ISSN:0975-9646

Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques

V.Krishnaiah [#1], Dr.G.Narsimha[*2], Dr.N.Subhash Chandra[#3]

[#1]*Associate Professor, Dept of CSE,*
CVR College of Engineering, Hyderabad, India
[#2]*Assistant Professor, Dept of CSE,*
JNTUH College of EngineeringKondagattu, Andrapradesh, India
[#3]*Professor of CSE & Principal, Holy Mary Institute of Technology and Science,*
Hyderabad, India

***Abstract*— Cancer is the most important cause of death for both men and women. The early detection of cancer can be helpful in curing the disease completely. So the requirement of techniques to detect the occurrence of cancer nodule in early stage is increasing. A disease that is commonly misdiagnosed is lung cancer. Earlier diagnosis of Lung Cancer saves enormous lives, failing which may lead to other severe problems causing sudden fatal end. Its cure rate and prediction depends mainly on the early detection and diagnosis of the disease. One of the most common forms of medical malpractices globally is an error in diagnosis. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this study, we briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information. For data preprocessing and effective decision making One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) are used. This is an extension of naïve Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with small or incomplete data sets. Discovery of hidden patterns and relationships often goes unexploited. Diagnosis of Lung Cancer Disease can answer complex "what if" queries which traditional decision support systems cannot. Using generic lung cancer symptoms such as age, sex, Wheezing, Shortness of breath, Pain in shoulder, chest, arm, it can predict the likelihood of patients getting a lung cancer disease. Aim of the paper is to propose a model for early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient.**

***Keywords*—Lung cancer, Naive Bayes, ODANB, NCC2, Data Mining, Classification.**

I. INTRODUCTION

Lung cancer is the one of the leading cause of cancer deaths in both women and men. Manifestation of Lung cancer in the body of the patient reveals through early symptoms in most of the cases. [1]. Treatment and prognosis depend on the histological type of cancer, the stage (degree of spread), and the patient's performance status. Possible treatments include surgery, chemotherapy, and radiotherapy Survival depends on stage, overall health, and other factors, but overall only 14% of people diagnosed with lung cancer survive five years after the diagnosis. Symptoms that may suggest lung cancer include:

- dyspnea (shortness of breath with activity),
- hemoptysis (coughing up blood),
- chronic coughing or change in regular coughing pattern,
- wheezing,
- chest pain or pain in the abdomen,
- cachexia (weight loss, fatigue, and loss of appetite),
- dysphonia (hoarse voice),
- clubbing of the fingernails(uncommon),
- dysphasia(difficulty swallowing),
- Pain in shoulder ,chest , arm,
- Bronchitis or pneumonia,
- Decline in Health and unexplained weight loss.

Mortality and morbidity due to tobacco use is very high. Usually lung cancer develops within the wall or epithelium of the bronchial tree. But it can start anywhere in the lungs and affect any part of the respiratory system. Lung cancer mostly affects people between the ages of 55 and 65 and often takes many years to develop [2].

There are two major types of lung cancer. They are Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) or oat cell cancer. Each type of lung cancer grows and spreads in different ways, and is treated differently. If the cancer has features of both types, it is called mixed small cell/large cell cancer.

Non-small cell lung cancer is more common than SCLC and it generally grows and spreads more slowly. SCLC is almost related with smoking and grows more quickly and form large tumors that can spread widely through the body. They often start in the bronchi near the center of the chest. Lung cancer death rate is related to total amount of cigarette smoked [3].

Smoking cessation, diet modification, and chemoprevention are primary prevention activities. Screening is a form of secondary prevention. Our method of finding the possible Lung cancer patients is based on the systematic study of symptoms and risk factors. Non-clinical symptoms and risk factors are some of the generic indicators of the cancer diseases. Environmental factors have an important role in human cancer. Many carcinogens are present in the air we breathe, the food we eat, and the water we drink. The constant and sometimes unavoidable exposure to environmental carcinogens complicates the investigation of cancer causes in human beings. The complexity of human cancer causes is especially challenging for cancers with long latency, which are

associated with exposure to ubiquitous environmental carcinogens.

Pre-diagnosis techniques

Pre-diagnosis helps to identify or narrow down the possibility of screening for lung cancer disease. Symptoms and risk factors (smoking, alcohol consumption, obesity, and insulin resistance) had a statistically significant effect in pre-diagnosis stage.[4]. The lung cancer diagnostic and prognostic problems are mainly in the scope of the widely discussed classification problems. These problems have attracted many researchers in computational intelligence, data mining, and statistics fields.

Cancer research is generally clinical and/or biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. As the use of computers powered with automated tools, large volumes of medical data are being collected and made available to the medical research groups. As a result, Knowledge Discovery in Databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical cases stored within datasets. The objective of this study is to summarize various review and technical articles on diagnosis of lung cancer. It gives an overview of the current research being carried out on various lung cancer datasets using the data mining techniques to enhance the lung cancer diagnosis.

II. LITERATURE FOR LUNG CANCER

The approach that is being followed here for the prediction technique is based on systematic study of the statistical factors, symptoms and risk factors associated with Lung cancer. Non-clinical symptoms and risk factors are some of the generic indicators of the cancer diseases. Initially the parameters for the pre-diagnosis are collected by interacting with the pathological, clinical and medical oncologists (Domain experts).

A. Statistical Incidence Factors:

i. Age-adjusted rate (ARR)
ii. Primary histology
iii. Area-related incidence chance
iv. Crude incidence rate

B. Lung cancer symptoms:

The following are the generic lung cancer symptoms [14].

i. A cough that does not go away and gets worse over time
ii. Coughing up blood (heamoptysis) or bloody mucus.
iii. Chest, shoulder, or back pain that doesn't go away and often is made worse by deep Hoarseness
iv. Weight loss and loss of appetite
v. Increase in volume of sputum
vi. Wheezing
vii. Shortness of breath
Viii.Repeated respiratory infections, such as bronchitis or pneumonia
Ix. Repeated problems with pneumonia or bronchitis
x. Fatigue and weakness
xi. New onset of wheezing
xii. Swelling of the neck and face
xiii. Clubbing of the fingers and toes. The nails appear to bulge out more than normal.
Xiv.Paraneoplastic syndromes which are caused by biologically active substances that are secreted by the tumor.
xv. Fever
xvi. Hoarseness of voice
Xvii.Puffiness of face
Xviii.Loss of appetite
Xix.Nausea and vomiting

C. Lung cancer risk factors:

a. Smoking:
i. Beedi
ii. Cigarette
iii. Hukka
b. Second-hand smoke
c. High dose of ionizing radiation
d. Radon exposure
e. Occupational exposure to mustard gas chloromethyl ether, inorganic arsenic, chromium, nickel, vinyl chloride, radon asbestos
f. Air pollution
g. Insufficient consumption of fruits & vegetables
h. Suffering with other types of malignancy

III. KNOWLEDGE DISCOVERY AND DATA MINING

This section provides an introduction to knowledge discovery and data mining. We list the various analysis tasks that can be goals of a discovery process and lists methods and research areas that are promising in solving these analysis tasks.

A. Knowledge Discovery Process

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process. The following figure. 1 shows data mining as a step in an iterative knowledge discovery process.

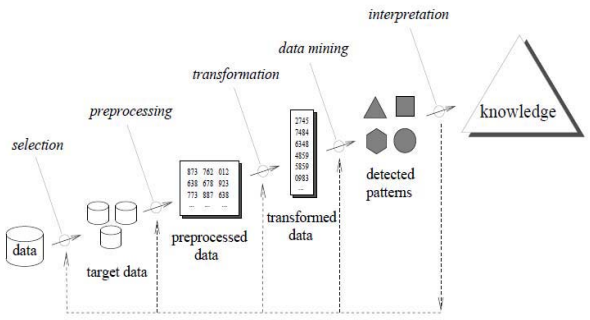

Figure 1. Steps in KDD

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge [5]. The iterative process consists of the following steps:

(1) *Data cleaning:* also known as data cleansing it is a phase in which noise data and irrelevant data are removed from the collection.

(2) *Data integration*: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

(3) *Data selection:* at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

(4) *Data transformation:* also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

(5) *Data mining:* it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

(6) *Pattern evaluation:* this step, strictly interesting patterns representing knowledge are identified based on given measures.

(7) *Knowledge representation:* is the final phase in which the discovered knowledge is visually represented to the user. In this step visualization techniques are used to help users understand and interpret the data mining results.

B. Data Mining Process

In the KDD process, the data mining methods are for extracting patterns from data. The patterns that can be discovered depend upon the data mining tasks applied. Generally, there are two types of data mining tasks: *descriptive data mining tasks* that describe the general properties of the existing data, and *predictive data mining tasks* that attempt to do predictions based on available data. Data mining can be done on data which are in quantitative, textual, or multimedia forms.

Data mining applications can use different kind of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects).Data mining involves some of the following key steps [6]-

(1) *Problem definition:* The first step is to identify goals. Based on the defined goal, the correct series of tools can be applied to the data to build the corresponding behavioral model.

(2) *Data exploration:* If the quality of data is not suitable for an accurate model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.

(3) *Data preparation:* The purpose of this step is to clean and transform the data so that missing and invalid values are treated and all known valid values are made consistent for more robust analysis.

(4) *Modeling:* Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next generation techniques such as decision trees, networks and rule based algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.

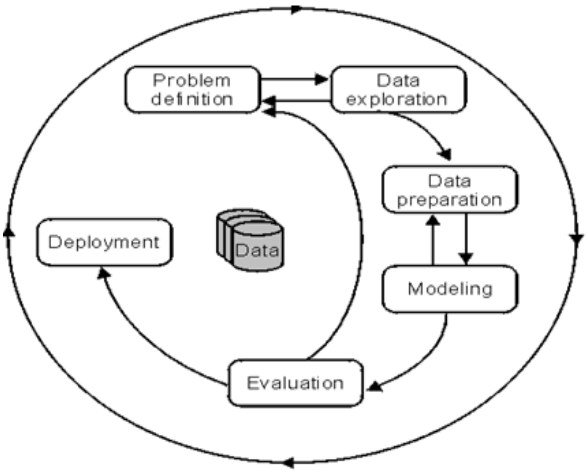

Figure 2. Data Mining Process Representation

(5) *Evaluation and Deployment:* Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

IV. DATA MINING CLASSIFICATION METHODS

The data mining consists of various methods. Different methods serve different purposes, each method offering its own advantages and disadvantages. In data mining, classification is one of the most important task. It maps the data in to predefined targets. It is a supervised learning as targets are predefined.

The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The most used classification algorithms exploited in the microarray analysis belong to four categories: IF-THEN Rule, Decision tree, Bayesian classifiers and Neural networks.

IF-THEN Rule:

Rule induction: is the process of extracting useful 'if then' rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. Knowledge represents has the form

IF conditions THEN conclusion:

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction. Rule Induction Method has the potential to use retrieved cases for predictions [7].

Decision Tree:

Decision tree derives from the simple divide-and-conquer algorithm. In these tree structures, leaves represent classes and branches represent conjunctions of features that lead to those classes. At each node of the tree, the attribute that most effectively splits samples into different classes is chosen. To predict the class label of an input, a path to a leaf from the root is found depending on the value of the predicate at each node that is visited. The most common algorithms of the decision trees are ID3 [8] and C4.5 [9]. An evolution of decision tree exploited for microarray data analysis is the random forest [10], which uses an ensemble of classification trees. [11] Showed the good performance of random forest for noisy and multi-class microarray data.

Bayesian classifiers and Naive Bayesian:

From a Bayesian viewpoint, a classification problem can be written as the problem of finding the class with maximum probability given a set of observed attribute values. Such probability is seen as the posterior probability of the class given the data, and is usually computed using the Bayes theorem. Estimating this probability distribution from a training dataset is a difficult problem, because it may require a very large dataset to significantly explore all the possible combinations.

Conversely, Naive Bayesian is a simple probabilistic classifier based on Bayesian theorem with the (naive) independence assumption. Based on that rule, using the joint probabilities of sample observations and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation. Despite its simplicity, the Naive Bayes classifier is known to be a robust method, which shows on average good performance in terms of classification accuracy, also when the independence assumption does not hold [12].

Artificial Neural Networks (ANN):

An artificial neural network is a mathematical model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. Neurons are organized into layers. The input layer consists simply of the original data, while the output layer nodes represent the classes. Then, there may be several hidden layers. A key feature of neural networks is an iterative learning process in which data samples are presented to the network one at a time, and the weights are adjusted in order to predict the correct class label. Advantages of neural networks include their high tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained. In [13] a review of advantages and disadvantages of neural networks in the context of microarray analysis is presented.

V. DATA MINING CLASSIFICATION METHODS

There are various data mining techniques available with their suitability dependent on the domain application. Statistics provide a strong fundamental background for quantification and evaluation of results. However, algorithms based on statistics need to be modified and scaled before they are applied to data mining. We now describe a few Classification data mining techniques with illustrations of their applications to healthcare.

A. *Rule set classifiers*

Complex decision trees can be difficult to understand, for instance because information about one class is usually distributed throughout the tree. C4.5 introduced an alternative formalism consisting of a list of rules of the form “if A and B and C and ... then class X”, where rules for each class are grouped together. A case is classified by finding the first rule whose conditions are satisfied by the case; if no rule is satisfied, the case is assigned to a default class.

IF conditions THEN conclusion

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction.

In the health care system it can be applied as follows:

(Symptoms) (Previous--- history) → (Cause—of---disease).

Example 1: If_then_rule induced in the diagnosis of level of alcohol in blood.

IF Sex = MALE AND Unit = 8.9 AND Meal = FULL THEN

Diagnosis=Blood_alcohol_content_HIGH.

B. *Decision Tree algorithm*

It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute). Tree based models which include classification and regression trees, are the common implementation of induction modeling [15]. Decision tree models are best suited for data mining. They are inexpensive to construct, easy to interpret, easy to integrate with database system and they have comparable or better accuracy in many applications. There are many Decision tree algorithms such as HUNTS algorithm (this is one of the earliest algorithm), CART, ID3, C4.5 (a later version ID3 algorithm), SLIQ, SPRINT [15].

The decision tree shown in Fig. 3 is built from the very small training set (Table 1). In this table each row corresponds to a patient record. We will refer to a row as a data instance. The data set contains three predictor attributes, namely Age, Gender, Intensity of symptoms and one goal attribute, namely disease whose values (to be predicted from symptoms) indicates whether the corresponding patient have a certain disease or not.

DECISION TREE

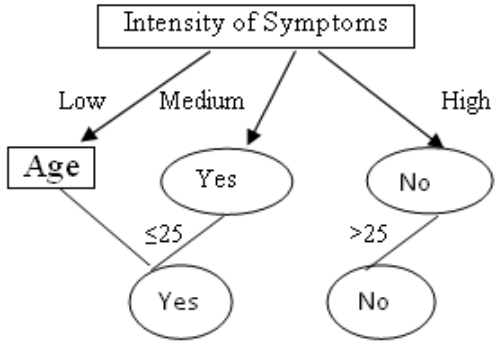

Figure 3. A decision tree built from the data in Table 1

Table 1: Data set used to build decision tree of Fig. 3

1) Age	Data Set		
	2) Gender	3) Intensity of Symptoms	4) Disease(goal)
25	Male	medium	yes
32	Male	high	yes
24	Female	medium	yes
44	Female	high	yes
30	Female	low	no
21	Male	low	no
18	Female	low	no
34	Male	medium	no
55	Male	medium	no

Decision tree can be used to classify an unknown class data instance with the help of the above data set given in the Table 1. The idea is to push the instance down the tree, following the branches whose attributes values match the instances attribute values, until the instance reaches a leaf node, whose class label is then assigned to the instance [15]. For example, the data instance to be classified is described by the tuple (Age=23, Gender=female, Intensity of symptoms = medium, Goal =?), where "?" denotes the unknown value of the goal instance. In this example, Gender attribute is irrelevant to a particular classification task. The tree tests the intensity of symptom value in the instance. If the answer is medium; the instance is pushed down through the corresponding branch and reaches the Age node. Then the tree tests the Age value in the instance. If the answer is 23, the instance is again pushed down through the corresponding branch. Now the instance reaches the leaf node, where it is classified as yes.

C. Neural Network Architecture

Especially, the neural network approach has been widely adopted in recent years. The neural network has several advantages, including its nonparametric nature, arbitrary decision boundary capability, easy adaptation to different types of data and input structures, fuzzy output values, and generalization for use with multiple images. Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions. (Actual biological neural networks are incomparably more complex.) Neural nets may used in classification problems (where the output is a categorical variable) or for regressions (where the output variable is continuous).

The architecture of the neural network shown in figure.4 consists of three layers such as input layer, hidden layer and output layer. The nodes in the input layer linked with a number of nodes in the hidden layer. Each input node joined to each node in the hidden layer. The nodes in the hidden layer may connect to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables [16].

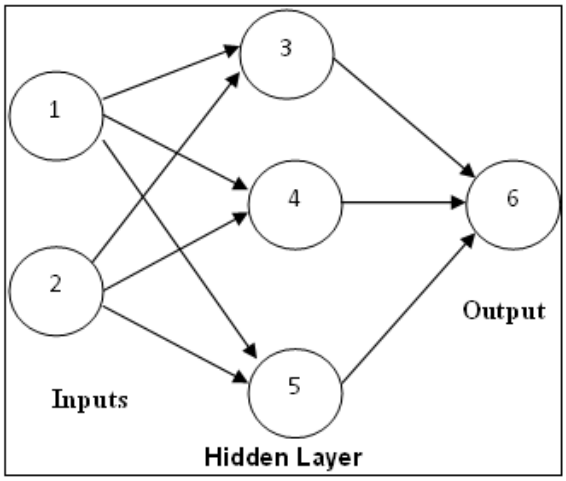

Figure 4. A neural network with one hidden layer.

A main concern of the training phase is to focus on the interior weights of the neural network, which adjusted according to the transactions used in the learning process. For each training transaction, the neural network receives in addition the expected output [17]. This concept drives us to modify the interior weights while trained neural network used to classify new images

D. . Bayesian Network Structure Discoveries

A conditional probability is the likelihood of some conclusion, *C*, given some evidence/observation, *E*, where a dependence relationship exists between *C* and *E*.

This probability is denoted as P*(C | E)* where

$$P\left(\frac{C}{E}\right) = \frac{P\left(\frac{E}{C}\right).P(C)}{P(E)} \quad (1)$$

Bayes' theorem is the method of finding the converse probability of the conditional,

$$P\left(\frac{E}{C}\right) = \frac{P\left(\frac{C}{E}\right).P(E)}{P(C)} = \frac{P(C,E)}{P(C)} \quad (2)$$

This conditional relationship allows an investigator to gain probability information about either *C* or *E* with the known outcome of the other. Now consider a complex problem with *n* binary variables, where the relationships among them are not clear for predicting a single class output variable (e.g., node 1 in Figure 5). If all variables

were related using a single joint distribution, the equivalent of all nodes being first level parents, the number of possible combinations of variables would be equal to ($2n$-1). This results in the need for a very large amount of data [18, 19]. If dependence relationships between these variables could be determined resulting in independent variables being removed, fewer nodes would be adjacent to the node of interest. This parent node removal leads to a significant reduction in the number of variable combinations, thereby reducing the amount of needed data. Furthermore, variables that are directly conditional, not to the node of interest but to the parents of the node of interest (as nodes 4 and 5 are with respect to node 1 in Figure 5), can be related, which allows for a more robust system when dealing with missing data points. This property of requiring less information based on pre-existing understanding of the system's variable dependencies is a major benefit of Bayesian Networks [20]. Some further theoretical underpinnings of the Bayesian approach for classification have been addressed in [21] and [22]. A Bayesian Network (BN) is a relatively new tool that identifies probabilistic correlations in order to make predictions or assessments of class membership.

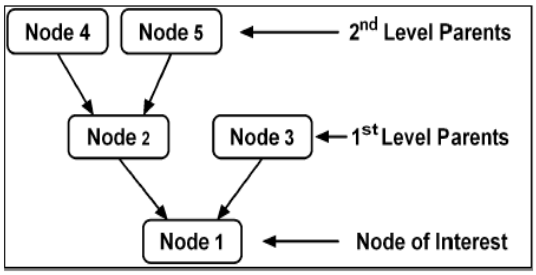

Figure 5. Basic Bayesian Network Structure an Terminology

While the independence assumption may seem as a simplifying one and would therefore lead to less accurate classification, this has not been true in many applications. For instance, several datasets are classified in [23] using the naïve Bayesian classifier, decision tree induction, instance based learning, and rule induction. These methods are compared showing the naïve classifier as the overall best method. To use a Bayesian Network as a classifier, first, one must assume that data correlation is equivalent to statistical dependence.

1) Bayesian Network Type

The kind of Bayesian Network (BN) retrieved by the algorithm is also called Augmented Naïve BN, characterized mainly by the points below.

- All attributes have certain influence on the class.
- The conditional dependency assumption is relaxed (certain attributes have been added a parent).

2) Pre-Processing Techniques

The following data pre-processing techniques applied to the data before running the ODANB [24] algorithm.

Replace Missing Values: This filter will scan all (or selected) nominal and numerical attributes and replace missing values with the modes and mean.

Discrimination: This filter is designed to convert numerical attributes into nominal ones; however the unsupervised version does not take class information into account when grouping instances together. There is always a risk that distinctions between the different instances in relation to the class can be wiped out when using such a filter.

E. Some Implementation Details

JNCC2 loads data from ARFF files this is a plain text format, originally developed for WEKA (Witten and Frank, 2005). A large number of ARFF data sets, including the data sets from the UCI repository are available from http://www.cs.waikato.ac.nz/ml/weka/index_datasets.html.2636. As a pre-processing step, JNCC2 [25] discreteness all the numerical features, using the supervised discretization algorithm of Fayyad and Irani (1993). The discretization intervals are computed on the training set, and then applied unchanged on the test set. NCC2 [25] is implemented exploiting the computationally efficient procedure.

Algorithm 1:

Pseudo code for validation via testing file.

```
ValidateTestFile ()
/*loads training and test file; reads list of non-Mar features;
discretizes features*/
parseArffFile ();
parseArffTestingFile();
parseNonMar();
discretizeNumFeatures();
/*learns and validates NBC*/
nbc = new NaiveBayes(trainingSet);
nbc.classifyInstances(testSet);
/*learns and validates NCC2; the list of non-Mar features in training and testing is required*/
ncc2 = new NaiveCredalClassifier2(trainingSet,
nonMarTraining, nonMarTesting);
ncc2.classifyInstances(testingSet);
/*writes output files*/
writePerfIndicators();
writePredictions();
```

JNCC2 can perform three kinds of experiments: training and testing, cross-validation, and classification of instances of the test set whose class is unknown. The pseudo code of the experiment with training and testing is described by Algorithm 1.

The ODANB has been compared with other existing methods that improves the Naïve Bayes and with the Naïve Bayes itself. The results of the comparison prove that the ODANB outperforms the other methods for the disease prediction not related to lung cancer.

The comparison criteria that have been introduced are

• Accuracy of prediction (measures defined from the confusion matrix outputs). The table-2 below recaps the benchmarked algorithms accuracy for each dataset consider. In each row in bold the best performing algorithm:

TABLE 2 COMPARISONS OF RESULTS

DATASETS	ODANB	NB
LUNG CANCER-C	80.46	**84.14**
LUNG CANCER-H	79.66	**84.05**
LUNG CANCER-STATLOG	80.00	**83.70**

We focus on the results which clearly states that TAN(Tree Augmented Naïve Bayes) [25] works efficiently for the comparison of data sets of general and regular things like vehicles, anneal(metallurgy) over ODANB, Naïve Bayes. But for Diagnosis of Lung Cancer Disease Naïve Bayes observes better results.

VI. CONCLUSION

A prototype lung cancer disease prediction system is developed using data mining classification techniques. The system extracts hidden knowledge from a historical lung cancer disease database. The most effective model to predict patients with Lung cancer disease appears to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network. Decision Trees results are easier to read and interpret. The drill through feature to access detailed patients' profiles is only available in Decision Trees. Naïve Bayes fared better than Decision Trees as it could identify all the significant medical predictors. The relationship between attributes produced by Neural Network is more difficult to understand.

In some cases even in the advanced level Lung cancer patients does not show the symptoms associated with the Lung cancer.

Prevalence of Lung cancer disease is high in India, especially in rural India, did not get noticed at the early stage, because of the lack of awareness. Also it is not possible for the voluntary agencies to carry out the screening for all the people. The emphasis of this work is to find the target group of people who needs further screening for Lung cancer disease, so that the prevalence and mortality rate could be brought down.

Lung cancer prediction system can be further enhanced and expanded. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining [26].

ACKNOWLEDGMENT

The authors would like thank CVR College of Engineering, Hyderabad, for providing its amenities.

REFERENCES

[1] Sang Min Park, Min Kyung Lim, Soon Ae Shin & Young Ho Yun 2006. Impact of prediagnosis smoking, Alcohol, Obesity and Insulin resistance on survival in Male cancer Patients: National Health Insurance corporation study. Journal of clinical Oncology, Vol 24 Number 31 November 2006.

[2] Yongqian Qiang, Youmin Guo, Xue Li, Qiuping Wang, Hao Chen, & Duwu Cuic 2007 .The Diagnostic Rules of Peripheral Lung cancer Preliminary study based on Data Mining Technique. Journal of Nanjing Medical University, 21(3):190-195

[3] Murat Karabhatak, M.Cevdet Ince 2008. Expert system for detection of breast cancer based on association rules and neural network. Journal: Expert systems with Applications.

[4] ICMR Report 2006. Cancer Research in ICMR Achievements in Nineties.

[5] Osmar R. Zaïane, Principles of Knowledge Discovery in Databases. [Online]. Available:webdocs.cs.ualberta.ca/~zaiane/courses/cmput690 /notes/Chapter1/ch1.pdf.

[6] [The Data Mining Process. [Online]. Available: http://publib.boulder.ibm.com/infocenter/db2luw/v9/index.jsp?topic=/com.ibm.im.easy.doc/c_dm_process.html.Shelly Gupta et al./ Indian Journal of Computer Science and Engineering (IJCSE).

[7] Harleen Kaur and Siri Krishan Wasan**,** Empirical Study on Applications of Data Mining Techniques in Healthcare, Journal of Computer Science 2 (2): 194-200, 2006ISSN 1549-3636.

[8] J.R. Quinlan. Induction of decision trees. Machine learning, 1(1):81–106, 1986.

[9] J.R. Quinlan. C4. 5: Programming for machine learning. *Morgan* Kauffmann, 1993.

[10] L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[11] R. D´ıaz-Uriarte and A. de André's. Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(1):3, 2006.

[12] R.S. Michal ski and K. Kaufman. Learning patterns in noisy data: TheAQ approach. Machine Learning and its Applications, Springer-Verlag, pages 22–38, 2001.

[13] R. Linder, T. Richards, and M. Wagner. Microarray data classified by artificial neural networks. METHODS IN MOLECULAR BIOLOGYCLIFTON THEN TOTOWA-, 382:345, 2007.

[14] Murat Karabhatak, M.Cevdet Ince 2008. Expert system for detection of breast cancer based on association rules and neural network. Journal: Expert systems with Applications.

[15] Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.

[16] [by Two Crows Corporation Introduction to Data Mining and Knowledge Discovery .Third Edition,2005. ISBN: 1-892095-02-5, Pages 10, 11.

[17] Maria-Luiza Antonie, Osmar R. Za¨ıane, Alexandru Coman Application of Data Mining Techniques for Medical Image Classification.Page 97

[18] Heckerman, D., A Tutorial on Learning with Bayesian Networks.1995, Microsoft Research.

[19] Neapolitan, R., Learning Bayesian Networks. 2004, London: Pearson Printice Hall.

[20] Neapolitan, R., Learning Bayesian Networks. 2004, London: Pearson Prentice Hall.

[21] Krishnapuram, B., et al., A Bayesian approach to joint feature selection and classifier design. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. **6**(9): p. 1105-1111.

[22] Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © Euro Journals Publishing, Inc. 2009.

[23] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.

[24] Juan Bernabé Moreno, One Dependence Augmented Naive Bayes, University of Granada, Department of Computer Science and Artificial Intelligence.

[25] Juan Bernabé Moreno, One Dependence Augmented Naive Bayes, University of Granada, Department of Computer Science and Artificial Intelligence.

[26] Weiguo, F., Wallace, L., Rich, S., Zhongju, Z.: "Tapping the Power of Text Mining", Communication of the ACM. 49(9), 77-82, 2006.