International Journal of Research Studies in Science, Engineering and Technology
Volume 4, Issue 1, 2017, PP 41-49
ISSN 2349-4751 (Print) & ISSN 2349-476X (Online)

Performance Analysis of Machine Learning Algorithms on Lung Cancer Disease

Dr. N. Rajasekhar

Professor, Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad-500090

Abstract: *The study of Lung cancer identification is the most challenging problem in the medical era. The identification of Lung cancer was studied by collecting various risk factors on a series of clinical trials. As of late, the investigation of Lung disease expectation utilizing Machine Learning calculations, for example, Neural Network, Support Vector Machines, Extreme Gradient Boosting, Random Forest and Linear Discriminant Analysis are expanded. In this examination we utilized the assessment measurements as Accuracy, Precision, Recall and F1 Score and ROC of these calculations. Our experimental work is administered on the info set of UCI carcinoma(lung cancer) data. The experimental results demonstrate that XGB algorithm perform better than the remaining algorithms within the context of carcinoma(lung cancer) data.*

Keywords: *Extreme Gradient Boosting, Supervised learning, Neural Networks and Boosting.*

1. Introduction

Lung malignant growth might be a malady of uncontrolled cell development in tissues of the lung; carcinoma is one among the premier normal and destructive ailments inside the world. Identification of carcinoma in its beginning time is that the key of its fix. In diagnosing lung infections, specialists need to manage numerous challenges, the patient's side effects are normally not satisfactory; the similitudes in some lung malady's side effects are hard to separate. Specialists consistently need to test over and over before settling on a decision. In this way the finding result relies upon not exclusively patient's side effects yet additionally the specialists' encounters. Wrong choice methods wrong treatment and in this way the patient would endure more. Men are increasingly vulnerable to Lung, prostate, stomach, disease of the liver, while ladies are progressively helpless to bosom, colorectal, lung, cervix , and stomach malignancy.

Right now, there are no specialized techniques to stop carcinoma, which is the reason early recognition speaks to an extremely significant consider malignant growth treatment and permits fulfillment a high endurance rate. Clinical information preparing might be a promising region of computational knowledge applied to a consequently examine patients' records focusing on the development of most recent data possibly supportive for clinical dynamic. Induced data is predicted not only to grow accurate finding and successful ailment treatment, yet notwithstanding reinforce security by diminishing solution related errors.

In generally speaking in excess of 200 kinds of malignancy are found. The name of the malignancy is given on the premise where the cell or tissue of the disease structures. For example in lung malignant growth the phones of lung is influenced so it is called lung disease, in like manner of cerebrum and other disease types. The kind of cell that shaped them are likewise answerable for their name and arrangement as epithelial or squamous cell. Coming up next are a few kinds of malignant growth that start in explicit sorts of cell [2] [6]. Carcinoma, Sarcoma, Lymphoma and Leukemia are the most effective types of cancer.

One of the primary driver of malignant growth is liquor. Practically all mixed beverages cause malignant growth. It causes various kinds of diseases including mouth, throat, liver, and bosom. Tobacco had such 80 substances that are the explanations behind disease. At the point when tobacco is breathed in as cigarette the synthetic compounds go into lungs arrives at the lungs and afterward it is shipped to circulation system and afterward all through the body. Because of increment in higher insulin level overweight individuals are increasingly inclined to of malignant growth. It additionally builds the odds of malignancy of food pipe, Kidney, gallbladders, bosom and belly disease in ladies.

Individuals who have frail resistant framework are more in danger of creating disease. Individuals having HIV or AIDS have more vulnerable resistant framework and they are progressively inclined to the malignant growth. Because of various kinds of disease there is a danger of malignant growth principally contamination brought about by infection.

2. Literature Review

A lot of work and explores have been never really out various techniques for analysis of different diseases types. It is an endeavor to foresee and analyze the malignant growth infection dependent on manifestations that happens at a beginning period.

This paper [1],[7] talks about the lung malignancy which is one of the fatal infections of lungs. In light of that, include determination process short arrangements of 20 such boundaries. A portion of those affecting boundaries are weight reduction, wicked bodily fluid, back torment and so forth. Here, scientists center around pre-finding which is considered as the most fundamental stage to know the helpless patients for experiencing unique determination process. They saw that managed learning ways are obviously better than to the cross approval approach. From this examination, it was discovered that irregular tree classifier, KNN, calculated, multilayer discernment, successive insignificant, enhancement has given substantially more better and dependable execution in this space.

A sub-atomic adjustments in carcinogenesis, one of the most well-known structures, is described by Aberrant DNA methylation cancer.[3][8] In this exploration, creators originally applied various approaches to uncover a few methylation designs that incorporate significant principles. These distinguished guidelines from methylation profile were proficient to precisely separate between tests speaking to tumor and typical case. Along these lines, the fundamental focal point of their methodology was DNA methylation profiles. These profiles were gathered from TCGA database. At that point after, a Rank-Based technique is applied to separate successfully an ordinary tissue from tumor tissues. This strategy utilized the relative methylation level inversion sets as up-and-comer markers sets. At that point recognizable proof of the most unique R-matches in all up-and-comers markers sets are finished by breaking down the CpG locales having most elevated appearance frequencies.

This examination managed [9] with bosom malignancy and the expectation of the sickness was done through Artificial Neural Network (ANN), Logistic relapse, Naive Bayes strategies. The goal of the examination targets giving the accompanying results; right off the bat, it assesses clinical informational collection as far as quality linguistically and also, it assesses information mining techniques concerning their appropriateness to the information. At long last, the information removed from the informational index is utilized for ailment forecast by applying Artificial Neural Network (ANN), Logistic Regression, Naïve Bayes. It is discovered that these methods had most elevated lifting factor for a large portion of class esteems.

The prime focal point of this work [10] is to describe the dispersion of colorectal malignant growth chance utilizing family ancestry of disease. In colon malignant growth libraries of populace disease were made for 10,066 colorectal malignant growth instances of families. Family ancestry is broke down utilizing information mining methods, for example, Novel Index ANN and Standardized Incidence Ratios (SIR's). These procedures are utilized fundamentally to know danger of the ailment.. This examination recognizes 5 major and 66 little classes of hazard.

The benefits of information mining and uncommon qualities of wellbeing information make information mining critical to be considered in wellbeing information examination. This examination [11] means to distinguish those data which exhibit the significance of information mining in social insurance. A basic examination of the malignant growth information is done in the work and some critical shrouded data are removed out of it. Valuable prescient arrangements are created utilizing bunching, characterization, Bio-medication and hereditary qualities.

Order based example investigation strategies are utilized in this work for diagnosing the disease. [12]. A few notable arrangement calculation, for example, DT (Decision Tree), SVM (Support Vector Machine), KNN (K-Nearest Neighbor) and NN (Neural Networks) are utilized for conclusions of disease. It is set up that the procedure of arrangement relies upon the estimation of different highlights in the gathered information. Here, the creators found that clinical malady information frequently have some commotion information just as limit esteem information. They recommended methods to manage such boisterous information. For improvement of precision they utilized Ant Colony Optimization strategy.

Creators [4] [13] in their work utilized a few information mining approaches like grouping, characterization rule mining, delicate processing procedures, Neural Networks and Fuzzy rationale for analysis of oral malignant growth. They indicated the viability of every one of the above procedures for the characterization task in clinical space. Aside from it, they demonstrated the significance of hereditary calculations in streamlining the information mining calculation as far as exactness for forecast.

Specialists [14] here tended to the issue of choice of little subset of qualities from wide examples of quality articulation information recorded on DNA smaller scale clusters from the accessible preparing dataset. In this work, writers proposed technique for quality choice by using Elitism Particle Swarm Optimization (ESPO) in light of Recursive Feature Reduction (RFR) is select. Thusly they segregated the examples of quality articulation of disease and ordinary patients. Further, creators assembled a classifier appropriate for hereditary conclusion.

The fundamental goal of this paper [15] is to locate the most significant things which are answerable for the demise of individuals experiencing gastric disease. Moreover, analysts presented Decision Tree model for research by applying arrangement approach. For this, two classes of patients were made one of dead and the other one of alive patients and chose 20% of informational index arbitrarily as test tests and remaining informational collection were considered as the preparation test.

The fundamental focal point of this work [16] is on the investigation of carcinoma characterization and expectation all together that preventive measures are frequently made at beginning time before the beginning of the carcinoma. Various information handling methods like Decision Tree, Clustering Algorithm are utilized to understand the objective. Perception mining procedure regular sexy on the information base fly inside the salve relationship and customs takes are useful in considering the movement of ailment.

The investigation [17] tended to the issue of applying clinical information mining and utilizing different information digging procedures for determination of Acute Myeloid Leukemia malignancy. Different methods utilized in this work are bunching, relapse, grouping and an endurance expectation model is built out of them. This paper talks about three significant viewpoints; right off the bat it presents hugeness of information mining approach in such manner, furthermore it gives an extensive overview identified with the chose task lastly it looks at the right exactness level of different models.

In this exploration work [18] in which the informational collection are taken from Institute of Portuguese for bosom malignancy, the methodology proposed by the analysts manage a dataset having high percent of obscure unmitigated data. Patients experiencing bosom malignant growth are for the most part stressed in this investigation and their endurance just as to help in improving the treatment of sickness. Diverse model, for example, EM-usage, KNN execution, order trees, strategic relapse are built.

The work proposed [19] here accentuates on deciding the elements which are liable for lung malignancy. It sorts people as the smokers and non smokers. Further, intends to group individuals influenced with lung malignant growth dependent on the smoking propensity through applying information mining methods. Various calculations used to accomplish the objective are Decision Tree and Ant Colony Optimization. Through this paper, we find that information mining procedures have rotate job in discovering the concealed data in clinical information and the quality if information is improved by information preparing

Focal topic of this work is [20] woven for medical caretakers and other human services experts who care for and instruct malignant growth patients and their family about lung disease side effects, way physiology and treatment. So as to accomplish better precision in the expectation of malady and improving survivability rate different huge procedures like Partitioned bunching grouping are utilized.

They encourages better medicos data framework and help to list kind of lung malignant growth, treatment alternatives for non little cell lung disease.

Creators of the work proposed [21] a system of blend of two information mining innovation to be specific grouping and arrangement highlights to foresee the distinction in manifestations of past situations where patients endure or kicked the bucket of oral malignant growth. For analysis, they received information mining procedures like Decision Tree, ANN, Logistic Regression and these strategies were adequately used to essentially dissect oral malignant growth. Also, it gives data to wellbeing doctors to take preventive measures

Remembering the need of forecast of medicinal services benefits in Abu Dhabi basically four models were worked by utilizing information mining strategies that help the organizers to take fitting choice in wellbeing authority and Abu Dhabi government that which sort of human services administrations ought to be completed either as emergency clinic or facility. To accomplish goal of the work, KNN, SMO (Sequential Minimum Optimization and, Naïve Bayes are utilized. [22]

3. Machine Learning Algorithms

We employed the following supervised algorithms which were implemented in caret R package.

- ✓ Feed-Forward NN
- ✓ SVM
- ✓ XG Boost
- ✓ Random Forest
- ✓ LDA

a).Feed-Forward NN: In recent times, Neural Networks has gained a lot of attention and it has become one of the most widely used techniques. It’s ability to learn multivariate data effectively. There are variety of NN models invented viz. Back propagation NN (BPNN), Feed-Forward NN (FFNN), Convolution NN (CNN) and Recurrent NN (RNN). BPNN and FFNN are arranged in three layers-input, hidden and the output layer. On the other hand CNN and RNN have multiple hidden layers. These networks are also known as deep networks or deep learning algorithms. In this work, we have used feed-forward neural network, as it is very simple, effective and easy to understand when compared with other models. In FFNN, a set of random weights are initialized to pass the data from one layer to the other. The features are supplied to input nodes which in turn are connected to hidden layer nodes. The hidden layer represents the relation between the input and output layer. Each hidden node learns by least squares to fit the model. The output layer has an activation function where it can classify the inputs to Outputs. In general the activation function can be sigmoid function which produces a binary outcome. The complete details can be found in chapter 5 of Ripley (2007). NN are used in real-time applications’ viz. time series forecasting Chakraborty et al. (1992), forecasting stock market returns Enke and Thawornwong (2005) and speech recognition Saon and Picheny (2017); Parascandolo et al. (2017).

b).Support Vector Machines (SVM): A support vector machine constructsa hyperplane or set of hyperplane during a high or infinite dimensional space, which might be utilized for Classification, relapse or different assignments. Instinctively, a legit detachment is accomplished by the hyperplane that has the most significant separation to the nearest preparing information purposes of any class, since for the most part the bigger the edge the lower the speculation mistake of the Classifier. On account of a standard hyperplane, limiting the VC measurement compares to amplifying the edge. As a result, for many applications, SVM have been conventional techniques. SVMs perform Classification, making use of two key ideas: maximal margin Classification and a “Kernel trick". The SVM tries to find a hyper plane that separates the differently classified data the most. It rises the insignificant separation between the hyper plane and every data class. Support vectors are adequately takes care of quadratic programming issues. In recent, SVMs are used in e-mail Spam filtering Sculley and Wachman (2007), gene prediction Brown et al. (2000), protein structures predictions Hua and Sun (2001), credit score prediction Huang et al. (2007a). The support vectors are those points which are close to the hyper plane. The practical guide to SVM can be found in Hsu et al. (2003).

c).Extreme Gradient Boost (XGBoost): Extreme Gradient Boosting (XGBoost) is a supervised Classification algorithms and it is very popular in various data science competitions. The term “gradient boosting” come from “greedy function approximation: A gradient boosting machine” Friedman (2001). It is similar to “gradient boosting” but more efficient. It supports various objective functions linear models, tree learning algorithms and ranking. The big prosperity and popularity of XGBoost is its scalability on a single machine by executing parallel computations which allow quicker model exploration. The detailed implementation can be found in Chen and Guestrin (2016). The usage of this algorithm is cited in Torlay et al. (2017); Nielsen (2016); Shi et al. (2018). XGBoost linear model is used in our experiments

d). Random Forest (RF): In recent times Random forestBreiman (2001) has gained a lot of importance as more data science problems are in place. Random forest is similar to “decision tree” models but have multiple decision trees constructs in training set. The drawback of single decision tree is over fitting of training examples due to highly irrelevant patterns and the tree grow very deep. It leads to low bias, but high variance. Random forest (RF) or random decision forest is an ensemble method of classification and regression. It is a supervised learning algorithm. It constructs several decision trees on training examples and outputs the mean prediction of all class labels. It reduces variance error. The RF splits the training set randomly with replacement and fit the trees by averaging multiple decision trees or majority vote. The woodland merges when the constraint of trees in the backwoods turns out to be enormous. Breiman (2001). Of course, RF finds the significance of factors in both classification and relapse issues. It allots the of-pack mistake (OOB) for each example and midpoints that over the timberland. The RF is utilized in land spread classification Gislason et al. (2006), picture classification Bosch et al. (2007) and nature Cutler et al. (2007). It is likewise utilized for quality appraisal of Wikipedia articles We˛cel and Lewoniewski (2015); Lewoniewski et al. (2016); Warncke-Wang et al. (2013). Deciding of variable choice has downsides. In the event that the information set incorporates all out highlights with different levels, RF is one-sided for these highlights with more levels. It are frequently overwhelmed by "fractional stages" of highlights.

e). Linear Discriminant Analysis (LDA): Linear Discriminant Analysis (LDA) James et al. (2013) is simple model to classify given dataset. LDA is closely associated with principal component analysis (PCA) where both techniques examine for linear combination of variables of dataset. LDA attempts to separate the classes of knowledge . PCA doesn’t take under consideration of classes. It computes the statistical properties namely mean and the covariance matrix of every feature/variable. These statistical properties are supplied to LDA to make predictions. It works based on two intuitions. i. The independent variables follow Gaussian distribution. ii. Every variable has the same variance. The dataset has independent variables denoted as X ={<x1,x2,...,xn>} and dependent variables (classes) as C ={<c1,c2,...,ck>}. The mean and variance is calculated from Equation 5 and Equation 6. Where µ is the mean of each class c, n number of instances and σ2 is variance calculated across all classes C.

$$\mu = \frac{1}{nc} * \epsilon(x)(5)$$

$$\sigma^2 = \frac{1}{n-c} * \epsilon((x\text{-}\mu)2)(6)$$

$$\mathrm{P}\left(\frac{Y=y}{X=x}\right) = \frac{f_{\frac{y}{X=x}}(y)P(X=x)}{f_{Y(y)}} \qquad (7)$$

LDA predicts based on the conditional probability density function of given classes i.e. p(x|y = 0) and p(x|y = 1) with mean and covariance respectively. Equation 7 estimates the probability of output classic and given input variables x, where f is a density function.

4. Model Performance Metrics

Metrics are quantitative estimates intended to help assess investigate yields. In this research two evaluation metrics are used.

i. Confusion Metrics

ii. Receiver Operating Characteristic Curve

i). Confusion Matrix: To evaluate the performance of a model, we use different metrics are computed from confusion matrix showed in below Table 1. A confusion matrix is a table that is frequently used to describe the performance of a classification model on a lot of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing [98].

Our aim to reduce the false positives i.e., if a patient has CHD should not be detected as non-CHD patient. To measure the performance of classifiers simple we defined as low classification error rate or higher accuracy i.e., the ratio of sum of true positive and true negative to total number of samples. However, our objective is to improve the precision not the accuracy instead. The Precision is the proportion of number of true positive to the sum of the true positives and false positives.

Where TP - True Positive,

FP - False Positive,

TN - True Negative and

FN - False Negative.

Table1. *Confusion Matrix*

	Reference	
Predicted	N	Y
N	TP	FP
Y	FN	TN

To evaluate the performance of a model, we use different metrics are computed from confusion matrix. Where TP - True Positive, FP - False Positive, TN - True Negative and FN - False Negative.

Accuracy = (TP + TN) / (TP + FP + TN + FN) (1)

Recall = TP / (TP + FN) (2)

Precision = TP / (TP + FP) (3)

F1 −Measure = 2 ∗ (Precision ∗ Recall)/ (Precision + Recall) (4)

ii). ROC Curve:

A ROC curve (Receiver Operating Characteristic Curve) is a chart demonstrating the execution of a classification model show at all grouping limits. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (TPR) is a synonym for Recall and is therefore defined as follows:

TPR=TP/(TP+FN) (5)

False Positive Rate (FPR) is defined as follows:

FPR=FP/(FP+TN) Or

FPR= 1- Specificity (6)

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. Area Under Curve (AUC) - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. By similarity, Higher the AUC, better the model is at recognizing patients with sickness and no disease. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

5. Results Analysis

All the experiments are implemented using “R-language” and executed on Intel i3 4-core machine with 4GB RAM PC.Lung Cancer Data was used by Hong and Young to illustrate the power of the optimal discriminant plane even in ill-posed settings. Applying the XGB method in the resulting plane gave 85.22% accuracy. However, these results are strongly biased. Results obtained by this are NN: 84.88%, LDA: 84.83%, SVM:84.75% and RF :83.45% . The data described 3 types of pathological lung cancers. The Authors give no information on the individual variables nor on where the data was originally used. In the original data 4 values for the fifth attribute were -1.These values have been changed to ? (unknown). (*)In the original data 1 value for the 39 attribute was 4. This value has been changed to ? (unknown). (*). Number of Instances: 218. Number of Attributes: 13. (1 class attribute, 13 predictive). Attribute Information: attribute 1 is the class label. All predictive attributes are nominal, taking on integer values 0-3. Missing Attribute Values: Attributes 5 and 39 (*).

The Lung cancer Disease dataset is normalized and k-fold cross-validation is performed on the data set, where k=10. The main focus is on the evaluation of the classifier with four metrics Accuracy, Precision, Recall and F1-Measure is presented in Table 2.

In-terms of accuracy, NN yielded second highest accuracy of 84.88%, LDA yields 84.83, SVM and RF 84.75 and 83.49 respectively. XGB yielded 85.22% of mean accuracy. The highest accuracy may be considered as the best metric to detect the lung cancer disease patients.

Table2. *The below figure 1 shows that ROC curve drawn for all the methods.*

Algorithm	Accuracy	Precision	Recall	F1	ROC
NN	84.88	75.46	99.52	91.95	0.5263
LDA	84.83	65.87	98.36	91.69	0.5105
SVM	84.75	82.82	99.80	91.79	0.5083
RF	83.49	81.96	99.65	91.72	0.5076
XGB	**85.22**	83.84	96.19	90.81	0.6536

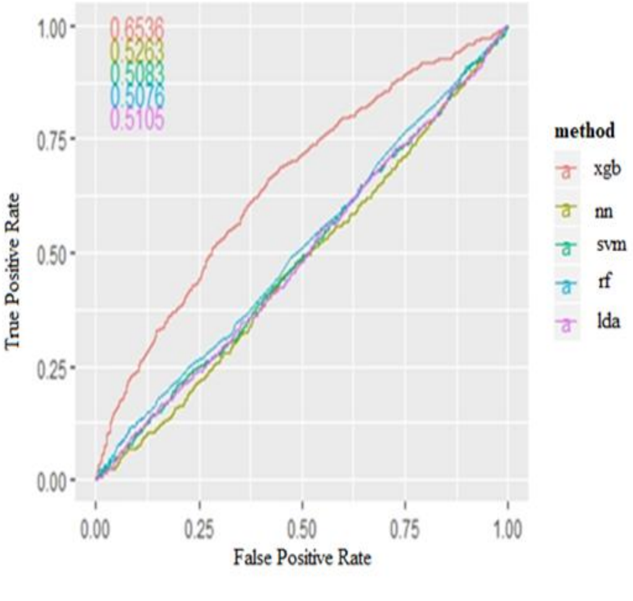

Figure1. *ROC curve for all the methods*

6. Conclusion

Identification of Lung Cancer assumes a fundamental job in clinical period. It gave a wide spread degree in software engineering field. There are a few dangers related with Lung Cancer. All in all to distinguish Lung Cancer, patients need to experience a few clinical assessments. In this paper we present discovery of Lung Cancer utilizing AI expectation models. Our work was completed on UCI standard informational index information identified with Lung malignancy by utilizing diversified classifiers viz., NN, LDA, SVM, RF and XGB in ear. We got normal correct nesses of these classifiers from 10-crease cross-approval on dataset. The precision saw as high in some classifiers.

REFERENCES

[1] D. Tomar and S. Agarwal, “A survey on Data Mining approaches for Healthcare”, International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, (2013), pp. 241-266.

[2] D. Tomar and S. Agarwal, “Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes”, Advances in Artificial Neural Systems, vol. 2015, no. 1.

[3] B. R. Prasad and S. Agarwal, “Modeling risk prediction of diabetes - A preventive measure”, 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, IEEE, (2014), pp. 1-6.

[4] D. Tomar, B. R. Prasad and S. Agarwal, “An efficient Parkinson disease diagnosis system based on Least Squares Twin Support Vector Machine and Particle Swarm Optimization”, 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, IEEE, (2014), pp. 1-6.

[5] A. K. Yadav, D. Tomar and S. Agarwal, “Clustering of lung cancer data using Foggy K-means”, 2013 International Conference on Recent Trends in Information Technology (ICRTIT), IEEE, (2013).

[6] WHO.http://www.who.int/mediacentre/factsheets/fs297/en/ Retrieved on May 20, (2016).

[7] K. Balachandran and R. Anitha, “Ensemble based optimal classification model for pre-diagnosis of lung cancer”, 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE, (2013).

[8] H. Li, G. Hong and Z. Guo, “Reversal DNA methylation patterns for cancer diagnosis”, 2014 8th International Conference on Systems Biology (ISB), IEEE, (2014).

[9] K. Shiny, “Implementation of Data Mining Algorithm to Analysis Breast Cancer”, International Journal for Innovative Research in Science and Technology, vol. 1, no. 9, (2015), pp. 207-212.

[10] R. Chau, “Determining the familial risk distribution of colorectal cancer: a data mining approach”, Familial cancer, (2015), pp. 1-11.

[11] N. Rathore, D. Tomar and S. Agarwal, “Predicting the survivability of breast cancer patients using ensemble approach”, 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), IEEE, (2014).

[12] S. S. Shrivastava, V. K. Choubey and A. Sant, “Classification Based Pattern Analysis on the Medical Data in Health Care Environment”, International Journal of Scientific Research in Science, Engineering and Technology, vol. 2, no. 1, (2016).

[13] R. Vidhu and S. Kiruthika, “A New Feature Selection Method for Oral Cancer Using Data Mining Techniques”, International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 1, (2016).

[14] R. Nagpal and R. Shrivastava, “Cancer Classification Using Elitism PSO Based Lezy IBK on Gene Expression Data”, Journal of Scientific and Technical Advancements, vol. 1, no. 4, (2015), pp. 19-23.

[15] F. Mohammadzadeh, “Predicting the probability of mortality of gastric cancer patients using decision tree”, Irish Journal of Medical Science, vol. 4, no. 2, (2015), pp. 277-284.

[16] M. Kumar, S. S. Tomar and B. Gaur, “Mining based Optimization for Breast Cancer Analysis: A Review”, International Journal of Computer Applications, vol. 19, no. 13, (2015).

[17] M. Duraira and R. Deepika, “Prediction of Acute Myeloid Leukemia Cancer Using Datamining-A Survey”, International Journal of Emerging Technology and Innovative Engineering, vo1. 2, (2015), pp. 94-98.

[18] P. J. García-Laencina, “Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values”, Computers in biology and medicine, vol. 59, no. 1, (2015), pp. 125-133.

[19] T. Christopher, “A Study on Mining Lung Cancer Data for Increasing or Decreasing Disease Prediction Value by Using Ant Colony Optimization Techniques”, Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications, (2015).

[20] K. Arutchelvan and R. Periasamy, “Analysis of Cancer Detection System Using Datamining Approach”, International Journal of Innovative Research in Advanced Engineering, vol. 2, no. 11, (2015).

[21] W. Tseng, “The Application of Data Mining Techniques to Oral Cancer Prognosis”, Journal of medical systems, vol. 39, no. 5, (2015), pp. 1-7.

[22] A. N., Noura, “Data mining approaches for predicting demand for healthcare services in Abu Dhabi”, 2014 10th International Conference on Innovations in Information Technology (INNOVATIONS), IEEE, (2014).