC	50	0.632	0.622	0.645
	100	0.664	0.635	0.672
	150	0.718	0.711	0.754
	200	0.854	0.867	0.863
	300	0.894	0.909	0.902
	400	0.905	0.913	0.941
	500	0.934	0.969	0.974

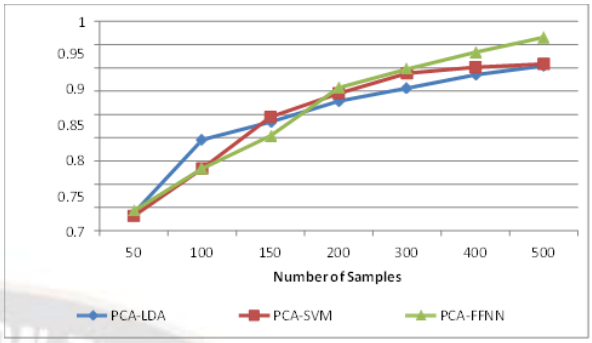

Figure 3 : TPR analysis involving PCA

Figure 3 shows the TPR comparison, when PCA based feature extraction is performed over the gene expression data set for leukaemia classification. The line graph represents the variations when number of samples on X-axis plotted against TPR values along Y-axis. It is observed that an average value of TPR when using PCA with LDA is 0.7992, PCA with SVM is 0.801 and PCA with FFNN is 0.8124. This means that the highest TPR is exhibited by when PCA is used in combination with FFNN as compared to SVM and LDA.

Similarly, TPR achieved when Canonical Correlation (CC) is used for feature extraction is shown in Figure 4. In this case, average TPR of 0.7808 is observed using CC with LDA, 0.8021 using CC-SVM and 0.8114 using CC with FFNN. However, the difference in the TPR in all the three scenarios is not very large; still it shows that FFNN demonstrated highest TPR due to being a multi layered architecture.

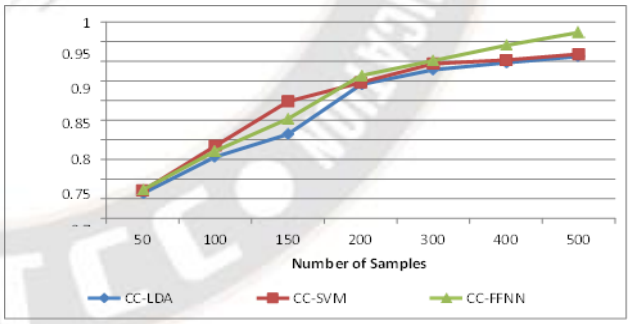

Figure 4: TPR analysis involving CC

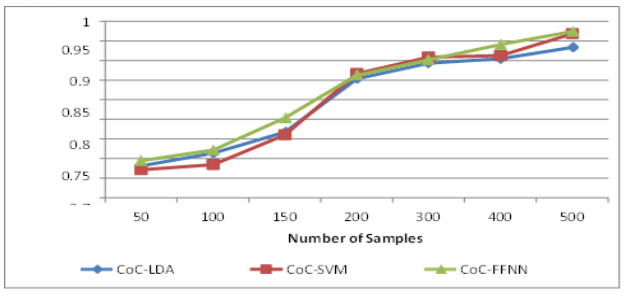

Figure 5: TPR analysis involving CoC

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
Article Received: *25 February 2023* ***Revised:*** *12 March 2023* ***Accepted:*** *30 April 2023*

The presented work also involved Cosine Correlation (CoC) based features extraction from the larger gene expression data. Figure 5 compares the TPR of the three combinations namely, CoC-LDA, CoC-SVM and CoC –FFNN. The line graphs shows that TPR increased with the increase in the number of samples with FFNN dominating the two other classification techniques in this case. An average TPR of 0.8001, 0.8037 and 0.8215 is observed using CoC-LDA, CoC-SVM and CoC–FFNN, respectively. Reflecting highest TPR demonstrated even when Cosine correlation is used for selecting features.

4.2. FPR comparison

The FPR values achieved when LDA, SVM and FFNN are used as machine learning classifiers were implemented in combination with feature extraction techniques is summarized in Table 2. The evaluation is performed against number of samples ranging from 50 to 500 for experimentation using different techniques for extracting features from gene expression data. Graphical comparison of these is further shown in Figure 6, Figure 7 and Figure 8.

Table 2 FPR comparison of LDA, SVM and FFNN

Feature Extraction	Number of Samples	FPR		
		LDA	SVM	FFNN
PCA	50	0.141	0.139	0.132
	100	0.128	0.133	0.125
	150	0.147	0.142	0.141
	200	0.166	0.172	0.157
	300	0.227	0.225	0.226
	400	0.287	0.279	0.264
	500	0.322	0.32	0.301
CC	50	0.158	0.152	0.132
	100	0.149	0.145	0.125
	150	0.195	0.167	0.141
	200	0.201	0.213	0.157
	300	0.229	0.225	0.226
	400	0.287	0.285	0.264
	500	0.312	0.299	0.291
CoC	50s	0.147	0.143	0.135
	100	0.175	0.184	0.122
	150	0.159	0.165	0.138
	200	0.199	0.218	0.159
	300	0.218	0.217	0.189
	400	0.254	0.238	0.264
	500	0.314	0.324	0.296

FPR analysis using PCA with LDA, SVM and FFNN is shown in Figure 6 with number of samples plotted along X-axis against FPR values along Y-axis. The line graph shows that FPR increases for large sample size for all the combinations with relatively lowest average FPR of 0.1922 demonstrated by PCA-FFNN followed by PCA-SVM of 0.2014 and PCA-LDA of 0.2025. This shows that PCA based feature extraction followed by FFNN exhibits least erroneous prediction using gene expression data.

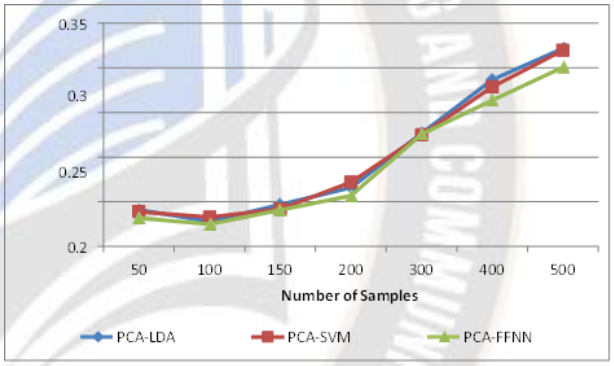

Figure 6 : FPR analysis involving PCA

Line graph in Figure 7 represents the FPR comparison of three classifiers when CC is used for extracting ALL and AML features from gene expression data. In this case average FPR using CC with LDA is 0.2187, CC with SVM is 0.2122 and CC with FFNN is 0.19085. This means that when CC is used for feature extraction FFNN again exhibits the least erroneous predictions. Following this, FPR of machine learning classifiers is further evaluated while utilizing CoC for extracting features of leukaemia classes from gene expression data. FPR values of the three combinations are graphical compared in Figure 8. It is observed that in this case too, FPR evaluation shows that when CoC based feature are classified using FFNN it exhibited least FPR of 0.1861 followed by Coc-SVM of 0.2127 and CoC-LDA of 0.2094. In other words, FPR analysis of the three scenarios shows that minimal erroneous classification is observed when FFNN is used as the classifier.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
Article Received: *25 February 2023* ***Revised:*** *12 March 2023* ***Accepted:*** *30 April 2023*

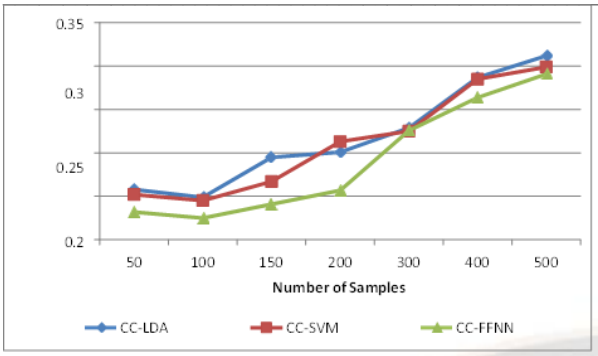

Figure 7 : FPR analysis involving CC

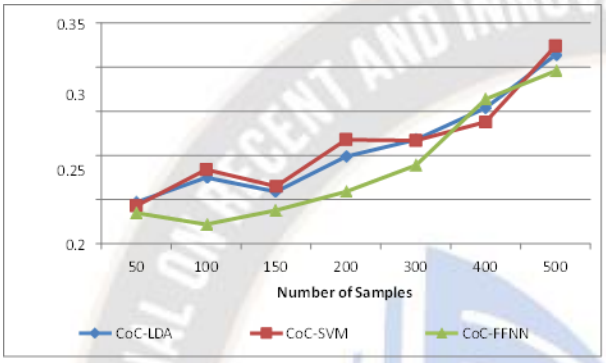

Figure 8 : FPR analysis involving CoC

4.3. Accuracy comparison

Accuracy of classification reflects the strength of the technique in predicting the true outcomes. Accuracy exhibited by each of the combination consisting of feature extraction technique and the classifiers is summarized in Table 3 for the number of samples varying from 50 to 500 for each case.

Table 3 : Accuracy comparison of LDA, SVM and FFNN

Feature Extraction	Number of Samples ACC	Accuracy (ACC)		
		LDA	SVM	FFNN
PCA	50	0.548	0.568	0.586
	100	0.564	0.557	0.531
	150	0.517	0.527	0.534
	200	0.645	0.597	0.618
	300	0.574	0.648	0.678
	400	0.679	0.675	0.694
	500	0.718	0.728	0.721
CC	50	0.512	0.524	0.516
	100	0.534	0.521	0.531
	150	0.501	0.512	0.528
	200	0.624	0.578	0.614
	300	0.562	0.624	0.654
	400	0.612	0.658	0.694
	500	0.681	0.701	0.703
CoC	50	0.501	0.513	0.516
	100	0.521	0.537	0.529
	150	0.512	0.512	0.528
	200	0.624	0.578	0.614
	300	0.542	0.624	0.654
	400	0.612	0.648	0.694
	500	0.671	0.704	0.715

Accuracy achieved when PCA is used as the feature extraction technique is shown in Figure 9. The line graph represents the accuracy graph exhibited by each classifier when used in combination with PCA. It is observed that highest average classification accuracy of 0.6231 is observed for PCA-FFNN followed by 0.6142 for PCA- SVM and 0.6064 for PCA-LDA despite of fluctuating accuracy in reference to sample size. This shows that highest PCA-FFNN based leukaemia classification is 0.885% and 1.671% higher than using PCA-SVM and PCA-LDA.

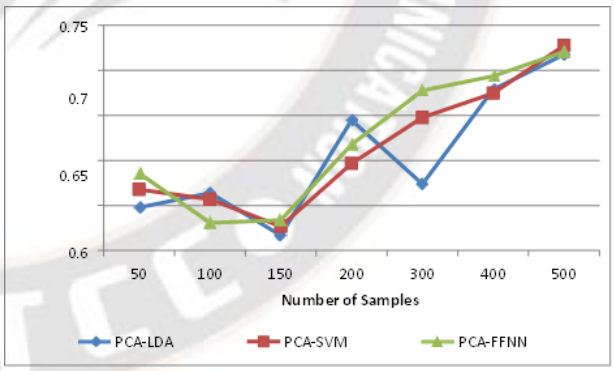

Figure 9: Accuracy analysis involving PCA

The next scenario comparing classifiers for leukaemia classification accuracy using CC for feature extraction is shown in Figure 10. In this case too, multiclass classifier FFNN outperformed the other classifiers namely SVM and LDA with an average accuracy of 0.6057 is observed for CC-FFNN, 0.5882 for CC-SVM and 0.5751 for CC-LDA. This reflects that classification accuracy using CC with FFNN is 1.742% and 3.057% higher than CC- SVM and CC-LDA.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
Article Received: *25 February 2023* ***Revised:*** *12 March 2023* ***Accepted:*** *30 April 2023*

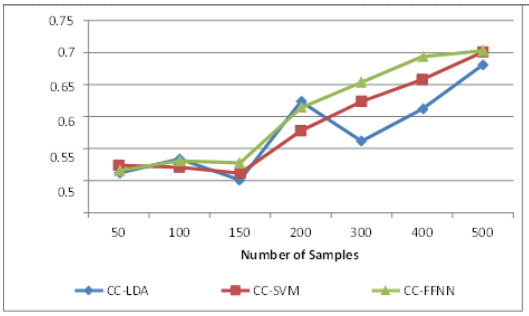

Figure 10: Accuracy analysis involving CC

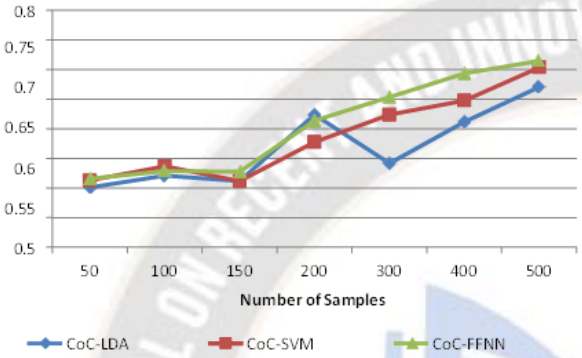

Figure 11 : Accuracy analysis involving CoC

Accuracy of leukaemia classification using CoC at feature extraction stage is shown in Figure 11. It is observed that accuracy values increases with the involvement of large number of samples for each of the classifiers. However, highest average classification accuracy of 0.6071 is observed for FFNN based classification followed by 0.588 for SVM and 0.569 for LDA over 500 samples which represents that implementation of FFNN for classifications improved the classification accuracy by 1.914% and 3.814% over SVM and LDA classifiers.

4.4. Kappa-Coefficient comparison

In statistical analysis, kappa coefficient is used to manage the occurrence of true predications that may arise due to chance factor. It is calculated by considering observed classification accuracy and the accuracy that is arises due to a change factor. The kappa coefficient observed for all the combinations used at feature extraction and classification stage is summarized in Table 4 to compare the effectiveness of the three classifiers.

Table 4 Kappa-Coefficient comparison of LDA, SVM and FFNN

Feature Extraction	Number of Samples	Kappa-Coefficient		
		LDA	SVM	FFNN
	50	0.0055	0.0057	0.0059
	100	0.0056	0.0056	0.0053
	150	0.0052	0.0053	0.0053
	200	0.0065	0.006	0.0062
	300	0.0057	0.0065	0.0068
	400	0.0068	0.0068	0.0069
PCA	500	0.0072	0.0073	0.0072
	50	0.0051	0.0052	0.0052
	100	0.0053	0.0052	0.0053
	150	0.005	0.0051	0.0053
	200	0.0062	0.0058	0.0061
	300	0.0056	0.0062	0.0065
	400	0.0061	0.0066	0.0069
CC	500	0.0068	0.007	0.007
	50	0.005	0.0051	0.0052
	100	0.0052	0.0054	0.0053
	150	0.0051	0.0051	0.0053
	200	0.0062	0.0058	0.0061
	300	0.0054	0.0062	0.0065
	400	0.0061	0.0065	0.0069
CoC	500	0.0067	0.007	0.0072

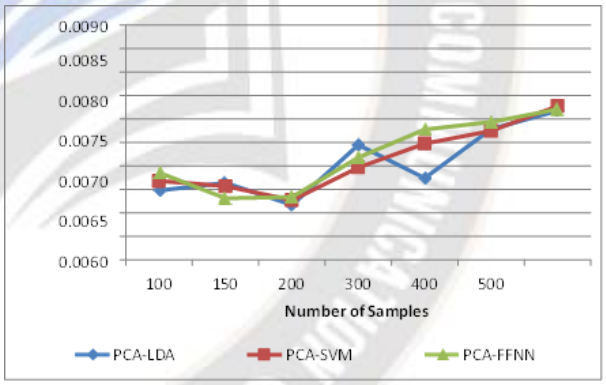

Figure 12 : Kappa-Coefficient analysis involving PCA

Kappa coefficient exhibited by the three classifiers when PCA is used at feature extraction stage is plotted together in Figure 12 in which number of samples are plotted on X-axis against the KC values on Y-axis. It is observed that an average KC of 0.0060 is demonstrated when LDA is used as the machine learning classifier. However, SVM and FFNN based classification achieved an average KC of 0.006142 and 0.00623, respectively. This means that highest KC is demonstrated when PCA is used with FFNN as compared to SVM and LDA.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
Article Received: *25 February 2023* ***Revised:*** *12 March 2023* ***Accepted:*** *30 April 2023*

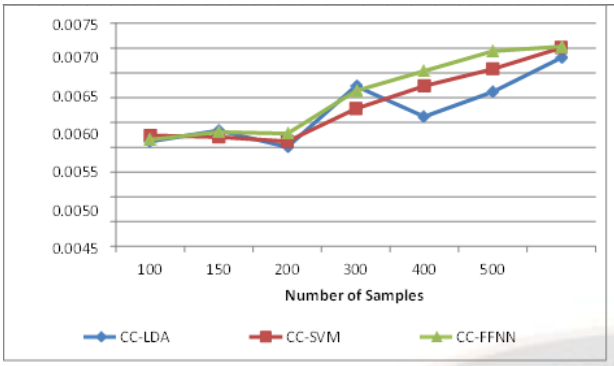

Figure 13 : Kappa-Coefficient analysis involving CC

Next, KC achieved by three machine learning approaches when CC is used at for feature extraction of leukaemia classes is compared in Figure 13. It is observed that an average KC of 0.00605 is observed for CC- FFNN, 0.0588 for CC-SVM and 0.0057 for CC-LDA. In other words, even when using CC at feature extraction stage FFNN exhibits highest KC values reflecting the effectiveness of multilayered classifier for leukaemia classification. Following this, KC of leukaemia classification of the three machine learning approaches is also compared when CoC is used at feature extraction stage. Figure 14 shows that using CoC with LDA exhibits an average KC of 0.00569 followed by CoC-SVM of 0.588 and CoC-FFNN of 0.00607.

Overall, it is observed that KC using CoC with FFNN demonstrated highest kappa coefficient among the three machine learning approaches.

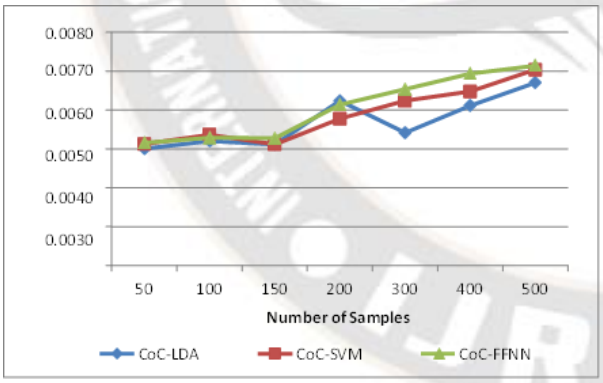

Figure 14 : Kappa-Coefficient analysis involving CoC

4.5. Overall Accuracy comparison of all scenarios

The last section showed the detailed performance comparison of the three machine learning classifiers in combination with three feature extraction techniques. The present section aims to determine the best combination of feature extraction and machine learning techniques to best performance for leukaemia classification based on gene expression data. The classification accuracies for all the combinations is summarized in Table 5 and Figure 16 represents it graphical analysis.

Table 5 Overall Accuracy comparison of various combinations

Feature Extraction Techniques	Machine Learning Classifiers		
	LDA	SVM	FFNN
PCA	0.569	0.588	0.607
CC	0.575	0.588	0.605
CoC	0.606	0.614	0.623

It is observed that classification accuracy increases when CC based feature extraction is performed as compared to CoC which is further increased when PCA is implemented at feature extraction stage. The graph reflects two interpretations; firstly, LDA and SVM based leukaemia classification exhibited least classification accuracy as compared to FFNN irrespective of the technique that is used for feature extraction from the gene expression data. Secondly, FFNN proved to be best classifier among the three machine learning approaches when PCA is used at feature extraction as compared to Canonical Coefficient and Cosine Coefficient. This shows that PCA- FFNN outperformed the other combinations due to involvement of multilayer and multiclass classifier as compared to the binary classifiers for microarray data analysis.

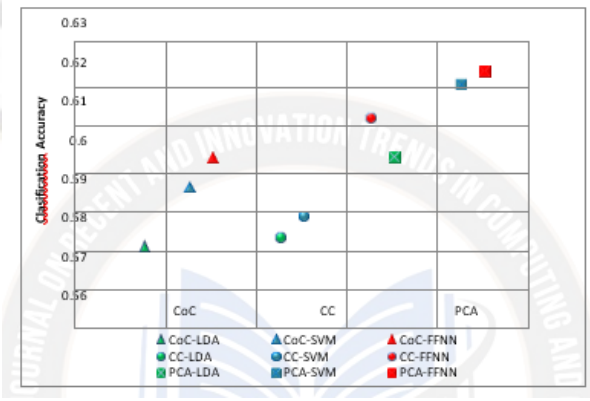

Figure 15: Overall Accuracy comparison of various combinations

1. Conclusion

The advent of microarray technology has emerged as a major advancement in biologic characterization of various diseases that even include identification of some unknown subtypes while leading to characterization of new ones too. The major challenge adjoining this technology is handling of high dimensional gene expression data which is costly and time-consuming while requiring expertise and centralized equipments. The present work aimed at identifying the best set of feature extraction and machine learning techniques to offer accurate leukaemia classification. The evaluation is performed over larger gene expression data records upto 500 samples with a combination of three techniques namely, PCA, CC and CoC at feature extraction stage and three machine learning techniques namely, LDA, SVM and FFNN at classification stage. TPR, FPR, accuracy and kappa coefficient analysis shows that PCA with FFNN outperformed all other combinations implemented for leukaemia classification with an average accuracy of 0.6231 which is 0.88% and 1.67% higher than PCA with SVM and PCA with LDA due to involvement of multilayered neural networks for classification. In future, substantial research based on gene expression profile of leukaemia will be focused on the optimising the features to further improve the classification performance of machine learning techniques to reduce financial and computational cost.

References

1. Round, Thomas. "Primary care and cancer: facing the challenge of early diagnosis and survivorship." *European Journal of Cancer Care* 26, no. 3 (2017): e12703.
2. World Cancer Research Fund, “Worldwide cancer data”, last accessed on November 4, 2018

   https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data
3. Khwaja, Asim, Magnus Bjorkholm, Rosemary E. Gale, Ross L. Levine, Craig T. Jordan, Gerhard Ehninger, Clara D. Bloomfield et al. "Acute myeloid leukaemia." *Nature reviews Disease primers* 2, no. 1 (2016): 1-22.
4. Pui, Ching-Hon, Kim E. Nichols, and Jun J. Yang. "Somatic and germline genomics in paediatric acute lymphoblastic leukaemia." *Nature reviews Clinical oncology* 16, no. 4 (2019): 227-240.
5. Seshi, Beerelli. "Gene expression analysis of pluri-differentiated mesenchymal progenitor cells and methods for diagnosing a leukemic disease state." U.S. Patent 7,049,072, issued May 23, 2006.
6. Saadatpour, Assieh, Guoji Guo, Stuart H. Orkin, and Guo-Cheng Yuan. "Characterizing heterogeneity in leukemic cells using single-cell gene expression analysis." *Genome biology* 15, no. 12 (2014): 525.
7. Narrandes, Shavira, and Wayne Xu. "Gene expression detection assay for cancer clinical use." *Journal of Cancer* 9, no. 13 (2018): 2249.
8. Abdeldaim, Ahmed M., Ahmed T. Sahlol, Mohamed Elhoseny, and Aboul Ella Hassanien. "Computer- aided acute lymphoblastic leukemia diagnosis system based on image analysis." In *Advances in Soft Computing and Machine Learning in Image Processing*, pp. 131-147. Springer, Cham, 2018.
9. Goossens, Nicolas, Shigeki Nakagawa, Xiaochen Sun, and Yujin Hoshida. "Cancer biomarker discovery and validation." *Translational cancer research* 4, no. 3 (2015): 256.
10. GaneshKumar P, Aruldoss Albert Victore T, Renukadevi P, Devaraj D. Design of Fuzzy Expert System for Microarray Data Classification using a Novel Genetic Swarm Algorithm. Expert Systems with Applications 2012; 39-2, p. 1811-1812.
11. Taylor, Sandra L., and Kyoungmi Kim. "A jackknife and voting classifier approach to feature selection and classification." *Cancer informatics* 10 (2011): CIN-S7111.
12. Alirezanejad, Mehdi, Rasul Enayatifar, Homayun Motameni, and Hossein Nematzadeh. "Heuristic filter feature selection methods for medical datasets." *Genomics* 112, no. 2 (2020): 1173-1181.
13. Santhakumar, D., and S. Santhakumar. "Efficient attribute selection technique for leukaemia prediction using microarray gene data." *Soft Computing* (2020): 1-10.
14. Algamal, Zakariya Yahya, and Muhammad Hisyam Lee. "A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification." *Advances in data analysis and classification* 13, no. 3 (2019): 753-771.
15. Sheikhpour, R., and M. Aghaseram. "Diagnosis of acute myeloid and lymphoblastic leukemia using gene selection of microarray data and data mining

algorithm." *Scientific Journal of Iran Blood Transfus Organ* 12, no. 4 (2016): 347-357.

16. Alshamlan, Hala M., Ghada H. Badr, and Yousef A. Alohali. "Abc-svm: artificial bee colony and svm method for microarray gene selection and multi class cancer classification." *Int. J. Mach. Learn. Comput* 6, no. 3 (2016): 184.
17. Dwivedi, Ashok Kumar. "Artificial neural network model for effective cancer classification using microarray gene expression data." *Neural Computing and Applications* 29, no. 12 (2018): 1545-1554.
18. Sharma, Rudrani, and Rakesh Kumar. "A Novel Approach for the Classification of Leukemia Using Artificial Bee Colony Optimization Technique and Back-Propagation Neural Networks." In *Proceedings of 2nd International Conference on Communication, Computing and Networking*, pp. 685-694. Springer, Singapore, 2019.
19. Aghamaleki, Fateme Shaabanpour, Behrouz Mollashahi, Mokhtar Nosrati, Afshin Moradi, Mojgan Sheikhpour, and Abolfazl Movafagh. "Application of an artificial neural network in the diagnosis of chronic lymphocytic leukemia." *Cureus* 11, no. 2 (2019).
20. Arif, Muhammad Azharuddin, and Zuraini Ali Shah. "Implementation of Statistical Feature Selection and Feature Extraction on Cancer Classification." *Academia of Intelligence Computing* 1, no. 1 (2020): 21-29.
21. Sarder, Md Alamgir, Md Maniruzzaman, and Benojir Ahammed. "Feature Selection and Classification of Leukemia Cancer Using Machine Learning Techniques.", (2020).
22. Santhakumar, D., and S. Logeswari. "Efficient attribute selection technique for leukaemia prediction using microarray gene data." *Soft Computing* (2020): 1-10.
23. Golub, Todd R., Donna K. Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P. Mesirov, Hilary Coller et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *science* 286, no. 5439 (1999): 531-537.
24. "Gene Expression Dataset", Accessed from https://www.kaggle.com/crawford/gene-expression
25. Arif, Muhammad Azharuddin, and Zuraini Ali Shah. "Implementation of Statistical Feature Selection and Feature Extraction on Cancer Classification." *Academia of Intelligence Computing* 1, no. 1 (2020): 21-29.
26. Adiwijaya, Wisesty UN, E. Lisnawati, A. Aditsania, and Dana S. Kusumo. "Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification." *Journal of Computer Science* 14, no. 11 (2018): 1521-1530.
27. Yan, Fangrong, Xiao Lin, and Xuelin Huang. "Dynamic prediction of disease progression for leukemia patients by functional principal component analysis of longitudinal expression levels of an oncogene." *The Annals of Applied Statistics* 11, no. 3 (2017): 1649-1670.
28. Knapp, Thomas R. "Canonical correlation analysis: A general parametric significance-testing system." *Psychological Bulletin* 85, no. 2 (1978): 410.
29. Won, Ji Hye, Mansu Kim, Jinyoung Youn, and Hyunjin Park. "prediction of age at onset in parkinson's disease using objective specific neuroimaging genetics based on a sparse canonical correlation analysis." *Scientific Reports* 10, no. 1 (2020): 1-12.
30. Fan, Ming, Zuhui Liu, Sudan Xie, Maosheng Xu, Shiwei Wang, Xin Gao, and Lihua Li. "Integration of dynamic contrast-enhanced magnetic resonance imaging and T2-weighted imaging radiomic features by a canonical correlation analysis-based feature fusion method to predict histological grade in ductal breast carcinoma." *Physics in Medicine & Biology* 64, no. 21 (2019): 215001.
31. Dubey, Vimal Kumar, and Amit Kumar Saxena. "A cosine-similarity mutual-information approach for feature selection on high dimensional datasets." *Journal of Information Technology Research (JITR)* 10, no. 1 (2017): 15-28.
32. Kumar, Luke, and Russell Greiner. "Gene expression based survival prediction for cancer patients—A topic modeling approach." *PloS one* 14, no. 11 (2019): e0224446.
33. Radakovich, Nathan, Matthew Cortese, and Aziz Nazha. "Acute myeloid leukemia and artificial intelligence, algorithms and new scores." *Best Practice & Research Clinical Haematology* (2020): 101192.

34. Mao, Yong, Xiaobo Zhou, Daoying Pi, Youxian Sun, and Stephen TC Wong. "Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection." *Journal of Biomedicine and Biotechnology* 2005, no. 2 (2005): 160.

35. Tharwat, Alaa, Tarek Gaber, Abdelhameed Ibrahim, and Aboul Ella Hassanien. "Linear discriminant analysis: A detailed tutorial." *AI communications* 30, no. 2 (2017): 169-190.

36. Hsieh, Sung-Huai, Zhenyu Wang, Po-Hsun Cheng, I-Shun Lee, Sheau-Ling Hsieh, and Feipei Lai. "Leukemia cancer classification based on Support Vector Machine." In *2010 8th IEEE International Conference on Industrial Informatics*, pp. 819-824. IEEE, 2010.

37. Cawley, Gavin C., and Nicola LC Talbot. "On over-fitting in model selection and subsequent selection bias in performance evaluation." *The Journal of Machine Learning Research* 11 (2010): 2079-2107.

38. Prieto, Alberto, Beatriz Prieto, Eva Martinez Ortigosa, Eduardo Ros, Francisco Pelayo, Julio Ortega, and Ignacio Rojas. "Neural networks: An overview of early research, current frameworks and new challenges." *Neurocomputing* 214 (2016): 242-268.

39. Maniruzzaman, Md, Md Jahanur Rahman, Benojir Ahammed, Md Menhazul Abedin, Harman S. Suri, Mainak Biswas, Ayman El-Baz, Petros Bangeas, Georgios Tsoulfas, and Jasjit S. Suri. "Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms." *Computer methods and programs in biomedicine* 176 (2019): 173-193.

40. Cohen, Jacob. "A coefficient of agreement for nominal scales." *Educational and psychological measurement* 20, no. 1 (1960): 37-46.