

An Empirical Analysis of Different Machine Learning Algorithms for Predicting Lung Cancer

M. Sharmila¹, R. Punithavathi^{2*}

¹ Assistant Professor, Dept. of Information Technology, M. Kumarasamy College of Engineering,
Karur, Tamil Nadu, India
sharmi28.it@gmail.com

² Professor and Head, Dept. of Information Technology, M. Kumarasamy College of Engineering,
Karur, Tamil Nadu, India
r.punithavathi@gmail.com

* Corresponding Author: sharmi28.it@gmail.com

Abstract: In the current scenario, cancer disease is substantial menace to human life globally. About 32 percent of people worldwide are affected by various types of cancer. But lung cancer depicts the highest ratio. Nowadays peoples are not having awareness about to detect the cancer in early stage. The survival rate of five year for lung cancer disease is 55 percent of the cases are affected most. However, only 14 percent of lung tumor cases are diagnosed at an early stage. For slight tumors the five-year survival rate is simply 3 percent. There are 4 stages in lung cancer. If we predict the disease in I and II stage, it is easy to cure by effortless operations. If it exceeds second stage, it may not be cured. So, diagnosing the cancer in earlier stage is the best solution to predict the patients from death. For that, the system uses the Decision Tree and K-Nearest Neighbor (KNN) Algorithms as preferred classification model. By using these algorithms, it becomes easier to diagnose the cancer in early stage. So, the survival rate of lung cancer patients becomes higher. This comparative analysis, calculates and compares the precision of Random Forest, Naive Bayes and KNN and the preliminary result reveals that ID3 furnish better precision for cancer dataset. The input has been accessed only in numeric format. The algorithms also maintain key stuffs of the dataset, which are predominant for extracting performance, and so it may warrant the correct defense and effective preservation. This leads to protection of any extracting works that depends on the sequence of distances between objects, such as Random Forest, Naive Bayes -search and classification, as well as many visualization techniques. In particular, it establishes a restricted isometric property, in which it is the tight leap on the shrinkage and enlargement of the original distances.

Keywords: Machine Learning; Unsupervised Learning; Naive Bayes classifiers; Decision Tree; Random forest; Decision Support system; Neural network.

I. INTRODUCTION

Big data is a term which is used to manage and analyze the large quantity of data which is not able to deal by the traditional software systems. Every day there is billions of data are generated from various factors such as e-commerce websites, social media, hospitals etc. Big data plays an important role to analyze these data with more effectiveness and provides the best result. Big data techniques are used to work with the effective performance of surgery strategies, other medical tests, and also to discover the relationships among very rushed medical, clinical and diagnosis data. In the field of health sector, the facility for doctors had introduced various data chassis with an enormous amount of information to regulate the patient information but regrettably, the effectiveness over hidden information discovering is less because of the data are not mined. In medical sector, the information mining is a term which means to overview and inspect the medical information to envision the condition of the patient's health. So the statistical analysis, machine learning, decision making and database techniques are applied to discover desiring pattern from healthcare data. The following Figure 1 explains the schema of disease prediction processing duct which contains various steps of gathering data, storing, reviewing and decision support. Distinct methods of data mining are applied for evaluating these data after gathering enormous amount of health complaint data from various sectors. This analyzing process followed by data preprocessing, then feature selection and certainly it makes the prediction by applying machine learning which may do classification and clustering approaches for those healthcare data.

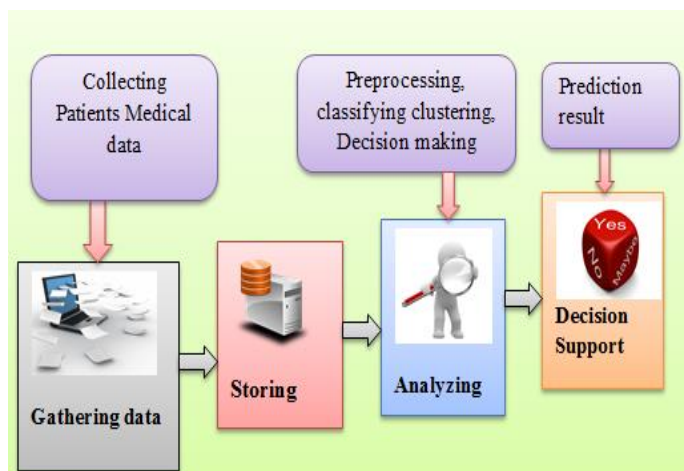


Fig. 1. Pipeline Process of Big Data Analytics

Lung cancer is the master cause of cancer deaths for both the gender in worldwide. Cigarette smoking, tobacco and genetics is the major risk factor for development of lung cancer. The prognosis of lung cancer is very much least because a doctor not able to find the disease until it reaches severe stage. Because the lung can function normally even it is partially damaged. Five-year survival rate in India is around 52% for early-stage lung cancer that is localized in the lungs, but only around 3% in advanced, is an inoperable lung cancer.

Table 1. Stages and Survival Rates of Lung Cancer Patients in India

Lung Cancer stages	Diagnosis Frequency	Five-year survival rate (%)
I	15%	>70%
II	25%	35%-40%
III A	20%	25%-40%
III B	20%	4%-7%
IV	45%	<2%

In the above Table 1 the stage I lung cancer symptoms include lingering or bleeding cough, coughing up or blood, rib ache while breathing deeply, barking cough, raucousness, shortened breathing, wheezing, weakness and dullness, trouble in appetite and weight loss. The stage II lung cancer symptoms are persistent worseness, coughing up blood, shortness in breath, chest pain and back pain, continuous infections like pneumonia and bronchitis. The stage III lung cancer includes pain in the chest, pain when breathing, wheezing, whirling or high pitched sound upon inhalation or exhalation, persistent cough, cough with blood, blood in saliva and mucus, hoarseness or altered voice.

The stage IV which is a final serious stage symptom may include a cough that will not get cleared and stacked, out coming of blood or rusty cold while coughing, chest rib ache which may frequently raucous with broad breathing, coughing, raucousness, sudden weight loss and troubles in appetite, compressed breath, tiredness or weak. The following Figure. 2 describe the four stages of cancer cell (tumor) development in lung. Stage A is a baseline which describes the minute development of tumor near the lung. In 4 months, the tumor reaches the rib. In 5 months it starts to get grow well and in 8 months it starts to decay the whole lung.

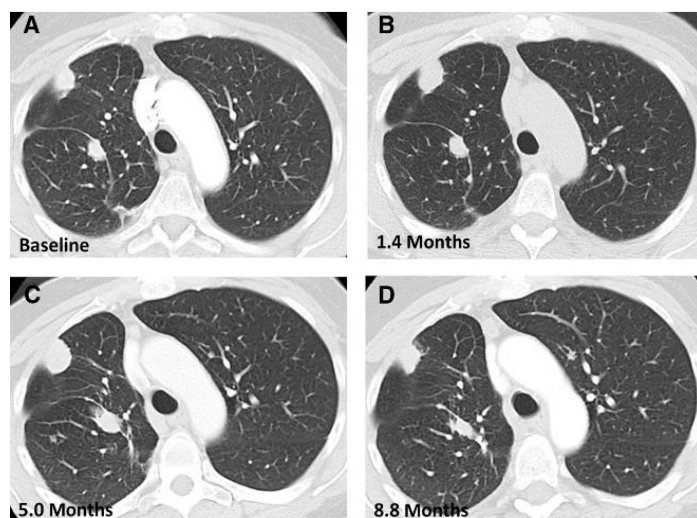


Fig.2. Lung Cancer Stages

II. RELATED WORKS

There are various explorations that may have benefit of machine learning and extracting techniques are explained beneath. The research of paper [1]-[2] developed the proposal as segmentation of lung for Computed Tomography data of a particular patient with Lung cancer. Using these regional conclusions, a wholly automated approach [3] of Lung collocation has been introduced. The collocation is based on 73- feature vectors calculated for the lung sector which consist of 3D edge, region, orientation and spatial relational information. For more accurate classification, paper [4] has implemented Naïve Bayes classifier and Support Vector Machines for accurate classification. The paper [5] has overviewed the classification and regression techniques and makes the comparison over them (SVM provided better accuracy-79.32%) for blood cancer disease prediction. In paper [6] system uses the classification models as Decision Tree and K-Nearest Neighbor (KNN) Algorithms. At the end, the proposed system which makes an analysis over it and correlate the certainty between C4.5 and KNN and the overviewed conclusion finalize that the C4.5 gives better certainty for diabetes mellitus diagnosis.

For the medical database, the Pima Indians Dataset is used in this exploration. For the automatic prediction, paper [7] proposed an automated process for identifying and fragmenting SRM (small renal mass) in contrast-enhanced CT images by using texture, pixel and context feature collocation and compared their result (accuracy- 77.45%). For classification of CT images of lung cancer, paper [8] has compared Multi segmented injection Forward Neural Network, J-48, Random Forest and Genetic programs are tested using ILPD (Indian Lung Patient Dataset) Data Set. The paper [9] uses, the covering methodology which is recently popularized and it offers an effortless and effective way to find the problem of selecting the variable, nevertheless of the chosen machine learning. Because, the machine learning is referred as a perfect black box and it lend to use of off-the-shelf machine learning software package. This approach consists in using the diagnosis and predicting valuation of a machine learning to evaluate the useful variable subset. In routine, one should define: (i) how to search all possible variable subset space; (ii) how to evaluate the prediction valuation of learning machine to supervise the search and kept it. To predict the type II blood cancer, paper [9] has make the comparison over the accuracy of C4.5 used to develop a decision tree. In paper [10], it introduces advanced algorithm for coaching the support vector machines: Sequential Minimal Optimization, or SMO.

III. DIFFERENT MACHINE LEARNING TECHNIQUES FOR LUNG CANCER PREDICTION

In this advanced analysis, the medical prediction has created on the basis of big data analyzing technique to detect the lung cancer disease. Decision tree, Random forest, Naïve Bayes and K-Nearest Neighbor (KNN) classifiers are used to make the system to train the dataset in more optimized manner.

A. Optimized KNN Approach

The projected KNN based classifier determines nearest values persistently from trained datasets and it works only with numeric and image vector of the Lung cancer dataset. The main advantage of this approach is the perfect predictive working based on symptoms and test diagnostic data of the Lung cancer patients. The proposed approach is used to detect the Lung cancer patients and their stage of cancer affected and the experimental application shows the conclusion of the efficiency of the prospective approach. With the combination to that for reviewing medical care data, important steps of extracting techniques like preprocessing data, change the missing values, feature selection, machine learning and decision support and also in addition naïve Bayes also used for the result accuracy are used in training dataset. Finally, the random forest method has been executed on the training dataset of cancer for the classification process.

• Step 1:

In this proposed system, user (doctor, physician etc.) can insert the attributes like their symptoms, their fantastic doubted symptoms, x-ray images and CT images and also the disease value and these values are added to the system with the help of admin through internet. Then, by applying the machine learning technique, the final result after various decisions making and pruning can be shown to user.

• Step 2:

On the server, there is an allowance for admin to load different disease datasets and suiting various extracting algorithms to train up the set of data. The inputs which are requested by users are assembled and refined on main server to predict the analysis result.

• Step 3:

For reviewing medical care data, the important steps of big data accessions like preprocess data, replace missing values, feature selection, machine learning, KNN and decision support are evolved for training the data. On the server there are numerous algorithms like ID3 and KNN gets accomplished over trained dataset and gets qualified to segment the data which are given for testing.

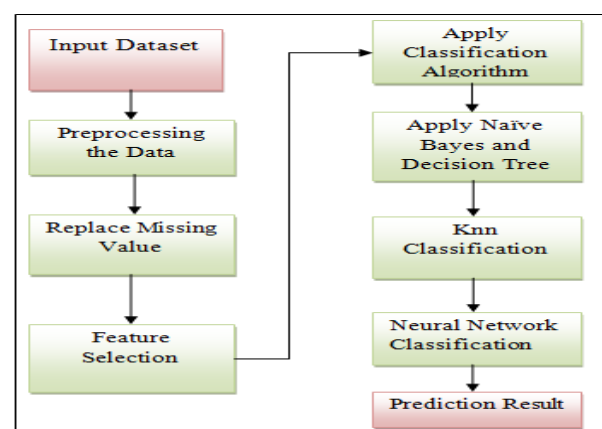


Fig.3. Proposed Architecture Work Flow

B. Naive Bayesian Algorithm

On the basis of Bayes' theorem, the self-subsistence estimates between the predictors done by Bayesian classifier. Naive Bayes classifiers [11] are a family of elementary feasibility classifiers on the basis of applying naïve Bayesian theorem. The anterior possibility $P(c/x)$, from $P(c)$, $P(x)$, and $P(x/c)$ has been thought out with the Bayes theorem. A self-subsistence value of another predictor has been generated by assumption of predictor(x) on the liable class(c) with the help of naïve Bayes classifier. The formula for predicting the tuple 'x' related to the class 'c' by naïve Bayesian classification is

$$P(c/x) = P(x/c) / P(x)$$

- $P(c/x)$ is the rear possibility of class target concluder.
- $P(c)$ is the earlier likelihood of a class.
- $P(x/c)$ is the probability which is the concluder given class.
- $P(x)$ is the prior probability of predictor.

C. Modified Decision Tree

Decision tree [3] is one of the strongest categorization technique. There are various techniques like ID3, C4.5, C5, J48, CART and CHAID algorithms are obtainable to estimate the dataset. Here, there is a continual value in this dataset, so we have chosen ID3 as our perfect classifier. By calculating the topmost information profit in all accredits we can find the decision points. Decision tree is modified to improve accuracy in classification. The algorithm is as follows,

```

Create a decision node
  if all tuple are in same Class
    Return a leaf;
  End if
For Each Attribute in the list
  Calculate information gain;
  Choose attribute with maximum gain as
  decision node;
  Divide the dataset for every branch;
End For
if attribute_list is Empty then
  attach leaf node to majority class;
else
  attach child node to decision tree;
Return the root node.

```

D. Modified K-Nearest Neighbor (MKNN)

MKNN is superintend learning algorithm that finds and recollect radical data based on marginal distance from radical data to the K nearest neighbor. The Euclidean distance is used as a proposal to explain the togetherness.

The algorithm for the modified KNN classifier is stated below:

```

Suppose if we have pairs  $\{(x1, y1), \dots, (xn, yn)\}$  then
 $x=(x1, \dots, xn)$  occurrence to be classified
For each occurrence  $(xi, yi)$  find out  $(xi, x)$ 
  Reorder  $d(xi, x)$ 
where  $i=1$  to  $n$ 
Select the K nearest instances to  $x$ 
Assign the value to  $x$ .

```

E. Neural Network

Mostly the regression type classification and grouping in data science are done perfectly by artificial neural networks. The feature vectors are grouped into classes and it gets ready for receiving the input from the user and find the label which suits most. This can be worked with its efficiency to label the things, like customer types, images and music genres. In medical science department, these classifiers are frequently used to explore the health of the patients, establish them as regular, unsure, or defective. The network is a consulting firm of artificial neurons, as same as neuron connection in human brain. It gets trained by experiencing various instances of every class and learns the likeness and contrast by making comparison. This is a routine which is similar as how the brain learnt, with repeated pattern that forms a better association over time. Visual inspection and k-means clustering of data from four aspects proposed seven various means of health. We can take color as an attribute for prediction result. Grey color is used to predict early stage of lung cancer. Green color is used to predict normal. Orange color is used to suspect that there may be lung cancer and further analysis required with more images. Red color is used to predict the final stage IV. Pink color is used to predict in between early stage and final stage.

IV. RESULTS AND DISCUSSION

A. Dataset

The dataset for predicting the cancer is taken from the major Centre for machine learning and artificial intelligent systems from university of California (UCI). There are 600 samples with 8 numerical valued attribute and 400 which gives the negative test and remaining 100 possess the positive instances in this dataset. The attributes which have been selected are explained briefly for cancer data analysis in Table 2. In the detection, sensitivity refers the fraud detection rate and it is defined as

$$\text{Sensitivity} = TP/N$$

False Alarm Rate: False alarm rate refers the section of actual negative instances which are predicted as positive instances and it is explained as

$$\text{False Alarm Rate} = FP/N$$

B. Result Comparison

The outcomes of final prediction using machine learning for lung cancer disease are discussed in this part. Here 600 instances of random chosen authentic set of data are used to train the machine and other 100 precedents are used for testing the data. This prediction system gets matured using three familiar algorithms i.e., Random forest, ID3 and KNN algorithm. The accomplishment of these three algorithms is discussed beneath.

Table 2. Analysis of Training Dataset

S. No	Training Dataset	Normal Image	Disease Image
1	150	55	95
2	300	74	226
3	450	116	334
4	600	124	476

Table 3. Analysis of Test Dataset

S. No	Test Dataset	Normal Image	Disease Image
1	100	39	61
2	200	65	135
3	300	92	208
4	400	115	285

C. Performance Assessment of Random Forest algorithm

While Applying Random Forest classifier, Table 4 has designed as the preliminary result of analysis. Random Forest classifier makes correct classification over 136 instances and made incorrect classification over 74 precedents. The certainty of perfectly classified precedent is 69.43% and imperfectly classified precedent is 30.57%.

Table 4. Analysis of Random Forest Classifier

Random Forest	Prediction result		Accuracy	
	Perfect classified precedent	Imperfect classified precedent	Perfect classified precedent	Imperfect classified precedent
Training instances: 600 Testing instances: 210	136	74	69.43%	30.57%

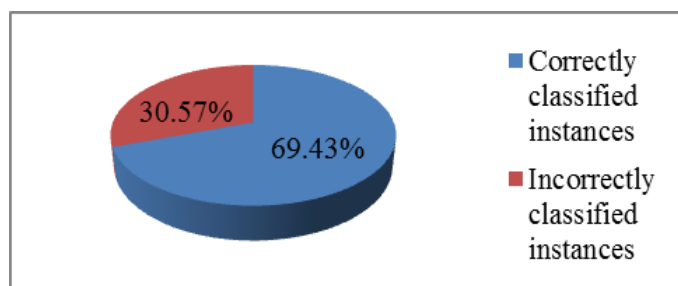


Fig.4. Accuracy chart of Random Forest Classifier

D. Performance Assessment of KNN algorithm

While applying KNN classifier, the prevision result provides, KNN classifier have the correct classification over 162 instances and incorrect classification over 48 instances. The certainty of perfectly classified precedent is 72.65% and imperfectly classified precedent is 27.45%.

Table 5: Analysis of KNN Classifier

KNN	Prediction result		Accuracy	
	Perfect classified precedent	Imperfect classified precedent	Perfect classified precedent	Imperfect classified precedent
Training instances: 600 Testing instances: 210	162	48	72.65%	27.45%

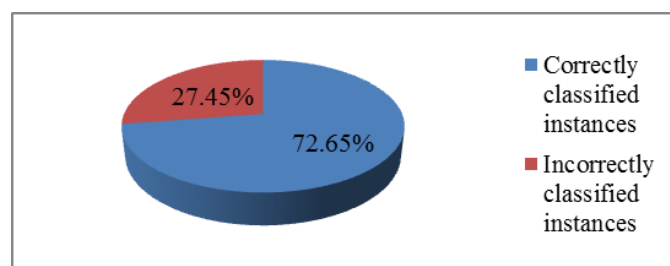


Fig.5: Accuracy chart of KNN Classifier

E. Performance Assessment of Decision Tree (ID3) algorithm

After execution of system using ID3 classifier, the machine has classified 198 instances in a correct manner and 12 instances are classified in incorrect manner. Finally, the certainty of perfectly classified precedent is 96.62% and imperfectly classified precedent is 3.38%.

Table 6. Analysis of ID3 Classifier

ID3	Prediction result		Accuracy	
	Perfect classified precedent	Imperfect classified precedent	Perfect classified precedent	Imperfect classified precedent
Training instances: 600 Testing instances: 210	198	12	96.62%	3.38%

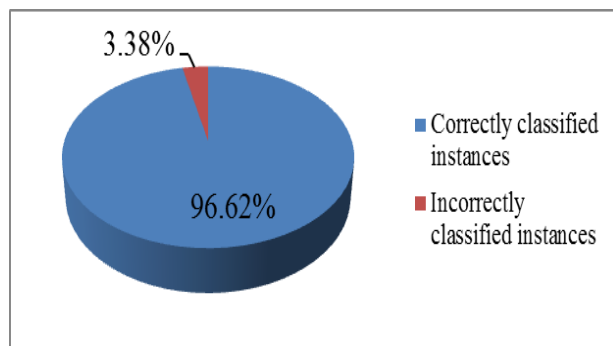


Fig.6. Accuracy chart of ID3 Classifier

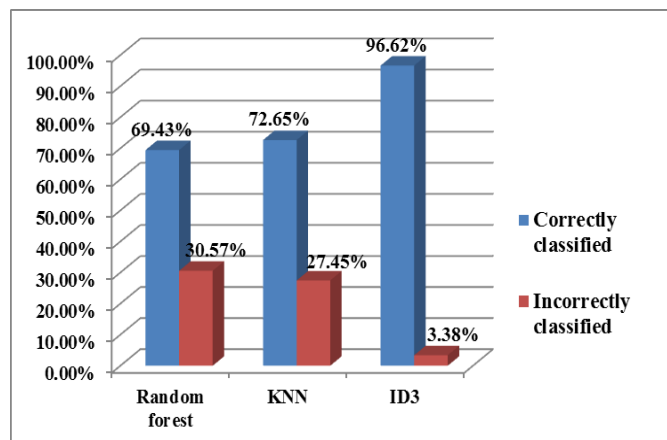


Fig. 7. Graph measure Representation of consummation allegory

F. Allegory of Classification certainty

The consummation of Machine learning system was overviewed with Lung cancer dataset using Random forest, KNN and ID3. In consonance with demonstration results, TABLE IV denotes the consummation allegory of random forest, KNN and ID3 classifier upon its allegory split (80:20) model.

Table 7: Accuracy Comparison of KNN, Random Forest and ID3

Classifier	No. of instances		Accuracy
Random Forest	Perfectly classified	136	69.43%
	Imperfectly classified	74	30.57%
KNN	Perfectly classified	162	72.65%
	Imperfectly classified	48	27.45%
ID3	Perfectly classified	198	96.62%
	Imperfectly classified	12	3.38%

In the below Fig. 7, the graph measure representation of consummation allegory is explained from the graphical model which has been proved that ID3 algorithm performs better with higher allegory.

V. CONCLUSION

In this comparative analysis, a diverse collection of machine learning techniques was used for predicting the lung cancer. The outcome of this system is beneficial to the doctors, medical learners and also the patients to take right resolution about the diagnosis and prediction of the Lung cancer disease. Wide use of neural network as an imaging approach enables early detection of tumor formation in the lung. So, this prediction provides great convenience for eliminate only the cancer tumors with a necessary deadline promoting conservation of healthy lung parenchyma volume as possible.

Finally, this research article uses the K-Nearest Neighbor (KNN) and decision tree (ID3) Algorithms as supervised classification and result prediction model. And the calculation accuracy of Random forest, Naive Bayes and KNN are compared and finalized the accurate prediction to solve the problem in hand.

REFERENCES

- [1]. Andrzej Skalski; Jacek Jakubowski; Tomasz Drewniak, "Lung tumor segmentation and detection on Computed Tomography data", *Imaging Systems and Techniques (IST)*, 2016 IEEE International Conference on 4-6 Oct. 2016
- [2]. Aneesh kumar and. A. C. Jothi Venkateswaran, "Estimating the Surveillance of Lung Disorder using Classification Algorithms", *International Journal of Computer Applications* (0975 – 8887), Volume 57– No.6, November 2012.
- [3]. Bard HSSINA, Abdel Karim MERBOUHA, "A comparative study of decision tree ID3 and C4.5", *International Journal of Computer Applications Beni-Mellal*, BP: 523

- [4]. Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu and Prof. N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Lung Disease Diagnosis", *International Journal of Database Management Systems (IJDBMS)*, Vol.3, No.2 (2011), PP.101-11.
- [5]. Cybenko.G, "Approximation by super positions of a sigmoidal function", *Mathematics of Control, Signals, and Systems*, Vol.2 (1989), PP. 303-314.
- [6]. Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan, "An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques", *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, February 16-18, 2017, Cox's Bazar, Bangladesh (2017).
- [7]. Han Sang Lee; Helen Hong; Junmo Kim, "Detection and segmentation of small renal masses in contrast-enhanced CT images using texture and context feature classification", *Biomedical Imaging (ISBI 2017)*, 2017 IEEE 14th International Symposium on 18-21 April 2017, ISSN: 1945-8452.
- [8]. Isabelle Guyon and Andr'eElisseeff," An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research* 3 (2013) 1157-1182.
- [9]. Jankishran Pahariyavohra, Jagdeesh makhijani and sanjay patsariya, "Lung patient classification using intelligence techniques", *International journal of advanced research in computer science and software engineering*, Volume 4, Issue 2, Pages 295-299,2013.
- [10]. John C. Platt," Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", *Technical Report*, April 21, 2010.
- [11].Rajeswari P and Sophia Reena.G, "Analysis of Lung Disorder Using Data Mining Algorithm", *Global Journal of Computer Science and Technology*, Vol. 10 Issue 14 (Ver. 1.0) November 2010.