ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Analysis of Breast Cancer data using Data Mining Techniques

[1]H. Lookman Sithic and [2]Dr. UmaRani

[1]*Research Scholar, Department of Computer Science, Bharathiar Univesity, Coimbatore, Tamilnadu, INDIA.*
[2]*Associate Professor, Dept. of Computer Science, Sri Saradha College for Women, Salem, Tamilnadu, INIDA*

ARTICLE INFO

Article history:
Received 12 March 2015 Accepted 28 April 2015 Available online 5 May 2015

Keywords:
Data Mining, Cancer Disease, Classification, Breast Cancer, Lung Cancer.

ABSTRACT

Cancer disease is major cause of death in the world over the past 10 years. By 2030 an estimated 30.5 million people will die to Cancer diseases if current trends are allow to continue. Data Mining plays an important role in healthcare organization because with the growth of population and dangerous deadly diseases like Cancer, SARS, Leprosy, HIV etc, cancer is one of the most dangerous disease. If the cancer is successfully detected and predicted in its early stages will reduce many treatment options and also reduce risk of invasive surgery and increase survival rate. Therefore cancer detection and prediction system will propose which is easy, cost effective and time saving. This will produce promising result for detection and prediction of cancer. Therefore early detection and prediction of cancer should play a vital role in the diagnosis process and also increase the survival rate of patient. Our research mainly focus the food path cancer is on the increase and oral cancer is decreasing in Erode district. But overall cancer cases are on the increased. Erode is located on the banks of Cauvery River and there are many villages on the banks of Kalingarayan Canal. Farmers and the public complain that owing to abundant use of chemicals and large-scale discharge of effluents into water sources many farmers and cattle are affected.

To Cite This Article: H. Lookman Sithic and Dr. UmaRani., Analysis of Breast Cancer data using Data Mining Techniques. ***Aust. J. Basic & Appl. Sci.,*** *9(7): 724-732, 2015*

INTRODUCTION

Cancer is one of the most common diseases in the world that results in majority of death. Cancer is caused by uncontrolled growth of cells in any of the tissues or parts of the body. Cancer may occur in any part of the body and may spread to several other parts. Only early detection of cancer at the benign stage and prevention from spreading to other parts in malignant stage could save a person's life. There are several factors that could affect a person's predisposition for cancer. Education is an important indicator of socioeconomic status through its association with occupation and life-style factors. A number of studies in developed countries have shown that cancer incidence varies between people with different levels of education. A high incidence of breast cancer has been found among those with high levels of education whereas an inverse association has been found for the incidence of cancers of the stomach, lung and uterine cervix. Such differences in cancer risks associated with education also reflect in the differences in life-style factors and exposure to both environmental and work related carcinogens. This study describes the association between cancer incidence pattern and risk levels of various factors by devising a risk prediction system for different types of cancer which helps in prognosis.

Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods in early detection of cancer. In classification learning, the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples. In association learning, any association among features is sought, not just ones that predict a particular class value. In clustering, groups of examples that belong together are sought. In numeric prediction, the outcome to be predicted is not a discrete class but a numeric quantity. In this study, to classify the data and to mine frequent patterns in data set Decision Tree algorithm is used. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. The top most node is the root node. The attribute value of the data is tested against a decision tree. A path is traced from root to leaf node, which holds the class

Corresponding Author: H. Lookman Sithic, Research Scholar, Bharathiar University, Coimbatore, Tamilnadu, INIDA.
E-mail: lookyas@gmail.com

prediction for that data. Decision trees can be easily converted into classification rules. This decision tree is used to generate frequent patterns in the dataset. The data and item sets that occur frequently in the data base are known as frequent patterns. The frequent patterns that is most significantly related to specific cancer types and are helpful in predicting the cancer and its type is known as Significant frequent pattern. Using this significant patterns generated by decision tree the data set is clustered accordingly and risk scores are given.

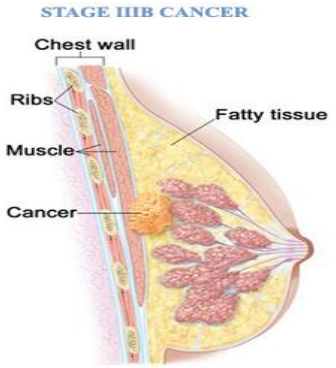

Fig. 1: Stage IIIB Breast Cancer.

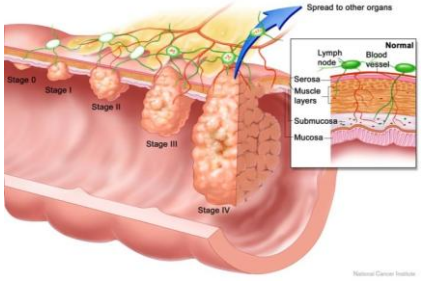

Fig. 2: Stages of Spreading the Cancer.

Clustering is a process of separating dataset into subgroups according to their unique features. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. In K-means clustering, the number of clusters needed is found out and then an algorithm is used to successively associate or dissociate instances with clusters until associations stabilize around k clusters. In our future research all the above mentioned Data Mining techniques are implemented together to create a novel method to diagnose the existence of cancer for a particular patient. When beginning to work on a data mining problem, it is first necessary to bring all the data together into a set of instances. Integrating data from different sources usually presents many challenges. The data must be assembled, integrated, and cleaned up. Only then it can be used for processing through machine learning techniques. This developed system can be used by physicians and patients alike to easily know a person's cancer status and severity without screening them for testing cancer. Also it is useful to record and save large volumes of sensitive information which can be used to gain knowledge about the disease and its treatment.

1.1 Symptoms and signs of Breast Cancer:

Possible Symptoms and signs of breast cancers can consist of:

- Presence of a lump or thickening in the breast;
- Change in the texture of the skin over your breasts
- Change in shape or appearance of the nipple
- Family History of Breast Disease
- Late stop of a menstrual cycle
- Pain in Breast

Several of these symptoms and signs can also be affected by benign problem, less serious, otherwise even by other cancers. Also still, it is very significant to see a dentist or doctor, suppose if any of these conditions lasts more than two weeks in order that the affect can be found and treated, if required. However, the oral cancer is to predict cancer and non-cancer of the each every patient by using various classification techniques.

2. *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 whereas 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.

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 and C4.5 [9]. An evolution of decision tree exploited for microarray data analysis is the random forest], which uses an ensemble of classification trees. 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.

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 a review of advantages and disadvantages of neural networks in the context of microarray analysis is presented.

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. 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. In the following 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.

3. *Related Works:*

Hiram Madero Orozco et.al presented a very simple but efficient methodology for lung nodule classification without the stage of segmentation. Eight texture features were extracted from the histogram and the gray level co-occurrence matrix (with four different angles) after image acquisition for each CT image. Support vector machine (SVM), used to classify lung tissues into two classes: with lung nodules and without lung nodules. The better reliability results were obtained with 90◦ and 135◦ of the GLCM.

Fatma Taher et.al presented a Bayesian classification and a Hopfield Neural Network algorithm for extracting and segmenting the sputum cells for the purpose of lung cancer early diagnosis. The HNN segmentation algorithm outperforms the Fuzzy C-Mean clustering, it allows the extraction of nuclei and cytoplasm regions successfully. Morphological processing on the segmented image improved the performance of HNN algorithm.

Kesav Kancherla et.al proposed an early lung cancer detection methodology using nucleus based features. Seeded region growing segmentation method is used to perform nucleus segmentation. An additional criterion like nucleus size to seeded region growing method is used for better accuracy.

Fan Zhang et.al presented a feature-based imaging classification method to classify the lung nodules in low dose computed tomography (LDCT) slides into four categories: well circumscribed, vascularized, juxta-pleural and pleural-tail. SVM classifier is used to conduct the classification. Specifically, a four-type SVM is trained with polynomial kernel by C-SVC from and the probability estimates upon the different types are predicted with the obtained SVM model, which is used to classify the feature descriptors into four categories.

S.Sivakumar et.al., develop an efficient lung nodule detection scheme by performing nodule segmentation through weighted fuzzy possibilistic based clustering is carried out for lung cancer images. Support Vector Machine (SVM), a machine learning technique is used for classification. The RBF kernel based SVM classifier performs better than linear and polynomial kernel based classifier.

Ritu Chauhan et.al., focuses on clustering algorithm such as HAC and K-Means in which, HAC is applied on K-means to determine the number of clusters. The quality of cluster is improved, if HAC is applied on K-means.

Dechang Chen et.al., algorithm EACCD developed which a two step clustering method. In the first step, a dissimilarity measure is learnt by using PAM, and in the second step, the learnt dissimilarity is used with a hierarchical clustering algorithm to obtain clusters of patients. These clusters of patients form a basis of a prognostic system.

S M Halawani et.al., suggests that probabilistic clustering algorithms performed well than hierarchical clustering algorithms in which almost all data points were clustered into one cluster, may be due to inappropriate choice of distance measure.

Zakaria Suliman zubi et.al., used some data mining techniques such as neural networks for detection and classification of lung cancers in X-ray chest films to classify problems aiming at identifying the characteristics that indicate the group to which each case belongs.

Labeed K Abdulgafoor et.al., wavelet transformation and K- means clustering algorithm have been used for intensity based segmentation.

M. Arfan Jaffar et.al describes a method for lung segmentation based on Genetic Algorithm (GA) and morphological image processing techniques. GA is applied on the normalized histogram determine the threshold to separate out background and object. After background removal morphological operation is performed in three operations: to filter noise, to smooth the image, to detect edges. Susan thinning algorithm is used to reduce the borders to the width of one pixel.

A study of Chou *et al*. the proposed an integrated approach outperforms the results using discriminant analysis, artificial neural networks and multivariate adaptive regression splines and hence provides an efficient alternative in handling breast cancer diagnostic problems.

Delen *et al*., took advantage of available technological advancements to develop prediction models for breast cancer survivability. They brought out some results about predict ability of different data mining algorithms.

In a study of Vald *et al*., despite no explicit preprocessing, exploration of high dimensional data sets is demonstrated. In particular, some of the visual

perspectives presented in that study may be useful for helping to understand breast cancer gene expressions or results from computational data mining procedures.

In a study, Kharbat *et al*. describe the use of a modern learning classifier system to a data mining task. They apply it to a primary breast cancer data set by collaboration with a medical specialist. Their results indicate more effective knowledge discovery.

A study of Liao *et al*. was designed to develop feasible clusters in DNA virus combinations that depend on the observed probability of breast cancer, fibroadenoma, and normal mammary tissue. The viral prerequisites for breast carcinogenesis showed a strongly protective effect on progression from fibroadenoma to breast cancer.

The data mining have shown significant improvement in medical industry in terms of prediction and decision making of lung cancer. Table 1 gives the summary of image processing and classification work, accuracy and sensitivity of various techniques.

Table 1:

Author	Images	Technique		Year	Accuracy	Sensitivity
		Segmentation Algorithm	Classifier			
Disha Sharma (2011)	CT	Edge detection(Sobel)	Diagnostic Indicators	2011	80%	NA
Anam Tariq (2013)	CT	Threshold Segmentation	Neuro-Fuzzy	2013	95%	NA
Atiyeh Hashemi (2013)	CT	Region Growing	FIS-Artificial Neural n/w	2013	NA	95%
Dansheng Song (2012)	CT	Entropy Threshold	SVM(bagging)	2012	85%	NA
S.K Vijai Anand (2010)	CT	Optimal Thresholding	Back propagation network classification	2010	86.30%	NA
Yang Liu (2010)	CT	Bounding box + Threshold Segmentation	SVM(GRBF kernel type)	2009	87.82%	93.75 %
Aparna Kanakatte (2008)	PET	Standard Uptake Values(SUV)	k-NN, SVM	2008	97%	NA
S.Sivakumar (2013)	CT	Weighted fuzzy possibilistic based clustering	SVM(RBF kernel type)	3013	80.36%	82.05 %
JIA Tong (2013)	CT	Optimal gray level threshold	Classification algo. based on medical knowledge	2007	NA	95%
Hiram Madero Orozco (2013)	CT	NA	SVM	2013	84%	NA
Fatma Taher (2014)	Sputum	Hopfield Neural Network(HNN)	Bayesian	2012	88.62	NA
Kesav Kancherla (2013)	Sputum	Seeded region growing	Random forest(bagging)	2013	87%	NA
Negar Memarian (2011)	CT	NA	Fuzzy C-Mean Clustering - iterative LDA	2006	NA	80.80 %

MATERIAL AND METHODS

Extensive literature reviews, case studies and discussions with medical experts show that there are number of factors influencing cancer. These factors are identified and taken as attributes for this study.

4.1 Data Source:

The data for this study was collected from a online survey questionnaire consisting of cancer and non cancer patients data and they are preprocessed to suit this research.

This data consists of more than 30 attributes such as Age, Marital status, Symptoms relating to cancer, occupational hazards, family history of cancer etc. These attributes are used to train and develop the system and a part is used to test the significance of the system. These attributes play an important role in diagnosing cancer in all the cases. This data is stored in a knowledge base which has the ability to expand itself as new data enters the system through front end from which new knowledge is gained and thus the system becomes intelligent.

In the dataset contains number of variables included all the fields depends on the standard medical record type. Here the dataset were prepared totally 23 variables (21 input variables and 2 output variables). There is two numerical variable i.e. Case id and Age and as a Categorical variable, we used Gender (Male, Female), History of Addiction (Alcohol, Smoking, Gutka, None, All), Co-Morbid Condition (Hypertension, Diabetes, Immuno-compromised, None) Symptoms (No, Burning, Ulcer, Mass, Loosening of tooth), Site (BM, LA, RMT, LIP, Tongue, UA, Palate), Gross Examination (Ulcero-proliferative, Infiltrative,Verrucous, Plaque Like, Polypoidal), Predisposing Factor (Leukoplakia, Submucous Fibrosis), Tumor Size (<2cm, 2 cm to 4 cm, >4 cm), Histopathology (Variant of SSCVerrucous, Papillary, Basaloid, Plaque Like, Sarcomatoid, acantholytic, Lymphoepithelioma like), Neck Node (Present, Absent), LFT (Normal,

Deranged), USG (Yes, No), FNAC of Neck Node (Yes, No), Diagnostic Biopsy (Squamous Cell Carcinoma, Variant of SCC, Benign), CT Scan / MRI (Bony Involvement, Normal) Diagnosis (SCC, Verrucous, Benign, Plaque Like, Sarcomatoid, Acantholytic, Adenoca, Lymphoepithelioma Like), Staging (I, II, III, IV), Surgery (Y,N), Radiotherapy (Y, N), Chemotherapy (Y, N).

4.2 Significant Pattern mined using Decision tree algorithm:

1. Age - gender - living area - family history- anemia- symptoms -> none- Cancer Type -> None. Weightage = 104.25
2. Age - gender- marital status-education-smoking-diet-symptoms-> Pain in chest, back, shoulder or arm->Shortness of breath and hoarseness-Cancer Type->Lung Weightage = 199.50
3. Gender-Education-Occupational hazards- Alcohol-Family history- Weight loss- symptoms-> severe abdominal pain or bloating-> abdominal pain with blood in stool- Cancer Type ->Stomach Weightage = 174.24
4. Age- gender- no of children- occupational hazards- Family history- relationship with cancer patient- symptoms-> swelling or mass in armpit -> discharge or pain in nipple -> Cancer Type -> Breast. Weightage = 172.00
5. Gender- education- living area- Smoking- Hot beverage- Diet- fast food addiction- Earlier cancer diagnosis- symptoms-> Ulcers in mouth or pain of teeth and jaw-> White or red patches in tongue, gums- Cancer Type -> Oral. Weightage = 188.50

Numerical values are given as risk scores to the attributes that have a direct link to the significant patterns mined.

4.3 Weightage for Significant Pattern:

The weightage is calculated for every frequent pattern based on the attributes to analyze its impact on the output. The frequent patterns mined which satisfies the below condition are taken as significant Frequent Pattern.

$Sw(i)= \Sigma(Wi*Fi)$ ----------(1)

Where Wi is the weightage of each attribute and Fi represents number of frequency for each rule. And significant Frequent Pattern is selected by using the following Equation (2)

$SFP=Sw(n) \geq \varphi$ ----------(2)

for all values of n (2). Where SFP denotes significant frequent pattern and φ denotes significant weightage.

Table 2: Risk scores for the attributes that represent the significant patterns.

Attributes	Values	Risk score
Age		
	x<30	3
	30 < x < 40	4
	40< x < 60	5
Education	Uneducated	5
	School	3
	College	2
Living Area	Urban	5
	Rural	3
Habits	Smoking	3
		5
	Alcohol	
	Chewing	3
		2
	Hot beverage	
Occupational Hazards	Radiation Exposure	3
	Chemical Exposure	3
	Sunlight Exposure	2
	Thermal Exposure	2
Anemia	Yes	3
	No	1
Weight Loss	Yes	2
	No	1
Family History of Cancer	Yes	5
	No	1

4.4 Rules for Decision Tree:

If symptoms = none and risk score = x <45 then result = you don't have cancer, tests = do simple clinical tests to confirm.

If symptoms = none and risk score = 45 < x < 60 then result = you may have cancer, tests = do blood test and x ray to confirm

Else if symptom= related to stomach and risk score = x > 45 then result = you have cancer, cancer type = stomach, tests = endoscopy of stomach

If symptom= related to breast and shoulder and risk score =x > 45 then result = you have cancer, cancer type = breast, tests= mammogram and PET scan of breast

If symptom= related to chest and shoulder and risk score =x > 40 then result = you have cancer, cancer type = lung, tests = take CT scan of chest.

If symptom= related to pelvis and lower hip and risk score= x >55 then result = you have cancer, cancer type = cervix, tests = do pap smear test

If symptom= related to head and throat and risk score =x > 40 then result = you have cancer, cancer type = oral, tests = biopsy of tongue and inner mouth.

Else symptom= other symptoms and risk score =x > 40 then result = you have cancer, cancer type = leukemia, tests = biopsy of bone marrow

Based on the above mentioned rules and the calculated risk scores the severity of cancer is known as well as some tests were prescribed to confirm the presence of cancer.

RESULT AND DISCUSSION

The data set consist of 320 patients' record. Among them, 81 or 25.5% are reported to have breast cancers while the remaining 239 or 74.5% are not. In order to validate the prediction results of the comparison of the six popular data mining techniques and the 10-fold crossover validation is used.

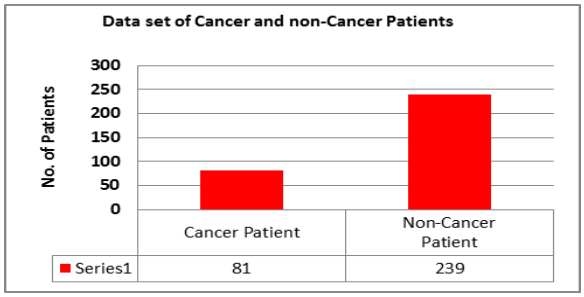

Fig. 3: Data set of Cancer and Non Cancer patients.

The k-fold crossover validation is usually used to reduce the error resulted from random sampling in the comparison of the accuracies of a number of prediction models. The entire set of data is randomly divided into k folds with the same number of cases in each fold. The training and testing are performed for k times and one fold is selected for further testing while the rest are selected for further training. The present study divided the data into 10 folds where 1 fold was for testing and 9 folds were for training for the 10-fold crossover validation. These diagnostic results of each patient's record in above dataset consist of ten variables that are summarized. One of the 10 variables is the response variable representing the diagnostic status of the patient with or without breast cancers (i.e. malignant or benign). The training data are selected from the whole dataset randomly and directly fed into the proposed mining approach.

We have used the Weka toolkit to experiment with these three data mining algorithms The Weka is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. WEKA version 3.6.9 was utilized as a data mining tool to evaluate the performance and effectiveness of the 6-breast cancer prediction models built from several techniques. This is because the WEKA program offers a well defined framework for experimenters and developers to build and evaluate their models. The performance of a chosen classifier is validated based on error rate and computation time. The classification accuracy is predicted in terms of Sensitivity and Specificity. The computation time is noted for each classifier is taken in to account. The evaluation parameters are the specificity, sensitivity, and overall accuracy. The sensitivity or the true positive rate (TPR) is defined by TP / (TP + FN); while the specificity or the true negative rate (TNR) is defined by TN / (TN + FP); and the accuracy is defined by (TP + TN) / (TP + FP + TN + FN)

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

These values are often displayed in a confusion matrix as be presented in Table 2. Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model.

Accuracy: Accuracy is the percentage of tuples that are correctly classified by the classifier.

Accuracy = (TP+TN)/ (TP +TN+FP+FN)

An efficient analysis of classification technique is utilized and it has investigated three data mining techniques: The multilayer perceptron neural network, Random Tree, and the C4.5 decision tree

algorithms. Finally, they concluded that C4.5 algorithm has a much better performance than other two techniques.

6. *Conclusion:*

In our study all the above data mining techniques has some drawbacks. So I have study and implement a better data mining technique to compare with other techniques.

7. *Future Work:*

Future work is to study on large database of workers of Erode district affected by cancer using other data mining techniques such as Logistic Regression, Clustering and Neural Network in order to determine similarities and relationship between multiple factors.

REFERENCES

Shaik Parveen, S., C. Kavitha, 2013. "Detection of lung cancer nodules using automatic region growing method", 4th ICCCNT.

Ada, Rajneet Kaur, 2013. "Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images", IJARCSSE, 3(3): 187-190.

Guruprasad Bhat, Vidyadevi, G. Biradar, H. Sarojadevi Nalini, 2012. " Artificial Neural Network based Cancer Cell Classification (ANN – C3)", Computer Engineering and Intelligent Systems, 3(2).

Disha Sharma, Gagandeep Jindal, 2011. "Identifying Lung Cancer Using Image Processing Techniques", International Conference on Computational Techniques and Artificial Intelligence (ICCTAI), 115-120.

Amin Mohammad Roozgard, Samuel Cheng and Hong Liu, 2012. "Malignant Nodule Detection on Lung CT Scan Images with Kernel RX –algorithm", Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2012) Hong Kong and Shenzhen, China, pp: 499-502.

National Cancer Institute, 2011. "The cancer imaging archive, "https:llwiki.cancerimagingarchive.net/display/Public/LlDC-lDRI.

Anam Tariq, M., Usman Akram and M. Younus Javed, 2013. "Lung Nodule Detection in CT Images using Neuro Fuzzy Classifier", Fourth International Workshop on Computational Intelligence in Medical Imaging (CIMI), pp: 49-53.

Anita chaudhary, Sonit Sukhraj Singh, 2012. "Lung cancer detection on CT images by using image processing", International Conference on Computing Sciences, pp: 143-146.

Atiyeh Hashemi, Abdol Hamid Pilevar, Reza Rafeh, 2013. "Mass Detection in Lung CT Images Using Region Growing Segmentation and Decision Making Based on Fuzzy Inference System and Artificial Neural Network", I.J. Image, Graphics and Signal Processing, 6: 16-24.

Dansheng Song, Tatyana, A. Zhukov, Olga Markov, Wei Qian3, Melvyn, S. Tockman, 2012. "Prognosis of stage i lung cancer patients through quantitative analysis of centrosomal features", ieee, 1607-1610.

Vijai Anand, S.K.,2010. "Segmentation coupled Textural Feature Classification for Lung Tumor Prediction", ICCCCT, 518-524.

Yang Liu, Jinzhu Yang, Dazhe Zhao, Jiren Liu, 2010. "A Method of Pulmonary Nodule Detection utilizing multiple Support Vector Machines", International Conference on Computer Application and System Modeling (ICCASM 2010), 118-121.

Aparna Kanakatte, Nallasamy Mani, Bala Srinivasan, Jayavardhana Gubbi, 2008. "Pulmonary Tumor Volume Detection from Positron Emission Tomography Images", International Conference on Biomedical Engineering and Informatics, 213-217.

Lee, S.L.A., A.Z. Kouzani and E.J. Hu, "A Random Forest for Lung Nodule Identification".

Yang Liu, Jinzhu Yang, Dazhe Zhao, Jiren Liu, 2009. "Computer Aided Detection of Lung Nodules Based on Voxel Analysis utilizing Support Vector Machines", International Conference on Future Biomedical Information Engineering, 90-93.

Fan Zhang, Yang Song, Weidong Cai, Yun Zhou, Shimin Shan and Dagan Feng, 2013. "Context Curves for Classification of Lung Nodule Images", ieee.

Chang, C.C. and C.J. Lin, 2011. "Libsvm: a library for support vector machines," ACM Trans. TIST, 2(3): 27.

Sivakumar, S., Dr.C.Chandrasekar, 2013. "Lung Nodule Detection Using Fuzzy Clustering and Support Vector Machines", International Journal of Engineering and Technology (IJET), 5(1): 179-185.

Arfan Jaffar, M., Ayyaz Hussain, M. Nazir, Anwar M. Mirza and Asmatullah Chaudhry, 2008. "GA and Morphology based automated Segmentation of Lungs from CTscan Images", CIMCA, IAWTIC, and ISE, 265-270.

Smith, S.M. and J.M. Brady, 1997. SUSAN, "a new approach to low level image processing", Int. Journal of Computer Vision, 23(1): 45-78.

Sivagowry, S., M. Dr. Durairaj and A. Persia, 2013. " an Empirical study on applying data mining techniques for the analysis and prediction of heart disease ", IEEE international and embedded system conferences on 21-22: 265-270.

Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal, Ashish Kumar Sen, 2013. "Data mining clustering Techniques in the prediction of heart disease using attribute selection method", International journal of science, engineering and technology research(IJSETR), 2-10.

Raj Mohan, K., Ilango Paramasivam, subhashinisathya narayan, 2014. " Prediction and Diagnosis of Cardio Vascular Disease- A Critical

Survey", WorldCongress on Computing and Communication Technologies, IEEE Page(s): 246–251.

Shamsher Bahadur Patel, Pramod Kumar Yadav, Dr D. P. Sukla., 2013. " Predict the Diagnosis of Heart Disease Patients using Classification mining techniques", IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), 4(2): 61-64.

Benish Fida, Muhammad Nazir, Nawazish Naveed,Sheeraz Akram, 2011. "Heart Disease Classification Ensemble Optimization using Genetic algorithm",Multitopic conference (INMIC) IEEE.

Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, 2013. " Genetic Neural Network Based Data Mining in prediction of Heart Disease using Risk Factors"IEEE Conference on Information and Communication Technologies (ICT 2013), Page(s): 1227-1231.

Miss. Chaitrali, S., Dangare, Dr. Mrs. Sulabha S. Apte, 2012. "A data Mining Approach for prediction of Heart Disease using Neural Network " International journal of computer engineering and technology, 3(3): 30-40.

Mai Shouman, Tim Turner, Rob Stocker, 2012. " Integrating naïve Bayes and K-Means clustering with different initial centroid selection methods in the diagnosis pf heart disease patients ",academia.edu.

John Peter, T., K. Somasundaram, 2012. "An Empirical study on prediction of heart disease using classification and data mining", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012), 30, 31.

Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data",

Computer research and development (ICCRD), 2011. 3rd International conference on volume 2,IEEE pages 480-485.

Mai Shouman, Tim Turner, Rob Stocker, 2012. " Integrating Decision tree and K-Means clustering with different initial centroid selection methods in the diagnosis of heart disease patients ", academia.edu.

Akhil jabbar, M., Priti Chandra, B.L Deekshatulu, 2012. "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature subset selection", Intelligent system design and application (ISDA) IEEE 12th international conference, pages, 628-634.

Akhil jabbar, M, Priti Chandra, B.L Deekshatulu, 2013. "Heart disease Prediction using Lazy Associative Classification" IEEE international multi conference, 40-46.