Parkinson Disease Classification using Data Mining Algorithms

Dr. R. Geetha Ramani
Professor & Head
Department of Computer Science and Engineering
Rajalakshmi Engineering College
Chennai, INDIA.

G. Sivagami
Master of Engineering
Department of Computer Science and Engineering
Rajalakshmi Engineering College
Chennai, INDIA.

ABSTRACT
Knowledge discovery in databases has established its success rate in various prominent fields such as e-business, marketing, retail and medical. Medical data mining has great potency for exploring the out of sight patterns in the respective medical data sets. This paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use today for the classification of Parkinson Disease. Parkinson Disease is a chronic malady of the central nervous system where the key indications can be captivated from the Mentation, Activities of Daily Life (ADL), Motor Examination and Complications of Therapy. The speech symptom which is an ADL is a common ground for the progress of the disease. The dataset for the disease is acquired from UCI, an online repository of large data sets. A comparative study on different classification methods is carried out to this dataset by applying the feature relevance analysis and the Accuracy Analysis to come up with the best classification rule. Also the intention is to sieve the data such that the healthy and people with Parkinson will be correctly classified.

General Terms
Data Mining, Healthcare Data, Parkinson Disease.

Keywords
Knowledge Data Discovery (KDD), Data Mining, Error Rate, Classification, mis-Classification Rate, Feature Relevance, Clinical Data, Parkinson Disease.

1. INTRODUCTION
Parkinson's disease (PD) [6] is chronic and progressive movement disorder, meaning that symptoms continue and worsen over time. Nearly one million people in the US are living with Parkinson's disease. The cause is unknown, and although there is presently no cure, there are treatment options such as medication and surgery to manage its symptoms. Parkinson's involves the malfunction and death of vital nerve cells in the brain, called neurons. Parkinson's primarily affects neurons in the area of the brain called the substantia nigra [6]. Some of these dying neurons produce dopamine, a chemical that sends messages to the part of the brain that controls movement and coordination. As Parkinson Disease progresses, the amount of dopamine produced in the brain decreases, leaving a person unable to control movement normally.

The four main symptoms of Parkinson Disease are tremor, rigidity, bradykinesia and postural instability. Tremor is an involuntary vibration to hands, arms, legs or jaws. The inflexibility to the limbs and trunk is referred as rigidity whereas the slowness in the movement is known as bradykinesia. Other symptoms may include depression and other emotional changes; difficulty in swallowing, chewing and speaking; urinary problems or constipation; skin problems and sleep disruptions[6].

The diagnosis is based on the medical history and neurological examination conducted by interviewing and observing the patient in person using the Unified Parkinson's Disease Rating Scale (UPDRS) [3]. Prior to UPDRS development, multiple scales, including the Webster, Columbia, King's College, Northwestern University Disability, New York University Parkinson's Disease Scale, and UCLA Rating Scales, were used in different centers. This made the comparative assessments difficult. The development of the UPDRS involved multiple trial versions, and the final published scale is officially known as UPDRS version 3.0 [8].

With rapid changes taking place in the field of health care, decision support systems play an increasingly important role. Huge data volume repositories, health care institutions deploy data warehouse and data mining solutions to extract relevant information.. Data Mining is defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use today in medical research and public health. Due to symptom overlap with other diseases, only 75% of clinical diagnosis of Parkinson Disease are confirmed to be idiopathic Parkinson Disease at autopsy [3]. The expert systems and different Artificial Intelligent (AI) techniques for classification have the potential of being good supportive for the expert. Classification systems can help in increasing accuracy and reliability of diagnosis and minimize the possible errors, as well as making the diagnosis more time efficient [1].

This paper aims at bringing the best accuracy classifier algorithm using the Tanagra Data mining tool, an open source project used for Data mining academic and research purpose.

1.1 Organization of the Paper
The paper is organized as follows. Section 2 defines the related works carried out in the Parkinson Disease area. Section 3 deals

with the Parkinson Dataset that is used in this research work. Section 4 handles the proposed classification model to classify the patient as PD or non-PD. Section 5 is dealt with the experimental results of the Classifier Algorithms. And Section 6 concludes research paper and proposes the future work.

2. RELATED WORKS

David Gil and Magnus Johnson [1] evaluated the performance of a classifier constructed by means of ANN and SVM. By the three methods i.e. by Multilayer Perceptrons (MLP) and SVM with two kernel types, they obtain a high precision level of the confusion matrix regarding the different measurement parameters like accuracy, sensitivity, specificity, positive predictive value and negative predictive value. They concluded by showing a high degree of certainty of above 90%. Some parameters reach very high accuracy such as "Sensitivity" and "Negative predictive value" with 99.32% and 97.06%.

Another paper published by Max A. Little, et.al [2] on suitability of Dysphonic measurements suggests that non standard measures in combination with traditional harmonics to noise ratios are best able to separate healthy people from PWP. The four features Harmonics to Noise Ratio, Recurrence Period Density Entropy, Detrended Fluctuation Analysis and Pitch Period Entropy are fed into the kernel support vector machine which generates an overall classification performance of 91.4%.

Marius Ene [3] applied the Probabilistic Neural Networks (PNN) types for the classification purpose. Incremental Search, Monte Carlo Search and Hybrid search are the three PNN types that have been used for classification purpose, related to smoothening factor. As an outcome, there is no major difference between the three techniques of searching the smoothening factor although the Hybrid search plays better of 81%.

The classification work on Parkinson Disease is carried out by several authors. The work done by Resul Das [4] emphasize on comparing the four classification methods. Various classifiers have been applied to recognize the Parkinson Disease by using SAS based software. Regression, DMNeural, Neural Network and Decision Tree are the four independent classification models used. As a result, Neural Network yields the best classification rate of 92.9%.

Mehmet Fatih CAGLAR, et.al [11] proposed two types of Artificial Neural Networks, Multilayer Perceptrons (MLP) and Radial Basis Function (RBF) Networks. And also Adaptive Neuro-Fuzzy Classifier (ANFC) with linguistic hedges is used. ANFC with linguistic hedges gave the best recognition results with 95.38% on training and 94.72% on test dataset.

3. PARKINSON DATASET

Voice Measurement has shown a great progress in the advancement of Parkinson Disease. About 90% of PWP (People with Parkinson's disease) present some kind of vocal deterioration [4]. And hence, this dataset is chosen, which mainly focus on the speech signals. The Parkinson Disease dataset used for classification purpose was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado. This organization recorded the speech signals. The dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table addresses a particular voice measure, and each row corresponds to one of 195 voice recording from these individuals. The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD. It is a two-decision classification problem. The characteristic features are shown in Table 1.

Table 1. Characteristic Features of Parkinson Dataset

Feature Number	Feature Name	Description
1	MDVP: Fo(Hz)	Average vocal fundamental Frequency
2	MDVP: Fhi(Hz)	Maximum vocal fundamental frequency
3	MDVP: Flo(Hz)	Minimum vocal fundamental frequency
4	MDVP: Jitter(%)	Kay Pentax MDVP jitter as percentage
5	MDVP:Jitter (Abs)	Kay Pentax MDVP absolute jitter in microseconds
6	MDVP: RAP	Key Pentax MDVP Relative Amplitude Perturbation
7	MDVP: PPQ	Kay Pentax MDVP five-point Period Perturbation Quotient
8	Jitter: DDP	Average absolute difference of differences between cycles, divided by the average period
9	MDVP: Shimmer	Key Pentax MDVP local shimmer
10	MDVP: Shimmer (dB)	Key Pentax MDVP local shimmer in decibels
11	Shimmer :APQ3	3 Point Amplitude Perturbation Quotient
12	Shimmer :APQ5	5 Point Amplitude Perturbation Quotient
13	MDVP: APQ	Kay Pentax MDVP eleven-point Amplitude Perturbation Quotient
14	Shimmer :DDA	Average absolute difference between consecutive differences between the amplitude of consecutive periods
15	NHR	Noise to Harmonic Ratio
16	HNR	Harmonics to Noise Ratio
17	RPDE	Recurrence Period Density Entropy
18	DFA	Detrended Fluctuation Analysis
19	Spread1	Non Linear measure of fundamental frequency
20	Spread2	Non Linear measure of fundamental frequency
21	D2	Correlation Dimension
22	PPE	Pitch Period Entropy
23	Status	Health Status 1- Parkinson ; 0- Healthy

Note: MDVP stands for (Kay Pentax) Multi Dimensional Voice Program.

4. PROPOSED WORK

In this research work, we intend to find the patterns formed by different classification algorithms. Various patterns help in picking the best one out of all. To the training dataset, the feature relevance is applied. Various Feature selection algorithm is worked in which Fisher filtering is found to be a good feature

ranking system. This filtering is applied to the algorithms for better classification purpose. Based on the Filtering algorithm, minimum number of attributes with which the better classification is selected and performed. The proposed architecture model flow is depicted in Fig 1.

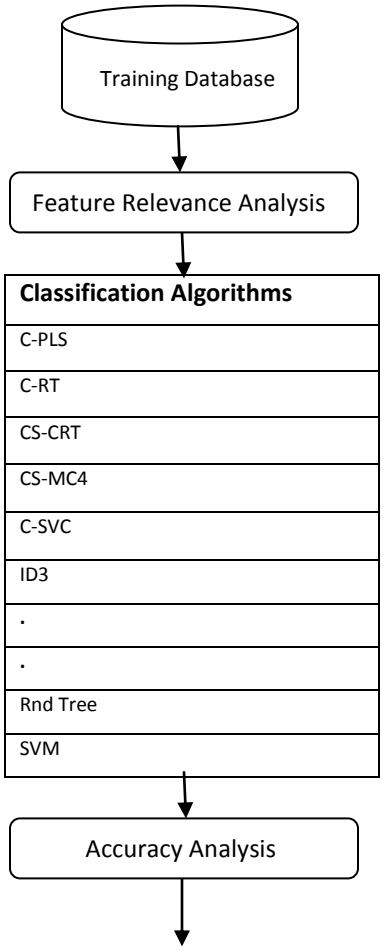

Fig 1: Proposed Architecture Model Flow

4.1 Training Dataset

The training dataset comprises of 197 instances with 22 characteristic features. These features constitute of a range of biomedical voice measurements. The features are discussed in section 3. The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado.

4.2 Feature Relevance Analysis

Before applying the data mining classification algorithms, it is necessary to pick up the right features. Adding irrelevant features will end up in negative effects on the prediction task. In general, data are better characterized usually with fewer variables. And hence feature relevance has to be done. The selection techniques can be done in three ways. Embedded approaches where feature selection is a part of the classification algorithm i.e. decision tree. Second is the filter approach where the features are chosen before performing the classification algorithms upon it. And the finale is the wrapper approach where the classification algorithm is used as a black box to find the best subset of attributes.

Diverse filtering algorithms are experimented. The Fisher filtering algorithm ranks the input attributes according to the relevance. A cutting rule enables to select a subset of these attributes. This approach does not take into consideration the redundancy of the input attributes. The Feature Relevance is performed on some of the classification algorithms. The main objective is to obtain the minimum error rate with the minimum characteristic features of the Parkinson Dataset. This is depicted in the Table 2. The Feature number in Table 2 is referred to the first column of the Table 1.

Table 2. Feature Relevance for Classification Algorithms

Algorithm	Total Number of Filtered Features	Feature Number
Binary Logistic Regression	3	19, 22, 20.
C4.5	15	19, 22, 20, 1, 3, 9, 13, 16, 12, 10, 11, 14, 21, 5, 17
C-PLS	3	19, 22, 20
C-RT	5	19, 22, 20, 1, 3
CS-CRT	5	19, 22, 20, 1, 3
CS-MC4	15	19, 22, 20, 1, 3, 9, 13, 16, 12, 10, 11, 14, 21, 5, 17
C-SVC	22	All Features Included
ID3	1	19
K-NN	20	19, 22, 20, 1, 3, 9, 13, 16, 12, 10, 11, 14, 21, 5, 17, 7, 4, 6, 8, 18.
LDA	21	19, 22, 20, 1, 3, 9, 13, 16, 12, 10, 11, 14, 21, 5, 17, 7, 4, 6, 8, 18, 16.
Log- Reg TRIRLS	14	19, 22, 20, 1, 3, 9, 13, 16, 12, 10, 11, 14, 21, 5
Rnd Tree	3	19, 22, 20
SVM	22	All Features included

.

4.3 Classification Process

An overview of the Algorithms used for the classification purpose of the Parkinson Dataset is discussed here.

4.3.1 Binary Logistic Regression

Logistic Regression is a predictive model where the class label or the target is categorical. This variable owns two categories, yes / no. For this Parkinson Disease dataset, the category will be whether the person is affected with Parkinson Disease or the person is a non-Parkinson Disease. Logistic regression can be used only with two types of target variables. 1. A categorical target variable that has exactly two categories (i.e., a binary or

dichotomous variable). 2. A continuous target variable that has values in the range 0.0 to 1.0 representing probability values or proportions.

4.3.2 ID3

ID3 is a mathematical algorithm used to construct the decision tree. It constructs the tree in a top-down fashion with no backtracking. A gist computation of ID3 algorithm is. 1. The entropy (formulae to calculate the homogeneity of the given sample) of all the unused attributes is compiled. 2. The minimum entropy attribute is chosen. 3. This attribute is built as a node.

4.3.3 C4.5

Decision trees are formed from the classification rules as a part of execution of the C4.5 program. The C4.5 [13] is an extension of the basic ID3 algorithm designed by Quinlan. It addresses some of the troubles that are not dealt with ID3. They are i. Choosing an appropriate attribute selection measure, ii. Handling continuous attributes, iii. Handling training data with missing attribute values, iv. Handling attributes with differing costs, v. Improving computational efficiency, vi. Reduced error pruning, vii. Avoiding over fitting of the data, vii. Rule post pruning.

4.3.4 Classification and regression Tree (C-RT)

Classification and Regression Tree [7] is a recursive partitioning method, builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). Regression-type problems are generally those where we attempt to predict the values of a continuous variable from one or more continuous and/or categorical predictor variables. Classification-type problems are generally those where we attempt to predict values of a categorical dependent variable (class, group membership, etc.) from one or more continuous and/or categorical predictor variables. The purpose of the analyses via tree-building algorithms is to determine a set of if-then logical (split) conditions that permit accurate prediction or classification of cases.

K- Nearest Neighbor (K-NN)

The k-nearest neighbor algorithm [14] classifies objects on closest training examples in the feature space. It is also known as instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor. The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. The k-nearest neighbor algorithm is sensitive to the local structure of the data.

4.3.5 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a classic method of classification. The algorithmic complexity is more in this method to produce the accuracy. The target variable can be of two or more categories. LDA [10] finds a linear transformation of two predictors that yields a new set of transformed values which provides a accurate discrimination than if predication done alone.

4.3.6 Random Tree (Rnd Tree)

Random forest (or random forests) [10] is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark. The term came from random decision a forest that was first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho and Amit and Geman in order to construct a collection of decision trees with controlled variation. The advantage of Random tree are listed below

- Decision tree forest models are as easy to create as single-tree models. By simply setting a control button, you can direct DTREG to create a single-tree model or a decision tree forest model or a TreeBoost model for the same analysis.
- Decision tree forest models often have a degree of accuracy that cannot be obtained using a large, single-tree model. Decision tree forest models are among the most accurate models yet invented.
- Decision tree forests use the out of bag data rows for validation of the model. This provides an independent test without requiring a separate data set or holding back rows from the tree construction.
- Decision tree forest models can handle hundreds or thousands of potential predictor variables.
- The sophisticated and accurate method of surrogate splitters is used for handling missing predictor values.
- The stochastic (randomization) element in the decision tree forest algorithm makes it highly resistant to over fitting.
- Decision tree forests can be applied to regression and classification models.

4.3.7 Partial Least Square Regression (PLS)

It is an extension of multiple linear regression model. In simpler manner, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the variables are projected to new spaces, the PLS family of methods are known as bilinear factor models. PLS regression is probably the least restrictive of the various multivariate extensions of the multiple linear regression model. This flexibility allows it to be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than predictor variables

4.3.8 Support Vector Machine (SVM)

A Support Vector Machine (SVM) performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. SVM models are closely related to neural networks. In fact, a SVM model using a sigmoid kernel function is equivalent to a two-layer, perceptron neural network [10].

Support Vector Machine (SVM) models are a close cousin to classical multilayer perceptron neural networks. Using a kernel function, SVM's are an alternative training method for polynomial, radial basis function and multi-layer perceptron classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training [10].

4.4 Accuracy Analysis

The supervised learning algorithms are applied one after the other. The confusion matrix gives the 2*2 matrix form of output. Confusion Matrix is a useful tool on how well the classifier recognizes the tuples of different classes. This shows the value of true positive, true negative, false positive and false negative. Based on the error rate the classifier accuracy is calculated and a comparative study is done to retrieve the best Classifier algorithm.

5. EXPERIMENTAL RESULTS

The experiment was performed on the above mentioned algorithms for both training and test dataset with the furnished feature relevance. The Random Tree Algorithm classifies the Parkinson Disease dataset accurately and provides the 100%. The Linear Discriminant Analysis, C4.5, CS-MC4 and K-NN yields the accuracy results above 90%. K-NN error rate is only 0.0256. Among all, the C-PLS algorithm classifies the dataset with least percentage of 69.74. The C-RT & CS-CRT produce the same error rate of 0.0462. The important observation to be made is the features that are selected. The feature relevance analysis shows that the three important features (spread1, PPE and spread2) are mainly aimed for better classification purpose. The Random Tree builds the classification rule based on this three characteristic features to obtain the zero error rate. Fig 2 gives an comparison of the experimented Classification Algorithms. Table 3 depicts the screenshot view of the confusion matrix of different classification algorithms and their corresponding Error Rate obtained.

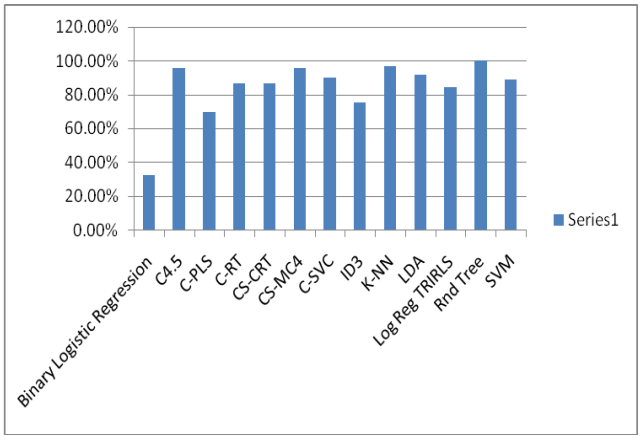

Fig 2: Comparison of Classification Algorithmic Accuracy

Table 3. Comparison of Error Rate of various Classification Algorithms

Algorithm	Confusion Matrix	Error Rate
Binary Logistic Regression	Confusion matrix Parkinson / Healthy / Sum Parkinson: 137 / 10 / 147 Healthy: 17 / 31 / 48 Sum: 154 / 41 / 195	0.1385
C4.5	Confusion matrix Parkinson / Healthy / Sum Parkinson: 143 / 4 / 147 Healthy: 4 / 44 / 48 Sum: 147 / 48 / 195	0.0410
C-PLS	Confusion matrix Parkinson / Healthy / Sum Parkinson: 91 / 56 / 147 Healthy: 1 / 47 / 48 Sum: 92 / 103 / 195	0.2923
C-RT	Confusion matrix Parkinson / Healthy / Sum Parkinson: 142 / 5 / 147 Healthy: 4 / 44 / 48 Sum: 146 / 49 / 195	0.0462
CS-CRT	Confusion matrix Parkinson / Healthy / Sum Parkinson: 142 / 5 / 147 Healthy: 4 / 44 / 48 Sum: 146 / 49 / 195	0.0462
CS-MC4	Confusion matrix Parkinson / Healthy / Sum Parkinson: 143 / 4 / 147 Healthy: 4 / 44 / 48 Sum: 147 / 48 / 195	0.0410
C-SVC	Confusion matrix Parkinson / Healthy / Sum Parkinson: 147 / 0 / 147 Healthy: 20 / 28 / 48 Sum: 167 / 28 / 195	0.1026
ID3	Confusion matrix Parkinson / Healthy / Sum Parkinson: 147 / 0 / 147 Healthy: 48 / 0 / 48 Sum: 195 / 0 / 195	0.2462
K-NN	Confusion matrix Parkinson / Healthy / Sum Parkinson: 144 / 3 / 147 Healthy: 2 / 46 / 48 Sum: 146 / 49 / 195	0.0256
LDA	Confusion matrix Parkinson / Healthy / Sum Parkinson: 144 / 3 / 147 Healthy: 13 / 35 / 48 Sum: 157 / 38 / 195	0.0821
Log- Reg TRIRLS	Confusion matrix Parkinson / Healthy / Sum Parkinson: 147 / 0 / 147 Healthy: 29 / 19 / 48 Sum: 176 / 19 / 195	0.1487
Random Tree	Confusion matrix Parkinson / Healthy / Sum Parkinson: 147 / 0 / 147 Healthy: 0 / 48 / 48 Sum: 147 / 48 / 195	0
SVM	Confusion matrix Parkinson / Healthy / Sum Parkinson: 146 / 1 / 147 Healthy: 21 / 27 / 48 Sum: 167 / 28 / 195	0.1128

6. CONCLUSION AND FUTURE WORK

The paper is intended to verify the effectiveness of the application of various classifiers to the Parkinson Dataset. This dataset comprises of 22 attributes with various range of values. A comparative study of several algorithms on the dataset is performed. This is done by first doing the feature relevance on the dataset. Then the classifiers are implemented upon the dataset.

Early detection of any kind of disease is an essential factor. This helps in treating the patient well ahead. In this research paper, Random Tree classifier yields the 100% accuracy.

Future work can be extended to classify the Parkinson Telemonitoring dataset. Accurate Classification eases the drug discovery process.

7. ACKNOWLEDGEMENTS

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.

Our sincere acknowledgement to Max Little of the University of Oxford, who has created the database, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals.

8. REFERENCES

[1] Gil, D., Manuel, D. 2009 Diagnosing Parkinson by using Artificial Neural Networks and Support Vector Machines.

[2] Little, M.A.; McSharry, P.E.; Hunter, E.J.; Spielman, J.; Ramig, L.O. 2009. Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease.

[3] Marius Ene. 2008. Neural Network based approach to discriminate healthy people from those with Parkinson's disease.

[4] Resul Das. 2010. A comparison of multiple classification methods for diagnosis of Parkinson disease.

[5] Varun Kumar; Luxmi Verma. 2010. Binary Classifiers for Health care Databases: A comparative study of Data Mining Classification Algorithms in the Diagnosis of Breast Cancer.

[6] Parkinson's Disease Foundation, http://www.pdf.org/en/about_pd, (2011.09.16).

[7] StatSoft Electronic Statistics Textbook http://www.statsoft.com/textbook/classification-and-regression-trees (2011.09.16).

[8] National Institute of Neurological Disorders and Stroke, http://www.ninds.nih.gov/disorders/parkinsons_disease/detail_parkinsons_disease.htm, (2011.09.16).

[9] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The Unified Parkinson's Disease Rating Scale (UPDRS): Status and Recommendations.

[10] DTREG. http://www.dtreg.com Software for Predictive Modeling and Forecasting.

[11] Mehmet Fatih CAGLAR, Bayram CETISLI, Inayet Burcu TOPRAK. 2010. Automatic Recognition of Parkinson 's disease from Sustained Phonation Tests using ANN and Adaptive Neuro-Fuzzy Classifier.

[12] UCI Machine Learning Repository- Center for Machine Learning and Intelligent System. http://mlr.cs.umass.edu/ml/datasets/Parkinsons+Telemonitoring

[13] Knowledge Discovery in Databases. http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html

[14] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques.

9. 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.

Ms.Sivagami. G has completed her B.Tech in Information Technology at Tagore Engineering College affiliated to Anna University, Chennai, India. Currently she is pursuing her M.E. in Computer Science and Engineering at Rajalakshmi College of Engineering, affiliated to Anna University of Technology, Chennai, India. She has a 5 years of industrial experience from JK Technosoft, a global software and solutions company from Bangalore. Starting from Trainee her career took a steadfast growth to system analyst. Her clients were Hindustan Unilever Limited (HUL), the India's Largest Fast Moving Consumer Goods Company with a heritage of over 75 years in India and touches the lives of two out of three Indians. Her areas of interest include ERP, Data Mining, and Databases.