International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-3, Issue-11, November 2017 Pages 88-93

A Study on Classification Algorithms and Performance Analysis of Data Mining using Cancer Data to Predict Lung Cancer Disease

E. SathiyaPriya, Dr. S. Mary venila

Abstract— **In the field of Healthcare, cancer diagnosis is the challenging problems and also many of the research has focused to improve the performance to get a satisfactory results in the particular area. To diagnose a Lung cancer is a difficult tasks in medical research. To overcome this challenging task, the many researchers use data mining techniques were applied to predict the many type of diseases. In this research we studied and make comparison of different classification to classify and predict the lung cancer disease. In this research work using Naïve Bayes classification algorithm, SVM (Support Vector Machine) algorithm, KNN algorithm, and J48 algorithm. By using a classification algorithms it produces a different results for the lung cancer datasets in this research. The quality of the result has measured depends on correct and incorrect instances that are correctly classified by a classification techniques.**

Index Terms— **Lung cancer, Two different dataset, classification algorithms are Naive Bayes , SVM, KNN, and J48 algorithms.**

I. INTRODUCTION

Data mining is process of extracting knowledge from huge voluminous amount of databases. The classification algorithms it has useful for classifying the data and it will predicting the various diseases in healthcare. Nowadays prediction is a important concept in many research fields. The many diseases predicting in data mining namely heart disease, lung cancer, diabetics etc., This paper analyzes the Lung cancer disease predictions using a different classification algorithms. Data mining is a best tool in predicting the disease which is useful for time and cost savings. The predicting is not taken considerably accurate results, it assigns a results to the physicians about the disease, so it has useful to physicians to visualize the diseases in advance stages to produce the best treatment for any type of diseases. The symptoms of lung cancer may include a Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, chronic Lung Disease, Balanced Diet, Obesity, Chest Pain, etc.

II. RELATED WORKS

Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System, Shubpreet Kaur [10] In their Paper has Applying data mining in the medical field is an incredibly challenging mission in the medical profession. They characterizes widespread process that demands thorough understanding of needs of the healthcare.

A Critical Study of Classification Algorithms for LungCancer Disease Detection and Diagnosis N.V. Ramana Murty[16], In their Paper they conducted an experiment to the analysis has been performed using WEKA tool with several data mining classification techniques and they found that the Naive Bayesian algorithm gives a better performance in all aspects over the other classification algorithms.

A Comparative Study of Data Mining Classification Techniques using Lung Cancer Data Tapas Ranjan Baitharu[1], In their Paper they conducted an experiment to find the impact of lung cancer data on the performance of different classifiers.

Comparative study of Recent Trends on cancer disease prediction using data mining techniques, Satyam shukla et. Al[30], in their paper they applying data mining techniques like Rank based method in which reversal pairs are rea and they can be easily to independent of samples by the help of which cancer diagnosis became easy.

A Study on Mining Lung Cancer Data for Increasing or Decreasing Disease Prediction Value by Using Ant Colony Optimization Techniques, J.Jamera banu[4] In their paper has successfully performed with several data mining classification techniques and they believed that the data mining can significantly help in the Lung Cancer research and ultimately improve the quality of health care of Lung Cancer patients.

A Survey On Early Detection And Prediction Of Lung Cancer, Neha Panpaliya, Neha Tada[5] in their paper they conclude that using the combination of neural network classifier along with binarization and GLCM will increase the accuracy of lung cancer detection process. By using this system will also decrease the cost and time required for cancer detection and also if the patient is not detected with the lung cancer the system will proceed further for the prediction process.

Early Detection of Lung Cancer Risk Using Data Mining, Kawsar Ahmed, Abdullah-Al-Emran[9] in their paper they show experimental results are separated into two sections. The significant frequent patterns discover and another is represents prediction tools to predict Lung Cancer. They using a data from data warehouse, the significant patterns are extracted for Lung cancer prediction.

III. DATA MINING TECHNIQUE

Data mining is the process of collecting data automatically from huge amount of data. The term `data mining' (often called as knowledge discovery) refers to the process of analyzing data from different purpose and it has useful

E. SathiyaPriya, M.Phil Scholar,PG and Research Department of Computer Science, presidency college, Chennai-05

Dr. S. Mary venila, PG and Research Department of Computer Science, presidency college Chennai

information by means of a number of tools and techniques, which in turn to increase the performance of a system. It has finding a hidden patterns and analyzing the relationships between different types of data to develop predictive models. Data mining is a best tool in predicting the disease which is almost valuable.

IV. DATASET DESCRIPTION

Lung Cancer Patient Dataset contains 2000 observations and 11 attributes that describes about Smoking, Yellow fingers, Anxiety, Peer pressure, Genetics, Attention Disorder, Car accident, Fatigue, Allergy, Coughing, Lung Cancer. In Lung cancer field it describes the True and False that is True means the patient have a lung cancer disease. The result will show False then it describes the patient have no lung cancer disease. An another Lung cancer data set it contains 309 observations and 16 attributes that describes about Gender, Age, Alcohol Consuming, Smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, coughing, shortness of breath, swallowing difficulty, chest pain, Results. In the result field it describes the Yes and No that is yes means the patient have a Lung cancer. The result will show No then it describes the patient is in normal. After collecting a different Lung Cancer Patient data , the data has already preprocessed and Normalized.

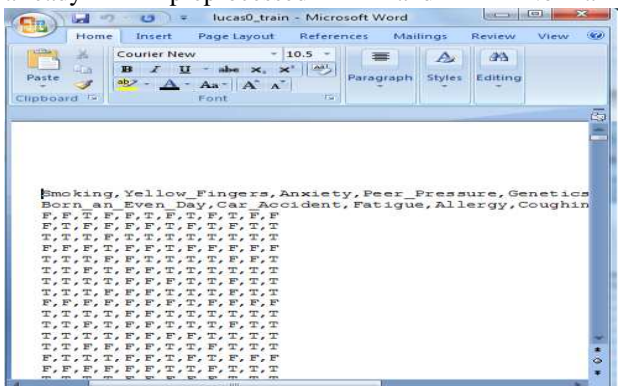

Figure 4.1 Lung cancer Dataset1

Lung Cancer Dataset2

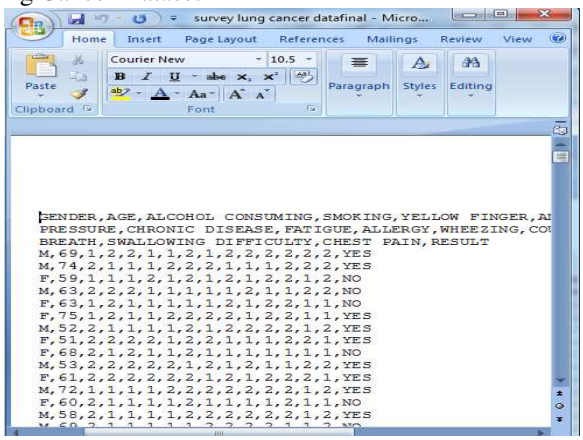

Figure 4.2 Lung Cancer Dataset2

V. PERFORMANCE ANALYSIS

Classification accuracy is generally calculated by the percentage of instances has correctly classified. The output include the result of the data values given in the data set. The result has compared based on the Accuracy , Speed of the each classification algorithms . In this study has been performed using WEKA tool with several data mining classification algorithms. Two Datasets used in this stud the accurate in order to improve the predictive accuracy of data mining algorithms. The different Attributes of symptoms is used to diagnosis of disease are to be handled efficiently to produce the best outcome from the data mining process.

Table 5.1 The different classification algorithms used in previous Lung Cancer prediction

Ref#	Classification Techniques	Best Classification Technique
[19]	J48	J48
	Random forest	
	Logit Boost	
	Random subspace	
	Alternating decision Tree	
[20]	Naïve Bayes	KNN
	KNN	
	RF	
	SVM	
	Bagging	
	Ada Boost	
[21]	REP Tree	REP Tree
	Random Tree	
[22]	Bayes Network	Bayes Network
	SVM	
[24]	ZeroR	Neural Network
	Neural Network	
[25]	ID3	ID3
[26]	Naïve Bayes	Naïve Bayes
[20]	SVM	SVM

The table 5.1 shows the Lung Cancer prediction system estimates risk of the lung cancers. This system was validated by comparing its predicted results with patient's prior medical information and analyzed using weka system.

Table 5.2 Accuracy level of Different Classification algorithms

Ref #	Classification Algorithm	No. of instances used	Accuracy
[19]	J48	5754 instances	91.4%
[20]	KNN	34 instances	84.6%
[21]	REP Tree	3777 instances	79.58%
[22]	Bayes Network	322 instances	77%
[24]	Neural Network	909 instances	96.04%
[25]	ID3	463 instances	100%
[26]	Naïve Bayes	303 instances	83.4%
[20]	SVM	34 instances	94%

The Table 5.2 shows that the high accuracy belongs to different classification Algorithms. Based on this accuracy many algorithms shows a better accuracy for a prediction.

Table 5.3 Different Classifier Evaluation for Lung Cancer Dataset 1

Classification Algorithms	Sensitivity	Specificity	Precision	Recall	F-Measure
Naïve Bayes	0.645	0.869	0.866	0.869	0.867
SVM	0.564	0.972	0.855	0.972	0.91
KNN	0.655	0.938	0.877	0.938	0.907
J48	0.609	0.955	0.866	0.955	0.908

The performance has evaluated of various classifiers and percentage split of Lung cancer data set1. In this table 5.3 shows the sensitivity (TP Rate), Specificity (TN Rate), Precision, Recall and F – Measure of the each Classification Algorithms. The Evaluation Result of the different classification algorithms Table5.11 has shows SVM, KNN and J48 performance as better .

Table 5.4 Different Classifier Evaluation for Lung Cancer Dataset2

Classification Algorithms	Sensitivity	Specificity	Precision	Recall	F – Measure
Naïve Bayes	0.945	0.57	0.945	0.945	0.945
SVM	0.945	0.857	0.981	0.945	0.963
KNN	0.927	0.714	0.962	0.927	0.944
J48	0.909	0.714	0.962	0.909	0.935

The performance has evaluated of various classifiers and percentage split of Lung cancer dataset2. In this table 5.4 shows the sensitivity (TP Rate), Specificity (TN Rate), Precision, Recall and F – Measure of the each Classification Algorithms. The Evaluation Result of the Table 5.4 has shows SVM performance as better than the other classification algorithms.

Table 5.5 Analysis of Different Classification algorithms for Lung Cancer Dataset1

Classification Algorithm	Correctly Classified Data	Incorrectly Classified data	Accuracy
Naïve Bayes	323	77	80.75%
SVM	344	56	86%
KNN	344	56	86%
J48	344	56	86%

The Table 5.5 shows the different classifiers such as Naïve Bayes it has correctly classified 323 data SVM it has correctly classified 344 data, KNN algorithms it has correctly classified 344 data, andJ48 has 344 data has correctly classified. when compare to these algorithms SVM, KNN and J48 shows a better performance.

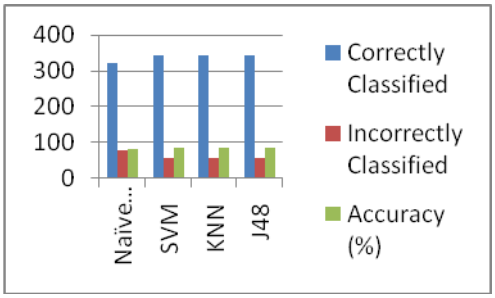

Figure 5.1 Performance of Different Classifier for Lung Cancer dataset1

The Figure 5.1 shows different algorithms performs better based on that SVM, KNN, J48 algorithms had better performance when compare to Naïve Bayes Algorithm.

Table 5.6 Time taken to build model of different Classifier for Lung Cancer dataset1

Classification Algorithms	Time Taken to build model (seconds)
Naïve Bayes	0.01
SVM	0.45
KNN	0.01
J48	0.02

The table 5.6 shows the Naïve Bayes takes 0.01 seconds to build the model, SVM takes 0.45 seconds to build the model, KNN takes 0.01 seconds to build the model and J48 takes 0.02 seconds to build the model based on the time many algorithms takes less time but SVM takes more time because to build a model for SVM algorithm its very complex

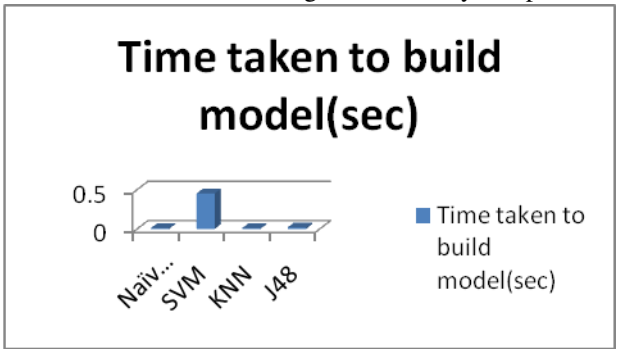

Figure 5.2 Time Taken to build model for Lung Cancer Dataset1

It is observed from figure 5.2 the SVM algorithms takes 0.45 seconds when compare to other algorithms for Lung cancer dataset1

Table 5.7 Analysis of Different Classification algorithms for Lung Cancer Dataset2

Classification Algorithm	Correctly Classified	Incorrectly Classified	Accuracy
Naïve Bayes	56	6	90.32%
SVM	58	4	93.54%
KNN	56	6	90.32%
J48	55	7	88.70%

The Table 5.7 shows the different classifiers such as Naïve Bayes it has correctly classified 56 data, SVM it has correctly classified 58 data, KNN algorithms it has correctly classified 56 data, and J48 has 55 data has correctly classified when compare to other algorithms, in these data set the SVM shows better performance

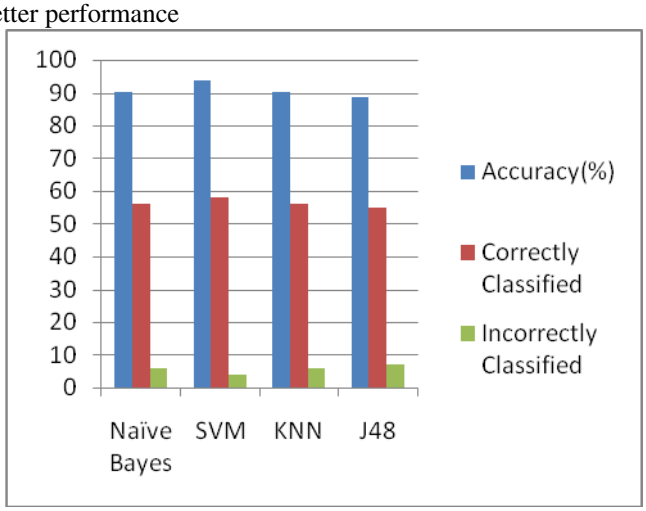

Figure 5.3 Performance of Different Classifier for Lung Cancer dataset2

The Figure 5.3 shows the overall all algorithms has a better performance. But SVM performance is little high when compare to other algorithms.

Table 5.8 Time taken to build model of different Classifier for Lung Cancer dataset2

Classification Algorithms	Time Taken to build model(seconds)
Naïve Bayes	0.01 seconds
SVM	0.02 seconds
KNN	0.01 seconds
J48	0.02 seconds

The table 5.8 shows all the algorithms takes less time to build model because it takes less no. of attribute for the testing prediction process.

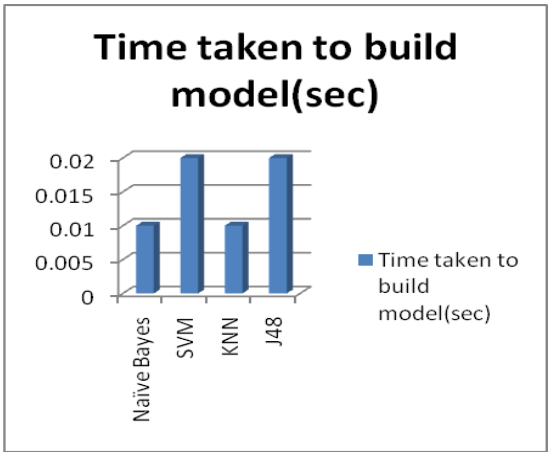

Figure 5.4 Time taken to build model for Lung Cancer Dataset2

The Figure 5.4 shows the SVM and J48 takes 0.02 seconds to build the model. When compare to other algorithms it has not more difference to take a time for build the model.

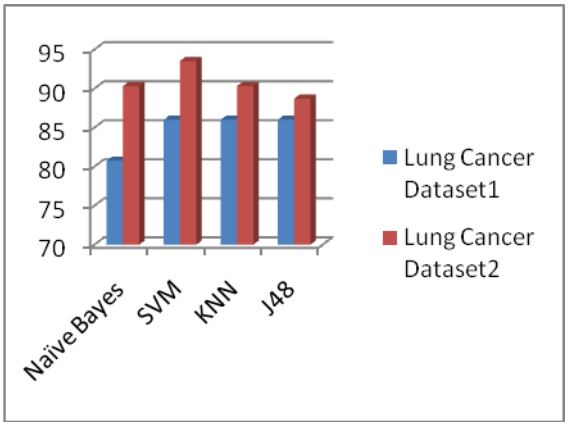

Figure 5.5 Accuracy of four different algorithms for Lung Cancer dataset1 and Lung Cancer dataset2

The Figure 5.5 shows the overall classification algorithms accuracy results for the both Lung Cancer Dataset1 and Lung Cancer Dataset2. Based on the graph all algorithm gives a better performance for a different attributes in a different datasets.

CONCLUSION

Lung cancer is one of the major causes of death in men and women. In Lung Cancer research ultimately to improve the quality of Healthcare and Lung Cancer patients. In some cases even in the advanced level Lung cancer patients does not show the symptoms associated with the Lung cancer. There are many patients did not know Lung cancer disease in early stage, because the lack of awareness. So, the prevention of lung cancer is needed in reducing life losses. By this experiment results we can clearly predict the lung cancer disease, which can be used to warn the people before to avoid the unwanted drinking habits, smoking, intake of contaminated food, obesity etc., for prevent from the lung cancer disease. In this study of classification Techniques used in several dataset is help to improve lack of awareness in the Lung Cancer patients. The experiment results has show three different classification algorithms gives a better performance on selected one Lung cancer dataset. When we applying the same classification techniques for another Lung

cancer dataset, the result has shows all algorithms performs significantly better performance. From this study, based on the results that no single classifier is better than other. But the results from this data mining tool , we cannot consider the results has surely as perfect. There is chance to get a results might changes and vary in a similar datasets related to cancer disease are classified on different tools like tanegra, rapid mining etc., because that are the latest data mining tools within the data mining. This experiment can be extended by applying additional range of classification algorithms or use a proposed algorithms on additional range of datasets from huge medical database and dataset has from various domains in future.

REFERENCES

[1] Er.Tapas Ranjan Baitharu, Dr.Subhendu Kumar Pani, A Comparative Study of Data Mining Classification Techniques Using Lung Cancer Data international Journal Of Computer Trends And Technology (IJCTT) – Volume 22 Number 2–April 2015

[2] Jaimini Majali, Rishikesh Niranjan, Vinamra Phatak, Omkar Tadakhe, Data Mining Techniques For Diagnosis And Prognosis Of Cancer, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 3, March 2015

[3] Hilal Almarabeh, A Study of Data Mining Techniques Accuracy for Healthcare, International Journal of Computer Applications, Volume 168 – No.3, June 2017.

[4] J.Jamera banu, A Study on Mining Lung Cancer Data for Increasing or Decreasing Disease PredictionValue by Using Ant Colony Optimization Techniques Special Issue Published in Int. Jnl. Of Advanced Networking and Applications (IJANA) Page 150.

[5] Neha Panpaliya, Neha Tadas, Surabhi Bobade, Rewti Aglawe, Akshay Gudadhe, A Survey On Early Detection And Prediction Of Lung Cancer, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.1, January- 2015, pg. 175-184.

[6] Mr. P. Thangaraju, R. Mehala, Novel Classification based approaches over Cancer Diseases, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 3, March 2015.

[7] K.Arutchelvan, Analysis Of Cancer Detection System Using Data mining Approach, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 11, Volume 2 (November 2015).

[8] Adnan Alam khan, Comparative analysis of data mining tools for lungs cancer patients, Journal of Information & Communication Technology Vol. 9, No. 1, (Spring2015) 33-40.

[9] Kawsar Ahmed, Abdullah-Al-Emran, Tasnuba Jesmi, Roushney Fatima Mukti, Md Zamilur Rahman, Farzana Ahmed Early Detection of Lung Cancer Risk Using Data Mining, Asian Pacific Journal of Cancer Prevention, Vol 14, 2013

[10] Shubpreet Kaur and Dr. R.K.Bawa, Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System, International Journal of Energy, Information and Communications Vol.6, Issue 4 (2015), pp.17-34

[11] Vidya R, Latha V and 3Venkatesan S, Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques, Special Issue Published in International Journal of Trend in Research and Development (IJTRD)

[12] Himani Sharma, Sunil Kumar, A Survey on Decision Tree Algorithms of Classification in Data Mining, International Journal of Science and Research (IJSR), Volume 5 Issue 4, April 2016.

[13] Priyanka D. ,Ms. S. Shahar Banu, Prediction on Lung Disease Using K means Algorithm ,2014 IJIRT | Volume 1 Issue 11

[14] .G. Krishnaveni, Prof. T.Sudha, A Novel Technique To Predict Diabetic Disease Using Data Mining – Classification Techniques, International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS, Volume.3,Special Issue.1,March.2017

[15] K. Arutchelvan, Prognosis of Lung Cancer Using Data Mining Techniques, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, March 2016

[16] N.V. Ramana Murty, A Critical Study of Classification Algorithms for LungCancer Disease Detection and Diagnosis, International Journal of Computational Intelligence Research,
pp. 1041-1048

[17] Durairaj M, Deepika R, Comparative Analysis of Classification Algorithms for the Prediction of Leukemia Cancer, International Journal of Advanced Research in Computer science and Software Engineering, Volume 5, Issue 8, August 2015.

[18] P. Thamilselvan, An Enhanced K Nearest Neighbor Method To Detecting And Classifying Mri Lung Cancer Images For Large Amount Data, International Journal Of Applied Engineering Research Issn 0973-4562 Volume 11, Number 6 (2016) Pp 4223-4229

[19] Ankit Agrawal, Sanchit Misra, En. At Al, A Lung Cancer Outcome Calculator Using Ensemble Data Mining On Seer Data, *Biokdd 2011,* August 2011, San Diego, Ca, Usa.

[20] Amit Bhola, Machine Learning Based Approaches For Cancer Classification Using Gene
Expression Data, Machine Learning And Applications: An International Journal (Mlaij) Vol.2, December 2015

[21] V. Kirubha, Comparison Of Classification Algorithms In Lung Cancer Risk Factor Analysis, International Journal Of Science And Research (Ijsr) Volume 6 Issue 2, February 2017.

[22] K. Jayasurya, G. Fung, S. Yu, C. Dehing-Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, P. Lambin, A. L. A. J. Dekker, Comparison Of Bayesian Network And Support Vector Machine Models For Two-Year Survival Prediction In Lung Cancer Patients Treated With Radiotherapy, T He International Journal Of Medical Physics And Research, Vol. 37, No, 4, (2010).

[23] Y.-C. Chen, W.-C. Ke, H.-W. Chiu Risk Classification Of Cancer Survival Using Ann With Gene Expression Data From Multiple Laboratories Compute Biol Med, Vol. 48, (2014), Pp. 1–7.

[24] Ada, Early Detection And Prediction Of Lung Cancer Survival Using Neural Network Classifier, International Journal Of Application Of Innovation In Engineering Of Management(Ijaiem), Volume 2, Issue 6, June 2013

[25] A.Priyanga, S.Prakasam, Ph.D, Effectiveness Of Data Mining - Based Cancer Prediction System (Dmbcps), International Journal Of Computer Applications , Vol. 83 – No 10, December (2013), Pp. 0975 – 8887.

[26] Thangaraju , Barkavi , Karthikeyan, Mining Lung Cancer Data For Smokers And Non-Smokers By Using Data Mining Techniques, International Journal Of Advanced Research In Computer And Communication Engineering Vol. 3,No. 7, July (2014).

[27] Shraddha Deshmukh et .al Hypothesis on Different Data Mining Algorithms, Int. Journal of Engineering Research and Applications, Vol. 5, Issue 12, (Part - 3) December 2015, pp.86-91

[28] Ms. Swati P. Tidke , Classification of Lung Tumor Using SVM, International Journal Of Computational Engineering Research, Vol. 2 Issue. 5

[29] J.Jamera banu, Study of Classification Algorithm for Lung Cancer Prediction, International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 2, February 2016.

[30] Satyam shukla et. Al, Comparative study of Recent Trends on cancer disease prediction using data mining techniques., International journal od database theory and application.

[31] V.Krishnaiah, Diagnosis of Lung Cancer Prediction System

Using Data Mining Classification Techniques, International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45.

[32] Vidyullata Pellakuri et. alPerformance Analysis of Classification Algorithms Using Healthcare Dataset, International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1103-1106