Classifying Cancer Patients Based on DNA Sequences Using Machine Learning

Article *in* Journal of Medical Imaging and Health Informatics · March 2019

DOI: 10.1166/jmihi.2019.2602

5 authors, including:

Fahad Hussain
Bahria University Karachi Campus
27 PUBLICATIONS **938** CITATIONS
SEE PROFILE

Umair Saeed
Sindh Madressatul Islam University
24 PUBLICATIONS **109** CITATIONS
SEE PROFILE

Ghulam Muhammad
Ghazi University
4 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Noman Islam
PAF Karachi Institute of Economics & Technology
89 PUBLICATIONS **1,284** CITATIONS
SEE PROFILE

Classifying cancer patients based on DNA sequences using machine learning

Fahad Hussain[1], Umair Saeed[2], Ghulam Muhammad[3], Noman Islam[4] and Ghazala Shafi Sheikh[5]
[1,2] Sind Madrassatul Islam University, Karachi
[3]Bahria University, Karachi
[4,5]Iqra University, Karachi
[1]fahadhussain1590@gmail.com,[2]umairsaeedmixit@gmail.com,[3]ghulammuhammad.bukc@bahria.edu.pk,[4]noman.islam@iuk.edu.pk,[5]ghazala.shafi@iuk.edu.pk

Abstract: **This paper analyzes the DNA sequences of cancer patients using machine learning techniques. Cancer is one of the most prevalent and painful diseases of modern era, in terms of yearly new cases and death rate. Every year, a huge population dies due to this deadly disease. Researchers from different domains are exploring the approaches to counter this disease. In recent years, machine learning has been used for analysis of DNA sequences of cancer patients. The authors of this paper believe that a machine learning solution provides a more elegant and effective solution to the identification of cancer patients. In the same direction, this paper analyzes the DNA sequences of cancer patients using various classification techniques. Based on this study, it can be concluded that cancer disease for people can be diagnosed easily based on their DNA sequences. Different classifiers have been used for the analysis purpose such as artificial neural network, k-nearest neighbors, decision trees, fuzzy classifier, Navies Bayes classifier, random forest and support vector machine. Based on the analysis of results, it has been found that these classifiers provide sufficient performance in terms of accuracy, recall, specificity and other parameters. The paper also highlights the best classifier in various scenarios and recommendations have also been provided.**

Keywords: cancer analysis, DNA, machine learning, classification, gene

1. Introduction:

Saving the life of patient is the biggest task for the doctor especially when the patient is suffering from chronic disease such as cancer. It is a heterogeneous disease posing serious challenges in early detection and management of the disease. Different research efforts are being done since last few decades to counter the diagnosis challenge. These techniques are based on medical science, engineering, psychology and information technology [12, 14]. This paper is focused towards diagnosing the cancer disease from DNA sequence. It provides an edge in treatment with early cancer classification of patients. This can be done at the gene level or by employing the imaging technology [1]. Different kinds of classification techniques based on classical statistics and machine leaning methods have been applied in cancer classification to detect the disease. This also includes further identifying the level of care need for the patient to recover. Classification of cancer disease is a non-trivial task due to a range of associated issues. First, the problem's data have very high dimensionally. The data normally comprises thousand to tens of thousands of genes. Second challenge is the difficulty in acquisition of dataset. Most of the publically available data contain less than 100 records [1]. Third challenge is that most of genes are irreverent for disease detection.

In this paper, different classification algorithms have been applied to classify the disease through genes. This includes support vector machine that is used to distinguish the two different kind of object using a hyper plane. Another classification approach we employed is Bayesian classifier that is used to find the probability of disease in the gene of DNA micro-array. To find the effected gene we also employ ANN Classifier. ANN is inspired from human neuron and consists of millions of neurons used to process and send signals in the form of electrical and chemical pulses. The paper also uses decision tree to predict the model from observation about an item (which is represented in the form of branches). Other useful classifiers which have been used in classification are fuzzy classifier, k-nearest neighbor and random forest classifier.

The paper has been organized into different sections as follows. First, an overview of related literature is provided in next section. Then the methodology is provided which is followed with the results and conclusion.

2. Literature Review

There has been plethora of research efforts being done on using machine learning for identification of diseases. For instance, [15] have used magnetic resonance imaging (MRI) for identification of Alzheimer's disease one to three years before the clinical diagnosis. The authors hypothesized that mild cognitive impairment is a transition phase during the progression to the disease. The authors employed a combination of semi-supervised and supervised learning techniques for identification of disease.

In [16], the authors have employed the convolutional neural network for multi-model disease risk prediction with the help of structured and unstructured data obtained live from hospitals of China. A convolutional neural network is a special type of neural network that is used for images. The authors primarily focused on the disease called cerebral infarction. An accuracy of 94.8% was obtained with a very good convergence speed. [17] employed machine learning to analyzed the 5-year mortality in patients undergoing angiography. [18] uses various learning models such as linear model, support vector machine, decision trees, neural network and random forest to predict the outcome of death, dialysis or doubling of serum creatinine based on the modification of diet. The focus of the study was patients having kidney disease. [19] uses wearable sensor data and applied machine learning models to measure the stage of Parkinson disease in patients.

There have also been a number of studies focused on analyzing cancer patients using machine and deep learning technology. These machine learning techniques can be primarily classified as clustering and classification [10]. Clustering is primarily concerned with grouping data that are similar to each other. Various techniques such as self organizing maps, hierarchical clustering and graph theory have been used in literature. In classification, various techniques such as neural network, k-nearest neighbor, decision trees and support vector machines can be used. The next section primarily discusses the techniques relevant to classification which is the core topic of discussion for this paper.

Several machine learning and data mining methods have been applied to solve the diagnosis problem of cancer patients [9]. An overview of these approaches to machine learning for cancer diagnosis has been provided in [12]. The authors have done a qualitative survey of research papers over the past five years and identified various studies such as those on cancer susceptibility, cancer recurrence and cancer survival.

[2] employed convolutional neural network on the MRI images of brain of cancer patients to accurately detect the disease in the patient. Different techniques (such as data augmentation and optimization of epochs) have been applied to address the overfitting problem of machine learning. In another approach, images of CT scan were analyzed using machine learning to detect bladder cancer [6]. The various morphological and texture features were extracted. 2-fold cross validation was performed. Different machine learning models were used such as LDA, SVM, ANN and RAF.

[20] have analyzed the complete blood count of patients to find the likelihood of colorectal cancer and need colonoscopy referral. [21] have used the Bayesian Hidden Markov Model and Gaussian Mixture clustering approach to model the DNA copy number change across the genome and hence the identification of cancer disease.

A novel approach based on deep learning was proposed in [3] for identification of skin cancer in patients. [4] attempted to analyze the effectiveness of machine learning algorithm for identification of lung cancer. The authors employed k-nearest neighbor, decision trees, support vector machine and Naïve Bayes algorithm. [5] provided a machine learning based approach for detection of breast cancer.

[13] provides an ensemble of classifiers based approach for detection of lung cancer in patients. In another research, machine learning models for breast cancer survival prediction has been provided in [8].

Table 1 provides a comparison of some of the ML approaches used for prediction of diseases. It has been identified that very few studies have analyzed the DNA sequences for identification of cancer disease. In addition there needs to be a thorough analysis of the multitude of machine learning models such as ANN, SVM, DT and Naiv Bayesian together to identify which learning model performs better for identification of cancer. Summarizing the discussion, identification of cancer based on machine learning has been an issue of prime concern, and an exclusive study is required to analyze them using the gene sequences. The next section provides a comparison of various classification techniques and validates that machine learning applied on gene sequences provides indeed a very high performance for detection of cancer.

Table 1: A comparison of ML techniques used for disease prediction/ classification

Approach	Dataset Type	Model	Disease	Results
Moradi et al. [15]	MRI image	LDS	Alzheimer's disease	Diagnosis 3 years earlier
Chen et al. [16]	Image	CNN	Cereberal infarction	94.8 % accuracy
Motwani et al. [17]	CCTA and clinical parameters	LogitBoost	Coronary disease	0.78 AUC
Khitan et al. [18]	MDRD study variables	SVM, DT, ANN	Kidney disease	66-77% accuracy
Kubota et al. [19]	Wearable sensor data	Qualitative Study	Parkinson's disease	Good
Sawant et al. [2]	MRI image	CNN	Brain Tumor	98.3%
Garapati et al. [6]	Morphological features	LDA, ANN, SVM, RAF	Urinary bladder cancer	Approx. 90%

Hornbrook et la. [20]	CBC	statistical detection modeling	Colorectal cancer	0.8 AUC
Manogaran et al. [21]	DNA sequences	HMM and GM	Cancer	Improved performance

3. Methodology

In order to analyze the DNA sequences of cancer patients, this paper employed the machine learning technique. The proposed work is based on supervised learning in which a set of training data $x_1, x_2, x_3, \ldots, x_n$ is passed to the classifier along with the result labels y. Here, x_i is the feature of dataset. The objective is to develop a mathematical model that adjusts its various parameters based on the training observations.

The data for the cancer patients DNA have been acquired from [11]. The gene's data contains 291 rows and 45 columns. The data is divided into different genes (1 to 45), and there are different gene sequences associated with each column. There are 5 output classes namely BRCA-1, KIRC-2, COAD-3, LUAD-4 and PRAD-5. These classes represent genes sequence for patients with five types of tumor.

In the next step, the paper applied various machine learning classifiers on the dataset. For this purpose, the models have been developed in the KNIME software. Figure 1 provides a glimpse of the experimental design used for classification purpose.

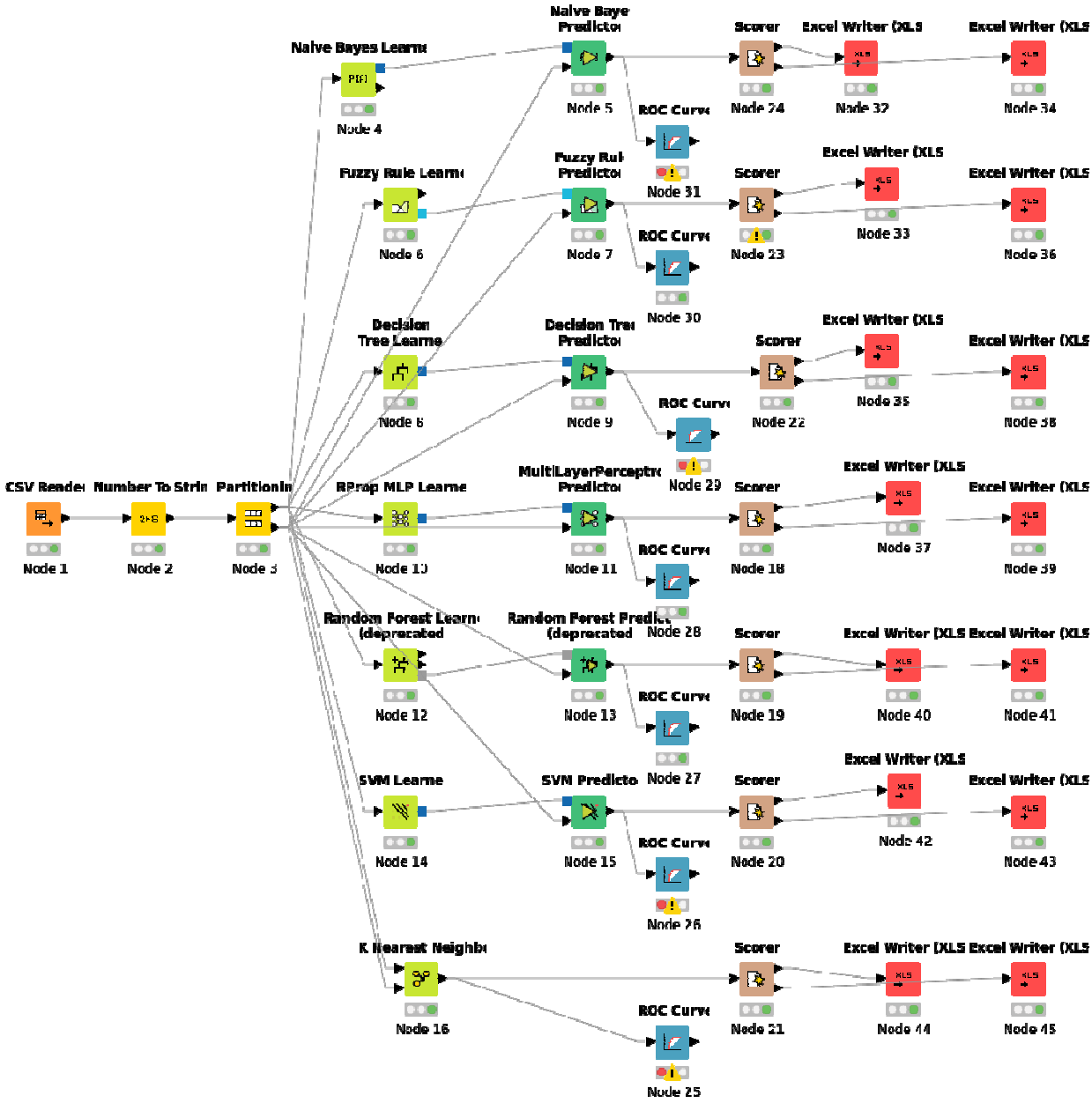


Figure 1: Experimental design for cancer classification

The data is read from the source file which is a comma separate values (CSV) file read via CSV reader, and data is converted into appropriate format for reading via different modules. The data is then passed through a partitioning module that splits the data into training and testing portion. The split ratio for training and testing is 80%-20% Different types of machine learning modules are then applied to the data. There is a scorer module to extract the results which are then written to an excel file using the Excel writer module. These results are then analyzed further to demonstrate the efficiency of various machine learning models.

The next section provides the results of the various classifiers applied for detection of cancer.

4. Results

Figure 2 shows the accuracy results of different classifiers for the five classes. All of the classifier provides reasonable accuracy greater than 90% in most of the cases. This proves the hypothesis that machine learning can be effectively used for detection of cancer disease. Figure 3 further compares the accuracy of various classifiers showing the average classification accuracy. It can be observed from the graph that random forest achieves the best accuracy for the classification of cancer DNA sequences. Figure 4 analyzes the true positive, true negative, false positive and false negative rates of various classifiers. The outcome predicted positively and correctly from the model is called true positive. Similarly, the outcome predicted negatively and correctly from the model is said to be a true negative. Also, the outcome predicted positively but incorrectly from the model is said to be a false positive. Similarly, the outcome predicted negatively and incorrectly from the model is said to be a false negative. The negative and false positive rates of all the classifiers are very low. This indicates that the margin of error in predicting a cancer disease as normal case and predicting a normal case as a cancer disease is very rare when machine learning is employed. This also asserts the hypothesis of using machine learning for predicting cancer disease.

Figure 5 summarizes the performance of classification's algorithm based on confusion matrix. Even though, accuracy is one of the main aspects of any classifier, it can often be misleading when you have unequal number of observations in each class. Calculating the confusion metric provide us about the better understanding of the result. An entry (i, j) represents a class i classified as belonging to class j by the classifier. The diagonal entry (i,i) in the figure therefore shows the correct classification made by a specific classifier. For instance a class BRCA is classified correctly 14 times, while it is incorrectly classified as COAD and LUAD one time each.

Figure 6 analyzes the sensitivity and recall values of the classifier. Analysis of the working performance of classifiers has been further performed using receiver operating characteristic (ROC) curve in Figure 7. In the graph two parameters curves are plotted i.e. true positive rate and false positive rate. True positive rate (TPR) or recall is defined as:

$$TPR = TP/TP + FN$$

False positive rate (FPR) is defined as:

$$FPR = FP/FP + TN$$

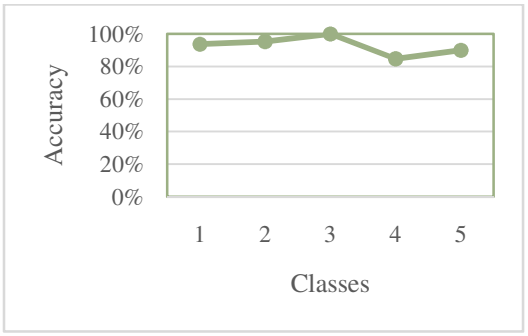
The ROC curve for the best classifier (random forest) is shown in Figure 7. It can be seen that ROC curve shows that performance of classifier is significantly above the baseline classifier.

a. ANN classifier

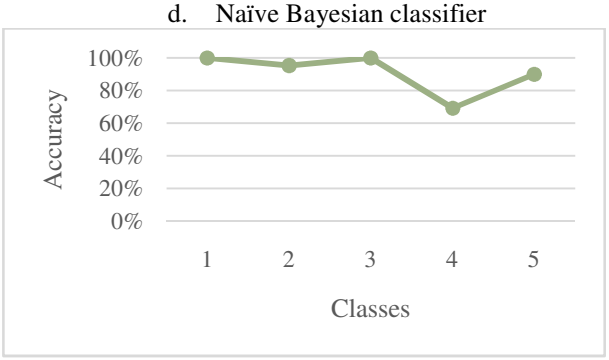
b. Decision Tree classifier

c. k-NN classifier

d. Naïve Bayesian classifier



e. Random forest classifier



f. SVM classifier

Figure 2: Results of accuracy of various classifiers

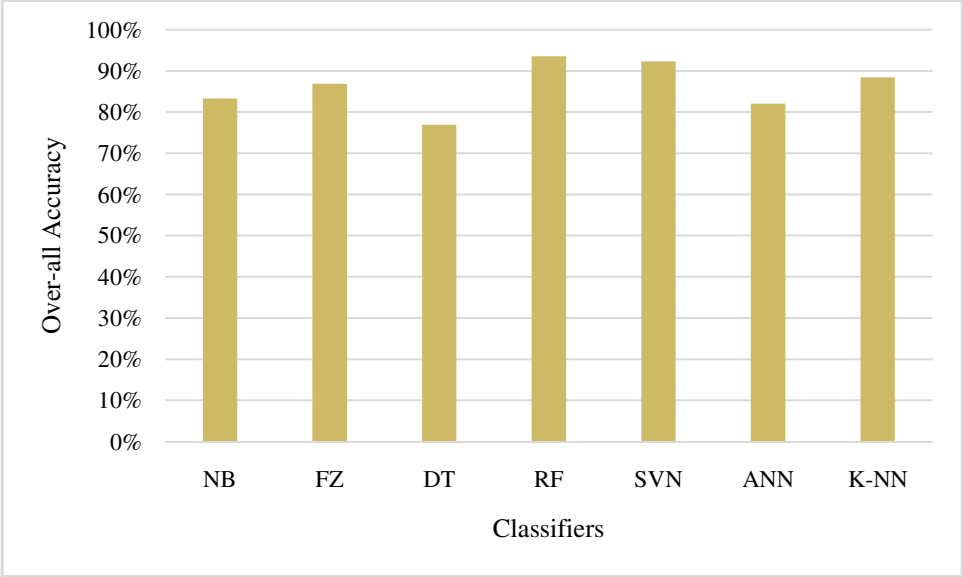
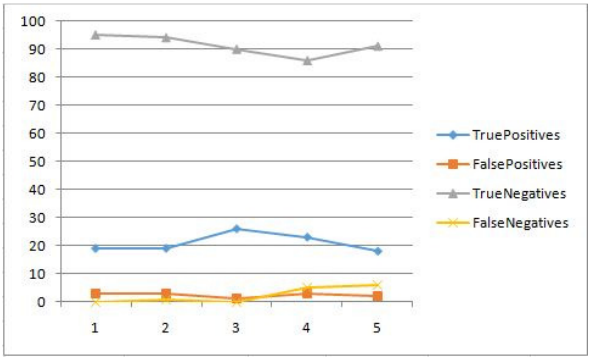
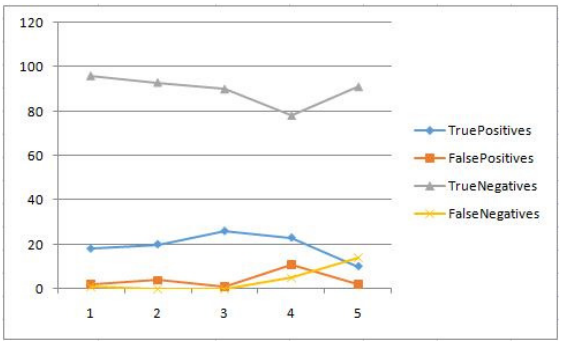


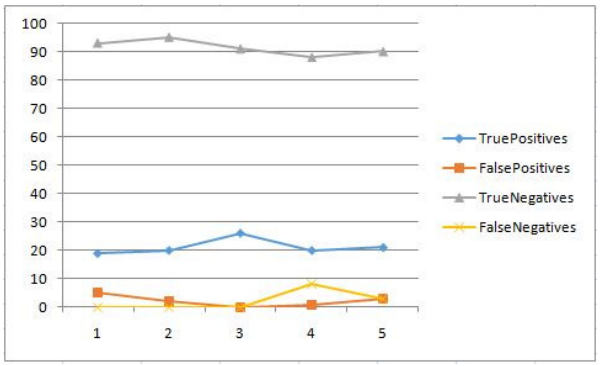
Figure 3: Comparison of accuracy of various classifiers



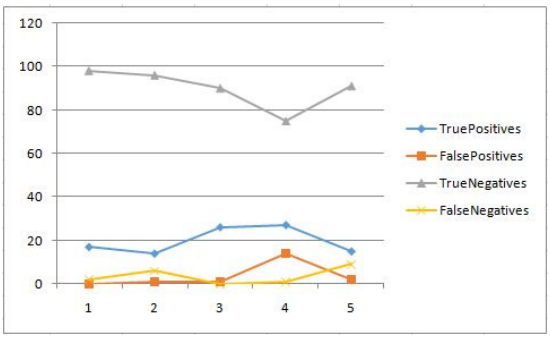
a. ANN classifier



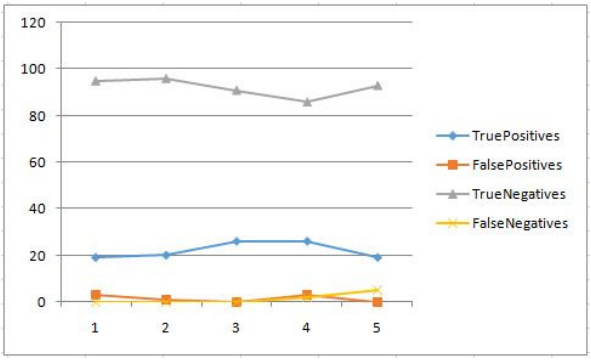
b. Decision Tree classifier



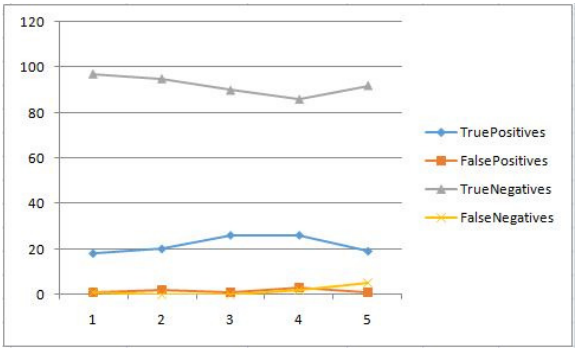
c. k-NN classifier



d. Naïve Bayesian classifier

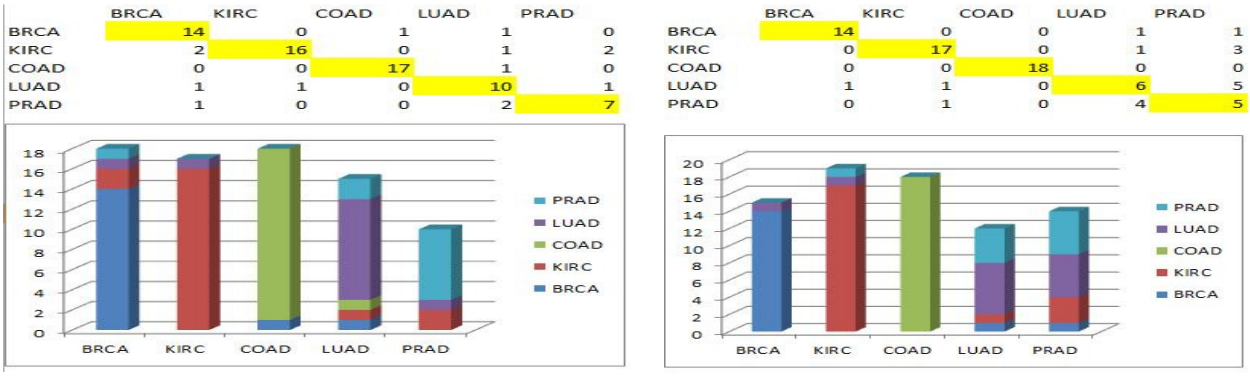


e. Random forest classifier



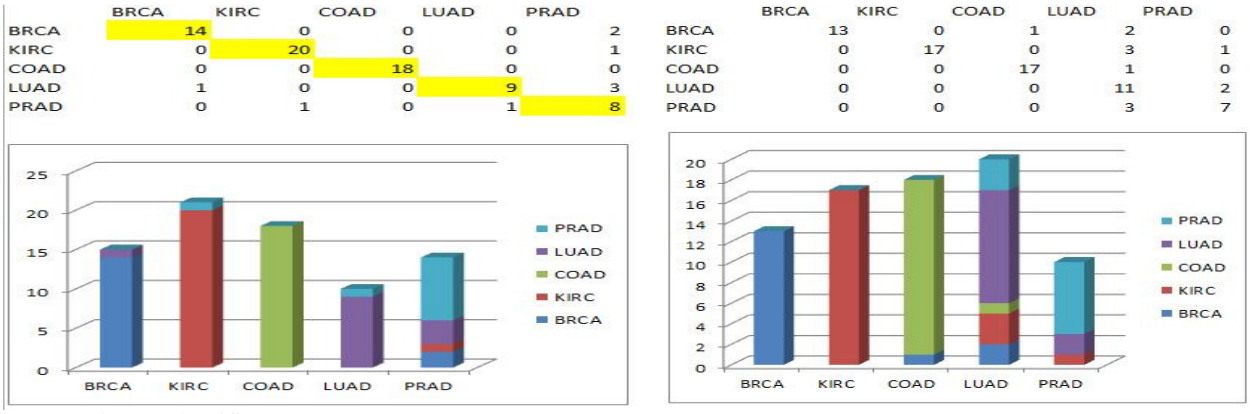
f. SVM classifier

Figure 4: TPR, TNR, FPR, FNR of the classifiers



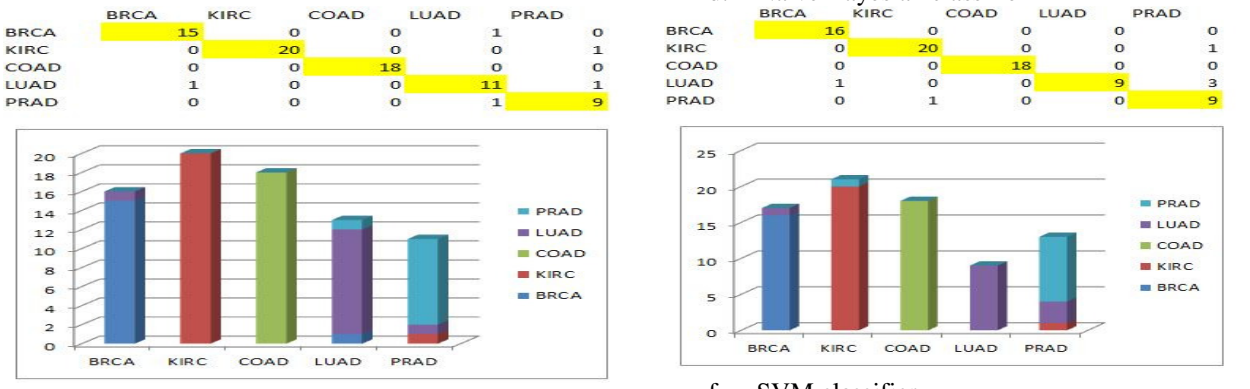
a. ANN classifier

b. Decision Tree classifier



c. k-NN classifier

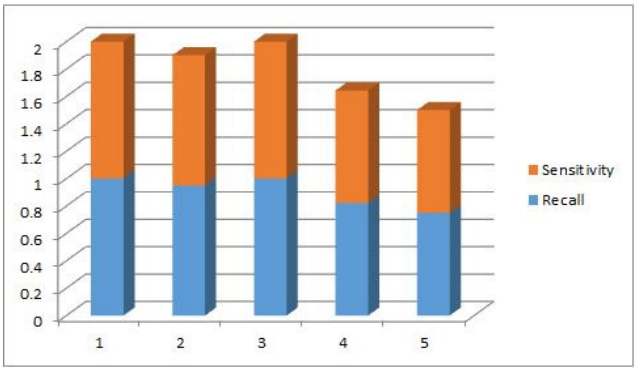
d. Naïve Bayesian classifier



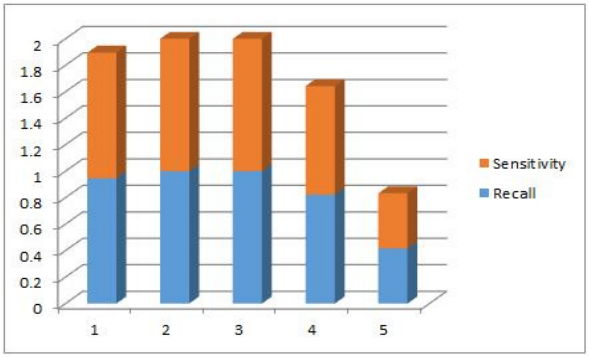
e. Random forest classifier

f. SVM classifier

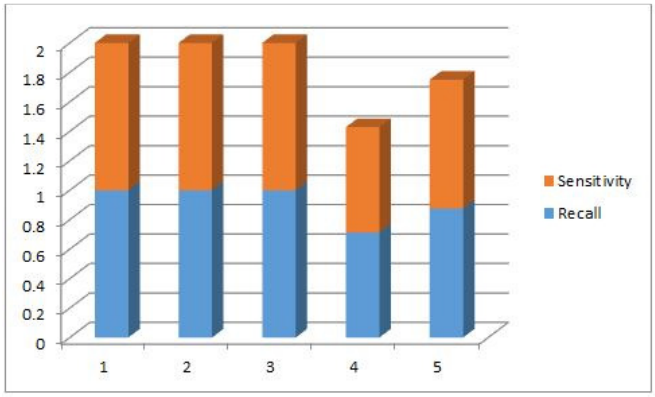
Figure 5: Confusion matrix of the classifiers



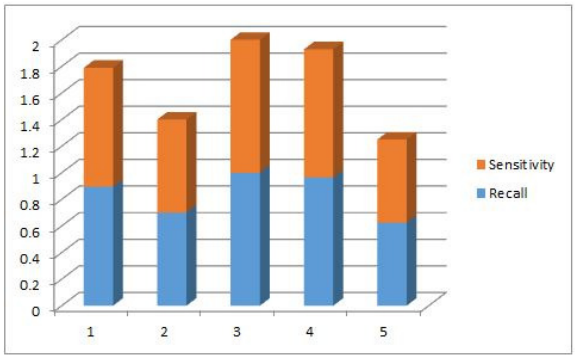
a. ANN classifier



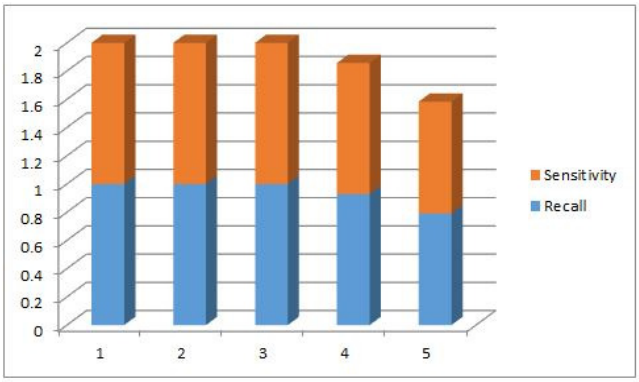
b. Decision Tree classifier



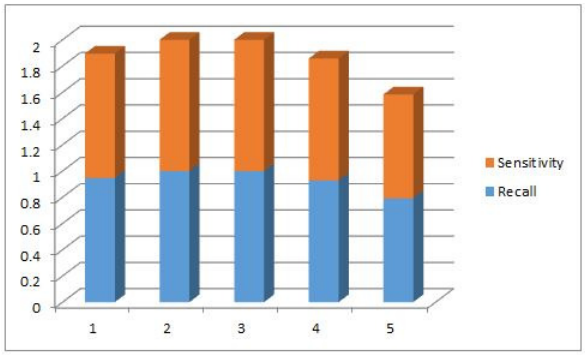
c. k-NN classifier



d. Naïve Bayesian classifier



e. Random forest classifier



f. SVM classifier

Figure 6: Sensitivity Vs Recall of classifier

row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Sensitivity	Specifity	F-measure	Accuracy	Cohen's kappa
1	19	3	95	0	1	1	0.969387755	0.926829268	0.863636364	
2	20	1	96	0	1	1	0.989690722	0.975609756	0.952380952	
3	26	0	91	0	1	1	1	1	1	
4	26	3	86	2	0.928571429	0.928571429	0.966292135	0.912280702	0.896551724	
5	19	0	93	5	0.791666667	0.791666667	1	0.88372093	1	
Overall									0.94017094	0.924910608

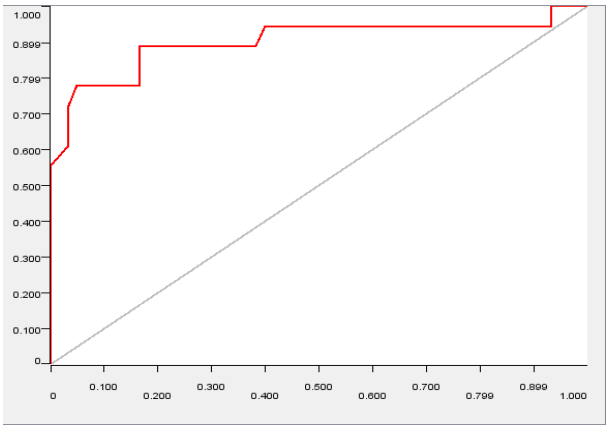


Figure 7: ROC curve for random forest.

5. Conclusion

This research employs the machine learning techniques to predict the cancer disease by using different state-of-the-art classifiers. Cancer is one of the chronic disease and death causing in most of the cases. We need intensive care of that effected person. By using DNA's gene and microarray, the paper predicts the cancer disease in the patient. By using gene dataset, the paper used different classifiers to analyze the accuracy. It is found that the random forest provide the best performance for the classification of cancer patients. The random forest actually employs an ensemble of classifiers for the learning and classification purpose. Hence the hypothesis space of the random forest is very profound and therefore outperforms other classifiers.

6. Acknowledgments

The authors would like to thank Sindh Madrassatul Islam University, Karachi and Iqra University, Karachi for their support in the accomplishment of this research.

References

[1] Ying LU, Jaiwei Han, "Cancer classification using gene expression data", Information System, 2003, vol. 28 (4), pp. 243-268

[2] Aaswad Sawant, Mayur Bhandari, Ravikumar Yadav, Rohan Yele,. Sneha Bendale, "Brain cancer detection from MRI: a machine learning approach (tensorflow)", International Research Journal of Engineering and Technology (IRJET), 2018, vol. 5(4)

[3]A Esteva, B Kuprel, RA Novoa, J Ko, SM Swetter, "Dermatologist-level classification of skin cancer with deep neural networks", Nature, 2017

[4] Maxim D Podolsky, Anton A Barchuk,, Vladimir I Kuznetcov, Natalia F. Gusarova1, Vadim S Gaidukov, Segrey A Tarakanov,"Evaluation of Machine Learning Algorithm Utilization forLung Cancer Classification Based on Gene Expression Levels", Asian Pacific Journal of Cancer Prevention, 2016, vol. 17, pp. 835-838

[5] Morteza Heidari, Abolfazl Zargari Khuzani, Alan B Hollingsworth, Gopichandh Danala, Seyedehnafiseh Mirniaharikandehei, Yuchen Qiu, Hong Liu1 and Bin Zheng, "Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm" ,Physics in Medicine & Biology, 2018

[6] Sankeerth S. Garapati, Lubomir Hadjiiski, Kenny H. Cha, Heang-Ping Chan , Elaine M. Caoili, Richard H. Cohan, Alon Weizer, Ajjai Alva Chintana, Paramagul Jun Wei Chuan Zhou,"Urinary bladder cancer staging in CT urography using machine learning", International Journal of Medical Physics Research and Particles, 2017

[7] Maxwell W. Libbrecht & William Stafford Noble g e, “Machine learning applications in genetics ann nbd genomics”, Nature Reviews Genetics, 2015, vol. 16, pp. 321-332

[8] Montazeri Mitraa, Montazeri Mohadesehc, Montazeri Mahdiehe, Beigzadeh Amin, “Machine learning models in breast cancer survival prediction”, Technology and Health Care, 2016, vol. 24(1), pp. 31-42

[9] Issam El NaqaRuijiang, LiMartin J. Murphy, Machine Learning in Radiation Oncology: Theory and Applications, 2015

[10] Sung-Bae Cho and Hong-Hee Won, “Machine Learning in DNA Microarray Analysis for Cancer Classification”, in proceedings of First Asia-Pacific bioinformatics conference on Bioinformatics, Adelaide, Australia, 2003

[11] “Gene expression cancer RNA-Seq Data Set”, retrieved from https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq, last accessed May, 2018

[12] KonstantinaKourou, Themis P.Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadisa, “Machine learning applications in cancer prognosis and prediction”, Computational and Structural Biotechnology Journal, 2015, vol. 13, pp. 8-17

[13] Zhihua Cai, Dong Xu, Qing Zhang, Jiexia Zhang, Sai-Ming Ngai and Jianlin Shao,“Classification of lung cancer using ensemble-based feature selection and machine learning methods”, Molecular BioSystems, 1(3), 2015

[14] Stefan Michiels, Serge Koscielny, Catherine Hill, “Prediction of cancer outcome with microarrays: a multiple random validation strategy”, Lancet, 2005, pp. 488–92

[15] Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, “Machine learning framework for early MRI-based Alzheimer’s conversion prediction in MCI subjects”, NeuroImage, 2014

[16] Min Chen , Yixue Hao , Kai Hwang , Lu Wang , Lin Wang, “Disease Prediction by Machine Learning Over Big Data From Healthcare Communities”, IEEE Access, 2017

[17] Manish Motwani, Damini Dey , Daniel S. Berman, Guido Germano , Stephan Achenbach et al., “Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis”, European Heart Journal, 2017, vol. 38(7), pp. 500-507

[18] Zeid Khitan, Anna P. Shapiro, Preeya T. Shah, Juan R. Sanabria, Prasanna Santhanam, Komal Sodhi, Nader G. Abraham, and Joseph I. Shapiro, “Predicting Adverse Outcomes in Chronic Kidney Disease Using Machine Learning Methods: Data from the Modification of Diet in Renal Disease”, Marshall Journal of Medicine, 2017, Vol. 3(4), pp. 68-80

[19] Ken J. Kubota, Jason A. Chen, Max A. Little, Machine learning for large-scale wearable sensor data in Parkinson's disease: Concepts, promises, pitfalls, and futures, Movement Disorders, 2016

[20] Mark C. Hornbrook, Ran Goshen, Eran Choman, Maureen O’Keeffe-Rosetti, Yaron Kinar, Elizabeth G. Liles, Kristal C. Rust, “Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data”, Digestive Diseases and Sciences, 2017, vol. 62(10), pp. 2719-2727

[21] Gunasekaran Manogaran, V. Vijayakumar R. Varatharajan, Priyan Malarvizhi Kumar, Revathi Sundarasekar, Ching-Hsien Hsu, “Machine Learning Based Big Data Processing Framework for Cancer Diagnosis Using Hidden Markov Model and GM Clustering”, Wireless Personal Communications, 2018, vol. 102 (3), pp. 2099-2116

[22] Konstantina Kourou, Themis P. Exarchosa, Konstantinos P. Exarchosa, Michalis V. Karamouzis, Dimitrios I. Fotiadisa, “Machine learning applications in cancer prognosis and prediction”, Computational and Structural Biotechnology Journal, 2015, vol. 13, pp. 8-17