VOL. 15, NO. 3, FEBRUARY 2020 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

IMPLEMENTATION OF RANDOM PROJECTION FILTER AND DECISION TREE J48 FOR LUNG CANCER DETECTION

Manju B. R. and Krishnapriya K. R.
Amrita Vishwa Vidyapeetham, Amritapuri Campus, Kollam, Kerala, India
E-Mail: manjubr@am.amrita.edu

ABSTRACT

One of the challenging tasks in this era is the early detection of cancer. The early detection helps to cure the disease completely. Random Projection (RP) is extensively used to reduce the high dimensional features to low dimensional features by projecting data onto a lower space while conserving most of the variation available in the data. J48 can handle both continuous and categorical features and is able to reduce misclassification errors. In this paper we have suggested a method for cancer prediction with the help of different data mining algorithms. The aim is to find out the best filter-classifier combination for the diagnosis. The competency of the algorithms can enhance the insight in to the problem and can thereby minimise the difficulty level in diagnosis.

Keywords: lung cancer, random projection, decision tree J48, random forest, Hoeffding tree, logistic model tree (LMT).

1. INTRODUCTION

Lung cancer is one of the main reasons for the death all around the word. The Worldwide data in 2012 alerts us that among all cancers the contribution of lung cancer is about 13%. It is found that the main reason for the cause of lung cancer is cigarette smoking. The smoke coming from cigarettes is able to damage the cells lying in the lungs and can lead to lung cancer. Inhaling smoke from other's cigarette may also lead to lung cancer. Those suffering from lung cancer have a five year survival rate of approximately 16% [American Cancer Society, 2012]. The primary method to inhibit lung cancer includes evading risk factors such as air pollution and smoking. The introductory symptoms of lung cancer include chest pain, cough, coughing up blood, weight reduction and shortness of breath. The methods commonly used in the detection of lung cancer include Imaging Tests, Sputum Cytology and Biopsy. The number of cigarette smokers reduced from 16.8% in 2014 to 15.1% in 2015 [Ahmed Jamal, MBBS, Brian A. King, PhD et.al. (2016)].

Random Projection (RP) has developed as a great strategy for diminishing the dimensionality. The most eminent property of RP is that it is a general dimensionality reduction method. In reducing the dimension RP is cost effective and gives precise result. The geometric transformations like translation, rotation and scaling make random projection more robust [Kun Liu and Hillol Kargupta (2006)]. RP works under the basic concept of Johnson-Lindenstrauss Theorem. The benefit of RP is, it is data independent and when compared to Principal Component Analysis (PCA) it functions well on a high dimensional dataset and is computationally cheaper [Dmitriy Fradkin, David Madigan (2002)]. J48 is a decision tree algorithm that is easy to handle and understand. It can be used as both filter and classifier. It is one of the most dynamic and popular classifier.

Data mining helps us to evaluate a huge amount of data and helps to extricate usable information from it. Machine learning is a rapid growing area where it learns from past experience to enhance future execution. It centres on extricating information from substantially vast sets of data, and after that distinguishes and recognizes underlying patterns utilizing different statistical measures to enhance its capability to elucidate new data and create more viable results. Waikato Environment for Knowledge Analysis (WEKA) is a modern device used for Machine learning [Pankaj Singh, Sudhakar Singh et.al.(2015)]. It includes a collection of Machine learning algorithms that helps in data mining process. The benefit of Weka is that it is simple to handle and is able to learn through graphs even in the absence of the knowledge of programming. The data mining tasks such as data pre-processing, classification, feature selection etc can be done in WEKA tool [Pankaj Singh, Sudhakar Singh et.al (2015)].

2. LITERATURE SURVEY

A method for detecting lung cancer applying image processing is proposed in [Abhigna B.S, Nishant Sai et.al. (2016)] so CT images can be used for identifying tumour in the lungs. The three steps involved in the procedure of recognizing lung cancer cell by the machine includes pre-processing, feature extraction and lung cancer cell recognition. With the help of numerous data mining techniques the most appropriate method for predicting lung cancer is proposed in [V. Krishnaiah, Dr. G. Narsimha et.al. (2013)]. The study shows that the most appropriate method for prediction is Naive Bayes followed by Decision Trees, If-Then rule, and Neural Network.. A method for the early detection and prediction of lung cancer is proposed in [Neha Panpaliya and Neha Tadas (2015)]. The common method utilized for early diagnosis and treatment is Image processing and recognizing genetic and environmental factors are essential for prediction. The study reveals that combination of NCC with binarization and Grey Level Co-occurrence method (GLMC) boosts the precision of lung cancer detection procedure. To analyze a specific disease different classification algorithms and medical datasets [Vanaja S. and K. Rameshkumar (2014)] are considered for comparing the performance.

With the help of image and text data a comparison between different dimensionality reduction techniques is

taken up here [Ella Bingham and Heikki Mannila (2001)]. In order to categorize the breast cancer dataset Random projection is employed for decreasing the dimension [Haozhe Xie and Jie Li (2016)]. The experimental study shows that the classification accuracy of Feature Selection with RP is superior to LDA and PCA. The comparison of efficiency between the dimensionality reduction techniques Latent Semantic Indexing (LSI) and RP is carried out here [Jessica Lin and Dimitrios Gunopulos (2003)]. Utilizing different datasets and Machine learning algorithms a comparison between RP and Principal Component Analysis (PCA) is presented here [Dmitriy Fradkin and David Madigan (2002)].

A comparative study on the two algorithms J48 and Adaboost [Poonam Pandey and Radhika Prabhakar (2016)] reveals that J48 is more suitable to datasets which have more than two class labels whereas Adaboost is appropriate where the class label is exactly two. The most admissible method [Anju Radhakrishnan and Vaidhehi V (2017)] for classifying an email as spam or non-spam was found to be J48 than Naive Bayes. In the experiment of credit card fraud detection the performance of J48 is preferable among the three classifiers taken [Farhad Alam and Sanjay Pachauri (2017)]. In this paper [CK Madhusudana, S Budati *et al*. (2016)] a method for fault diagnosis with the help of machine learning algorithms is presented. The result shows that the combination of Decision tree and Naive method can be suggested for the fault diagnosis.

3. METHODOLOGY

The methodology of the work is clearly explained in figure.

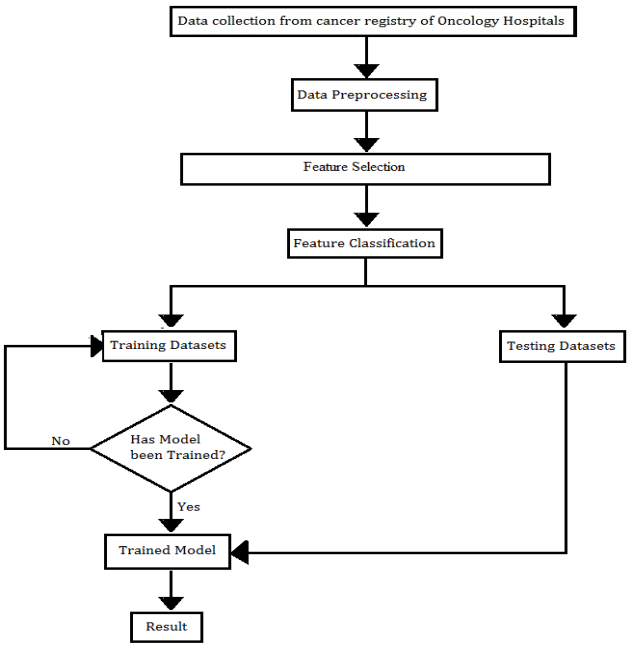

Figure-1. Methodology.

3.1 Data collection and Data pre-processing

The data is gathered from the Oncology department of various hospitals, for the period 2005-2010 for conducting a study on the lung cancer patients. Data pre-processing is a critical step in data mining process. Pre-processing is done in order to avoid noise, missing values and inconsistency. The original data contains 24 attributes, 599 instances and 3 classes. The irrelevant features such as patient-id, age and gender are removed from the data. Now the data comprises of 21 attributes, 599 instances and 3 classes. The classes are low, medium and high.

3.2 Feature selection

Feature selection helps to select the features which carry the most important information by eliminating the features which are irrelevant or which does not increase the classification accuracy. Here RP and J48 are used for feature selection.

3.2.1Feature selection (FS) using RP

RP is a dimensionality reduction method where, the original data which lies in high dimension is projected into a subspace in lower dimension using an arbitrary matrix. A matrix A that has $c$dimensions and d instances can be reduced to lower dimension p by

$$B = CA$$

Where,
B is a $p \times d$ matrix
C is a $p \times c$ matrix and
A is a $c \times d$ matrix

The mathematical theory behind Random Projection is Johnson-Lindenstrauss lemma: [Jessica Lin and Dimitrios Gunopulos (2003)] which states that any point in a high dimensional space can be projected to a lower dimensional subspace by conserving the distance between the points.

3.2.2 Feature selection (FS) using Decision tree J48

J48 algorithm can be used for both feature selection and classification. J48 is an improved version of C4.5 and is used in WEKA data mining tool. It is a flow chart like pattern where each node represents an attribute, the branches indicate an outcome, and the leaf node represents class label [Anshul Goyal and Rajni Mehta (2012)]. Splitting criteria is used in J48 inorder to choose the attribute that gives best split at a particular stage. Entropy measures the disorder involved in the data. It helps to select the most helpful features in the data. So the attributes which reduces the entropy are taken for the further process.

3.3 Feature classification

In Machine learning and pattern recognition a feature is an individual measurable trait of an event being noticed. Selecting the most relevant, differentiating and independent features is a very important step in

classification. The next step is feature classification. The classification is done using 3 algorithms.

3.3.1 Random forest

Random forest is an ensemble procedure that can be utilized for both classification and regression. It was suggested by Breiman in 2001. Random forest creates multiple decision trees and combines them together in order to get a better accuracy and reliable prediction. In Random forest the accuracy increases with increase in number of trees. To strengthen the precision of random forest the correlation should be minimized while maintaining their strength [Eesha Goel and Er. Abhilasha (2017)]. In Random forest the final decision is taken by calculating the majority vote given by the decision trees.

3.3.2 Logistic model trees (LMT)

Logistic model tree (LMT) is a supervised machine learning algorithm that assimilates logistic regression and decision tree learning. The benefit of utilizing logistic regression is that unequivocal class probability estimates are created instead of just a classification. LMT builds a single tree comprising of logistic regression function at the leaves, binary splits on numeric attributes and multiway splits on nominal ones [N. Saravanan and Dr. V. Gayathri (2015)]. Structurally, a model tree takes the form of a decision tree with linear regression functions instead of class values at its terminal node [Eibe Frank, Yong Wang et.al. (1998)]. It can deal with both continuous and numeric values.

3.3.3 Hoeffding tree

Hoeffding tree is a decision tree algorithm that uses Hoeffding bound for the construction and analysis of decision tree. The quality of the trees produced by Hoeffding tree is not influenced by the incremental nature of it. Hoeffding bound determines the number of instances required for attaining an acceptable level of confidence. . The Hoeffding bound confines that with probability $1-\mu$, a random variable of range R will not change from the predicted mean after k interpretations by more than

$$\varepsilon = \sqrt{\frac{R^2 \ln(1/\mu)}{2k}}$$

[Manju, bB.R, A. Joshuva *et al*. (2018)]. The interesting property of Hoeffding tree is that it is independent of the probability distribution generating the observation [Pedro Domingos].

4 RESULTS AND DISCUSSIONS

4.1 Classification using Random forest

Table-1 shows the stratified cross validation details of the classifier whose features were selected using RP and J48, Tables 2 and 3 gives the detailed accuracy by class, Tables 4 and 5 shows the confusion matrix and Table 6 gives values for objects of the trained Random forest.

Table-1. Stratified cross validation.

Parameters	Feature selection using RP	Feature selection using J48
Correctly Classified Instances Incorrectly Classified Instances	578 21	586 13
Kappa statistic	0.9474	0.9674
Mean absolute error	0.0599	0.0641
Root mean squared error	0.1625	0.1493
Relative absolute error	13.4924%	14.4442%
Root relative squared error	34.5069%	31.6844%
Total number of instances	599	599

Table-2. Detailed accuracy by class (FS using RP).

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.979	0.012	0.974	0.979	0.976	0.965	0.997	0.998	Low
0.969	0.032	0.935	0.969	0.952	0.929	0.982	0.947	Medium
0.949	0.008	0.986	0.949	0.967	0.949	0.982	0.975	High
0.965	0.017	0.966	0.965	0.965	0.948	0.986	0.970	Weighted Avg

Table-3. Detailed accuracy by class (FS using J48).

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.984	0.005	0.989	0.984	0.987	0.981	0.998	0.989	Low
0.979	0.025	0.955	0.979	0.967	0.951	0.979	0.952	Medium
0.972	0.005	0.991	0.972	0.981	0.971	0.984	0.973	High
0.978	0.011	0.979	0.978	0.978	0.968	0.987	0.972	Weighted Avg

Table-4. Confusion matrix (FS using RP).

Classified as	Low	medium	High
Low	186	3	1
Medium	4	188	2
High	1	10	204

Table-5. Confusion matrix (FS using J48).

Classified as	Low	medium	High
Low	187	3	0
Medium	2	190	2
High	0	6	209

Table-6. Values for objects of the trained Random forest.

Attribute	Values (FS using RP)	Values (FS using J48)
Number of iterations(I)	10	18
Number of features(K)	0	1
Seed(S)	1	12

The classifier depends on three variables which are number of iterations, number of features and seeds. The variation of these parameters vs. the algorithms classification accuracy is plotted in Figure 2-4.

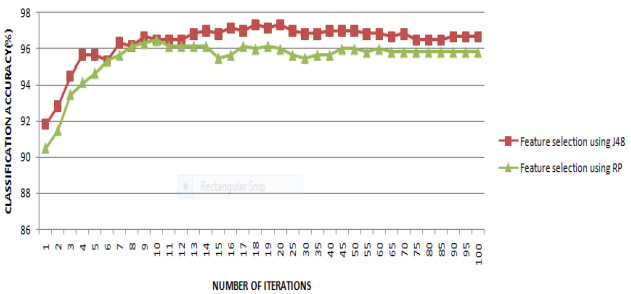

Figure-2. Number of iterations vs. classification accuracy.

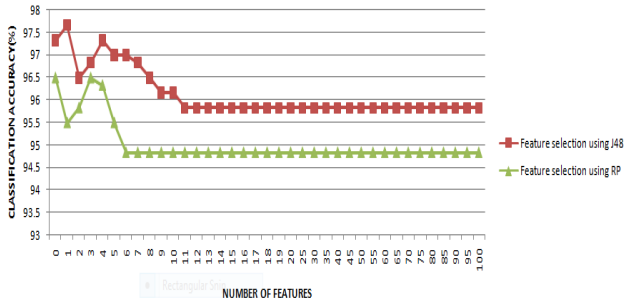

Figure-3. Number of features vs. Classification accuracy.

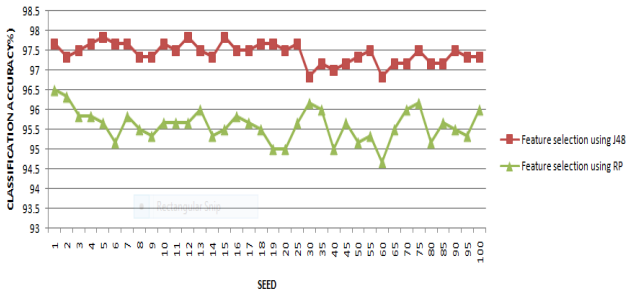

Figure-4. Seed vs. Classification accuracy.

Varying the parameter titled 'number of iteration' (Figure-2) from 1 to 100 increases the classification accuracy. However on further changes in the variable, it was noted that the classification accuracy got fluctuated and then remained constant in the case of both FS using J48 and RP. But the accuracy of J48 stood higher. In both the cases, on varying the parameter titled 'Number of features' (Figure-3) from 1 to 100 after certain stages, classification accuracy became a constant. On varying the parameters 'seed' (Figure-4) from 1 to 100, a fluctuation in classification accuracy is visible in both cases. Table 3 and 4 gives the values of objects for the trained algorithm. Confusion matrix table (Table-4) indicates that 186/190 samples were correctly classified as low, 188/194 samples were correctly classified as medium, and 204/215 samples were correctly classified as high. Confusion matrix table (Table-5) indicates that 187/190 samples were correctly classified as low, 190/194 samples were correctly classified as medium, and 207/215 samples were correctly classified as high. In the case of feature selection using RP the classifier attained a maximum classification accuracy of 96.4942% after training and in the case of feature selection using J48 the classifier attained a maximum classification accuracy of 97.8297 % after training.

4.2 Classification using logistic model tree (LMT)

Table-7 shows the stratified cross validation details of the classifier. Tables 8 and 9 gives the detailed accuracy by class, Tables 10 and 11 shows the confusion matrix and Table-12 gives values for objects of the trained LMT.

Table-7. Stratified cross validation.

Parameters	FS using RP	FS using J48
Correctly Classified Instances Incorrectly Classified Instances	577 22	582 17
Kappa statistic	0.9449	0.9574
Mean absolute error	0.06	0.039
Root mean squared error	0.1619	0.1403
Relative absolute error	13.5246	8.7932%
Root relative squared error	34.3606	29.7754%
Total number of instances	599	599

Table-8. Detailed accuracy by class (FS using RP).

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.974	0.010	0.979	0.974	0.976	0.965	0.994	0.985	Low
0.964	0.035	0.930	0.964	0.947	0.921	0.978	0.957	Medium
0.953	0.010	0.981	0.953	0.967	0.949	0.983	0.980	High
0.963	0.018	0.964	0.963	0.963	0.945	0.985	0.974	Weighted Avg

Table-9. Detailed accuracy by class (FS using J48).

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.979	0.002	0.995	0.979	0.987	0.981	0.997	0.984	Low
0.969	0.022	0.954	0.969	0.962	0.943	0.982	0.967	Medium
0.967	0.018	0.967	0.967	0.967	0.949	0.986	0.982	High
0.972	0.015	0.972	0.972	0.972	0.957	0.988	0.978	Weighted Avg

Table-10. Confusion matrix (FS using RP).

Classified as	Low	Medium	High
Low	185	4	1
Medium	4	187	3
high	0	10	205

Table-11. Confusion matrix (FS using J48).

Classified as	Low	Medium	High
Low	186	2	2
Medium	1	188	5
high	0	7	208

Table-12. Values for objects of the trained LMT.

Attribute	Values (FS using RP)	Values (FS using J48)
Number of Boosting Iterations (I)	1	19
Minimum Number of Instances (M)	13	1
Weight Trim Beta(W)	0.0	0

The classifier depends on three variables which are number of boosting iterations, minimum number of instances and weight trim beta. The variation of these parameters vs. the algorithms classification accuracy is plotted in Figures 5-7.

Figure-5. Number of boosting iterations vs. Classification accuracy.

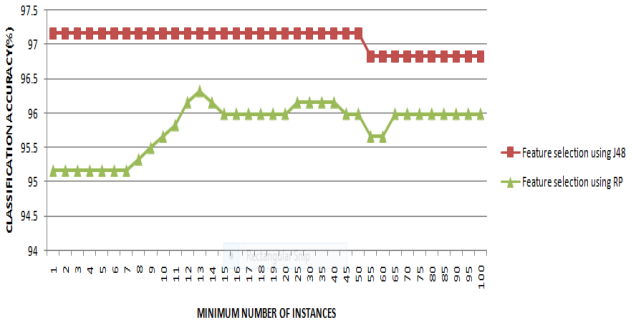

Figure-6. Minimum number of instances vs Classification accuracy.

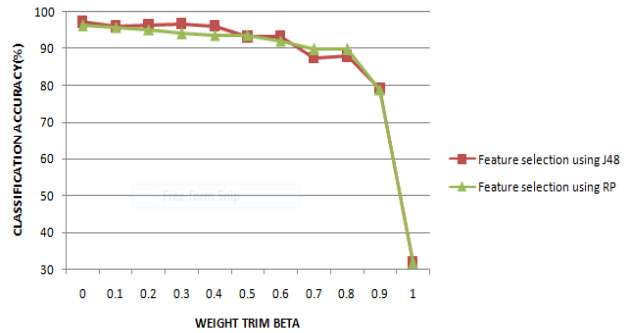

Figure-7. Weight trim beta vs. Classification accuracy.

Varying the parameter titled 'Number of boosting iterations' (Figure-5) from -3 to 100 a fluctuation in classification accuracy is visible in the case of feature selection by RP and J48. In the case of feature selection using RP varying the parameter titled 'minimum number of instances' (Figure-6) from 1 to 100 remains constant for some time and then increases gradually. Consequently as the variable increases the accuracy becomes constant. In the case of feature selection using J48 varying the parameter titled 'minimum number of instances' (Figure-6) from 1 to 100 the classification accuracy remains constant for some time then decreases and on further variation it becomes constant. In both the cases on varying the parameters 'weight trim beta' (Figure-7) from 0 to 1 a fluctuation in classification accuracy is noted and after certain stage the accuracy suddenly drops. Table-12 gives the values of objects for the trained algorithm. Confusion matrix table (Table-10) indicates that 185/190 samples were correctly classified as low, 187/194 samples were correctly classified as medium, and 205/215 samples were correctly classified as high. Confusion matrix table (Table-11) indicates that 186/190 samples were correctly classified as low, 188/194 samples were correctly classified as medium, and 208/215 samples were correctly classified as high. In the case of feature selection using RP the classifier attained a maximum classification accuracy of 96.3272% after training and in the case of feature selection using J48 the classifier attained a maximum classification accuracy of 97.1619% after training.

4.3 Classification using Hoeffding tree

Table-13 shows the stratified cross validation details of the classifier. Tables 14 and 15 gives the detailed accuracy by class, Tables 16 and 17 shows the confusion matrix and Table 18 gives values for objects of the trained Hoeffding tree.

Table-13. Stratified cross validation.

Parameters	FS using RP	FS using J48
Correctly Classified Instances Incorrectly Classified Instances	521 78	556 43
Kappa statistic	0.8042	0.8922
Mean absolute error	0.1028	0.0576
Root mean squared error	0.245	0.2059
Relative absolute error	23.1722%	12.98%
Root relative squared error	52.001%	43.7188%
Total number of instances	599	599

Table-14. Detailed accuracy by class (FS using RP).

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.868	0.051	0.887	0.868	0.878	0.822	0.968	0.952	Low
0.809	0.074	0.840	0.809	0.824	0.742	0.959	0.924	Medium
0.926	0.070	0.881	0.926	0.902	0.846	0.980	0.971	High
0.870	0.066	0.869	0.870	0.869	0.805	0.970	0.950	Weighted Avg

Table-15. Detailed accuracy by class (FS using J48).

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.884	0.005	0.988	0.884	0.933	0.908	0.974	0.964	Low
0.964	0.086	0.842	0.964	0.899	0.850	0.958	0.873	Medium
0.935	0.016	0.971	0.935	0.953	0.927	0.971	0.974	High
0.928	0.035	0.935	0.928	0.929	0.896	0.968	0.938	Weighted Avg

Table-16. Confusion matrix (FS using RP).

Classified as	Low	Medium	High
Low	165	18	7
Medium	17	157	20
high	4	12	199

Table-17. Confusion matrix (FS using J48).

Classified as	Low	medium	High
Low	168	22	0
Medium	1	187	6
High	1	13	201

Table-18. Values for objects of the trained Hoeffding tree.

Attribute	Value of feature selection using RP	Value of feature selection using J48
Hoeffding Tie Threshold	0.1	0.25

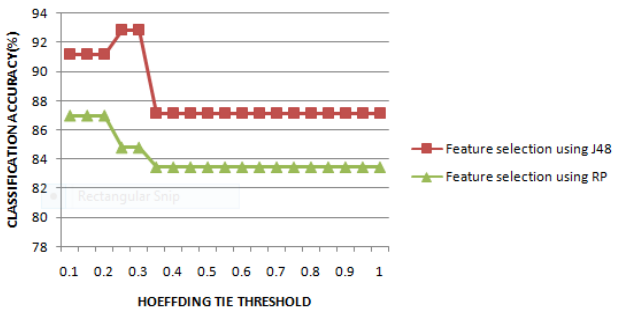

Figure-8. Hoeffding tie threshold vs. Classification accuracy.

In the case of feature selection using RP varying the parameter titled 'hoeffding tie threshold' (Figure-8) from 0.1 to 1 in the step of '0.2' the classification accuracy remained constant to a certain stage and beyond 0.35 the accuracy remains consistent. In the case of feature selection using J48 varying the parameter titled 'hoeffding tie threshold' (Figure-8) from 0.1 to 1 in the step of '0.2' the classification accuracy remained constant to a certain stage then the classification accuracy increases and decreases. Beyond 0.35 the accuracy remains consistent. Table-18 gives the values of objects for the trained algorithm. Confusion matrix table (Table-16) indicates that 165/190 samples were correctly classified as low, 157/194 samples were correctly classified as medium, and 199/215 samples were correctly classified as high. Confusion matrix table (Table-17) indicates that 168/190 samples were correctly classified as low, 187/194 samples were correctly classified as medium, and 201/215 samples were correctly classified as high. In the case of feature selection using RP the classifier attained a maximum classification accuracy of 86.9783% after training and in the case of feature selection using J48 the classifier attained a maximum classification accuracy of 92.8214% after training.

5. CONCLUSIONS

For the detection of lung cancer, Random Projection and J48 algorithm were used for feature selection and different classifiers were used for feature classification. The aim for finding out the most suitable algorithm for the detection well in advance to enhance the efficiency for doctor diagnosis is thus made possible. The accuracy of Random forest whose features were selected using J48 gave an accuracy of 97.8297% and that of RP gave an accuracy of 96.4942%. Hence, Random forest can be combined with J48 to give a good performance for the lung cancer detection. The combination of LMT and J48 is equally good when compared to the performance

accuracy of J48 Random forest combination. Hoeffding tree shows a relatively lesser performance when compared to other combinations.

REFERENCE

Abhigna B. S., Nishant Sai, Prof. Tapas Kumar. 2016. Lung Cancer Detection through Image Processing. International Journal of Scientific & Engineering Research, 7(12), ISSN 2229-5518.

Ahmed Jamal MBBS, Brian A. King PhD, Linda J. Neff PhD, Jennifer Whitmill MPH, Stephen D. Babb MPH, Corinne M. Graffunder DrPH. 2016. Cigarette smoking among adults-United States, 2005-2015, Centers for Disease Control and Prevention.

American Cancer Society. 2012. Cancer Facts & Figures 2012. Atlanta: American Cancer Society.

Anju Radhakrishnan, Vaidhehi V. 2017. Email Classification Using Machine Learning Algorithms, International Journal of Engineering and Technology (IJET). 9(2).

Anshul Goyal, Rajni Mehta. 2012. Performance Comparison of Naïve Bayes and J48 Classification Algorithms, International Journal of Applied Engineering Research, ISSN 0973-4562, 7(11).

Dmitriy Fradkin, David Madigan. 2002. Experiments with Random Projections for Machine Learning.

Eesha Goel, Er. Abhilasha. 2017. Random Forest: A Review. International Journal of Advanced Research in Computer Science and Software Engineering. 7(1).

Eibe Frank, Yong Wang, Stuart Inglis, Geoffrey and Ian H. Witten (1998): Technical Note: Using Model Trees for Classification, Machine Learning. 32, 63-76.

Ella Bingham, Heikki Mannila. 2001. Random projection in dimensionality reduction: Application to image and text data, Laboratory of Computer and Information Science, Helsinki University of Technology.

Farhad Alam, Sanjay Pachauri. 2017. Comparative Study of J48, Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA, Advances in Computational Sciences and Technology, ISSN 0973-6107, 10(6): 1731-1743

Haozhe Xie, Jie Li, Qiaosheng Zhang, Yadong Wang. 2016. Comparison among dimensionality reduction techniques based on Random Projection for cancer classification, Computational Biology and Chemistry. 65, 165-172.

Jessica Lin, Dimitrios Gunopulos. 2003. Dimensionality Reduction by Random Projection and Latent Semantic Indexing, Department of Computer Science & Engineering, University of California, Riverside.

Krishnaiah V., Dr. Narsimha G., Dr. Subhash Chandra N. 2013. Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques, International Journal of Computer Science and Information Technologies. 4(1): 39-45.

Kun Liu; Hillol Kargupta; Senior Member, IEEE, and Jessica Ryan. 2006. Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining, IEEE Transactions on Knowledge and Data Engineering. 18(1).

Madhusudana C. K., Budati S., Gangadhar N., Kumar H., Narendranath S. 2016. Fault diagnosis studies of face milling cutter using machine learning approach, Journal of Low Frequency Noise, Vibration and Active Control. 35(2): 128-138.

Manju B.R., Joshua, A., Sugumaran V. 2018. A Data Mining Study for Condition Monitoring on Wind Turbine Blades Using Hoeffding Tree Algorithm through Statistical and Histogram Features. International Journal of Mechanical Engineering and Technology (IJMET), 9(1): 1061-1079 Article ID: IJMET_09_01_113.

Neha Panpaliya, Neha Tadas, Surabhi Bobade, Rewti Aglawe, Akshay Gudadhe. 2015. A Survey on Early Detection and Prediction of Lung Cancer, International Journal of Computer Science and Mobile Computing, IJCSMC. 4(1): 175-184.

Pankaj Singh, Sudhakar Singh, Rakhi Garg, Devisha Singh. 2015. Comparative Study of Data Mining Algorithms through Weka. International Journal of Emerging Research in Management &Technology ISSN: 2278-9359, 4(9).

Pedro Domingos, Geoff Hulten: Mining High-Speed Data Streams.

Poonam Pandey, Radhika Prabhakar. 2016. An Analysis of Machine Learning Techniques (J48 & AdaBoost) -for Classification, 1st India International Conference on Information Processing (IICIP), DOI: 10.1109/IICIP.2016.7975394

Saravanan N., Dr. Gayathri V. 2015. Comparing Analysis of Decision Tree Algorithms Based on Healthcare, International Journal of Advanced Research in Computer Science and Software Engineering. 5(9).

Vanaja S., Rameshkumar K. 2014. Performance Analysis of Classification Algorithms on Medical Diagonosis – a survey, Journal of Computer Science.