Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multi-class Categorization of Breast Tissue Data

Mrs.Shomona Gracia Jacob
Ph.D Research Scholar
Department of Computer Science and Engineering
Rajalakshmi Engineering. College
(Affiliated to Anna University, Chennai)
Thandalam, Chennai
Tamilnadu, India.

Dr. R.Geetha Ramani
Professor & Head
Department of Computer Science and Engineering
Rajalakshmi Engineering College
(Affiliated to Anna University, Chennai)
Thandalam, Chennai
Tamilnadu, India.

ABSTRACT
This paper highlights the significance of classification in data mining and knowledge discovery. In this paper we investigate the performance of various data mining classification algorithms viz. Rnd Tree, Quinlan decision tree algorithm (C4.5), K-Nearest Neighbor algorithm etc., on a large dataset from the 'Wisconsin Breast tissue dataset' (derived from the UCI Machine Learning Repository) that comprises of 11 attributes and 106 instances. The results of this study indicate the level of accuracy and other performance measures of the algorithms in detecting the presence of breast cancer and the associated breast tissue conditions that increase the risk of developing cancer in future. Moreover the importance of feature selection/reduction in improving the performance of classification algorithms is also described. The classification algorithm Rnd Tree produced 100 percent accuracy for classification of all the training data under multiple classes. The classification algorithm was also applied to verify it's correctness in classifying test data.

General Terms
Data Mining, Breast tissue research

Keywords
Knowledge Patterns, Pattern Recognition, Clinical Data, Healthcare, Breast Cancer, Breast Tissue, Classification

1. INTRODUCTION
Data mining [1] is the process of analyzing data from various perspectives and summarizing it into useful, meaningful and related information. Technically, Data Mining [2] is the process of finding correlations, associations or patterns among a large number of fields in huge relational databases. Knowledge discovery [2] in database and data mining play an important role in exploring data and revealing important data patterns. These patterns can then be seen as a kind of outline of the input data, and can be used in further investigation or can be applied in the field of machine learning and predictive analytics. For instance, the data mining step could be used to detect multiple groups in the data. Identification of these groups can then aid in obtaining more precise and accurate prediction results by a decision support system. There are many data mining algorithms and tools that have been developed for feature selection, clustering, rule framing and classification. These algorithms are used to discern and uncover knowledge patterns and make out significant and meaningful information associated with the application domain. Classification and prediction [2] [3] [4] are the data analysis techniques that are used to identify important data classes and predict probable trends.

Clinical data mining [4] is the application of data mining techniques with clinical data. In machine learning and pattern recognition, classification refers to an algorithmic method or process for assigning a given portion of input data into one of a suggested number of categories. The term "classifier", refers to the mathematical operation, implemented by a classification algorithm, which maps the keyed-in data to a class or category.

Cancer [5] [6] [7] is a vast and diverse class of diseases in which a set of cells exhibit unrestrained growth, invasion that intrudes upon and damages adjacent tissues, and often metastasizes. Breast cancer is one of the most prominent types of cancer among women. Detecting early stage breast cancer with high sensitivity and specificity has proven to be a demanding task for the existing state of clinical research. These challenges, combined with the massive toll that this disease takes, have been the driving factor for continuing research and study to develop techniques with better and accurate implementation results for early stage breast cancer detection and treatment. Breast cancer is cancer stemming from breast tissue. Breast tissue is an intricate assembly of tissues closely tied to nerves, blood vessels and fatty tissues, also called adipose tissue.

Realization and consciousness about the causes, indications and mental trauma associated with breast cancer has scientifically augmented during recent years [8]. By 2015, there is expected to be nearly 2.5 Lac new cases in India. According to the World Cancer Report in the year 2000, malignant tumors were accountable for roughly 12 percent of the nearly 56 million deaths worldwide from all causes. Breast cancer is the second leading cause of death in women, next only to lung cancer [8]. Approximately 2.4 million women residing in the U.S. are currently under treatment for breast cancer.

The escalating threat of breast cancer in women all over the world is attributed to the changing lifestyle and work pattern of women. The societal impact and personal distress caused by the occurrence of this cancer, that is known to have a strong propagating nature from the victim to the future generations has been the rationale for this research.

In this research work we compare the error rates and related performance measures produced by the various classification algorithms on the Wisconsin breast tissue dataset (latest dataset, updated in 2010) [9] and the effect of feature selection algorithms on improving the accuracy of classification of carcinoma in the breast tissue dataset. The existence of other tissue features like Fibro- adenoma, Mastopathy that indicate a higher risk of developing cancer in future is also classified.

1.1 Organization of the paper

The rest of the paper is organized as follows: Section 2 reviews the related work in this area. Section 3 presents the proposed system design and details of the breast tissue data set while Section 4 briefly describes the feature reduction algorithms used in our study. The classification algorithms used in the experimental analysis are briefed about in Section 5. Experimental results and performance evaluation are given in Section 6 whereas the conclusion is narrated in Section 7.

2. RELATED WORK

The literature survey of the work related to our study is presented. A few researchers have worked on the breast cancer dataset, the details of which are given below. Ressom et.al [3] gives an overview of statistical and machine learning-based feature selection and pattern classification algorithms and their application in molecular cancer classification or phenotype prediction. Their work does not involve experimental results. C.Y.V Watanabe et.al [4], have devised a method called SACMiner aimed at breast cancer detection using statistical association rules. The method employs statistical association rules to build a classification model. Their work classifies medical images and is not applicable to textual medical data. Siegfried Nijssen et al., [10] have presented their work on multi-class co-related pattern mining. Their work resulted in the design of a new approach for item set mining on data from the UCI repository. Their comparison included only the new approach designed and the extension of the Apriori algorithm. Their results reveal comparison mainly on the runtime of the mining approaches. T. Cover and P. Hart [11] performed classification task using K- Nearest Neighbor classification method. Their work shows that K-NN can be very accurate in classification tasks under certain specific circumstances. Their results reveal that for any number of categories, the probability of error of the Nearest Neighbor rule is bounded above by twice the Bayes probability of error. Aruna et.al [6] presented a comparison of classification algorithms on the Wisconsin Breast Cancer and Breast tissue dataset but has not provided feature selection as a pre-classification condition. Moreover they have analyzed the classification results of only five classification algorithms namely Naïve Bayes, Support Vector Machines (SVM), Radial Basis Neural Networks (RB-NN), Decision trees J48 and simple CART. Luxmi et. al., [12] have performed a comparative study on the performance of binary classifiers. They have used the Wisconsin breast cancer dataset with 10 attributes and not the breast tissue dataset. Moreover they have not brought out the effect of feature selection in classification. Their experimental study was restricted to four classification algorithms viz. ID3, C4.5, K-NN and SVM. Their results did not reveal complete accuracy for any of the classification algorithms.

3. PROPOSED SYSTEM DESIGN

Classification [2] [3] [13] is a data mining function that designates items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. In the model build (training) process, a classification algorithm finds relationships or associations between the values of the predictors and the values of the target.

3.1 System Description

The diagrammatic representation of the proposed system design in presented in Figure 1.

3.1.1 Training Data

The data set used in our experimental study is taken from the UCI Machine Learning Repository Wisconsin Breast Tissue dataset which is the most recently updated dataset (updated in 2010) [6][9]. The details of the attributes in the dataset are given in Table 1.

Table 1. Description of the attributes in the dataset

	Attributes	Description
1	I0	Impedivity (ohm) at zero frequency
2	PA500	Phase angle at 500 KHz
3	HFS	High-frequency slope of phase angle
4	DA	Impedance distance between spectral ends
5	AREA	Area under spectrum
6	A/DA	Area normalized by DA
7	MAX IP	Maximum of the spectrum
8	DR	Distance between I0 and real part of the maximum frequency point
9.	P	Length of the spectral curve

Carcinoma [5] is the medical term that indicates cancer. Fibro-adenoma is the most common benign tumor of the breast. Any disease, pain or disorder of the mammary glands is referred to as Mastopathy [5]. It is epitomized by the occurrence of pain and seals in the breast. Its presence raises the risk of developing cancer.

Breasts [5] are made up of fat, connective tissue, glandular tissue and ducts. The glandular tissue is prearranged in lobes that are connected to the nipple by ducts to secrete and deliver milk during lactation. The percentage of glandular, connective and adipose tissue in the mammograms may have some indication of abnormality or malfunction in the patient's breast tissue characteristics that could be indicative of developing cancer in future.

This is a dataset with electrical impedance measurements in samples of freshly excised tissue from the breast. It consists of 106 instances that are described by 10 attributes that includes 9 features and 1 class attribute [6]. Figure 1 portrays the proposed classification model.

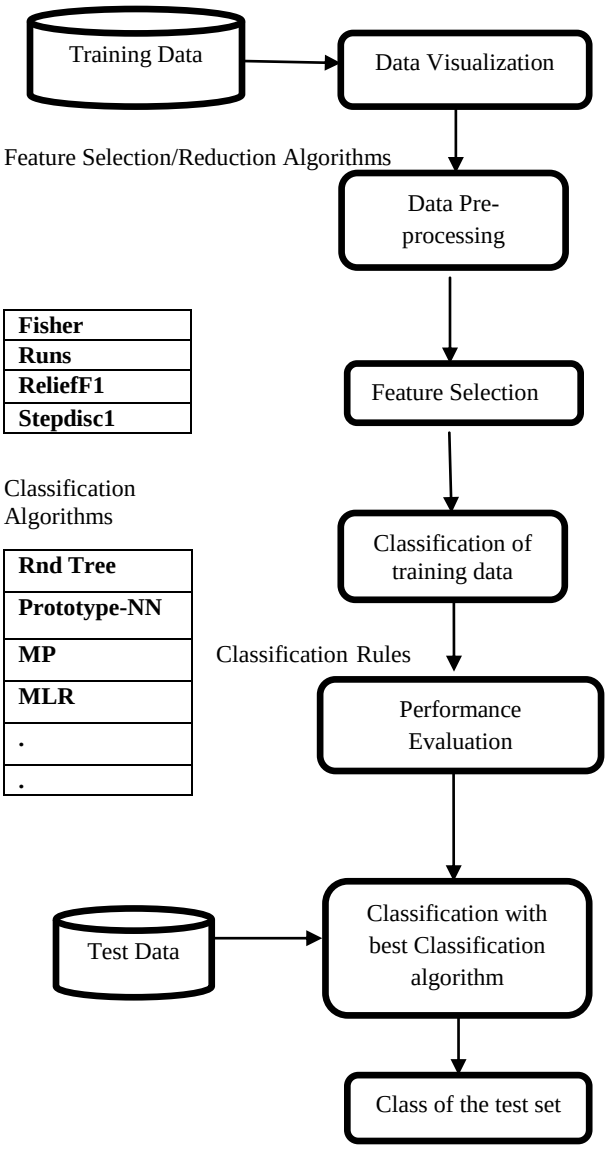

Figure 1: Proposed System Design

Six classes of freshly expurgated tissue were studied using electrical impedance measurements. Though this dataset has been previously used in a classification study, the accuracy of five classification algorithms alone has been explored. This paper compares the performance of fifteen classification algorithms with important attributes ranked prior to the classification phase by a well performing feature selection algorithm, the results of which are discussed in Section 4.

3.1.2 Data Visualization

The Wisconsin Breast Tissue dataset is stored as a Microsoft Excel 97 spreadsheet (.xls) [9]. This is loaded into the Data Mining tool after which a verification of the dataset is performed. This is done by viewing the dataset within the TANAGRA tool to ensure the correctness of the loaded training data. A sample of the training data is shown in Table 2. The dataset comprises of 11 attributes with 106 examples. The data is originally present in the breast tissue mammogram images. Textual information is derived from the mammograms and is stored as a public repository for study and research purpose.

Table 2: Dataset description

Attribute	*Category*	*Information*
Case #	Continue	-
I0	Continue	-
PA500	Continue	-
HFS	Continue	-
DA	Continue	-
Area	Continue	-
A/DA	Continue	-
Max IP	Continue	-
DR	Continue	-
P	Continue	-
Class	Discrete	6 values

3.1.3 Data Pre-processing

The data was downloaded in the form of an Excel spreadsheet. The downloaded files were analyzed [14]. According to the download instructions, the missing values have been replaced with zero. The case attribute is not considered for classification as it is only a representation of the patient identity.

Once the dataset is obtained in the form to be loaded into the data mining tool, the feature selection/reduction algorithms have to be applied on the dataset to weigh the attributes as explained in the following section.

4. FEATURE SELECTION

Feature selection [15] has been a lively and effective research area in pattern recognition, statistics, and data mining communities. The main idea behind feature selection is to identify and choose a subset of the input variables by analyzing and eliminating features with little or no predictive information. Feature selection can appreciably improve the comprehensibility and lucidity of the resulting classifier and often construct a model that generalizes better to unseen points. Further, it is often the case that finding the precise and exact subset of predictive attributes is an important problem in its own sense [16] [17]. The feature selection algorithms used for this comparison are briefly explained in the following subsections.

4.1 Fisher Filtering

Univariate Fisher's ANOVA ranking [7]. It is a supervised feature selection algorithm based upon a filtering approach. It processes the selection algorithm independently from the learning algorithm. This component ranks the input attributes according to their importance and relevance. A cutting rule facilitates the selection of subset of these attributes. This algorithm does not take into account the redundancy of the input attributes.

4.2 ReliefF Filtering

ReliefF is an extension of the popular Relief algorithm [7] [8]. A key idea of the ReliefF algorithm is to evaluate, estimate and assess the quality of features according to how well their values distinguish between sample points that are near to each other.

4.3 Stepwise Discriminant Analysis

STEPDISC (Stepwise Discriminant Analysis) [14] [18] is always associated with discriminant analysis because it functions on the same criterion i.e. the WILKS' partial lambda. So it is often presented as a method especially intended for the discriminant analysis. In the FORWARD approach, at each step, we determine the variable that really contributes to the discrimination between the groups. We add this variable if its contribution is significant. The process stops when there is no attribute to add in the model. In the BACKWARD approach, we begin with the complete model with all descriptors. We search to identify the less relevant variable. We remove this variable if its removal does not notably damage the discrimination between groups. The process stops when there is no variable to remove.

4.4 Runs Filtering

Univariate attribute ranking [14] from Runs test. It is a supervised feature selection algorithm based upon a filtering approach. It processes the selection algorithm independently from the learning algorithm. This component identifies the significance of the input attributes according to their relevance and gives their ranking.

Once the attributes are weighted by the feature selection method, we will classify the training set by applying the fifteen classification algorithms. We have performed all the fifteen classification algorithms on the training set with unfiltered attributes, attributes weighted by ReliefF algorithm and attributes weighted by Stepwise Discriminant Analysis algorithm.

5. CLASSIFICATION

Classification [13] [19] is the process of finding a set of models that describe and distinguish data classes. This is done to achieve the goal of being able to use the model to predict the class whose label is unknown. Some of the algorithms used in our experimental study are briefed in the following sections.

5.1 Iterative Dichomotiser (ID3)

The aim of the ID3 algorithm [12] is to generate a decision tree that predicts correctly the value of the category attribute, based on the answers to questions about the non-category attributes.ID3 algorithm uses a fixed group of examples to build a decision tree and then uses this tree to classify given data samples.

5.2 C4.5

C4.5 constructs decision trees [13] from a set of training data in the same way as ID3, using the concept of Information Entropy. The training data is a set S = s1, s2, of already classified samples. Each sample s_i = x1, x2, is a vector where x1, x2, represent attributes or features of the sample. The training data is augmented with a vector C = c1, c2, where c1, c2, represent the class of each sample.

5.3 Linear Discriminant Analysis

Linear discriminant analysis (LDA)[13] and the related Fisher's linear discriminant are methods used in statistics, pattern recognition and machine learning to identify a linear combination of features which characterize or separate two or more classes of objects or events. The resulting combination may be used as a linear classifier, or for dimensionality reduction before later classification.

5.4 Multinomial Logistic Regression

In statistics, economics, and genetics, a multinomial logistic model, also known as multinomial logistic regression[13], is a regression model which generalizes logistic regression by permitting more than two discrete outcomes. This suggests that, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.).

5.5 Naïve Bayes Continuous

Naive Bayes classifier [6] is a probabilistic classifier based on the Bayes theorem, considering strong (Naive) independence assumption. Thus, a Naive Bayes classifier believes that all attributes (features) independently contribute to the probability of a certain decision. Considering the characteristics of the underlying probability model, the Naive Bayes classifier can be trained very efficiently in a supervised learning setting. This could yield much better results in many complex real-world situations, especially in the field of computer-aided diagnosis [5] [6]. Here it is assumed that all variables are independent. Hence only the variances of the variables for each class need to be determined and not the entire covariance matrix.

5.6 Rnd Tree

The Rnd tree [13] algorithm can be applied to both classification and regression problems. Random trees are a collection or assembly of tree predictors that is called **forest** [13]. The classification works as follows: the random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". In the case of regression the classifier response is the average of the responses over all the trees in the forest.

5.7 Partial Least Squares -Discriminant Analysis

Partial least squares regression (PLS regression)[13] is a statistical method that bears some relation to principal components regression; Here it develops a linear regression model by launching the predicted variables and the observable variables to a new space. Because both the *X* and *Y* data are projected to new spaces, the PLS family of methods are called as bilinear factor models. Partial least squares Discriminant Analysis (PLS-DA) is a variant used when the Y is binary.

5.8 Multilayer Perceptron

The perceptron [13] is a multiclass classification algorithm. Here, the input *x* and the output *y* are drawn from random sets. A feature representation function *f*(*x*, *y*) maps each possible input/output pair to a finite-dimensional real-valued feature vector. The feature vector is multiplied by a weight vector *w*, but now the resulting score is used to choose among many possible outputs:

$$\hat{y} = \text{argmax}_y f(x, y).w.$$

Repeated learning iterates over the examples, calculating and forecasting an output for each. The output involves two scenarios. The first case lets the weights remain unaltered when the predicted output matches the target. The second case changes the weights when the predicted output does not tone up with the target. The update becomes:

$$w_{t+1} = w_t + f(x,y) - f(x,\hat{y}).$$

Here, w_{t+1} is the new weight which is formed by adding the existing weight to the difference between the predicted output and the target.This multiclass formulation reduces to the original perceptron when x is a real-valued vector, y is chosen from {0, 1}, and $f(x, y) = yx$.

5.9 K-Nearest Neighbor

The K-Nearest Neighbour [12] algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbours. The object is then designated to the class most common amongst its k nearest neighbours. K is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbor.

5.10 Prototype- Nearest Neighbor

Prototype selection [14] is primarily effective in improving the classification performance of Nearest Neighbor (NN) classifier. It carries with it the advantage of partially minimizing NN classifier storage and computational requirements.

6. PERFORMANCE EVALUATION

The training data of all the 106 cases with all the attributes were loaded as an Excel spreadsheet in the data mining tool. All the attributes were continuous and the target class was discrete. It is to be noted that Rnd Tree algorithm have produced 100% accuracy and complete precision in classifying the breast tissue dataset. Four feature selection algorithms viz. Fisher Filtering, Runs Filtering, ReliefF and Stepdisc filtering were executed on the dataset. ReliefF was found to produce the best results for three algorithms but Stepdisc filtering produced better results for five algorithms as given in Table 3.Fisher's filtering did not filter any attributes and Runs Filtering selected 8 attributes out of 9 but did not produce any significant change in results

The ReliefF algorithm has improved the accuracy of Naïve Bayes Continuous classification, Multinomial Logistic Regression and PLS-LDA whereas the Stepdisc filtering algorithm has produced better accuracy for C4.5,Naïve Bayes Continuous, PLS-DA, PLS-LDA and Prototype –NN.

6.1 Performance Measures

The measures and their exact meaning have been given as stated by Han and Kamber [1].

6.1.1 Confusion Matrix

Given m classes, a confusion matrix [1] is a table of at least size m*m. An entry, $CM_{i,j}$ in the first m rows and m columns indicates the number of tuples of class I that were named by the classifier as j. It is a valuable tool for analyzing how well your classifier can recognize tuples of different classes.

6.1.2 Precision and Recall

Precision [2] and Recall [2] are two basic measures for assessing the performance of text retrieval. In our study Precision refers to the data that is correctly classified by the classification algorithm. 1.000 precision indicates 100% accuracy. Recall is the percentage of information relevant to the class and is correctly classified.

6.1.3 Accuracy

The accuracy [1] of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

6.1.4 Error- Rate

The error rate [1, 20] is also called the misclassification rate. It is simply 1-Acc (M), where Acc (M) is the accuracy of M.

The experimental results of performing ReliefF filtering and Stepwise Discriminant analysis is shown in the following subsection.

6.2 Experimental Results

The sample results obtained by applying feature reduction/selection algorithms on the Wisconsin Breast Tissue dataset are given in Figure 2, Figure 3, Figure 4 and Figure 5.

6.2.1 Feature Selection Results

The ReliefF and the Stepwise Discriminant Analysis algorithm have shown more accurate filtering and classification results. The results obtained by applying these algorithms on the dataset are tabulated in Table 3.

Table 3: Comparison of Feature Selection Algorithms

S.No	Feature Selection/ Reduction Algorithms	No. of Attributes	
		Before filtering	After filtering
1	Fisher Filtering	9	9
2	Runs Filtering	9	8
3	ReliefF1	9	3
4	Stepwise Discriminant Analysis	9	5

The results of the ReliefF and Stepwise Discriminant Analysis feature selection/reduction algorithms are displayed in Figure 2 and Figure 4 respectively. The ReliefF also displays the weight assigned to the attributes and selects only those attributes that are relevant to the current classification.

Parameter	Value
Neighbors	10
Use sampling	0
Selection mode	2
INPUT selection	
Before filtering	**9**
After filtering	**3**

Fig 2: ReliefF-Parameter and Attribute Selection

The weights assigned to the attributes by ReliefF are given in Figure 3.

N°	Attribute	Weight
1	I0	0.238329
2	P	0.228159
3	PA500	0.104178
4	DA	0.089638
5	Max IP	0.084564
6	DR	0.075999
7	A/DA	0.050489
8	HFS	0.025455
9	Area	0.018711

Figure 3: ReliefF Feature Selection Algorithm Ranked Attributes (Weight)

The attributes selected by the Stepdisc feature selection algorithm are given in Figure 4.

N°	Selected attributes
1	I0
2	PA500
3	P
4	DA
5	Area

Figure 4: Stepdisc Filtering Results

The comparative analysis of the feature selection algorithms are given in the form of a graph in Figure 5.

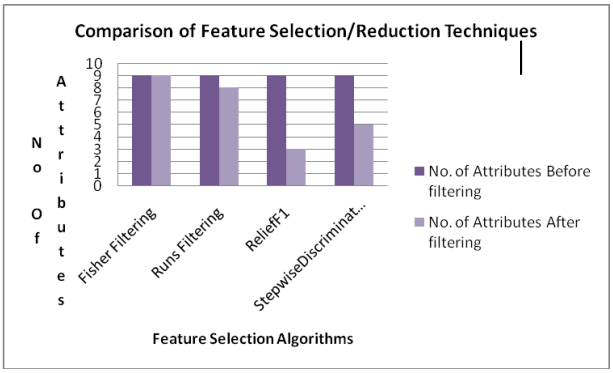

Figure 5: Comparison of Feature Selection Algorithms

After feature reduction, the classification algorithms are applied on the training data with the weighted attributes. The results of classification are described in the following section.

6.2.2 Classification Results

The comparative study of fifteen classification algorithms on all the attributes, attributes weighted by ReliefF and Stepwise Discriminant feature selection algorithms are tabulated and given in Table 4. The classification results of Rnd Tree, Prototype NN, Naïve Bayes Continuous and C4.5 classification algorithms are given in Figure 6, Figure 7, Figure 8 and Figure 9 respectively.

Table 4. Comparison of Error Rate Produced by Classification Algorithms

Classification Algorithms	Error rate after Classification		
	Without Feature selection	Feature Selection Algorithms	
		ReliefF1	Step disc1
C4.5	0.1887	0.2075	0.1792
C-RT	0.3491	0.3774	0.3491
CS-CRT	0.3491	0.3774	0.3491
CS-MC4	0.2075	0.2736	0.2075
C-SVC	0.3396	0.3868	0.3774
ID3	0.7925	0.7925	0.7925
KNN	0.1792	0.2547	0.217
LDA	0.2547	0.3396	0.3302
MP	0.3113	0.4057	0.3208
MLR	1	0.2264	1
Naïve Bayes	0.3396	0.3302	0.2925
PLS-DA	0.3774	0.3774	0.3679
PLS-LDA	0.3679	0.3396	0.3396
P-NN	0.2453	0.2736	0.2358
RND TREE	**0.000**	**0.000**	**0.000**

The number of attributes selected for split in Rnd tree classification = 5.

Error rate	0.0000

Values prediction

Value	Recall	1-Precision
car	1.0000	0.0000
fad	1.0000	0.0000
mas	1.0000	0.0000
gla	1.0000	0.0000
con	1.0000	0.0000
adi	1.0000	0.0000

Confusion matrix

	car	fad	mas	gla	con	adi	Sum
car	21	0	0	0	0	0	21
fad	0	15	0	0	0	0	15
mas	0	0	18	0	0	0	18
gla	0	0	0	16	0	0	16
con	0	0	0	0	14	0	14
adi	0	0	0	0	0	22	22
Sum	21	15	18	16	14	22	106

Figure 6: Rnd Tree Classification Algorithm Results

Error rate	0.2358							
Values prediction			**Confusion matrix**					
Valu	Recall	1-		**ca**	**fa**	**ma**	**gl**	**co**	**a**	**Sum**
car	0.857	0.1000	**car**	18	0	3	0	0	0	21
fad	0.466	0.2222	**fad**	0	7	5	3	0	0	15
mas	0.444	0.5556	**ma**	1	2	8	6	1	0	18
gla	0.812	0.4091	**gla**	1	0	2	13	0	0	16
con	0.928	0.0714	**con**	0	0	0	0	13	1	14
adi	1.000	0.0435	**adi**	0	0	0	0	0	22	22
			Su	20	9	18	22	14	23	106

Figure 7: Prototype NN Classification Algorithm results

Error rate	0.2925							
Values prediction			**Confusion matrix**					
Value	Recall	1-Precision		**car**	**fad**	**mas**	**gla**	**con**	**adi**	**Sum**
car	0.8571	0.1429	**car**	18	0	3	0	0	0	21
fad	0.5333	0.5556	**fad**	1	8	5	1	0	0	15
mas	0.6111	0.5417	**mas**	2	3	11	2	0	0	18
gla	0.4375	0.3000	**gla**	0	5	4	7	0	0	16
con	0.7857	0.1538	**con**	0	2	1	0	11	0	14
adi	0.9091	0.0000	**adi**	0	0	0	0	2	20	22
			Sum	21	18	24	10	13	20	106

Figure 8: Naïve Bayes Continuous Classification Results

Error rate	0.1792							
Values prediction			**Confusion matrix**					
Valu e	Recall	1-Precisio n		**ca r**	**fa d**	**ma s**	**gl a**	**co n**	**ad i**	**Sum**
car	0.952	0.1304	**car**	20	0	1	0	0	0	21
fad	0.666	0.3333	**fad**	0	10	2	3	0	0	15
mas	0.388	0.3000	**mas**	2	5	7	4	0	0	18
gla	0.937	0.3182	**gla**	1	0	0	15	0	0	16
con	1.000	0.0667	**con**	0	0	0	0	14	0	14
adi	0.954	0.0000	**adi**	0	0	0	0	1	21	22
			Su m	23	15	10	22	15	21	106

Figure 9: C4.5 Classification Algorithm Results

The comparison of the classifcation algorithms's performance is given in Figure 10. Rnd Tree classification algorithm gives 100 percent accuracy for classifying the Wisconsin Breast Tissue dataset.Multilayer Logistic Regrssion shows 80 percent improved result when attributes are weighted by ReliefF1 and then classsfied. Naïve Bayes Continuous and PLS-LDA also show decreased error rates when classified after ReleifF1 algorithm. However Stewise Discriminant filtering has given more accurate classification result by C4.5, Naïve Bayes Continuous, PLS-LDA, Prototype-NN and PLS-DA. The performance of the classification algorithms have been rated based on the error rate and displayed in the form of a graph in Figure 10.

The classification algorithm that performed best was the Rnd Tree and this when apllied to the test data set without the class label resulted in accurate classifcation.

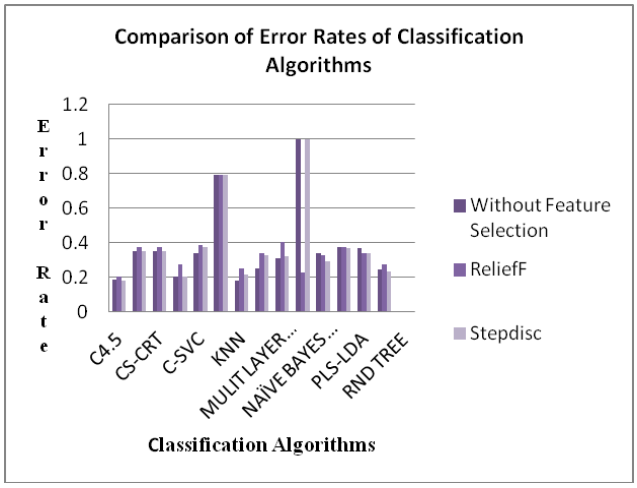

Figure 10: Comparison of Error Rates of Classification Algorithms on the Wisconsin Breast Tissue Dataset

7. CONCLUSION

The classification phase of data mining is one of the most significant phases that have a wide range of applications in different domains. Hence it becomes essential to rate the performance of these algorithms and compare their performance on the test data sets. Applications of data mining in the medical field are challenging and there exist umpteen avenues for exploration and expansion. A comparative study of classification algorithms will definitely be a boon in improving the state of decision making in the field of medicine. This study is the case of a multi-class classification. The Rnd Tree classification algorithm has given 100% accuracy in classifying the data sets. This will improve the current state of breast cancer detection from the breast tissue characteristics. Moreover the features from the breast tissue data can be weighed and analyzed to predict the risk of developing cancer in future. The classification algorithms namely C4.5, Prototype-NN, PLS-LDA, Naïve Bayes continuous and PLS-DA show improved accuracy after the attributes are weighted and selected by Stepdisc filtering algorithm. Multinomial Logistic Regression, Naïve Bayes continuous and PLS-LDA classification algorithms have shown reduced error rates after the attributes are filtered through ReliefF algorithm. The Rnd Tree classification

algorithm classified the sample test data accurately, thus achieving the objective of our study.

8. ACKNOWLEDGEMENT

This research work is a part of the All India Council for Technical Education(AICTE), India funded Research Promotion Scheme project titled "Efficient Classifier for clinical life data (Parkinson, Breast Cancer and P53 mutants) through feature relevance analysis and classification" with Reference No:8023/RID/RPS-56/2010-11, No:200-62/FIN/04/05/1624.

9. REFERENCES

[1] J. Han and M. Kamber, ―Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers, 2000.

[2] K. Cios, W. Pedrycz, and R. Swiniarski. Data Mining Methods for Knowledge Discovery. Boston: Kluwer Academic Publishers, 1998

[3] W. Ressom, Rency S. Varghese, Zhen Zhang, Jianhua Xuan, and Robert Clarke. 2008 Classification Algorithms for phenotype prediction in genomic and Proteomics Front BioScience.

[4] C. Y. V. Watanabe, M. X. Ribeiro, C. Traina, and A. J. M. Traina. 1997 SACMiner: A New Classification Method Based on Statistical Association Rules to Mine Medical Images," in Enterprise Information Systems, vol. 73.

[5] Breast Cancer Statistics from Centers for Disease Control and Prevention, http://www.cdc.gov/cancer/breast/statistics/

[6] S. Aruna, Dr S.P. Rajagopalan and L.V. Nandakishore, 2011 Knowledge Based Analysis Of Various Statistical Tools In Detecting Breast Cancer

[7] MedlinePlus:Breast Diseases

[8] Wennberg J, Cooper MM, editors. The Dartmouth atlas of medical care in the United States: a report on the Medicare program. Chicago, IL:AHA Press; 1999

[9] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[10] Siegfried Nijssen and Joost N.Kok Multi-Class Correlated Pattern Mining.

[11] T. M. Cover, Member, IEEE, and P. E. Hart, Member, IEEE, "Nearest Neighbour Pattern Classification", IEEE Transactions on Information Theory, 1967.

[12] Luxmi Verma, Dr.Varun Kumar, "Binary Classifiers for Health Care Databases: A ComparativeStudy of Data Mining Classification Algorithms in the Diagnosis of Breast Cancer", IJCST, Vol 1, Issue 2, 2011.

[13] M. James. Classification Algorithms. John Wiley, 1985.

[14] Tanagra Data Mining tutorials, http://data-mining-tutorials.blogspot.com/ This website provides detailed information on the basics of Data Mining Algorithms

[15] K. Kira, L. Rendel, The feature selection problem: Traditional methods and a new algorithm, in: M. Press (Ed.), Proceedings of Tenth National Conference on Artificial Intelligence, 1992, pp. 129–134.

[16] I. Kononenko, Estimating attributes: Analysis and extensions of relief, in: Machine Learning:ECML-94, Vol. 784 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 1994,pp. 171–182..

[17] Yong Seog Kim, W. Nick Street, and Filippo Menczer, University of Iowa, USA, "Feature Selection in Data Mining".

[18] Jean S. Whitaker, 1997. Use of Stepwise Methodology in Discriminant Analysis.

[19] D.Lavanya, Dr.K. Usha Rani, Performance Evaluation of Decision Tree Classifiers on Medical Data Sets", International Journal of Computer Application, 2011

[20] C. Laredo, F. Austerlitz, O. David, B. Schaeffer, K. Bleakley,N. Vergne1, M. Veuille, "Error rates of phylogenetic and supervised classification algorithms in DNA Barcoding" Barcode Conference, Mexico, 7-12 Nov. 2009

10. AUTHORS PROFILE

Dr.R. Geetha Ramani is Professor & Head in Department of Computer Science and Engineering, Rajalakshmi Engineering College, India. She has more than 15 years of teaching and research experience. Her areas of specialization include Data mining, Evolutionary Algorithms and Network Security. She has over 50 publications in International Conferences and Journals to her credit. She has also published a couple of books in the field of Data Mining and Evolutionary Algorithms. She has completed an External Agency Project in the field of Robotic Soccer and is currently working on projects in the field of Data Mining. She has served as a Member in the Board of Studies of Pondicherry Central University. She is presently a member in the Editorial Board of various reputed International Journals.

Mrs.Shomona Gracia Jacob completed her M.E. in Computer Science and Engineering at Jerusalem College of Engineering, affiliated to Anna University, Chennai, India. She has more than 3 years of teaching experience. Presently she is pursuing her Ph.D in Computer Science and Engineering at Rajalakshmi Engineering College, affiliated to Anna University of Technology, Chennai. Her areas of interest include Data Mining, Artificial Intelligence and Software Engineering. She has attended and presented few papers at National and International Conferences.