Machine Learning Model for Cancer Diagnosis based on RNAseq Microarray

Hanaa Torkey
dept. computer science and engineeing
Faculty of Eleronic Engineering, Menoufia University
Menoufia, Menouf
htorkey@el-eng.menofia.edu.eg

Mostafa Atlam
dept. computer science and engineeing
Faculty of Eleronic Engineering, Menoufia University
Menoufia, Menouf
mostafasami768@el-eng.menofia.edu.eg

Nawal El-Fishawy
dept. computer science and engineeing
Faculty of Eleronic Engineering, Menoufia University
Menoufia, Menouf
nelfishawy@hotmail.com

Hanaa Salem
Communications and Computers Engineering Department, Faculty of Engineering, Delta University for Science and Tecnology, Gamasa, Egypt.
hana.salem@deltauniv.edu.eg

Abstract— **Microarray technology is one of the most important recent breakthroughs in experimental molecular biology. This novel technology for thousands of genes concurrently allows the supervising of expression levels in cells and has been increasingly used in cancer research to understand more of the molecular variations among tumors so that a more reliable classification becomes attainable. Machine learning techniques are loosely used to create substantial and precise classification models. In this paper, a function called Feature Reduction Classification Optimization (FeRCO) is proposed. FeRCO function uses machine learning techniques applied upon RNAseq microarray data for predicting whether the patient is diseased or not.**

The main purpose of FeRCO function is to define the minimum number of features using the most fitting reduction technique along with classification technique that give the highest classification accuracy. These techniques include Support Vector Machine (SVM) both linear and kernel, Decision Trees (DT), Random Forest (RF), K-Nearest Neighbours (KNN) and Naïve Bayes (NB). Principle Component Analysis (PCA) both linear and kernel, Linear Discriminant Analysis (LDA) and Factor Analysis (FA) along with different machine learning techniques were used to find a lower-dimensional subspace with better discriminatory features for better classification. The major outcomes of this research can be considered as a roadmap for interesting researchers in this field to be able to choose the most suitable machine learning algorithm whatever classification or reduction. The results show that FA and LPCA are the best reduction techniques to be used with the three datasets providing an accuracy up to 100% with TCGA and simulation datasets and accuracy up to 97.86% with WDBC datasets. LSVM is the best classification technique to be used with Linear PCA (LPCA), FA and LDA. RF is the best classification technique to be used with Kernel PCA (KPCA).

Keywords— Cancer Classification, Diagnosis, Gene Expression, Gene Reduction, Machine learning.

I. Introduction

Cancer is considered to be one of the deadliest illnesses in the world. In 2016, it is estimated that more than 1,500,00 new cancer cases have been diagnosed only in the United States, so what about the whole world, and that nearly five hundred thousand people have died of the disease. [2]. By early diagnosis of cancer better clinical management for the patients could be facilitated. Machine learning and data mining are both widely used in many fields [3]. Machine Learning is a process of extracting the interesting knowledge or information from the available data. These techniques have been utilized to detect and model the treatment of various cancer conditions. Also, Machine learning tools have the ability to reveal the key features from complex datasets.

It is notoriously difficult and tedious task to extract the most significant and meaningful information from the high dimensional data. The curse of dimensionality is a common way to stigmatize the whole set of issues that are found in high-dimensional data analysis; find appropriate projections, choose meaningful dimensions, and get rid of noise, being just a few. This problem could be solved using dimensionality reduction [4]. By reducing the number of features to the most variance features which the classification is affected most. Generally, there are two reasons for using dimensionality reduction techniques; the first is for data compression and the second reason is for data visualization [5]. Data compression does not only allow us to compress the data but also using up less computer memory or disk space and speeding up the learning process. It would be nearly impossible for high dimensional data to be visualized, therefore by using dimensionality reduction techniques features would be reduced to the most related features to the classification problem. By reducing the number of features, it would be easier for the data to be visualized. Dimensionality reduction techniques could be divided into two major approaches named supervised and unsupervised.

During the usage of the unsupervised approach, classes' labels of the data are not needed such as Independent Component Analysis (ICA) and Principal Component Analysis (PCA), which is considered as one of commonly used dimensionality reduction techniques of unsupervised approach [6]. This approach is suitable for various applications such as visualization and noise removal.

Whereas in the supervised approach, the class labels of data are taken into consideration such as Mixture Discriminant Analysis (MDA) [7] and Linear Discriminant Analysis (LDA) which is considered as one of commonly used dimensionality reduction techniques of supervised approach and this approach is suitable for various applications such as bioinformatics, biometrics and chemistry.

Many dimensionality reduction techniques had been applied in this paper which include, Principle Component Analysis [8]. PCA allow mapping data into a coordinate system in an unsupervised way where components that have the highest variance between them is represented by the basis vectors, Linear Discriminant Analysis [3], and Factor Analysis (FA) [9]. LDA is like PCA, maximizing the separation among known categories is the main goal of LDA and this is done in a supervised way. The information about the correlations between observed features can be used by FA to reduce the dimensions. Using these reduction techniques would be problematic needing large computational resources and may not fit time requirements.

In this paper, machine-learning techniques are applied upon RNAseq dataset (gene expression data) of breast cancer for prediction whether the sample is diseased or not. These techniques include Linear SVM (LSVM), kernel SVM (KSVM), Decision Trees (DT), Random Forest (RF), K-Nearest-Neighbours (KNN) and Naïve Bayes (NB). Linear PCA (LPCA), kernel PCA (KPCA), LDA and FA along with different machine learning techniques were used for finding a lower-dimensional subspace with better discriminatory features to classify datasets for breast cancer that have a large number of genes and small samples.

This paper is organized as follows. In Section II, DNA microarray and gene expression and cancer classification problem are briefly introduced. Then cancer classification techniques are in section III. Then, recent research in microarray tumor classification is introduced in section IV. Section V describes, the proposed system and the proposed system workflow. A description of the datasets is introduced, performance metrics are described, and finally experimental results are illustrated, in section VI. At last, conclusion and future work are introduced in Section VII.

II. DNA Microarray and gene expression

Gene expression is a very important technology. All human cells carry the same number and the same type of genes, only the expression varies from a cell to another and from type to another during different developmental stages in response to environmental changes [10, 11]. Varying gene expression can be analysed using microarrays or what is called DNA chip. DNA chip is nothing but a simple slide where there is a simple region in between. All the genes of a particular organism are placed in that region organised in different grooves. Gene expression can be measured by the amount of proteins produced by DNA, but this can be very complex, so instead of proteins gene expression can be measured by the amount of mRNA.

By using microarray technology thousands of genes can be analysed simultaneously, so that a good chance has been provided for researchers in different fields especially in medicine. Using this technology makes it easier for patients to be categorized or classified. Microarray gene expression dataset is represented in a tabular form where a one particular gene is represented by a single row, a sample for each column and all features are represented numerically.

A. *The Cancer Classification Problem*

Many deaths are caused by cancer all over the world. the earlier the diagnosis is the more the chances of being cured. Many methods had been developed for earlier diagnosis and prediction of cancer. Machine learning techniques are commonly used by researchers for prediction and cancer diagnosis [12].

Accurate prediction and dealing with datasets with large number of features that is up to tens of thousands of features would be considered as a challenging task. Therefore, classification techniques along with dimensionality reduction techniques are used together for providing the most accurate prediction removing irrelevant features and using only features by which the classification is affected the most.

III. Classifiction techniques and dimensionality redction techniques

Figure 1 [13, 14] shows general approaches of machine learning. As shown in the figure machine learning techniques can be divided into two main approaches supervised learning and unsupervised learning. For supervised learning, it is divided into classification techniques, regression techniques and dimensionality reduction techniques. For unsupervised learning, it is divided into clustering techniques and dimensionality reduction techniques.

A. *Dimensionality Reduction Techniques*

1) Principle Component Analysis: It is a method by which the most variance features and the features that affects the result of the classification are extracted. It is one of the most popular dimensionality reduction techniques that is based on covariance matrix. It is considered as unsupervised learning. A lower dimensional surface onto which the data to be projected on condition minimizing the squared projection error is obtained.

Reduce from n-dimension to k-dimension: where n-dimensional vectors $x_1, \dots, x_p$or, equivalently, an n × p data matrix X, whose jth column is the vector x_j of features on the jth variable. The main purpose is to find linear combination of the columns of matrix X with maximum variance. Such linear combinations are given by equation (1) [8]:

$$\sum_{j=1}^{p} a_j x_j = Xa \qquad (1)$$

Where a is a vector of constants $a_1, \dots, a_p$.

Therefore, dimensionality reduction works so that: the variance retained is maximized and the least square reconstruction error is minimized. A principal component is considered as a linear combination of the original variables. Principal components are extracted in such a way that maximum variance in the dataset is explained by the first component. Then, the rest of variance in the dataset can be explained by the second component and isn't related to the first component. The variance that is not explained by the first two principal components is attempted to be explained by the third principal component and so on.

2) Linear Discriminant Analysis: LDA is like PCA, maximizing the separation among known categories is the main goal of LDA. LDA is considered as supervised learning

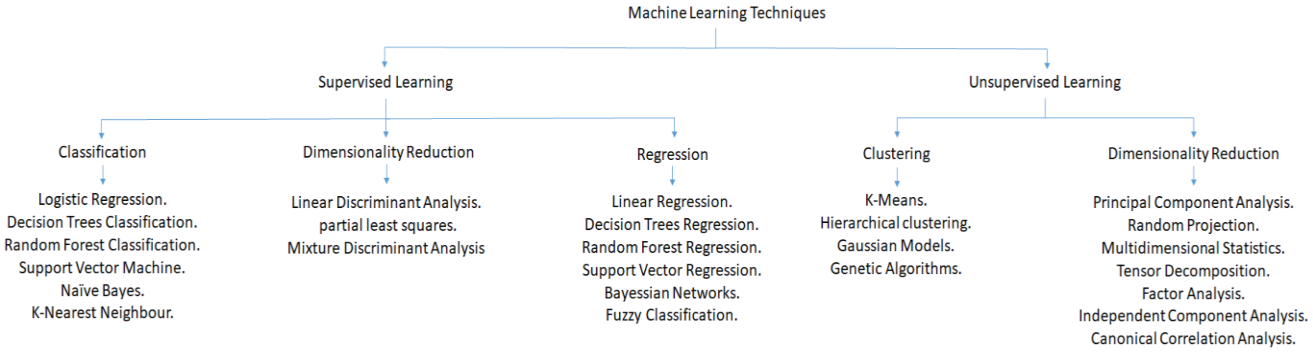

Figure 1: Machine learning techniques for classification and reduction

method. Therefore, a new dimension is picked by LDA based on: Maximizing separation between projected classes means in addition to minimal variance within each projected class. LDA is not assumed to be better for classification: it is assumed by LDA that classes are unimodal Gaussians and fails when discriminative information is in the variance of the data, and not in the mean.

3) Factor Analysis: In FA, features are grouped according to relationships between them, i.e. a high correlation will be found among all features in specific group, on the contrary a lower correlation with features from other groups. Each group is known as Factor these factors if compared to original dimensions would be smaller. FA is considered as unsupervised learning method.

B. Classification Techniques

1) Support Vector Machine: SVM is a very popular classification technique to deal with complex and non-linear problems by construction N-dimensional hyperplane [15]. Maximizing margin is considered as one of the main goals of SVM providing the best accuracy as much as possible. Margin can be defined as the separation area between two classes. It can be calculated by the distance between the closest point to the hyperplane and a data point on that plane. For maximizing the margin, the function implemented by equation 2 [15] is attempted to be maximized by SVM classifier with respect to vector w and a.

$$L_p = \frac{1}{2} \| \bar{w} \| - \sum_{i=1}^{t} \alpha_i y_i \left(\bar{w} \,.\, \bar{x_i} + a \right) + \sum_{i=1}^{t} \alpha_i \qquad (2)$$

Where t is the number of training samples, α_i is the Lagrange multipliers, L_p is called the Lagrangian and i ranges from 1 to t representing non-negative numbers such that the derivatives of L_p with respect to α_i are zero. In this equation, w is the weight vector, and x is the input vector. The weight vector, w and constant a define the hyperplane.

2) K-Nearest Neighbours: The KNN is a similarity-based learning technique that uses the nearest neighbours to determine the class for the new data point by knowing the category for the most counted neighbours [16]. The nearest neighbours for the data points is determined traditionally using Euclidean distance and it is shown by equation 3 [17].

$$Y = d(p, q) = \sqrt{(x_1 - x_3)^2 + (x_2 - x_4)^2} \qquad (3)$$

Where (x_1, x_2) and (x_3, x_4) are the indices of data points p and q respectively.

There are distances rather than Euclidean distance that can be used with KNN to obtain the similarity look to table.

Table 1: Similarity measurements functions

Manhattan distance	$(x_{i1} - x_{j1}) + (x_{i2} - x_{j2}) + \ldots + (x_{in} - x_{jn})$ (4) Where $i = x_{i1}, x_{i2}, \ldots, x_{in}$ and $j = x_{j1}, x_{j2}, \ldots, x_{jn}$
Minkowski distance	$\mid (x_{i1} - x_{j1})^p + (x_{i2} - x_{j2})^p + \cdots + (x_{in} - x_{jn})^{\frac{1}{p}} \mid$ (5) Where $i = x_{i1}, x_{i2}, \ldots, x_{in}$, $j = x_{j1}, x_{j2}, \ldots, x_{jn}$ and p can be defined as the positive integer.
Jaccard coefficient	$J(A, B) = (A \cap B) / (A\, U\, B)$ (6) where A and B are documents

3) Naïve Bayes: Naïve Bayes is a well-known supervised learning classification technique that is well suited for textual data, where each feature corresponds to an observation for a particular word [18]. It can be considered as a pattern classifier that determine likelihood and class restricted likelihood. These classifiers are based on some probabilistic models of how the data in each class might have been generated. They are called ‘Naive’ because it is assumed that features are conditionally independent given the class. highly efficient learning and prediction can be provided by Naïve Bayes so a competitive performance for high dimension datasets is provided, but generalization performance may be worse than algorithms that are more sophisticated. Predicting the class of a new data point corresponds mathematically to estimating the probability that each classes Gaussian distribution was most likely to have generated the data point. Classifier then picks the class that has the highest probability. So that, the conditional independence assumption is used and the probability of observed data (likelihoods) $p\,(x_i / c)$ are independently identified. The new instance class is computed using equation 7 [19] as follows:

$$p(f_1, f_2, \ldots f_n / c) = p(f_1 / c) \,.\, p(f_2 / c) \,.\, p(f_n / c) \qquad (7)$$

Where $f_1, f_2, \ldots f_n$ are the features, c is the class label and $p\,(f_i / c)$ is the likelihood.

The correct class label C_{NB} is identified by a Naïve Bayes classifier using the equation 8 [20]:

$$C_{NB} = arg_{c \in C} \max p(c) \prod_{f \in F} P(f / c) \qquad (8)$$

4) Decision Tree: DT is one of the most popular techniques, dealing with both classification and regression problems. It can be constructed as an acyclic graph. Human level thinking can be mimicked via decision trees so that understanding the data and making some good interpretations would be easier. Therefore, it is a tree-structured plan of a set of attributes to be tested so that the output would be predicted [20]. In Decision Tree technique each feature (attribute) is represented by a node, each decision (rule) is represented by a link (branch) and an outcome (categorical or continues value) is represented by a leaf. It is all about that, a tree like this is to be created for the entire data and a single outcome is processed at every leaf (or minimize the error in every leaf).

There are couple of algorithms that can be used for building a decision tree:

Classification and Regression Trees (CART) → uses Gini Index (Classification) as metric.

Iterative Dichotomiser 3 (ID3) → uses Entropy function and Information gain as metrics.

Therefore, for choosing the best attribute that can be used to classify the training data is the one with the highest information gain.

5) Random Forests: Random forest is considered as ensemble learning algorithm [16, 21]. In ensemble learning multiple models, such as classifiers, are combined in such a way that a particular computational intelligence problem is solved with increased accuracy. Large collections of uncorrelated decision trees are generated in order to solve a specific problem. Random forest is like decision trees as they are used for both classification and regression problems. However, random forest provides a good job in classification problems; they do not perform with the same quality in regression problems, as it does not give precise continuous nature predictions.

IV. RECENT RESEARCH IN MICROARRAY CLASSIFICATION TECHNOLOGY

It is essential to efficiently analyse DNA microarray data because the amount of DNA microarray data is usually very large. The analysis of DNA microarray data can be divided into four branches: clustering, classification, gene identification, and gene regulatory network modelling. Many machine learning and data mining techniques have been applied to resolve them. Table 2 shows a comparison of various area research for cancer diagnosis.

Table 2. Comparison of various earlier research

Author(s)	**Dataset**	**Methods**		**Accuracy (%)**	
		Feature Reduction	**Classifier**		
M. J. Rani and D. Devaraj [22]	Colon Cancer	MI-GA gene selection approach applied with 10 genes and 20 genes.	Kernel SVM	10 genes	20 genes
				96.77 %	100 %
	Lung Cancer			81.37 %	80.39 %
	Ovarian Cancer			98.43 %	99.21 %
B. Sahu, S. N. Mohanty and S. K. Rout [23]	Wisconsin Breast Cancer (WBC)	PCA	Random Forest	92 %	
			Artificial Neural Network	95 %	
M. Sadhana, A. Sankareswari and M.C.A., M.Phil. [24]	WBC	NONE	SVM	92.27 %	
			Decision tree	94.54 %	
	Wisconsin Diagnostic Breast Cancer (WDBC)		SVM	84.34 %	
			Decision tree	85.1 %	
	Wisconsin Prognostic Breast Cancer (WPBC)		SVM	79 %	
			Decision tree	82%	
H. Xie, J. Li, Q. Zhang and Y. Wang [25]	Breast Cancer- The Cancer Genome Atlas (BC-TCGA)	Random Projection (RP)	SVM	Training	Testing
				100 %	86.23 %
		Feature Selection (FS) + RP		99.95 %	98.97 %
		FS + RP +LDA		99.93%	98.67 %
		FS + RP + Post Feature Selection (PFS)		99.97%	98.95%
		RP +FS		100 %	92.70 %
		RP + PCA		100 %	76.70 %
		RP + LDA		98.68 %	98 %
	GSE2034 Wang et al. (2005)	RP		Training	Testing
				97.85 %	59.59 %
		FS + RP		99.88 %	61.25 %
		FS + RP +LDA		80.71 %	60.56 %
		FS + RP + PFS		99.99 %	61.55 %
		RP +FS		100 %	60.89 %
		RP + PCA		100 %	54.25 %
		RP + LDA		73.65 %	58.87 %
	GSE25066 Hatzis et al. (2011)	RP		Training	Testing
				100 %	66.90 %
		FS + RP		96.48 %	71.07 %
		FS + RP +LDA		87.81 %	69.56 %
		FS + RP + PFS		68.50 %	67.06 %
		RP +FS		99.98 %	68.65 %
		RP + PCA		100 %	60.44 %
		RP + LDA		80.23 %	69.21 %
Hanaa Salem, Gamal Attiya and Nawal El-Fishawy [26]	Breast Cancer	Information Gain (IG) Dynamic Threshold + Genetic Algorithm (GA)	Genetic Algorithm (GA)	Threshold	Accuracy
				0.2	99.94%

V. Proposed System

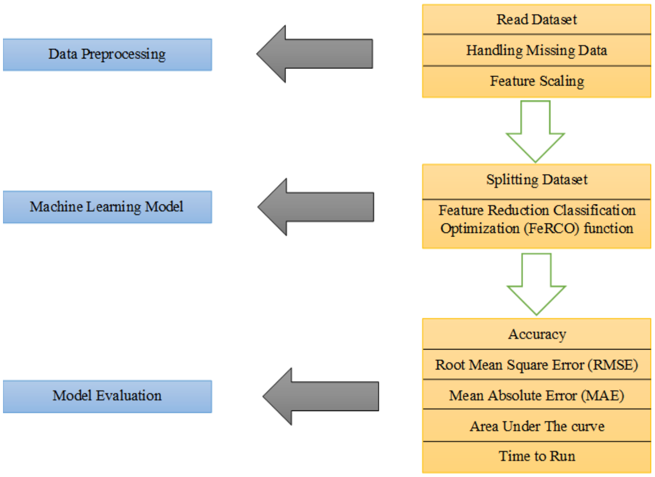

Figure 2: The general framework of proposed system

A. *Proposed Model Workflow*

The major purpose of this model is that after the dataset is being processed the dataset is passed to a function that apply different dimensionality reduction techniques along with different classification techniques. The main contribution in this paper is focused in designing a new function that apply different dimensionality reduction techniques with different classification techniques. The output of that function will define the minimum number of features with the most suitable reduction techniques and classification techniques that give the highest accuracy. Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

1) The first stage is data preprocessing: As shown in figure 2 the data preprocessing stage consists of four steps. The first one is to read the dataset to be classified. Then, searching for missing data that are non-numeric values and can be NaN or empty string to be replaced. Missing data can be handled by various ways but here missing data is replaced by the mean value of the column in which it resides, and this is done by using imputer class. Finally, the last step is that out of the range data is scaled and normalized so that all features would be in the same range ranging from -1 to 1. In order to scale data, the shape of the distribution is ignored, and the data is transformed in order to be centered, by removing the mean value of each feature, then scaling data by dividing non-constant features by their standard deviation. Here in this paper, scaling is performed via StandardScaler class. StandardScaler is a utility that uses the Transformer API to compute a training subset's mean and standard deviation to be able to reapply the same transformation on the test subset later.

2) The second stage is classificatio model: After the dataset is processed, the data is split to training data for training the classification model and testing data to be used in evaluating machine learning model. All of this is done using the train_test_split function in SKlearn cross validation by making random partitions for the two subsets one for training and the other for testing. Then, the training samples are passed to a Feature Reduction Classification Optimization function (FeRCO function).

- The range of components to be tested represented by two variables a and b.
- The step for moving between components represented by c, X_Train, X_Test, Y_Train and Y_Test indicating the sets to be tested are passed as input arguments to the function.
- The output of FeRCO function will define the minimum number of features represented by Number_Of_Components with the most suitable dimensionality reduction techniques and classification techniques that give the highest accuracy and represented by Techniques array.
- The highest accuracy is represented by Accuracy_max.
- An array of accuracy for each component for each dimensionality reduction technique along with each classification technique is provided by FeRCO function and represented by an array that is called array Accuracy.

Table 2: The pseudocode of FeRCO function.

```
Input: a, b, c, X_Train, X_Test, Y_Train and Y_Test.

Output: Accuracy array, Number_Of_Components array and Techniques
        array.
        A variable called Accuracy_max.

Steps:
  - Initialize an array called Accuracy.
  - Initialize an array called Number_Of_components.
  - Initialize an array called Techniques.
  - Initialize x_train. y_train, x_test, y_test.
  - x_train ← X_Train.
  - x_test  ← X_Test.
  - y_train ← Y_Train.
  - y_test  ← Y_Test.
  - For i = a to b, step = c do:
        1-  Apply the first reduction technique that is one of the four-
            reduction techniques previously stated in that paper with
            number of components equal to i.
        2-  Apply the six classification techniques previously stated
            in that paper.
        3-  For each classification technique along with the reduction
            technique accuracy is calculated.
        4-  The accuracy is appended to the Accuracy array.
        5-  The corresponding number of components and the
            reduction and the classification techniques are then
            appended to Number_Of_components Techniques arrays
            respectively
        6-  x_train ← X_Train.
            x_test  ← X_Test.
            y_train ← Y_Train
            y_test  ← Y_Test.
        7-  Apply the second reduction technique that is one of the
            four-reduction techniques previously stated in that paper
            with number of components equal to i.
        8-  Repeat from step 2 to step 6.
        9-  Apply the third reduction technique that is one of the
            four-reduction techniques previously stated in that paper
            with number of components equal to i.
        10-  Repeat from step 2 to step 6.
        11-  Apply the fourth reduction technique that is one of the
             four-reduction techniques previously stated in that paper
             with number of components equal to i.
        12-  Repeat from step 2 to step 6.
- End for
- Accuracy_max ← Search for the maximum value in Accuracy array.
- For j in accuracy range do:
     if a value in Accuracy array is equal to Accuracy_max
        Outputs the value of Accuracy_max.
        Outputs the value from Number_Of_components array of index j.
        Outputs the value from Techniques array of index j.
     End if
- End for.
```

3) The third and last stage is model evaluation: This stage is all about measuring Accuracy, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Area Under the Curve (AUC) and time to run.

VI. Experimental Results and Discussion

A. *Microarray Datasets*

The microarray dataset is arranged as an array. The rows of the array represent the features (genes) while the columns represent the instances. The microarray contains two arrays; one for training data and second for testing data. Dimensionality reduction techniques will be applied upon the training data array and the resulting decreased train subgroup will be used for training the proposed framework.

The testing data array is used for evaluating the proposed framework, by noting the number of test samples that the system will classify correctly.

All information about the three datasets is summarized in table 3. The first dataset is The Cancer Genome Atlas (TCGA), which consists of 17213 features and 482 sample, 61 of samples are benign and the other 421 are malignant. The second dataset is Wisconsin Diagnostic Breast Cancer (WDBC), which consists of 10 features and 699 sample, 459 of samples are benign and the other 240 are malignant. The third and last dataset is the simulation dataset, which consists of 10000 features and 200 sample, 100 of samples are benign and the other 100 are malignant.

Table 3: Dataset details

Datasets	Classes	Genes	Train Samples	Test samples
TCGA [27]	Two classes: benign or malignant	17213	337	145
WDBC [28]		10	559	140
Simulation Dataset [29]		10000	140	60

B. *Performance Metrics*

Table 4 summarizes the various performance metrics. The diagnostic implementation procedures and results are measured in contradiction of the following: P positive instances and N negative instances True Positive (TP): positive instances diagnosed correctly numbers. True Negative (TN): negative instances diagnosed correctly numbers. False Positive (FP): negative instances detected as positive numbers (Type I error). False Negative (FN): positive instances detected as negative numbers (Type II error).

Table 4: Diagnostic implementation procedures breast cancer

	Total Population	Condition Positive	Condition Negative	Total
Predicted Condition	Diagnostic positive	TP	FP	(TP + FP)
	Diagnostic negative	FN	TN	(TN + FN)
	Total	(TP + FN)	(TN + FP)	(TP+FN+ TN+FP)

These performance metrics are first computed and then used to compute Classification Accuracy (CA) of the algorithm according to equation (9) [26].

$$CA = \frac{No.\ of\ correct\ classified\ sampels}{Total\ no.\ of\ samples} = \frac{TP+TN}{TP+FN+TN+FP} \qquad (9)$$

Root Mean Square Error (RMSE): represents the standard deviation of residuals (i.e.: Here, the differences between the target variable to be predicted and the predicted variable). RMSE can be calculated by taking the root of Mean Square Error (MSE) by which the average of the squares of the errors is measured as shown in equation 10 [30].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{\sim})^2} \qquad (10)$$

Where y is the target variable to be predicted, $y^{\sim}$ is the predicted variable and n is the total number of samples ($TP + FN + TN + FP$).

Mean Absolute Error (MAE): is obtained by calculating the absolute difference between the target variable to be predicted and the predicted variable as shown in equation 11 [30].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^{\sim}| \qquad (11)$$

Area Under The curve (AUC): the whole two-dimensional area below the entire receiver operating characteristic (ROC) curve is calculated via AUC. ROC is a graph that show the true positive rate (sensitivity) plotting as a function of the false positive rate (1- specificity) for different cut off points of a parameter. AUC can be used as a measure of how well a parameter can distinguish between two groups (as in the used datasets diseased and normal cases).

C. *Experimental Results*

1) Evaluating TCGA Dataset: Figure 3 compares the accuracy of different dimensionality reduction techniques including PCA, LDA, and FA with different classification techniques applied on TCGA dataset. It has been found that the best accuracy can be obtained via LSVM and RF with LPCA and LSVM with FA up to 100% with MAE and RMSE up to 0.

The worst dimensionality reduction technique applied on TCGA dataset is KPCA and NB with accuracy up to 15.86%, MAE up to 0.841 and RMSE up to 0.917 and the number of features is ranged from 120 to 1000. For LDA, DT and RF have the same accuracy that is up to 98.62%, LSVM, KSVM and KNN have the same accuracy that is up to 99.31% and NB has accuracy that is up to 84.14% as shown in figure 3.

Table 5 shows MAE along with RMSE for the highest accuracy classification techniques with each dimensionality reduction technique applied on TCGA dataset. The highest accuracy with LPCA that is up to 100% can be obtained via RF with number of features equals 15 and LSVM with many numbers of features but the least number of features with highest accuracy up to 15. The best classification techniques to work with KPCA are NB with number of features up to 40 and RF with many numbers of features but the least number of features with highest accuracy that is up to 99.31% is 150. The best classification techniques to work with LDA reduction technique are LSVM, KSVM and KNN for all numbers of features and accuracy up to 99.31%. LSVM technique works well with FA method and having accuracy up to 100% with many numbers of features but the least number of features with the highest accuracy is up to 10.

Table 5: Evaluation of the best classification techniques working with each reduction technique applied on TCGA dataset

	LPCA	KPCA	LDA	FA
Classification Technique	RF and LSVM	NB and RF	LSVM, KSVM and KNN	LSVM
MAE	0	0.007	0.007	0
RMSE	0	0.083	0.083	0
Accuracy	100%	99.31%	99.31%	100%

2) Evaluating WDBC Dataset: Figure 4 compares the accuracy of different dimensionality reduction techniques including PCA (Kernel and linear), LDA, and FA with different classification techniques applied on WDBC dataset. For LPCA reduction technique, the best accuracy can be obtained via RF, LSVM and KNN classification techniques. While NB and LSVM work well with LDA. For FA method, the highest accuracy that is up to 97.86% can be obtained via RF, LSVM and KSVM.

The worst dimensionality reduction technique applied on WDBC dataset is KPCA along with DT with accuracy up to 91.29%, MAE up to 0.175 and RMSE up to 0.590 with number of features up to 5.

For LDA, LSVM and NB techniques have the same accuracy that is up to 97.86%, KSVM and KNN have the same accuracy that is up to 97.14%, DT has accuracy that is up to 94.29% and RF has accuracy that is up to 95%.

Table 6 shows MAE along with RMSE for the highest accuracy classification techniques with each dimensionality reduction technique. The best classification techniques to work with LPCA are KNN with number of features up to 3 ,8 and 9, LSVM with number of features up to 3 and RF with number of features up to 7 and accuracy up to 97.86%. While RF technique works well with KPCA with number of features up to 4 and accuracy up to 96.57%. It has been found that LSVM and NB are the best techniques to work with LDA with all number of features and accuracy up to 97.86%. LSVM, KSVM and RF are the best classification techniques to work with FA with number of features up to 5, 6 and 9 for the first technique, 9 for the second technique and 7 for the third technique and accuracy up to 97.86%.

Table 6: Evaluation of the best classification techniques working with each reduction technique applied on WDBC dataset

	LPCA	KPCA	LDA	FA
Classification Technique	KNN, LSVM and RF	RF	LSVM and NB	LSVM, KSVM and RF
MAE	0.043	0.069	0.043	0.043
RMSE	0.293	0.370	0.293	0.293
Accuracy%	97.86%	96.57%	97.86%	97.86%

3) Evaluating simulation Dataset: Figure 5 compares the accuracy of different dimensionality reduction techniques with different classification techniques applied on simulation dataset. It has been found that the best accuracy is obtained via LSVM with LPCA, KNN with KPCA. For FA, the highest accuracy is 100% and is obtained via LSVM, KSVM, DT, NB and RF with MAE and RMSE up to 0.

The worst accuracy for LPCA can be obtained with KSVM classification technique with number of features ranging from 10 to 1000. For KPCA, the worst accuracy can be obtained via RF with number of features up to 110 and KSVM with number of features ranging from 5 to 1000. For LDA, the worst accuracy can be obtained via NB, LSVM and KSVM for all number of features. That worst accuracy is up to 48.334% with MAE up to 0.51667 and RMSE up to 0.71880. For all dimensionality reduction techniques applied on simulation dataset after the number of features reaches 120 the accuracy still with the same value for most of classification techniques. For LDA, both DT and RF have an accuracy of 60%.

Table 7 shows MAE along with RMSE for the highest accuracy classification techniques with each dimensionality reduction techniques. It has been found that LSVM is the best classification technique to work with LPCA with many numbers of features but the least number of features with highest accuracy is 90 having MAE and RMSE up to 0 and accuracy up to 100%. It has been found that KNN is the best classification technique to work with KPCA with number of features equals to 2 having MAE and RMSE equal to 0 and accuracy up to 100%. It has been found that DT and RF are the best classification techniques to work with LDA for all number of features with MAE up to 0.4, RMSE up to 0.632 and accuracy up to 60%. The best accuracy that can be obtained with FA is 100% with LSVM with minimum number of features up to 5, KSVM with 25 features, DT with minimum number of features up to 25, NB with minimum number of features up to 10 and RF with minimum number of features up to 25.

Table 7: Evaluation of the best classification techniques working with each reduction technique applied on Simulation dataset

	LPCA	KPC A	LDA	FA
Classification Technique	LSVM	KNN	DT and RF	LSVM, KSVM, DT, NB and RF
MAE	0	0	0.4	0
RMSE	0	0	0.632	0
Accuracy%	100%	100%	60%	100%

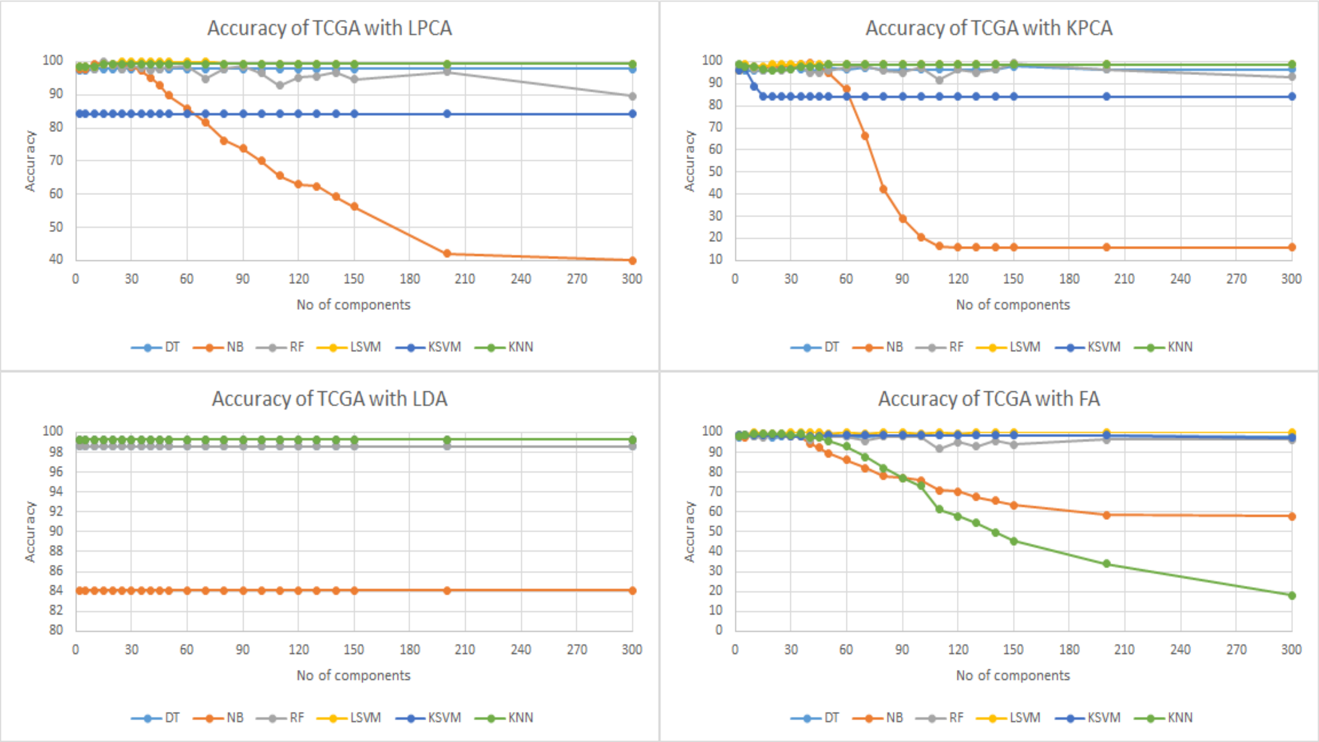

Figure 3: The accuracy of different classification techniques applied with each reduction technique on TCGA dataset

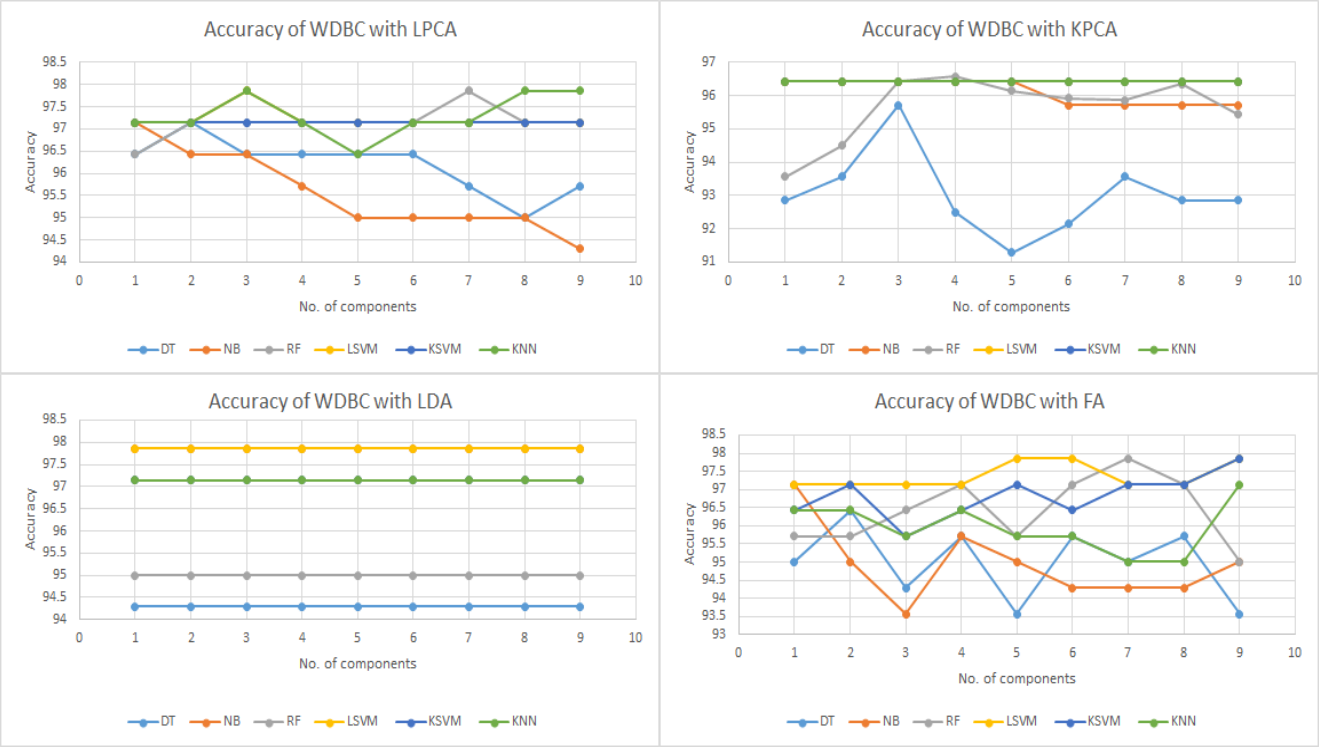

Figure 4: The accuracy of different classification techniques applied with each reduction technique on WDBC dataset

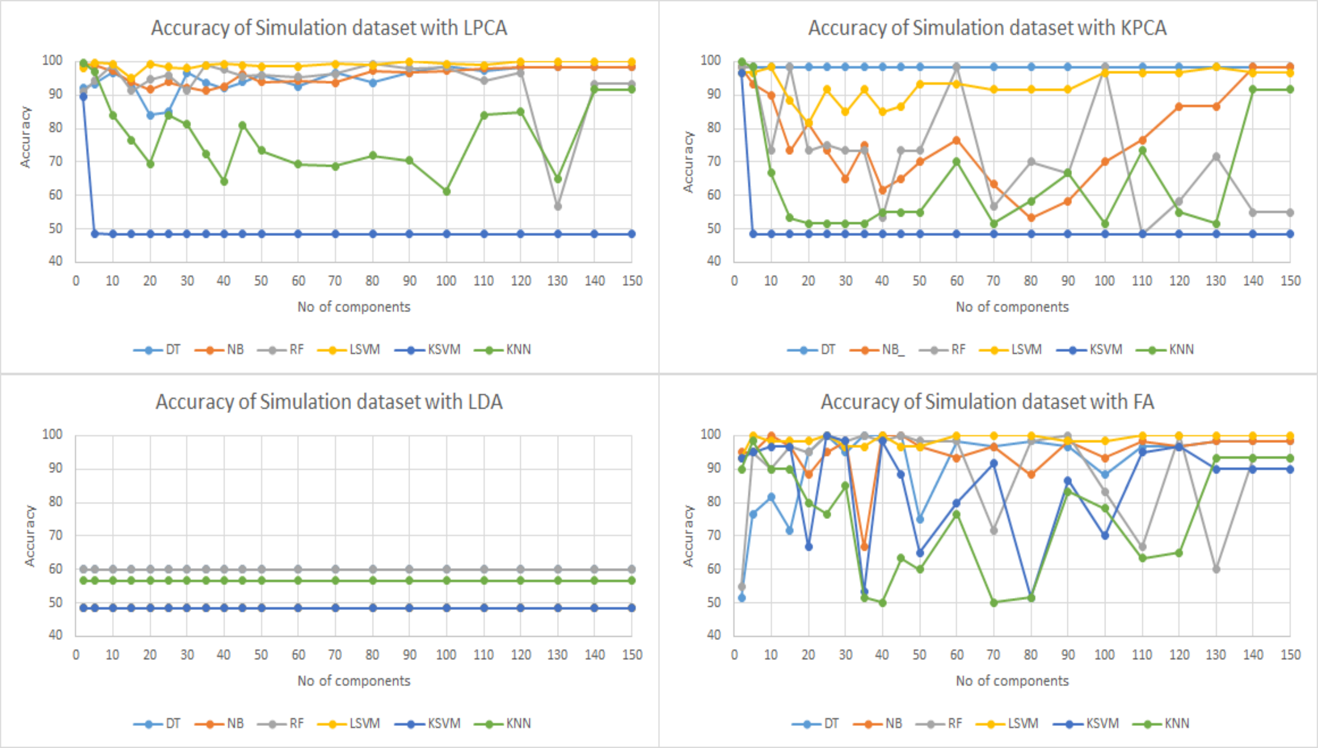

Figure 5: The accuracy of different classification techniques applied with each reduction technique on Simulation dataset

Table 8 shows the best accuracy for each dataset along with the corresponding classification technique with and without reduction techniques are grouped. As shown in the table the best accuracy with WDBC dataset is up to 97.86%, the best accuracy with TCGA dataset is up to 100% and the best accuracy with Simulation dataset is up to 100%.

As shown in Table 9, AUC and Accuracy for the highest accuracy classification technique with each dataset with and without reduction are represented. If there are many reduction techniques applied with different classification techniques on the same dataset and providing the highest accuracy, the one with minimum number of features is provided in table 9. If minimum number of features is the same, the one with the maximum AUC is added to table 9.

Table 9: AUC and accuracy for the best classification techniques with each Dataset with and without reduction

Classification Technique	Dataset	Reduction Technique	No of Features	AUC	Accuracy
KNN	WDBC	No Reduction Technique	All features	0.979	97.86%
RF	TCGA	No Reduction Technique	All features	1.0	100%
LSVM	Simulation Dataset	No Reduction Technique	All features	1.0	100%
LSVM & NB	WDBC	LDA	1	0.997	97.86%
LSVM	TCGA	Factor Analysis	10	1.0	100%
KNN	Simulation Dataset	Kernel PCA	2	1.0	100%

Table 10 shows the best reduction techniques for each classification technique that give a good accuracy in all datasets. As shown in the table, the best results are provided by LPCA and FA with most classification techniques. The highest accuracy is provided by LPCA when used with NB, RF, LSVM and KNN and the highest accuracy is provided by FA when used with DT, NB, LSVM and KSVM.

Table 10: The best dimensionality reduction techniques for each classification technique with all datasets

Classification Technique	The Best dimensionality Reduction Technique
DT	FA
NB	LPCA and FA
RF	LPCA
LSVM	LPCA and FA
KSVM	FA
KNN	LPCA

Table 11 shows a comparison between the run time for FeRCO function applied on the three datasets, TCGA, WDBC and simulation.

Table 11: The run time for FeRCO function applied on TCGA, WDBC and Simulation datasets

Dataset	Time to Run (Sec.)
TCGA	514.74
WDBC	0.28
Simulation	218.07

Table 12 shows the run time for the highest accuracy reduction techniques applied with each Classification Technique on the three datasets. If there are many reduction techniques applied with different classification techniques on the same dataset and providing the highest accuracy, the run time for the reduction technique applied with the classification technique with minimum number of features is provided in table 12.

Table 8: The highest accuracy for each dataset with and without reduction

Reduction Technique	WDBC		TCGA		Simulation Dataset	
	Classification Technique	Accuracy	Classification Technique	Accuracy	Classification Technique	Accuracy
Without Reduction	KNN	97.86%	RF	100%	NB	98.33%
LPCA	KNN, RF and LSVM	97.86%	LSVM and RF	100%	LSVM	100%
KPCA	All classification techniques except Decision Tree	96.43%	RF	99.31%	KNN	100%
LDA	LSVM and NB	97.86%	LSVM and KSVM and KNN	99.31	DT and RF	60%
FA	RF, LSVM and KSVM	97.86%	LSVM	100%	DT, RF, LSVM and KSVM	100%

Table 12: The run time for theHighest Accuracy Classification Technique with Each Dataset with and without Reduction

Dataset	Reduction Technique	Classification Technique	Number of Features	Time to Run (Sec.)
WDBC	No Reduction Technique	KNN	All features	0.052
TCGA	No Reduction Technique	RF	All features	2.56
Simulation	No Reduction Technique	LSVM	All features	2.69
TCGA	FA	LSVM	10	7.24
WDBC	LDA	LSVM	1	0.017
		NB	1	0.11
Simulation	KPCA	KNN	2	1.42

As shown in table 13, the model is evaluated using the Matthews correlation coefficient (MCC) and K-fold cross validation with k equals 10. MCC is a correlation coefficient value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction. Table 13 shows that the prediction accuracy is perfect. K-fold cross validation is considered as a resampling method for model evaluation. K-fold cross validation is performed by splitting the data into k random splits and here k equals 10 in order to evaluate how good is the prediction.

Table 13: MCC and 10-fold cross validation for best results in order to evaluate model

Dataset	Reduction Technique	Classificat-ion Technique	Number of Features	Cross Validation Accuracy	MCC	CA
WDBC	No Reduction Technique	KNN	All features	96.3%	0.96	97.86 %
TCGA	No Reduction Technique	RF	All features	99.2%	1	100%
Simulat-ion	No Reduction Technique	LSVM	All features	100%	1	100%
TCGA	FA	LSVM	10	99.6%	1	100%
WDBC	LDA	LSVM	1	96.72%	0.96	97.86 %
		NB	1	97.01%	0.96	97.86 %
Simulat-ion	KPCA	KNN	2	100%	1	100%

VII. CONCLUSION

DNA microarray and gene expression datasets contain too many features that is up to thousands of features. Thus, reducing features into low-dimensional subspace with better discriminative features by which the classification is affected the most is the main purpose in this paper. In this paper, a function that is called FeRCO function is developed to search through different dimensionality reduction techniques along with classification techniques producing the minimum number of features with the most suitable reduction and classification techniques that give the highest accuracy. The results show that FA and LPCA are the best reduction techniques to be used with the three datasets providing an accuracy up to 100% with TCGA and simulation datasets and accuracy up to 97.86% with WDBC dataset. LSVM is the best classification technique to be used with Linear PCA (LPCA), FA and LDA. RF is the best classification technique to be used with Kernel PCA (KPCA). For enhancing the results optimization techniques such as genetic and swarm techniques are intended to be used in future work.

REFERENCES

[1] R. Siegel, C. DeSantis, K. Virgo, et al., “Cancer Treatment and Survivorship Statistics”, CA: A Cancer Journal for Clinicians, Vol. 62, No. 4, pp. 220-41, Jul-Aug 2012.

[2] “Cancer Statistics". National Cancer Institute. Retrieved 2016-11-17.

[3] R. Duda, P. Hart and D. Stork, “Pattern Classification, 2nd Edition”, Wiley, 2012.

[4] C. Burges., “Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, chapter Geometric Methods for Feature Selection and Dimensional Reduction: A Guided Tour.”, Kluwer Academic Publishers, 2005.

[5] W. Müller, T. Nocke and H. Schumann, “Enhancing the visualization process with principal component analysis to support the exploration of trends”, Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation, Australian Computer Society, Vol. 60, pp. 121–130, Inc., 2006.

[6] A. Hyvärinen, J. Karhunen and E. Oja, “Independent Component Analysis”, Wiley-Interscience Publication, Vol. 46, 2004.

[7] T. Hastie and R. Tibshirani, “Discriminant analysis by Gaussian mixtures”, Journal of the Royal Statistical Society. Series B(Methodological), Vol 58, No. 1, pp. 155–176, 1996.

[9] I. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments”, Philosophical Transactions A Mathematical, Physical And Engineering Sciences, Vol. 374, Issue 2065, pp. 202, 2016.

[9] C. Spearman, “General intelligence objectively determined and measured”, American Journal of Psychology, Vol. 15, No. 2, pp. 206–221, 1904.

[10] M. Fajila and N. Fasmie, “CWIG: Consecutive Wrappers for Informative Gene Selection from Microarray Analysis in Cancer Detection and Classification”, Current Genomics, 2019.

[11] R. Singh and M. Sivabalakrishnan, “Feature Selection of Gene Expression Data for Cancer Classification: A Review”, Procedia Computer Science, Vol. 50, pp. 52 – 57, 2015.

[12] K. Kourou, T. Exarchos, K. Exarchos, M. Karamouzis and D. Fotiadis, “Machine learning applications in cancer prognosis and prediction”, Computational and Structural Biotechnology Journal Vol. 13, pp. 8–17, 2015.

[13] P. Hall, J. Dean, I. Kabul, J. Silva, “An Overview of Machine Learning with SAS® Enterprise Miner™”, SAS Institute Inc., 2014.

[14] A. Dey, “Machine Learning Algorithms: A Review”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7, pp. 1174-1179, 2016.

[15] K. Rajput and B. Oza, “A Comparative Study of Classification Techniques in Data Mining”. International Journal of Creative Research Thoughts (IJCRT), Vol. 5, Issue 3, pp. 154-163, 2017.

[16] A. Kadhim, “Survey on supervised machine learning techniques for automatic text classification”, Artificial Intelligence Review, No. 1, pp. 273–292, 2019.

[17] K. Chomboon, P. Chujai, P. Teerarassamee, K. Kerdprasop, N. Kerdprasop, “An Empirical Study of Distance Metrics
for k-Nearest Neighbor Algorithm”, Proceedings of the 3rd International Conference on Industrial Application Engineering, pp. 280-285, 2015.

[18] S. Vani.M, S. Uma, Sherin.A and Saranya.K, “Survey on Classification Techniques Used in Data Mining and their Recent Advancements”, International Journal of Science, Engineering and Technology Research, Vol. 3, Issue 9, pp. 2380-2385, September 2014.

[19] P. Kaviani1, S. Dhotre, “Short Survey on Naive Bayes Algorithm”, International Journal of Advance Engineering and Research Development (IJAERD), Vol. 4, pp. 607-611, 2017.

[20] E. Zimányi, R. Kutsche, “Business Intelligence: 4th European Summer School, eBISS 2014, Berlin, Germany, July 6-11, 2014, Tutorial Lectures (Lecture Notes in Business Information Processing)”, Springer, Vol. 205, 2015.

[21] L. Breiman, “Random forests”, Machine Learning, Vol. 45, Issue 1, pp. 5–32, 2001.

[22] M. Rani and D. Devaraj, “Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification”, Journal of Medical Systems, Vol. 43: 235, Issue 8, pp. 1-11, 2019.

[23] B. Sahu1, S. Mohanty and S. Rout, “A Hybrid Approach for Breast Cancer Classificationand Diagnosis”, EAI Endorsed Transactions on Scalable Information Systems, Vol. 6, Issue 20, 2019.

[24] M. Sadhana, A. Sankareswari, M.C.A. and M. Phil., “A PROPORTIONAL LEARNING OF CLASSIFIERS USING BREAST CANCER DATASETS”, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.11, pp. 223-232, November 2014.

[25] H. Xiea, J. Lia, Q. Zhanga and Y. Wanga, “Comparison among dimensionality reduction techniques based on Random Projection for cancer classification”, Computational biology and chemistry, Vol. 65, pp. 165-172, 2016.

[26] H. Salem, G. Attiya and N. El-Fishawy, “Intelligent Decision Support System for Breast Cancer Diagnosis by Gene Expression Profiles”, 33rd NATIONAL RADIO SCIENCE CONFERENCE (NRSC) Arab Academy for Science, Technology & Maritime Transport, pp. 421-430, Feb 22- 25, 2016.

[27] http://www.cbioportal.org/study?id=brca_tcga_pub2015#summary

[28] https://archive.ics.uci.edu/ml/machine-learning databases/breast cancer-wisconsin/

[29]https://data.mendeley.com/datasets/v3cc2p38hb/1/files/892a740d-ed29-44a8-b591-8284e68fb9f6

[30] W. Wang and Y. Lu, “Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model”, IOP Conference Series: Materials Science and Engineering, Vol. 324, pp. 012049-012058, 2018