ISSN:2229-6093

THORACIC SURGERY ANALYSIS USING DATA MINING TECHNIQUES

Sindhu.V, S.A.Sathya Prabha, S.Veni and M.Hemalatha
Department of Computer Science, Karpagam University
Coimbatore, TamilNadu, India.
sinnuveluu@gmail.com
sathyaselvaraj08@gmail.com
venikarthik04@gmail.com
csresearchhema@gmail.com

ABSTRACT

Data Mining is taking out of hidden pattern from a huge database. In data mining, machine learning is mainly focused as research which is mechanically learnt to identify complex patterns and make intelligent decisions based on data. These days Lung Tumor is one of the major causes of death in the developing countries. Today, lung tumor is the most frequent indication for thoracic surgery. By classification, general thoracic surgery includes knowledge, methodological skill and judgment to diagnose and treat diseases of the chest. In this paper the data classification is ***Thoracic Surgery (Lung Cancer)*** *patients' data set which is consists of* ***470 instances*** *with* ***14 different attributes*** *is collected retrospectively. Traditionally, more standard Dm algorithms has presented to the society past decades. Among them we cannot access all Dm algorithms. So, in this paper presents, choosen best one algorithm among the Standard Dm algorithms for reduce the search and time complexity.*

Keywords: Lung Tumor, Thoracic Surgery, Diagnosis, NaiveBayes, RandomForest, OneR, PART, DecisionStump, J48.

1. INTRODUCTION

A vital configuration process in the exertion to offer a valid Medical Decision Support System (MDSS) is to observe, process, and select different structures to represent the information of the acquired results and the distinctive patterns of clients, which may be authorized clients, patients, medical practitioners and students. It is comprehended that different kinds of clients has expanded necessities for representation of the extracted learning. Because of high therapeutic lab requests, serious time imperatives, and the current status of the data base in thoracic surgery [15], it is barely conceivable to predict the exact thoracic severe by means of discovering from the existing patient information and patterns in a complete and auspicious way. Presenting new helpful therapeutic equipment considerably more necessities on the clinics to understand, apply, survey the execution and securely develop the new devices.

Working on MDSS with a capacity to assemble and combine patient information, and predictive learning of the patterns are captured with the data, to give an appraisal of a patient situations, results and conclusions when applying mechanical circulatory medical practitioners and by patients at the end. Therefore it is mandatory to represent more contained structures information.

The data mining is considered as recognizing "good, novel, possibly functional, and ultimately reasonable patterns regards the data". With a challenging end goal to understand these regularities several strategies can be utilized. For example machine learning, statistical analysis, modelling techniques, database technology or human-computer interaction. These data mining strategies begin in the AI (Artificial Intelligence) [19] and the machine learning. Despite the statement that the data mining is somewhat a young discipline (about 25 years old) [10], it is prominent because of great requisitions in telecommunication, marketing and tourism. In recent years, the usefulness of the methods also been proven in medicine. Data mining aims at portraying particular patterns (dependencies, interrelations, various regularities) which may be available in the data. Such knowledge may additionally have enormous value for executed in treatment planning, risk analysis and different forecasts. Prior to the mining procedure, it is essential to gain a sufficient amount of information.

Current health centres not only focuses on doctors, patients and medical staff, but also in different methodologies, including the patient's treatment. In recent years, advanced frameworks and techniques have been aware in healthcare institutions to make possible their operations possible [18]. A tremendous measure of medical records is archived in databases and data warehouses. The more advanced one is therapeutic staff record of patients' visits and detailed information relating to their health condition. A few frameworks like encouraging patients' enrollment, units' accounts and booking of visits. As of latest another type of restorative framework has been risen

[24], [18]: medicinal choice help supportive network. It starts in the business discrimination and is to help medicinal choices. The information which is saved in such a framework may hold significant learning covered up in medicinal records [10]. The situation depicted above is the explanation behind a nearby joint effort between the researchers and the medicinal staff. It aims at the development of the most suitable strategy for data preparing which might enable discovering nontrivial rules and conditions in data. The test outcomes may enhance the methodology of diagnosing and treatment, in addition to diminish the danger of a restorative mix-up or the time of a finding diagnosis delivery. This can be turned out to be critical, particularly in emergency incidents. These undertaking points at recognizing and assessing the most well-known data mining algorithms executed in present day Medical Decision Support Systems (MDSS's) [23]. Assessment of different data mining strategies has been now displayed in various exploration papers [8], [27], [17]. However, they concentrate on a small amount of restorative data sets [20], the calculations utilized are not balanced (tried just one parameters' settings) or the c algorithms compared are not regular in the MDSS's [27].

2. RELATED WORKS

During a make contact in a healthcare entity a physician evaluates a patient’s condition. Symptoms are the basis for a diagnosis. This information might be stored either in a medical unit’s classification or in patient’s files. This data may include nontrivial dependencies, which may revolve out to be expensive. There are many methods and algorithms used to extract data for hidden information using artificial neural networks, decision trees, association rules and Naïve Bayes, support vector machines, cluster, logistic regression to name just a few[]. Studying the text, it turns out to be the most repeated choices for the Medical Decision Support Systems are the decision trees (C4.5 algorithm), Multilayer Perceptron [9] and the Naïve Bayes. These algorithms are very useful in remedy because they can decrease the time spent for processing symptoms and producing diagnoses, making them more particular at the same time. This paper aims at satisfying this gap in the body of awareness.

The authors of Cosic D., Loncaric S., Duch W. Et.al., and Richards G., have worked on remedial rules induction. Cosic D., et al., has presented a study on unsupervised fuzzy clustering algorithms and rule based systems, which are useful in classification of tomography images. However, in other applications their helpfulness is much lower. In additional, their independence allows for changing one rule does not affect the others. Duch W. Et al., the rule extraction is achieved with the use of a Multilayer Perceptron. The authors have proposed an algorithm C-MLP2LN. It generates supplementary nodes, deletes the connections among them, and optimizes the rules. Such solution leads to simpler and more accurate rules. Richards G., present a study on generation of rules which describe associations among attributes. The experiments are conducted on real medical data and their correctness is verified with the use of statistical procedures and surgeon evaluation. These projects present an examination of real data from St. Thomas’ Hospital in London. It also provides a clarification of all the steps performed from pre-processing, through data mining experiments to the verification of accuracy of the results. Another way to classify instance is with the usage of an artificial neural network.

Comak E.et.al., introduces artificial neural networks with back transmission for classification of heart disease cases. This solution is implemented in a medical system to support the classification of the Doppler signals in cardiology. The prediction yield by the way were more accurate than similar presented in Turkoglu I.et.al, West D., et.al, claim that Multilayer Perceptron is one of the most frequently employed neural networks algorithms in modern MDSS’s. They discuss applications of this algorithm to the categorization of different cancers (lung and breast cancers) and other diseases.

A common problem with data sets is they results from their small cardinality. Studies relating ways to overcome this problem in case of medical data has been offered in [4]. The author Duch .et.al, make use of an artificial neural network. The experiment has revealed the poor performance of the method which has yielded low-accuracy models, and the most popular include the probability distributions, estimated values, 15 variances and one or two sided intervals. The problem of statistical estimation of the algorithm’s performance is repeatedly brought up in professional literature. The authors of Choi, Y.S., Shim, Y.M., Kim, [6] discuss the difficulties that accompany proportional classification studies. They put effort to find a solution on how to choose the best machine learning method to reduce the bias while classifying unusual types of cancer.

The arithmetic comparisons of various classifiers of particular data are conducted. The authors of M. A. Meziane, and E. A. Zerhouni et al. [18] decided to use k-fold cross-validation [18] and repeated random sampling. The author Mitchell in [22] claims that it is important to consider assurance intervals, specially while comparing small data sets like for example microarray or other biological data. It is difficult to objectively evaluate the results obtained in different studies. This results from various pre-processing techniques, sampling strategies or learning methods that are applied earlier to the actual analyses. This can make a judgment difficult. Finally, yet importantly, insufficient testing strategy also leads to false conclusions about selected method [20]. These projects evaluate data mining algorithms under the same conditions on the same databases. This is to enable association of the algorithms. For assessing learning method's performance, various strategies are selected [6] leave-one out cross-validation (LOOCV), k-fold cross-validation [21], recurring random sub sampling (repeated hold-out method) and bootstrapping [22].

3. CLASSIFICATION TECHNIQUES

A major focus of machine learning research is automatically learning to recognize complex patterns and make intellectual decisions based on records. Hence, machine learning is strictly related to fields such as statistics, probability theory, data mining, pattern recognition, artificial intellect, adaptive organization, and abstract computer science.

3.1 Naive Bayesian Classifier

A Naive Bayesian classifier is a simple probabilistic classifier based on applying Bayesian theorem (from Bayesian statistics) with strong (naive) self-determination assumptions. By using the Bayesian theorem if can written

$$p(C|F_1 \ldots . F_n) = \frac{p(C)\, p(F_1 \ldots . F_n|C)}{p(F_{1 \ldots\ldots} F_n)}$$

Advantages

- It is fast, highly scalable model building and scoring
- Balance linearly with the number of predictors as well as rows
- The build procedure for Naive Bayes is parallelized
- Uses evidence from several attributes, the Naïve Bayes can be used for both binary and multiclass classification efforts.

3.2 PART (Partial Decision Trees) Classifier

PART is a rule based algorithm and produces a set of if-then rules that can be used to classify data. It is a modification of C4.5 and RIPPER algorithms and draws strategy from both. PART adopts the divide-and-conquer strategy of RIPPER and combines it with the decision tree approach of C4.5.

PART generates a set of rules according to the divide-and-conquer strategy, removes all instances from the training collection that are covered by this rule and takings recursively until no instance remains.

To generate a single rule, PART builds a partial decision tree for the current set of instances and chooses the leaf with the largest reporting as the new rule. It is different from C4.5 because the trees built for each rule are one-sided, based on the remaining set of examples and not complete as in the case of C4.5.

Advantages

- It is simpler and has been created to give sufficiently tough rules.

3.3 J48 Decision Tree Classifier

J48 is a simple tree [18], it creates a binary tree. C4.5 builds decision trees from a set of training data which is like an ID3, using the model of information entropy.

The algorithm

- Check for bottom cases
- For each attribute 'c' finds the normalized in sequence gain from splitting on 'c'
- Let c_best be the attribute with the maximum normalized information gain
- Generate a decision node that splits on a c_best
- Alternative on the sub lists obtained by splitting on c_best, and add those nodes as children of the node

Advantages

- Gains the balanced flexibility and correctness
- Limits the number of achievable decision points
- Provides higher accuracy

3.4 OneR (One Rule) Classifier

"One Rule" known as OneR is a simple, yet accurate classification algorithm that generates one rule for each analyst in the data, and then selects the rule with the smallest total error as its "one rule". To create a rule for a predictor, the researchers construct a frequency table for each predictor next to the target. It has been shown that OneR produces rules only a little less accuracy than the state-of-the-art classification algorithms while producing rules that are simple for humans to understand.

The algorithm

For each value of that interpreter, make a rule as follows

- Count how repeatedly each value of the target (class) appears
- Find the most frequent class

- Calculate the total error of the rules for each predictor
- Choose the predictor with the smallest total error.

3.5 Random Forest Tree Classifier

A random forest [14] consisting of a collected works of tree structured classifiers (h (x, _k), k = 1,), Where the _k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

The algorithm

- Choose T number of trees to produce
- Choose m number of variables used to split each node. m<<M, where M is the number of input variables, m is holed constant while growing the forest
- Construct a bootstrap sample of size n sampled from Sn with the replacement and grow a tree from this bootstrap sample
- When growing a tree at each node select m variables at random and use them to find the best split
- Grow the tree to a maximal amount and there is no pruning

3.6 DecisionStump Classifier

A decision stump is a machine learning model with one level decision tree also called as 1-rules. It includes one root node that is associated to the terminal nodes. A decision stump makes a prediction based on the value of a single input feature. This is a small decision tree with a single split. This decision stump works on both numerical and nominal attributes. In nominal features, it could have a stump with two leaves that corresponds to some exacting category, and remaining leaf trends to all the other categories. In the case of binary features two schemas are identical and missing value is taken as another category. In continuous description, some threshold value is selected, and the stump contains two leaves for values below and above the threshold value. However, multiple thresholds are selected rarely; therefore the stump contains three or more levels.

WEKA Tool

Weka (Waikato Environment for Knowledge Analysis) is a collected works of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also compatible for developing new machine learning schemes.

Weka Data Format (ARFF)

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instance sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato to be used with the Weka machine learning software. This document describes the version of ARFF used with Weka versions 3.2 to 3.3; this is an extension of the ARFF format as described in the data mining book written by Ian H. Witten and Eibe Frank (the new additions are string attributes, date, attributes).

4. DATA SET DESCRIPTION

The data were collected retrospectively at the Wroclaw Thoracic Surgery Centre for patients who have undergone major lung resections for primary lung cancer in the years 2007 to 2011. **Thoracic Surgery (Lung Cancer)** patients' data set is developed by collecting data from the hospital repository consists of **470 instances** with **14 different attributes**. The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, while the research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland.

Table 4.1 Attribute details in the dataset

Attributes Name	Type	Attribute description
PRE4	Numeric	Forced vital capacity - FVC (numeric)
PRE5	True, False	A volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
PRE6	PRZ2, PRZ1, PRZ0	Performance status - Zubrod scale
PRE7	True, False	Pain before surgery
PRE8	True, False	Haemoptysis before surgery
PRE9	True, False	Dyspnoea before surgery
PRE10	True, False	Cough, before surgery
PRE11	True, False	Weakness before surgery
PRE14	OC11, OC14,	T in clinical TNM - size of the original tumor, from

	OC12, OC13	OC11 (smallest) to OC14 (largest)
PRE17	True, False	Type 2 DM - diabetes mellitus
PRE19	True, False	MI up to 6 months
PRE25	True, False	PAD - peripheral arterial diseases
PRE30	True, False	Smoking
PRE32	True, False	Asthma
AGE	Numeric	Age at surgery
RISK1Y	True, False	1 year survival period - (T)rue value if dead

5. EXPERIMENTAL RESULTS

5.1 Data Preparation

The variables are already categorized and represented by numbers. The manner in which the collision occurred has been categorised as three.

5.1.1 Based on Diagnosis

Diagnosis is the detailed combination of ICD-10 codes for primary and secondary as well as multiple tumor if some (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1)

Table 5.1 & Fig5.1 shows the evaluation results for the above six methods reveals that the performance using the researchers proposed prediction to improve the overall performance of all the six classification methods, and the noticeable improvement is that logistic regression produces a very significant change in its performance represents 100 and 95.65 which outperforms the other classification methods. The feature ranking techniques are based on three different subsets of attributes such as Diagnosis, Performance and Tumor size. The remaining 14 attributes subset consists of performance of the tumor patients' details. From the results obtained it is observed that the attribute with the six different classifier algorithms have the Random Forest classifier is the best algorithm and higher accuracy than the other set of algorithms. So hereby it is concluded that the three main attributes with six algorithms, the random forest is the future processing instead of using the other attributes.

.

5.1.2 Based on Performance

Performance status - Zubrod scale (PRZ2, PRZ1, and PRZ0). There are different ways of assessing general health. The World Health Organization designed the scale that doctors use most often. It has categories from 0 to 5.

- 0 – you are fully active and more or less as you were before your illness
- 1 – you cannot carry out heavy physical work, but can do anything else
- 2 – you are up and about more than half of the day and can look after yourself, but are not well enough to work
- 3 – you are in bed or sitting in a chair for more than half of the day and you need some help in looking after yourself
- 4 – you are in bed or a chair all the time and need a lot of looking after yourself
- 5 – death

5.1.3 Based on Size of the tumor

T in clinical TNM - size of the original tumor, from OC11 (smallest) to OC14 (largest) (OC11, OC14, OC12, OC13)

Table 5.1 Accuracy and testing for existing algorithm

Algorithm	Accuracy of Testing Methods			
	Correctly Classified Instances	Testing Percentage split 70%	Testing Percentage split 80%	Testing Percentage split 90%
Naïve Bayes	82.34	82.26	80.85	76.59
PART	91.91	76.59	71.27	72.34
J48	85.10	83.68	81.91	76.59

OneR	85.31	82.97	81.91	76.59
Random Forest	95.65	85.97	83.53	77.21
Decision Stump	85.10	84.39	82.97	76.59

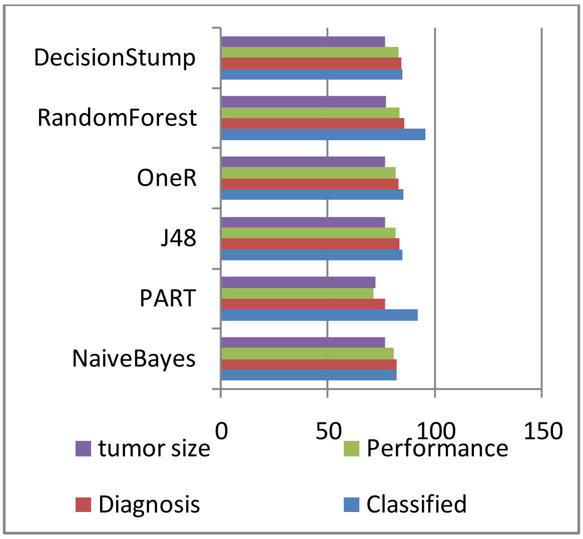

Figure 5.1 Overall Analysis of Classification, Diagnosis, and Performance and Tumor size

To aid clinicians in the diagnosis of lung cancer, recent researchers has looked into the development of computer aided diagnostic tools. Different data mining techniques have been widely used for lung cancer analysis. This paper focuses on the patient improvements based on their diagnosis in the thoracic surgery Lung cancer patient's dataset. The result reveals the accuracy of their testing methods based on their percentage. The evaluation results show that the four accuracy testing methods reveal that the performance using currently classified instance, drastically improves the overall performance of all the six classification methods. And the noticeable improvement is that Random forest produces (95.65) a very significant impact on the classification method. Here, it is concluded that the proposed approach given has the best classification accuracy compare to the then remaining approaches.

Table 5.2 Accuracy, Precision, Recall, F measure for existing algorithm

Attributes	Naïve Bayes	PART	J48	OneR	Random Forest	DecisionStump
Accuracy	82.34	91.91	85.1	85.31	95.65	85.1
Precision	0.763	0.718	0.711	0.771	0.81	0.712
Recall	0.823	0.766	0.837	0.83	0.855	0.844
F-Measure	0.782	0.74	0.769	0.787	0.813	0.773

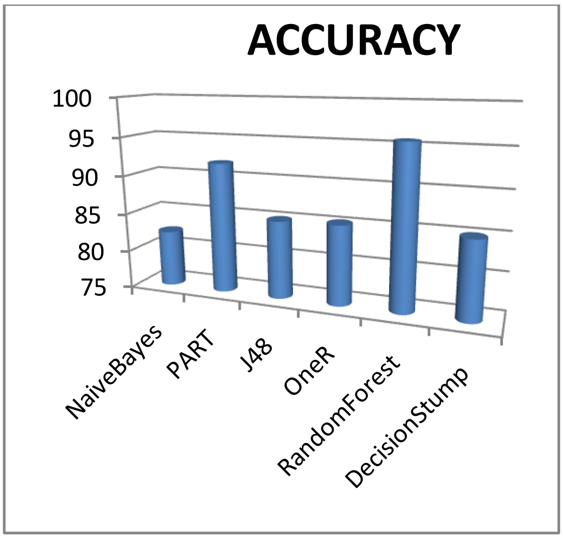

Figure 5.2 Analysis of Accuracy values

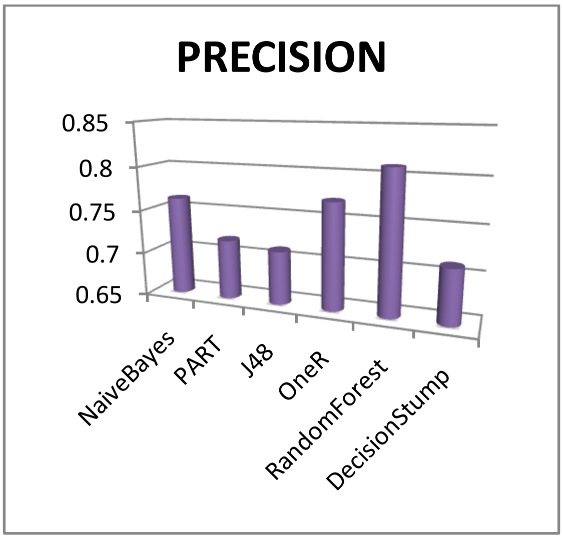

Figure 5.3 Analysis of Precision values

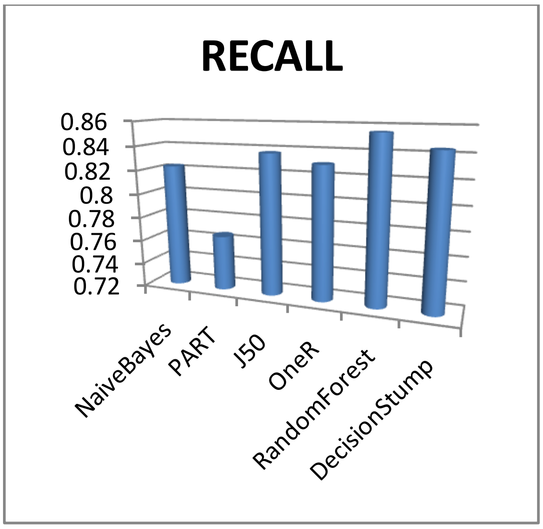

Figure 5.4 Analysis of Recall values

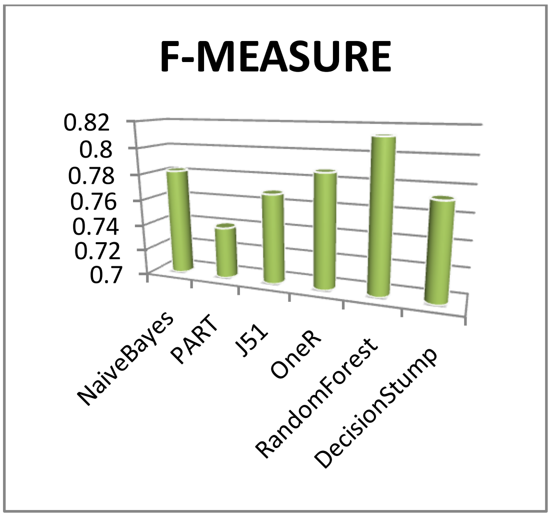

Figure 5.5 Analysis of F measure values

5.2 K-Fold Cross-validation

In data mining, in order to minimize the bias associated with the random sampling of the training and holdout (testing) data samples in comparing the predictive accuracy of two or more methods, researchers tend to use k-fold cross-validation (Kohavi, 1995). In k-fold cross-validation, also called rotation estimation, the complete dataset (D) is randomly split into k mutually exclusive subsets (the folds: 1 2... DD Dk) of approximately equal size [26] [27]. The classification model is trained and tested k times. Each time (t k ∈{1, 2,..., }), it is trained with all but one fold (Dt) and tested on the remaining single fold (Dt). The cross-validation estimate of the overall performance criteria is calculated as simply as the average of the k individual performance measured as follows,

$$CV = \frac{1}{k}\sum_{i=1}^{k} PM_i$$

Where CV stands for the cross-validation, k is the number of folds used, and PM is the performance measure for each fold (Olson & Delen, 2008). In this study, a stratified 10-fold cross-validation approach was used to estimate the performance of classifiers. Empirical studies have shown that 10 is the optimal number of folds that optimizes the time it takes to complete the test while minimizing the bias and variance associated with the validation process .In 10-fold cross-validation, the entire dataset is divided into 10 mutually exclusive subsets (or folds) with approximately the same class distribution as the original dataset (stratified). Each fold is used once to test the performance of the classifier that is generated from the combined data of the remaining nine folds, leading to 10 independent performance estimates.

CONCLUSION

The aim of this paper is to detect the causes of Lung Tumor. The dataset for the study contains Thoracic Surgery and it has totally 17 attributes of the year 2013 produced by the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland and investigates the performance of Naive Bayes, J48, PART, OneR, DecisionStump and Random Forest classifiers for predicting classification accuracy. The classification accuracy of the test result reveals the following three cases such as Diagnosis, Performance and Tumor size. When using the Naive Bayes algorithm, it shows the correctly classified percentage values for all varieties. J48 gives more accuracy than a Naïve Bayes classification algorithm. One rule induction gives results based upon any one of the varieties (attribute) and determines the best splitting terms of minimizing the training error. The rule induction generates only correct rules based on the accuracy. The Decision Stump gives the overall view (values and classes). Random Forest algorithm shows the different models and each model gives different results. Random Forest outperforms the other classification algorithms instead of selecting all the attributes for classification. In this analysis, it has been found that the Random Forest is the best in the thoracic surgery for predicting the algorithms.

Future Enhancement

The project reported in this paper opens new avenues for medical decision making. The primary decision-making and confirmation algorithms when combined together generate decisions of high accuracy. The diagnosis by the Random Forest algorithms has perfect accuracy for the clinical data reported in the paper. Additional developments of the algorithms and large-scale testing will be the ultimate proof of diagnostic accuracy for lung cancer and other diseases. The number of features necessary for high-accuracy autonomous diagnosis is smaller than in the original data set. This reduced number of features would lower the testing costs.

ACKNOWLEDGEMENT

We thank the Karpagam University and our guide Dr.M.Hemalatha for the Motivation and Encouragement to make this paper as successful one.

REFERENCES

[1] A .Kusiak., Computational Intelligence in Design and Manufacturing. New York: Wiley, 2000.

[2] B.Kavitha.,P.Sheeba Maybell, Dr.S.Karthikeyan and Dr,M.Hemalatha, et al. (2012) “An Emerging Method of Intutionistic Fuzzy Set for Breast Cancer Diagnosis”.

[3] D.M .Nathan., Cleary P.A., Backlund J.Y., Genuth S.M., Lachin J.M., Orchard T.J., Raskin P. and Zinman B., Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) Study Research Group. Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. The New England Journal of Medicine, 2005, vol. 353, 2643- 2653.

[4] D.M.Shahian, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. Ann Thorac Surg 2001;72:2155–68.

[5] G .Richards., Rayward-Smith V. J., Sönksen P. H., Carey S., Weng C., Data mining for indicators of early mortality in a database of clinical records. Artificial Intelligence in Medicine, 2000, vol. 22, no. 3, 215–231.

[6] G.Ruhe., “Qualitative analysis of software engineering data using rough sets,” in Proc. Fourth Int. Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery, Tokyo, Japan, 1996, pp. 292–299.

[7] J. W. Grzymala-Busse., “A new version of the rule induction system LEERS,” Fundamental Informaticae, vol. 31, pp. 27–39, 1997.

[8] J .Lodhia., T. Christensen, S. Trotter and E. Bishay, "Glomus Tumors of the Lung: Diagnostic Considerations and Treatment," *Open Journal of Thoracic Surgery*, Vol. 4 No. 1, 2014, pp. 5-8. doi: 10.4236/ojts.2014.41002.

[9] L.Kaiser, Singhal S. Surgical Foundations: Essential of Thoracic Surgery. CV Mosby;2004.

[10] M.A .Steegers, Snik DM, Verhagen AF, van der Drift MA, Wilder-Smith OH (2008) Only half of the chronic pain after thoracic surgery shows a neuropathic component. J Pain 9: 955–961. doi: 10.1016/j.jpain.2008.05.009

[11] M.Hunt, von Konsky B., Venkatesh S., Petros P., Bayesian networks and decision trees in the diagnosis of female urinary incontinence. Engineering in Medicine and Biology Society, Proceedings of the 22nd Annual International Conference of the IEEE, 2000, vol. 1, 551-554.

[12] M.A.Kinney, Hooten WM, Cassivi SD, Allen MS, Passe MA, et al. (2012) Chronic postthoracotomy pain and health-related quality of life. Ann Thorac Surg 93: 1242–1247. Do: 10.1016/j. athoracsur. 2012.01.031

[13] N. C Gupta., Maloof J., and E. Gunel, "Probability of malignancy in solitary pulmonary nodules using fluorine-18-FDG and PET," J. Nucl. Medicine, vol. 37, no. 6, pp. 943–948, 1996.

[14] P. Adriaans and D. Zantinge, Data Mining. New York: Addison Wesley, 1996.

[15] R.D.Searle, Simpson MP, Simpson KH, Milton R, Bennett MI (2009) Can chronic neuropathic pain following thoracic surgery be predicted during the postoperative period? Interact Cardiovasc Thorac Surg 9: 999–1002. doi: 10.1510/icvts.2009.216887

[16] S.Krishnaveni., Dr.M.Hemalatha, et al.,A Perspective Analysis of Traffic Accident using Data Mining Techniques".2012.

[17] S.Mizuguchi., M. Kaji, T. Yoshida, T. Iwasaki, T. Kamimori and H. Fujiwara, "Severe Mediastinal Emphysema and Tension Pneumothorax Caused by Cough-Induced Intercostal Lung Herniation," *Open Journal of Thoracic Surgery*, Vol. 4 No. 1, 2014, pp. 1-4. doi: 10.4236/ojts.2014.41001.

[18] .S.Senthamilaru and M.Hemalatha.et.al. "Dynamically Adaptive Count Bloom Filter for Handling Duplicates in Data Stream" 2013

[19] T. Chandra, S. R. Leclair, J. A. Meech, B. Varma, M. Smith, and B. Balachandran, Eds., "Rough sets and data mining," in Proc. Australasian-Pacific Forum on Intelligent Processing and Manufacturing of Materials, vol. 1, Gold Coast, Australia, 1997, pp. 663–667.

[20] T.W .Shields, LoCicero J, Ponn R B, Rusch VW. General Thoracic Surgery. 6th ed. Vols 1 and 2 Lippincott Williams & Wilkins;2004.

[21] T. W .Shields., C. T. Drake and J. C. Sherick, "Bilateral Primary Bronchogenic Carcinoma," The Journal of Thoracic and Cardiovascular Surgery, Vol. 48, 1964, pp. 401-417.

[22] W. Kowalczyk and F. Slisser, "Modeling customer retention with rough data models," in Proc. First Eur. Symp. PKDD '97, Trondheim, Norway, 1997, pp. 4–13.

[23] Y.S Choi,., Shim, Y.M., Kim, K. and Kim, J. (2004) Pattern of Recurrence after Curative Resection of Local (Stage I and II) Non-Small Cell Lung Cancer: Difference According to the Histologic Type. Journal of Korean Medical Science, 19, 674-676. http://dx.doi.org/10.3346/jkms.2004.19.5.674

[24] AJCC (American Joint Committee on Cancer). Cancer Staging Manual, 7th edition, Edge SB, Byrd DR, Compton CC, et al (Eds), Springer-Verlag, New York 2010. p.347.

[25] National Pilot to Reduce the Cancer Waiting Times www.acl.icnet.uk/lab/docs/ext/era.pdf (retrieved on 1.05.2007)

[26] http://dbigbear.blogspot.com/2007/10/k-fold-cross-validation.html

[27] http://shawndra.pbworks.com/f/A+study+of+cross-validation+and+bootstrap+for+accuracy+estimation+and+model+selection.pdf