Prediction of Parkinson's disease using Machine Learning Techniques

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Software Engineering

Submitted By

Kirti Sharma

(Roll No. 801631010)

Under the supervision of:

Dr. Ashutosh Mishra

Designation of Supervisors

Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY

PATIALA – 147004

June 2018

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Prediction of Parkinson's disease using Machine Learning Techniques*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Software Engineering/ Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of Dr. Ashutosh Mishra and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Signature:

(Kirti Sharma)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Ashutosh Mishra)

Assistant Professor, CSE

ACKNOWLEDGEMENT

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life.

This work would not have been possible without the encouragement and able guidance of my supervisor **Dr. Ashutosh Mishra**. I thank my supervisor for their time, patience, discussions and valuable comments. Their enthusiasm and optimism made this experience both rewarding and enjoyable.

I am equally grateful to **Dr. Maninder Singh**,Professor and Head, Computer Science & Engineering Department, a nice person, an excellent teacher and a well – credited researcher, who always encouraged me to keep going with work and always advised me with his invaluable suggestions.

I am again thankful to **Dr. Ashutosh Mishra**, P.G. Coordinator, and Assistant Professor of Computer Science & Engineering Department of TIET for the motivation and inspiration for the completion of thesis.

I will be failing in my duty if I don't express my gratitude to **Dr. S.S. Bhatia,** Professor and Dean of Academic Affairs, TIET, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection, which made my stay at TIET memorable.

Last but not least, I would like to thank my family whom I dearly miss and without whose blessings none of this would have been possible. To my parents, I own thanks for their wonderful love and encouragement. I would also like to thank my brother, since he insisted that I should do so. I would also like to thank my close friends for their constant support.

(Kirti Sharma)

ABSTRACT

Parkinson's disease (PD) is one of the major public health disease in the world which is progressively increasing day by day and had its effect on many countries. Thus, it is very important to predict it in early age which has been challenging task among researchers because the symptoms of disease come into existence in either middle or late middle age. So this thesis focuses on the speech articulation difficulty symptoms of PD affected people and formulates the model using various machine learning techniques such as adaptive boosting, bagging, neural networks, support vector machine, decision tree, random forest and linear regression. Performance of these classifiers is evaluated using various metrics i.e. accuracy, receiver operating characteristic curve (ROC), Sensitivity, precision, specificity. At last, Boruta feature selection technique is used to find the most important features among all the feature to predict the Parkinson's disease.

TABLE OF CONTENTS

List of Tables

List of Figures

ABBREVIATIONS

ND's – Neurodegenerative disorders

PD – Parkinson's disease

ML – Machine learning

NN – Neural Network

RF – Random forest

SVM – Support vector machine

LR – Linear regression

ROC – Receiver operating operating characteristic curve

CHAPTER 1
INTRODUCTION

Neurodegenerative disorders are the results of the progressive tearing and neurons loss in different areas of the nervous system. Neurons are the functional unit of brain .They are contiguous rather than continuous. A good healthy looking neuron as shown in fig 1 has extensions called dendrites or axons, a cell body and a nucleus that contains our DNA. DNA is our genome and hundred billion neurons contains our entire genome which is packaged into it .When a neuron get sick, it loses its extension and hence its ability to communicate which is not good for it and its metabolism become low so it starts to accumulate junk and it tries to contain the junk in the little packages in little pockets .When things become worse and if the neuron is a cell culture it completely loses its extension, becomes round and full of the vacuoles.

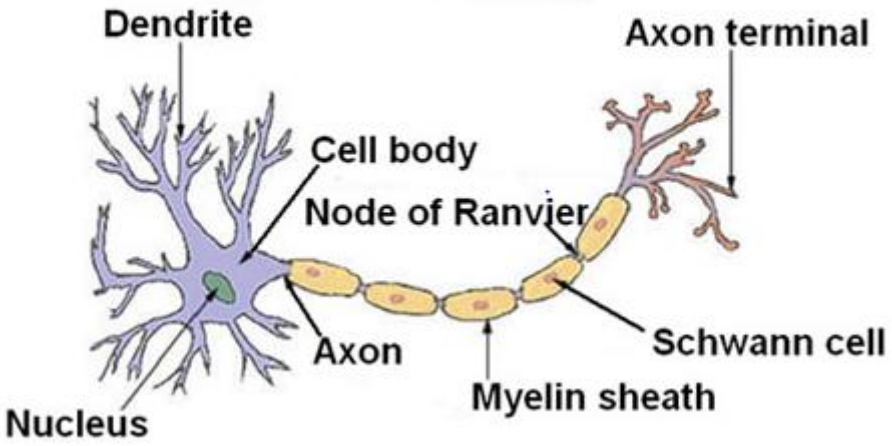

Fig 1: Structure of neuron present in human brain

This work deals with the prediction of Parkinson's disorder which is now a days is a tremendously increasing incurable disease. Parkinson's disease is most spreading disease [19] which get its

name from James Parkinson who earlier described it as a paralysis agitans and later gave his surname was known as a PD. It generally affects the neurons which is responsible for overall body movements. Main chemicals are dopamine and acetylcholine which affects human brain.

There are various environmental factor which have been implicated in PD [20].below are the listed factor which caused Parkinson's disease in an individual.

A. **Environmental factors**: Environment is defined as the surrounding or the place in which an individual lives .So environment is the major factor that will not only affects the human's brain but also affects all the living organism who lives in the vicinity of it. Many researches and evidences has proved that the environment has big hand in the development of neurodegenerative disorders mainly Alzheimer's and Parkinson's.

   There are certain environmental factors that are influencing neurodegenerative disorder with high pace are:-

   a. Exposure to heavy metals (like lead and aluminum) and pesticides.
   b. Air Quality: Pollution results in respiratory diseases.
   c. Water quality: Biotic and Abiotic contaminants present in water leads to water pollution.
   d. Unhealthy lifestyle: It leads to obesity and sedentary lifestyle.
   e. Psychological stress: It increases level of stress hormone that depletes the functions of neurons.

B. **Brain injuries or Biochemical Factors**: Brain is a control central of our complete body. Due to certain trauma people have brain injuries which leads some biochemical enzymes to come into picture which provides neurons a stability and provide support to some chromosomes and genes in maintenance.

C. **Aging Factor:** Aging is one of the reason for the development of the Parkinson's disease. Wi. According to the author [22], in India, 11,747,102 number of people out of 1, 065, 070, 6072 are affected from Parkinson's disease.

D. **Genetic factors:** Genetic factor is consider as the main molecular physiological cause which leads to neurodegenerative disorders. The size, depth and effect of actions of different genes

defines the status or level of neurodegenerative disease which increases itself gradually overtime. Mainly the genetic factors which leads to Neurodegenerative disorders are categorized into pharmacodynamics and pharmacokinetics [21].

E. **Speech Articulation factors**: Due to the condition associated with Parkinson's disease (rigidity and bradykinesia), some speech language pathology such voice , articulation and swallowing alterations are found . There are various ways in which Parkinson's disease (PD) might affect the individual.

(i) The voice get breathy and softer.

(ii) Speech may be smeared.

(iii) The person finds difficulty in finding the right words due to which speech becomes slower.

1.1 Motivation

Ten percent of people aged 65 or more do have a neurodegenerative disease, and there are no cures for them. Almost 30% of the people are facing this incurable disease[23]. Current treatment, if available at all,only reduces symptoms and that too for limited period of time. The main cause for the parkinson's disease is the accumulation of protein molecules in the neuron which gets misfolded and hence causing Parkinson's disease. So till now researchers got the symptoms and the root cause i.e. from where this disease had evolved. But very few have come to its cure.So in this era where parkinson's disease is progressing with double pace ,it is very important to find the solution which can detect it in its early phases.

1.2 Parkinson's Disease Symptoms

The symptoms of the Parkinson disease broadly divided into two categories.

- Motor symptoms
- Non-motor symptoms

1.2.1 Motor symptoms

This is a symptoms where any voluntary action involved. It's indicates the movement related disorder like tremor, rigidity, freezing, Bradykinesia or any voluntary muscle movement.

1.2.2 Non-motor symptoms

Non motor symptoms include disorders of mood and affect with apathy, cognitive dysfunction as well as complex behavioral disorders. There are two other categories of PD which are divided by doctors: Primary symptom and Secondary symptom.

1.2.3 Primary symptoms

It is the most important symptom. Primary symptoms are rigidity, tremor and slowness of movement.

1.2.4 Secondary symptoms

It is a symptom which directly impact life of an individual. These can be either motor or non-motor. Its effect depends on person to person.. A very wide range of symptoms is associated with Parkinson‘s,.

Besides these symptoms, there are some other symptoms found that leads to Parkinson’s disease. These symptoms are micrographia, decreased olfaction & postural instability, slowing of the digestive system, constipation, fatigue, weakness and Hypotension [24]. Speech difficulties i.e. dysphonia (impaired speech production) and dysarthria (speech articulation difficulties) [25] are found in patients of parkinsons..

1.3 Objective

The main objective is to predict the prediction efficiency that would be beneficial for the patients who are suffering from Parkinson and the percentage ratio will be reduced. Generally in the first stage Parkinson can be cured by the proper treatment. So it‘s important to identify the PD at the early stage for the betterment of the patients. The main purpose of this research work is to find the best prediction model i.e. the best machine learning technique which will distinguishes the Parkinson’s patient from the healthy person. The techniques investigated are Neural Network, SVM, Adaboost, Bagging, Linear Regression, Random Forest, Decision trees. We have found that Neural network ,SVM, Linear Regression have been reported in various researches, whereas it has been found that only few researchers have explored Adaboost and bagging. The experimental study is performed on the biomedical voice measurement from 31 people, 23 with Parkinson’s

disease. The prediction is evaluated using error rates.. Further the Feature selection technique has been implemented with the aim to get the important features that can detect the Parkinson's disease.

1.4 Thesis Organization

This research work proposes an effective methodology for identify the Parkinson's disease. Different results of the classification algorithms of machine learning are analyzed.

The rest of the thesis is as organized below: The first chapter provides a basic introduction of Neurodegenerative Disorders, Parkinson disorder and different types of symptoms on it. The second chapter gives an account of the review on pre-processing techniques, machine learning techniques and various technologies using which prediction have been done. The third chapter deals with the problem statement. This chapter deals with the main aim of carrying this research work and the objectives of the thesis. The fourth chapter deals with the material and the machine learning prediction techniques used in the paper. The fifth chapter identifies the tools and techniques employed to do the proposed work. It explains the R studio on which we have implemented all the models and the workflow .Sixth chapter illustrates the comparative analysis of all the techniques used . Seventh chapter contains Conclusion and Future scope .

CHAPTER 2

LITERATURE REVIEW

Prediction of Parkinson disorder is one of the most important problem that has to be detected in the early phases of the commencement of the disease so as to reduce the disease progression rate among the individuals .Various researches have been made to find the basic cause and some have reached to the heights by proposing a system which differentiates the healthy people from those with any ND'S using various machine learning techniques. Lots of pre-processing, feature selection and classification techniques have been implemented and developed in the past decades. Following is the given work done in the prediction of Parkinson's disorders. We have categories the review into three parts i.e.

(i) Review on Pre-processing techniques.
(ii) Review on classification methods.
(iii)Review on different computational methodology.

2.1 Review on pre-processing techniques used in Parkinson's disease prediction

Sahoo *et. al* (2012) [2] reported a study for the prediction of Parkinson's disease using Data mining techniques. The three methods used i.e. Decision stump, Logistic Regression and Sequential Minimization Optimization.The results inferred, support vector machine model outstands among the other with accuracy of 76%, sensitivity 0.97 while in terms of specificity statistical model has done well with 0.62 as compare with two other.

Bonato *et. al* (2004) [3] have proposed evidences that data mining and artificial intelligent may help in recognizing the severity of motor fluctuations in PD patient .They collected the data using ACC (accelerometer) and EMG (electromyography) signals which was recorded while execution of standardized sets of motor assessment tasks.

In another study, Saritha .k et al (2017) [4] have implemented javascript program to record the voice of the patient and later used Praat to convert that accepts input in .wav file and using a script

yields a voice report .Decision tree gave the best results among the applied algorithm with the accuracy of 100% without feature selection and with feature selection it is 94%.

2.2 Review on classification techniques used in prediction of Neurodegenerative disorders.

Nayan reddy challa *et al* (2016) [1] have discussed the importance of non-motor systems which was neglected by many doctors over motor systems. In the study Rapid eye movement (REM), sleep behavior distortion and olfactory loss were considered and using four machine learning techniques i.e. Multilayer Perceptron, Bayes Net, RF and Boosted Logistic Regression, prediction is performed. Among which Boosted logistic regression with an accuracy of 97.159% and area under ROC curve was 98.9%.is considered as a better method.

Chandrayan *et al.* (2016) [5] proposed extreme learning machines to predict PD..Using ELM they have done a comparative analysis and inferred that unlike conventional Neural Network elm doesn't require iterative variation of hidden neurons. So the simple architecture make elm a reliable method than others for prediction.

Jennifer He *et al.* (2017) [6] observed that the best feature for the prediction of Parkinson's disease is fundamental frequency among all voice recording features. They have tested various machine learning methods which includes Boosted decision tree, Decision jungle, Locally Deep SVM, Logistic regression ,Neural Networks and SVM on Microsoft azure machine learning studio amongst which the best is Two-class Boosted decision trees, which is an ensemble technique.

Weitschek *et al.* (2014) [7] uses EEG Electroencephalography to diagnose brain abnormalities. They have given an automatic patients classification from the EEG biomedical signals involved in Alzheimer's disease and MCI in order to support medical doctors. The authors performed preprocessing using time-frequency transforms and then applied classification using machine learning.

Rodrigues *et al.* (2012) [8] uses K-mean which obtain (EEG) temporal events in order to improve AD diagnosis . They achieved the sequence of EEG energy variation that are found more frequent in AD patient than in any healthy person.

Fernandez-Ruiz *et al.* [9] found that Alzheimer disease had shown a volume reduction at some region of the brain. Some areas like precuneus start showing changes when measured through the Magnetic Resonance Imaging. So in their study they took precuneus as a biomarker to identify defects in brain using machine learning techniques.

Johnstone *et al.* [10] took the dataset which was collected from ADNI (Alzheimer's disease Neuroimaging Initiative) of plasma proteome. They applied combinatorial optimization ie k-feature selection. So they differentiated the MCI patient and the AD patient depending on whether APOE was included or some other factors were there. At last they get an accuracy of 90% by generating the signature longitudinal rather that cross sectional data which further improved the classification.

Rathore *et al.(*2016) [16] used various machine learning techniques mainly regression techniques. After comparing the ML techniques ,error rates have been calculated ie AAE and ARE . K fold validation is applied to validate the results. At last, Kruskal Wallis test and Dunn's multiple comparison test is used to do comparative analysis of techniques used

In 2016, Kumar Tiwari [17] proposed minimum redundancy maximum relevance feature selection algorithm to select the most important feature which alone can predict Parkinson's disease. He observed that the random forest provided the overall accuracy of 90.3% which is better in comparison to all other machine learning based approaches such as bagging, rotation forest, random subspace, support vector machines etc.

Mamoshina *et al* (2016) [18] represented his work using deep learning as he states that it is different from traditional feature learning technique. He use deep learning with multiple hidden layers so as to provide meaningful and higher level of abstraction. Fig illustrates the approach in three steps.

(i) Started by preprocessing raw data to overcome main issues such as missing values, outliers and data quality.

(ii) Second step is to apply unsupervised deep learning for producing higher level of abstraction of input data.

(iii) Finally supervised learning method is implemented for predicting the target value and model evaluation.

Using the unsupervised learning before the supervised learning helped the author to get the high accuracy in predictive value as all ML techniques depends on feature representation and extraction.

2.3 Review on different methodologies used in Neurodegenerative disorder's prediction

Bioinformatics is emerging day by day and lots of researchers are now inclined towards this branch of science as Bioinformatics deals with the biological aspects of individual like health ,nutrition, environment .One of the most trending disorder is neurological disorders which has shown a tremendous increase in recent years. So by analyzing all the disorders, we found that different researchers have used different technologies to distinguish the ND patient with the healthy one. The technologies used are big data processing, Virtual reality, facial and emotion recognition, handwriting recognition and artificial intelligence .The table 1 given below illustrate the work of various researchers in the field of bioinformatics.

Table 1: Different technologies used in the prediction of Parkinson's disease.

Sno.	Paper Title	Description	Methods	Ref no
1.	Emulation of Physician Tasks in Eye-tracked Virtual Reality for Remote Diagnosis of Neurodegenerative Disease	• A virtual reality system which tracks how the eye movement of the individual has reduced the work for physicians. • They have taken a step to make a remote diagnosis a reality using Virtual reality	Virtual Reality	[11]
2.	Deep Machine Learning Application to the Detection of Preclinical Neurodegenerative Diseases of Aging	• The author of the paper had argued for a fundamentally different approach: using AI models applied to large datasets derived from single individuals, in order to detect preclinical decline.	Artificial intelligence (used in categorizing the health states.)	[12]

3.	Facial expression recognition in Alzheimer's disease: a longitudinal study	• The author has taken facial expression features for differentiating AD PATIENTS . • They observed a significant difference in the situational recognition task so cognition come up with the most suitable method to read expression in subtle cases also.	Facial /Emotion recognition	[13]
4.	An exploratory study on Big data processing: a case study from a biomedical informatics	• Big data processing is used in terms of biomedical informatics • They have considered medical imaging and bioinformatics. • It has been found that the data of these two field should be processed using big data technology	Big Data	[14]
5.	Machine learning-based classification of simple drawing movements in Parkinson's disease	• Based on the handwriting markers the author has differentiates the PD patient from healthy one. • They asked the patients to draw straight line, as doing this will involve certain muscular movements which was tracked.	Movements recognition	[15]

CHAPTER 3

PROBLEM DEFINITION AND OBJECTIVES

3.1 Research Gap and Problem Definition

Most of the studies reported in the literature survey focused on the usage of machine learning techniques like Logistic regression, Decision Tree, Support vector machine ,Random Forest .Very few studies performed Adaptive boosting ,Bagging and neural network . The study evaluated and compared various machine learning techniques for the early prediction of Parkinson's disease[17]. Our study is proposed with the aim to perform feature selection and to provide the comparative study of machine learning technique algorithms i.e. adaptive boosting, Bagging, Neural Network, Support vector machine, Random Forest, Decision Tree. So our study will focus on finding the best model to provide an automated method to extract the necessary biomarkers which will help in the prediction of Parkinson's disease.

3.2 Thesis Objectives

Various objectives that are needed to be fulfilled to solve the problem in hand are listed as below:

- To study and review various machine learning that could enhance the process of prediction of Parkinson's disease
- To find out the error rate using the predicted and actual values using different error techniques.
- To find various performance evaluation metrics and providing the comparative analysis to find the best method among them.
- To compute the performance of different ML techniques with various features selected by Boruta feature selection method

CHAPTER 4

MACHINE LEARNING METHODS FOR DISEASE PREDICTION

This chapter deals with the description of the dataset used and the approaches taken to achieve the early prediction of Parkinson's disease in a PD patient .The approaches taken were selected with the aim to distinguish a Parkinson's disease patient from those who are healthy patient. The idea is to do a comparative analysis of different machine learning technique by implementing different models on the selected dataset and finding the best machine learning technique among them by evaluating some performance metrics like accuracy, ROC, AAE, and ARE etc. Further the work is extended by implementing Boruta feature selection technique.

4.1 Dataset Description

The dataset was created by Max little of the University of Oxford, in collaboration with the national Centre for voice and speech, Denver, Colorado, who recorded the speech signals. This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). The parameters are classified into 6 categories i.e. Amplitude parameters, Pulse parameters , Frequency Parameters, Voicing Parameters, Pitch parameters, Harmonicity parameters as shown in the table 2. The datasets has 195 instances .Each column in the table is a particular voice measure, and each row corresponds one of the 195 voice recordings from these individuals. The 'Status' parameter is the most importance among all other parameter as it is the only parameter which will differentiate healthy people from those with Parkinson's disease. 0 states that the person is healthy while 1 states that the person has Parkinson's disease. The fig 2 illustrates the sample of data set used.

Table 2: Extracted Features From Speech Recordings

Feature	Group
Shimmer (dda) Shimmer (local) Shimmer (apq3) Shimmer (apq11) Shimmer (apq5) Shimmer (local,dB)	**Amplitude Parameters**
Number of pulses Mean period Number of periods Standard deviation of period	**Pulse Parameters**
Jitter (ddp) Jitter (local) Jitter (rap) Jitter (local, absolute) Jitter (ppq5)	**Frequency Parameters**
Number of voice breaks Fraction of locally unvoiced frames Degree of voice breaks	**Voicing Parameters**
Mean pitch Median pitch Standard Deviation Maximum pitch Minimum pitch	**Pitch Parameters**
Harmonic-to-Noise Noise-to-Harmonic Autocorrelation	**Harmonicity Parameters**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	MDVP:Fo	MDVP:Fhi	MDVP:Flo	MDVP:Jitt	MDVP:Jitt	MDVP:RA	MDVP:PP	Jitter:DDP	MDVP:Shi	MDVP:Shi	Shimmer:	Shimmer:	MDVP:AP	Shimmer:	NHR	HNR	RPDE	DFA	spread1	spread2
2	119.992	157.302	74.997	0.00784	0.00007	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	0.414783	0.815285	-4.81303	0.266482
3	122.4	148.65	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134	0.626	0.03134	0.04518	0.04368	0.09403	0.01929	19.085	0.458359	0.819521	-4.07519	0.33559
4	116.682	131.111	111.555	0.0105	0.00009	0.00544	0.00781	0.01633	0.05233	0.482	0.02757	0.03858	0.0359	0.0827	0.01309	20.651	0.429895	0.825288	-4.44318	0.311173
5	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492	0.517	0.02924	0.04005	0.03772	0.08771	0.01353	20.644	0.434969	0.819235	-4.1175	0.334147
6	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425	0.584	0.0349	0.04825	0.04465	0.1047	0.01767	19.649	0.417356	0.823484	-3.74779	0.234513
7	120.552	131.162	113.787	0.00968	0.00008	0.00463	0.0075	0.01388	0.04701	0.456	0.02328	0.03526	0.03243	0.06985	0.01222	21.378	0.415564	0.825069	-4.24287	0.299111
8	120.267	137.244	114.82	0.00333	0.00003	0.00155	0.00202	0.00466	0.01608	0.14	0.00779	0.00937	0.01351	0.02337	0.00607	24.886	0.59604	0.764112	-5.63432	0.257682
9	107.332	113.84	104.315	0.0029	0.00003	0.00144	0.00182	0.00431	0.01567	0.134	0.00829	0.00946	0.01256	0.02487	0.00344	26.892	0.63742	0.763262	-6.1676	0.183721
10	95.73	132.068	91.754	0.00551	0.00006	0.00293	0.00332	0.0088	0.02093	0.191	0.01073	0.01277	0.01717	0.03218	0.0107	21.812	0.615551	0.773587	-5.49868	0.327769
11	95.056	120.103	91.226	0.00532	0.00006	0.00268	0.00332	0.00803	0.02838	0.255	0.01441	0.01725	0.02444	0.04324	0.01022	21.862	0.547037	0.798463	-5.01188	0.325996

Fig 2: Sample dataset of biomedical voice measurements of 31 people

4.2 Prediction Techniques

4.2.1 Neural Network

Neural Network had its base as that of biological neuron which is used for prediction.. Let's understand the single neuron. In the fig 3 you can see a diagram of single neuron with single input. The given equation will explain the single input neuron where O is the output ,σ is the sigmoid function or transformed function ,ξ is the input to the neuron and ω is the weight that connects that input to the neuron

$O = \sigma(\xi\ \omega)$ **…… 1**

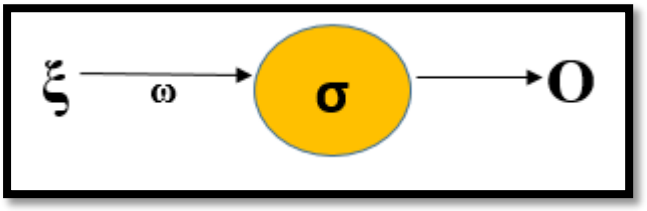

Fig 3: A Single input neuron0

So when multiple inputs are given to a neuron as mentioned in fig 4, it will form a MLP.hich consists of inputs connected through the weights in the form of layers. So the neuron takes multiple inputs and generates output which is known as Multilayer perceptron. The diagram below demonstrates a multilayer perceptron.

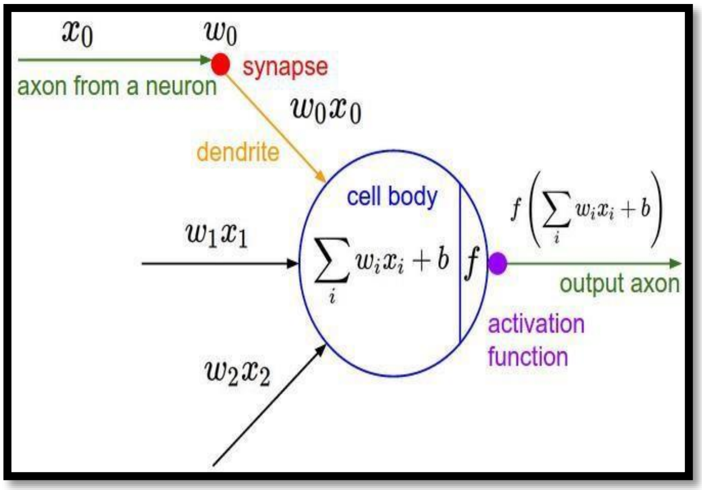

Fig 4 : Multilayer perceptron

$$O = \boldsymbol{\sigma}(\boldsymbol{\xi 1\ \omega 1} + \boldsymbol{\xi 2\ \omega 2} + \cdots \ldots \ldots + \boldsymbol{\xi k \omega k}) + \boldsymbol{\Theta} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\mathbf{2}$$

where O is the output

σ is the sigmoid function or transformed function

ξ is the input to the neuron

ω is the weight of input (1 to k)

Θ is the bias

4.2.2 Linear Regression

This model is used to find relationship between two continuous variable. One variable is called the dependent or response and the other one is called the independent or predictor using a best fit straight line known as regression line. The purpose of linear regression model is that it looks for a statistic relationship between the two variable and not the deterministic variable .By deterministic relationship we mean that if one variable can be accurately expressed by the other one.

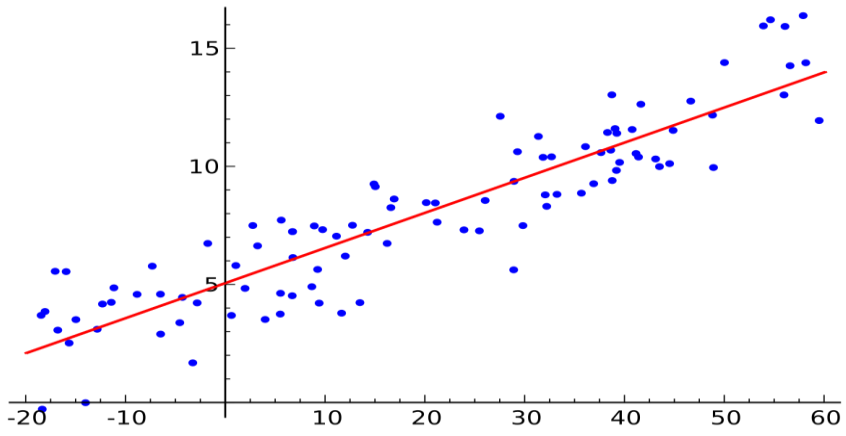

Fig 5: Straight line plot in Linear regression

The mathematical representation of Linear Regression:

$$Y=[X][W]+[B] \quad \ldots\ldots 3$$

$$Y=b_0+b_1X_1+b_2X_2 \quad \ldots\ldots 4$$

In eq 3 and eq 4, Y is the the dependent variable and X_1, X_2 represents the independent variables b_1, b_2 are the coefficients of the independent variables and b_0 is the intercept .

4.2.3 Random Forest

Random Forest is one of the machine learning method which is used for both classification as well as regression tasks. It is a type of ensemble method with which a group of weak model when combines turns into a powerful model. In random forest, multiple tress are created .To classify every tree gives a classification, are supposed to vote for that class. The forest selects the classification having the highest votes. The selection process by random forest is shown in fig 6.

Random Forest Prediction Pseudo code:

1. Takes the test sets features and make decision trees to predict the outcomes and stores the predicted outcomes.
2. Calculate the votes for each predicted outcome.
3. Consider the high voted predicted outcome as the final prediction.

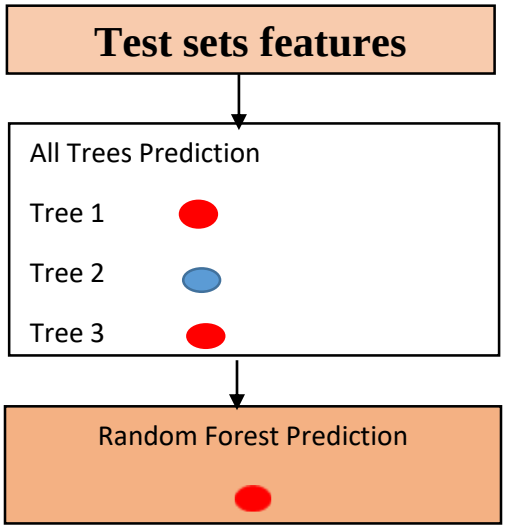

Fig 6: Prediction process taken by random forest

4.2.4. Decision Tree

Decision tree algorithm is a supervised learning algorithm which is used for the classification as well as regression problems. The main objective of using Decision tree is to create a training model which can be used for prediction of Parkinson's by learning decision rules inferred from training datasets. It tries to resolve the problem by using tree representation or tree hierarchy.It has three nodes:

1. Root
2. Internal Nodes
3. Leaf nodes

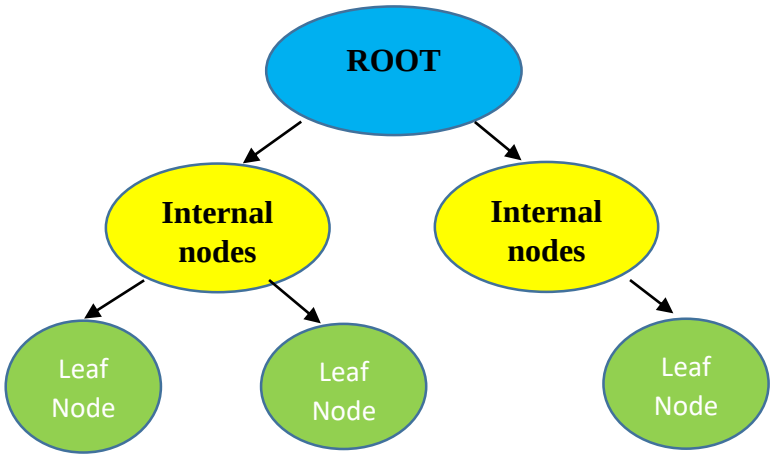

Fig 7 : Representation of decision tree

Root node represents the entire sample which is further splits into nodes known as leaf nodes which represents the attribute which is further divided into leaf nodes which represents the class labels.

4.2.5 Adaboost

Adaboost like random forest classifier is another ensemble classifier. AdaBoost which is known as adaptive boosting which is used for classification rather than regression .It is a best algorithm for predicting. It is used to boost the performance of decision tree or binary classification problems. . For the new input we are providing to adaboost, each weak learner calculates a predicted value .the vaue can be either 1.0 or -1.0. Each weak learner weights the predicted values. The prediction for the ensemble model is calculated by taking the sum of the weighted predictions.If the Sum is positive it will be assigned First predicted class,if Sum is negative it comes under Second predicted class.

Mathematics involved in Adaboost :

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} a_t h(x)\right)$$

h(x) is the output of weak classifier t for input x

a is the weight assigned to classifier.

$a = 0.5 * \ln((1-E)/E)$

4.2.6 Support Vector Machine

Support vector machine is defined by separating hyperplane. The output of the approach is an optimal hyper-plane which categorizes new examples. In 2 dimensional space, this new hyper plane in a line dividing a plane in two parts where each class lies in one side.It gives better result for complex classification problems. Each data item is plotted as a point in n- dimensional space with value of every feature reflecting the coordinates of the plane. The SVM is performed classification that differentiating the two classes very efficiently.

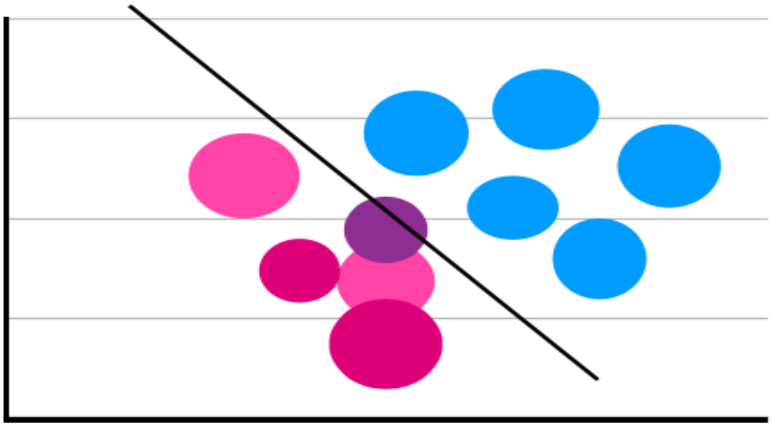

Fig 8: Hyper plane classifying two classes.

4.2.7 Bagging

Bagging is an ensemble algorithm, bagging methods forms an efficient class of algorithms which bring together several instances of black box estimators on random subsets of the original data set and then efficiently aggregate their individual predictions to process and formulate the final prediction. The bagging methods make immense efforts to reduce the variance of the base estimators by efficiently introducing the randomization into its construction and then makes an ensemble from it. Let’s take an instance where you have a learner for example The Decision Tree. Many times you have made efforts to improve its accuracy and variance by applying Bootstrap technique.

1. You end up generating multiple number of samples from your data set that has been classified as training set using an approach of next scheme: you can take randomly any element from your training set and then can pull it back. This results in a scenario where some of the elements of training set will be present multiple times in the generated new sample and some will be accidentally be absent. These samples should have the same size as the train set.

2. You can train your learner on each generated sample to gain the efficient results and improve the model better.

3. When you apply the algorithm you are just doing an average predictions of learners in case of regression or make the voting in case of classification.

CHAPTER 5

TOOLS AND METHODOLOGY

5.1 Tools

Tools that are used for implementation of the problem solution are as follows:

- R Studio: Version 0.99.473 - © 2009-2015 R Studio, Inc.
- Microsoft Excel 2013

Table 3: H/w and S/W requirements

1.	CPU	64 bit
2.	Random access memory	4 Gigabyte
3.	Operating system	Windows 10
4.	Programming Language	R
5..	Platform	R Studio

5.2 Methodology

This section explains the steps taken to achieve the prediction of Parkinson's disease using various machine learning. The various steps taken are Data gathering , Data Preprocessing, Model Selection, Training, Evaluation, prediction .

5.2.1 Data Gathering

The first step is Data gathering .This step is very important because the quality and quantity of the data you gather will directly affects the level of your prediction model. So we have taken data of different voice recordings of the patient.

5.2.1 Data preparation

In this step the data is visualized well to spot the relationship between the parameters present in the data so as to take the advantage of as well as to get the data imbalances. With this ,we need to split the data into two parts .The first part for training the model like in our model we have used 70 percent of data for training and 30 percentage for testing. Which is the second part of the data

5.2.2 Model Selection

The next step in our workflow is model selection. There are various models that have been used till date by researchers and scientist. Some are meant for image processing ,some for sequences like text, numbers or patterns. In our case we have 26 features which defines the voice recording

of various patients so we have chosen such models which will classify or differentiates the unhealthy patient with the healthy one.

5.2.3 Training

Training the dataset is one of the main task of machine learning .we will apply the data to progressively improve the selected model's ability to predict better ie the actual result should be approx. to predict one.

5.2.4 Evaluation

The metrics we have calculated are ROC, Accuracy , Specificity , Precision etc. which will highlights the best algorithm among all.

5.2.5 Prediction

In this phase we finally get the model ready to detect the prediction of Parkinson's disease based on the given dataset.

5.3 Using R Tool on Standalone machine Environment

The R computer programs are an essential tool for progression in the numeric examination and machine learning spaces. R is a perfect way to deal with make reproducible, extraordinary examination. R is extensible and offers rich value for architects to manufacture their own specific gadgets and procedures for examining data. With machines winding up recognizably more basic as data generators, the noticeable quality of the dialects must be depended upon to create.

In this module, the accuracy of different machine learning algorithms has been explored using R Tool on the Standalone machine. Here initial analysis has been done using Microsoft excel. A csv file has been provided as an input for R-Studio. Analysis has been done using programming language R as illustrated in fig 9.

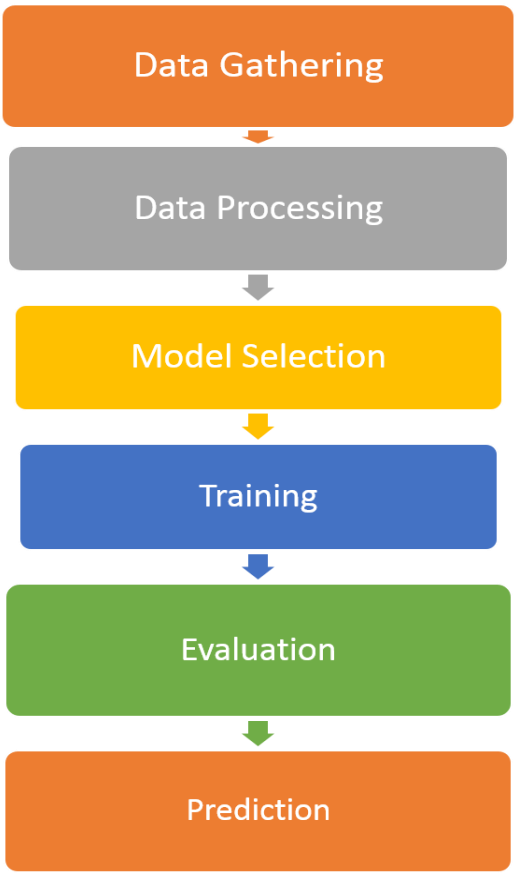

Fig 9: Workflow of training the models of ML in R

In any case, R has both upsides and downsides that designers ought to know. With enthusiasm for the programming developing, as appeared on language notoriety files, for example, TIobe, Redmond and PyPL, R initially showed up in the 1990s and has filled in as an execution of the S measurable programming languages.

"R is the most mainstream dialect utilized as a part of the field of statistics."It has all the adaptability and power. R is in reality only accumulations of scripts that are sorted out into projects."

Data purifying/cleaning is a term identified with getting the significant data from the crude information and noisy data removal (information not profitable to us). This should be possible effectively in Microsoft Excel and is a generally utilized strategy for each information researcher.

5.4 Evaluation Criteria Used for Classification

Performance evaluations measures are the parameters which helps in comparative analysis of different machine learning techniques i.e. it tells the best algorithm among all other algorithms or method which can be used by medical science in the early prediction of neurodegenerative diseases.

We have used several measures to evaluate the predictive results. These measures are average absolute error (AAE), average related error (ARE), accuracy (ACC), Precision, Receiver Operating Characteristics (ROC) , Area under ROC curve (AUC) ,sensitivity and specificity. Let's understand the performance evaluation measures.

5.4.1 Correlation Matrix

The confusion matrix is also called as Error matrix. It is a table that is often used to describe the performance of a classification method on a set of test data for which actual value are known. Each column of the matrix represents the instances in a predicted class. the correlation matrix is represented as given

		Predicted	
		No	**Yes**
Actual	**No**	**TN**	**FP**
	Yes	**FN**	**TP**

True Positive: is the count of healthy patients predicted accurately as healthy

True Negative: is the count of diseased subjects accurately predicted diseased.

False Positive: is the count of diseased patients predicted as healthy

False Negative: is the count of healthy patients predicted to be diseased

5.4.2 Accuracy and Precision

In classification, accuracy and precision are two important evaluation parameters. Accuracy is the proportion of the total number of predictions that were correct. It can be obtained by the sum of true positive and true negative instances divided by 100.And Precision is fraction of true positive and predicted yes instances. The formula of Accuracy and Precision are given below:

$$\text{Accuracy} = \frac{\text{TP+TN}}{100} \qquad \text{Precision} = \frac{\text{TP}}{TP+FP}$$

5.4.3 Recall and F-Square

Recall is defined as the fraction between True Positive instances and Actual yes instances whereas F-Square is the fraction between product of the recall and precision to the summation of recall and precision parameter of classification. The formula of recall and precision given below:

$$\text{Recall} = \frac{\text{TP}}{\text{Actual Yes}} \qquad \text{F-Square} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

5.4.4 Sensitivity, Specificity and ROC

Sensitivity is defined as the fraction of true positive and actual yes instances whereas specificity is the difference between one and false positive rate value. .ROC is defined as the fraction between true positive rate and false positive rate.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \qquad \text{Specificity} = \frac{\text{TN}}{FP+TN} \qquad \text{ROC} = \frac{\text{TPR}}{\text{FPR}}$$

CHAPTER 6

IMPLEMENTATION AND RESULTS

In chapter 6, we have explained the implementation of different machine learning techniques and the feature selection method ie Boruta method. We have divided our work in two parts:

1. In the first part, we have run seven machine learning models and evaluated their error rates i.e. AAE and ARE as given in table 4.
2. After that, based on their different error rates, we have evaluated and compared all the methods on the basis of their accuracy, ROC, sensitivity, specificity and precision values
3. We found that random forest is the most efficient algorithm with the accuracy of 87%, Precision 85.0%, ROC 96.4%.
4. In the second part , we are trying to selected the most important and minimum number of features from the speech articulation data of 31 people where we have 23 features as explained in chapter 4 in dataset description .
5. For that we have used Boruta feature selection whose working is shown in fig 12 by changing the number of features selected in multiples of 5 ie firstly we check over 20 features than 15 features, 10 features and lastly 5 features.
6. From all the experiments random forest with 20 features selection outstands from all the other ML techniques as it is giving the overall accuracy 96.6%, ROC value 93.6 and precision of 88.7 which is better from all other machine learning techniques when compared with 5,10 and 15 feature's performance metrics.

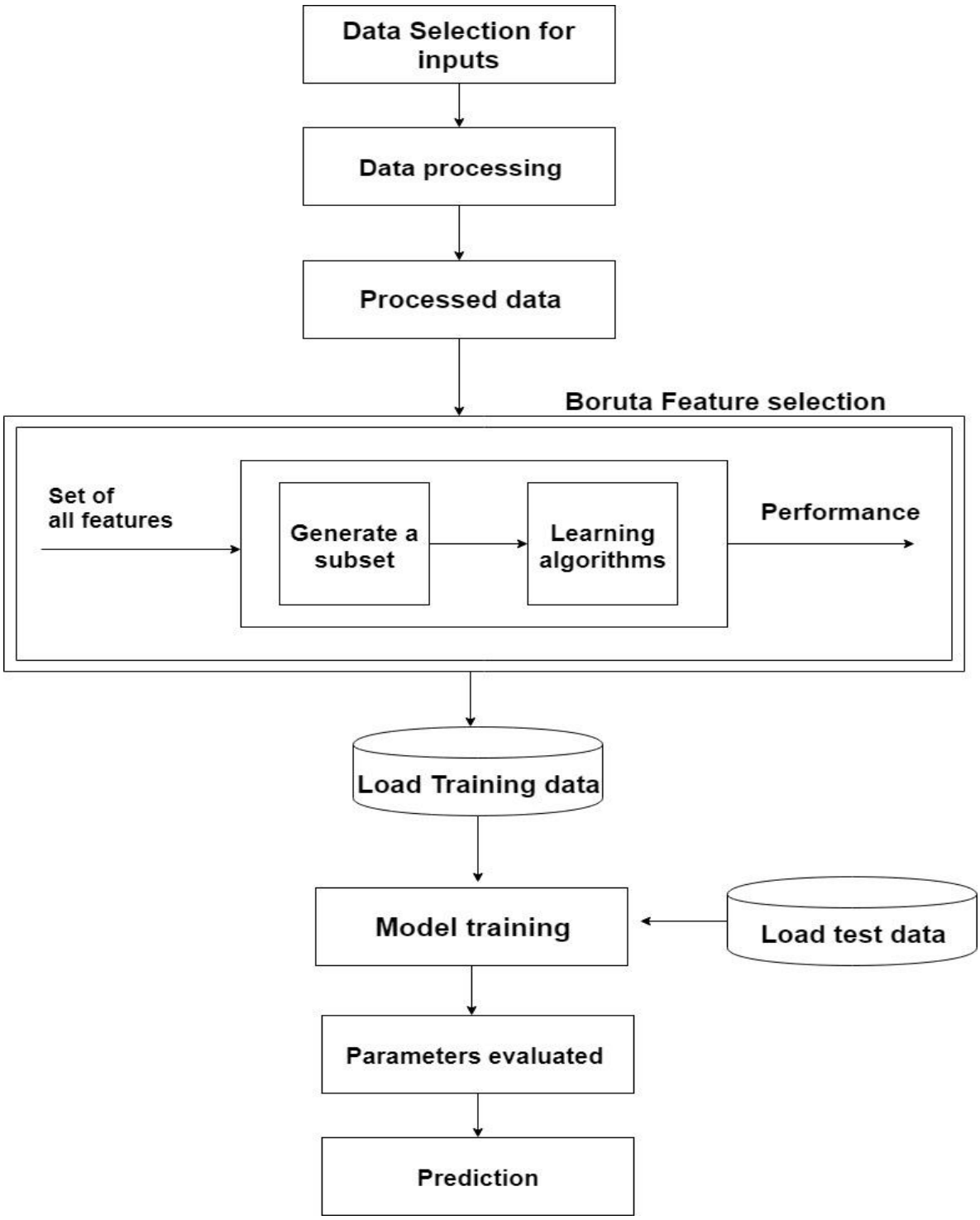

Fig 10: Feature selection by boruta method

6.1 AAE and ARE

1. Average absolute error (AAE): AAE measure is calculated by taking the difference of predicted value and the actual value .It can be understand average absolute error by the below equation

$$AAE = (1/n) \sum_{I=1}^{n} \left| (\bar{Y}i - Y) \right| \quad \text{.....5}$$

In eq 5, $\bar{Y}i$ is the predicted status value of the patient.

Y is the corresponding actual status value of the patient. .

n is the total number of parameters/columns.

2. Average related error (ARE): ARE measures how large the absolute error is compared with the total size of the object measured. It is defined by the below equation

$$ARE = (1/n) \sum_{I=1}^{n} \left| (\bar{Y}i - Y) \right| / (Yi+1) \quad \text{.....6}$$

In eq 6, $\bar{Y}i$ = predicted status value of the patient.

Y = actual status value of the patient.

n = total number of parameters/columns.

In Average related error, sometimes the value of Yi can be zero, to make the definition we defined , we need to add '1' with the value of Yi at the denominator . A small value of AAE and ARE measures indicates that we have a good classification model. The calculated values are shown in Table 4.

Table 4: Error rates analysis of seven classification methods

Sno.	Classifier	AAE	ARE
1.	**Adaboost**	0.040609	0.035533
2.	**Bagging**	0.055838	0.048223
3.	**Decision tree**	0.060914	0.055838
4.	**Neural Networks**	0.274112	0.187817
5.	**Random Forest**	**0.040609**	**0.035533**
6.	**Support vector machine**	0.045685	0.045685
7.	**Linear Regression**	0.050761	0.043147

From the above analysis, we can conclude that random forest is the better model as we know lesser the error rates, more efficient is the algorithm.

6.2 AAE and ARE analysis

AAE and ARE analysis as given in table 4 that Random Forest and Adaboost techniques come up with the lower error rate values and hence outstands best from all the other techniques used. Support vector machine and bagging are better classification techniques. While neural network and decision tree have produced relatively lower accuracy as AAE and ARE values are higher as compared to the other techniques.

In Fig 11 and 12, X-axis shows the prediction techniques and Y-axis shows the values of error produced by the classification techniques.

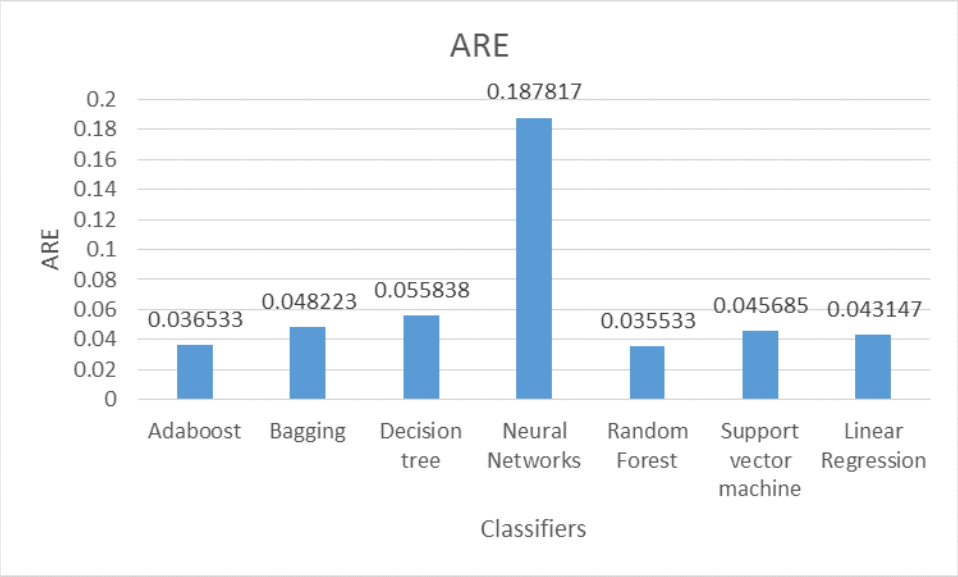

Fig 11: ARE analysis of different ML techniques

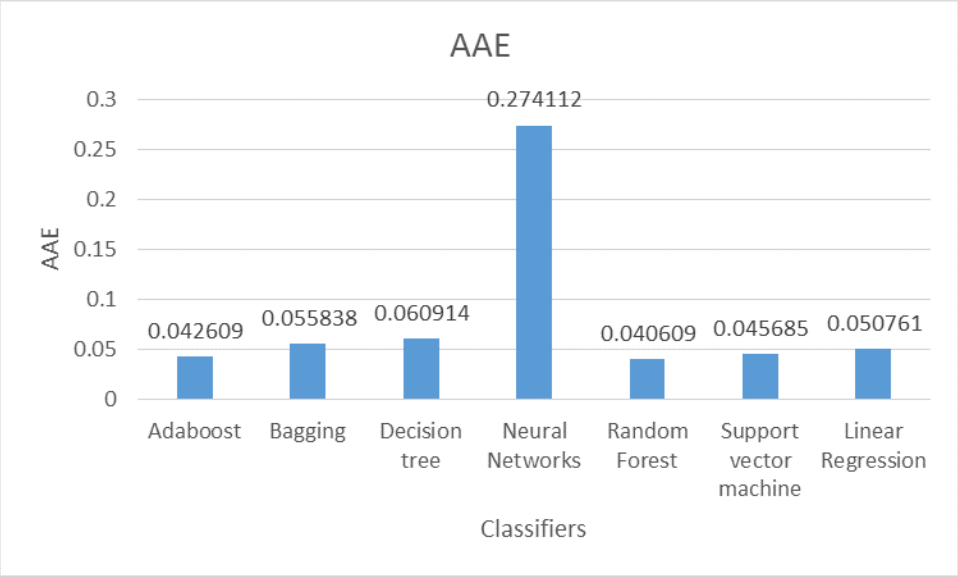

Fig 12: AAE analysis of different ML techniques

6.3 Comparative Analysis of classification techniques.

We have performed a comparative analysis of seven classification techniques and we have calculated the performance metrics such as accuracy, ROC, Precision, Confusion Matrix, Sensitivity, Specificity. Based on which we have predicted the best machine learning technique It is observed that Random Forest is better in comparison to all other ML techniques with accuracy

87%, Precision 85 % , ROC 96.4% that can be see through the shaded table 5. After the comparative analysis of ML techniques, each performance metrics has been observed and the major metrics i.e. accuracy and ROC (receiver operating characteristic curve) has been analyzed individually. Fig 14 shows the Accuracy analysis graph.

Table 5: Comparison between all the ML techniques using performance metrics

SNO.	Classifier	Confusion Matrix	ROC	ACC	Precision	Sensitivity	Specificity
1.	ADABOOST	TP 40 FP 6 TN 11 FN 2	95	86.4	87	0.95	0.64
2.	BAGGING	TP 38 FP 8 TN 10 FN 3	74	81.3	82	0.92	0.55
3.	NEURAL NETWORK	TP 28 FP 9 TN 10 FN 12	69	75	30.51	0.52	0.7
4.	DECISION TREE	TP 40 FP 10 TN 7 FN 2	68	79.6	80	0.95	0.41

5.	**RANDOM FOREST**	TP 42 FP 7 TN 9 FN 1	**97**	**87**	**85**	**0.97**	**0.56**
6.	SVM	TP 39 FP 9 TN 11 FN 0	95	84.7	81.2	1	0.55
7.	LINEAR REGRESSION	TP 44 FP 7 TN 5 FN 3	0.791	86	83.05	0.417	0.936

6.3.1 Accuracy analysis

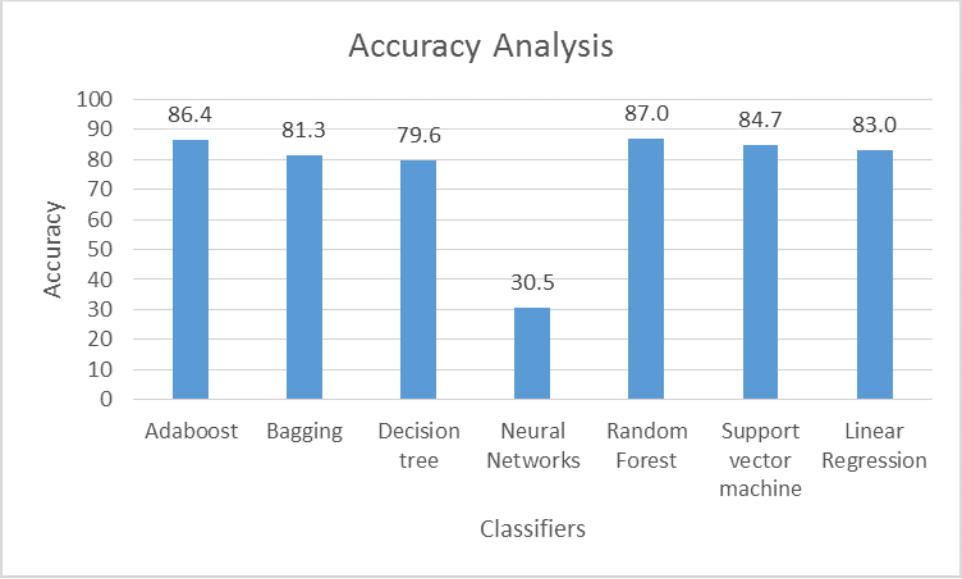

Fig 13: Accuracy analysis of different ML Techniques

With respect to accuracy, Random forest shows better results when compared to other techniques which can be seen by fig 13. After random forest, Adaptive boosting outstands from support vector machine and linear regression. Bagging has shown better results than decision tree and worst results were shown by neural networks.

6.3.2 ROC analysis

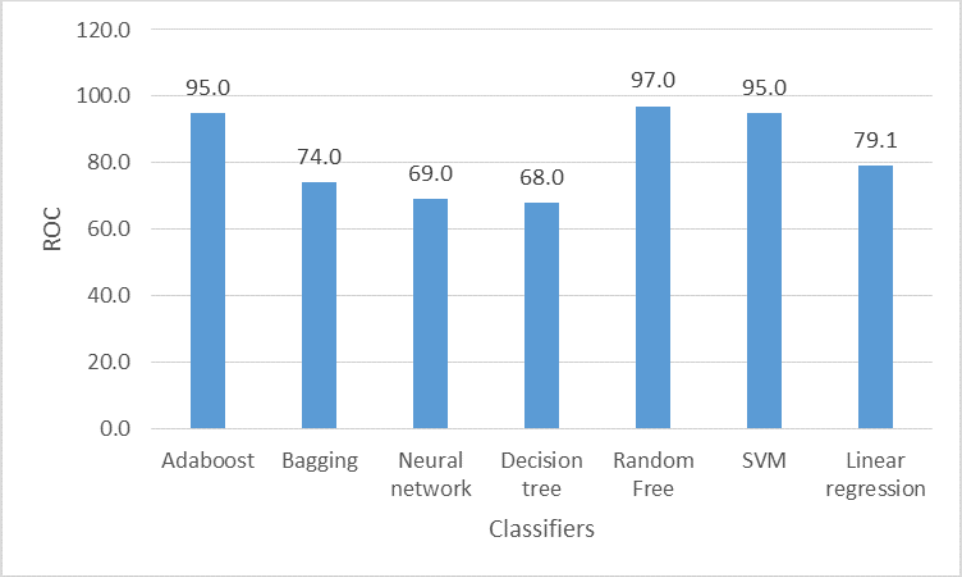

Fig 14: ROC analysis of different ML Techniques

AUC stands for "Area under the ROC curve" i.e. it measures the performance across all possible classifications thresholds values or we can say it is the probability that the model identifies the random positive example more highly than random negative example. It is used for visualize binary classifier performance ie with two output classes. With respect to the fig 14, random forest outstands when compared to other methods with ROC curve value of 97%.

6.4 Boruta feature selection

So far, in this work we have found the accuracies of different models on all the 23 features extracted by Patients. Also, we have selected the minimum features which alone can predict the PD. Here, the Boruta feature selection technique has been used to select the 5 number of features, 10 number of features, 15 number of features and 20 number of features among all the features. After that performance of various machine learning techniques are evaluated with different features selected by boruta method as illustrated in table 6,7,8 and 9.

Table 6: 20 Features selected by Boruta

S.No.	CLASSIFIERS	ACCURACY	ROC	PRECISION
1	AdaBoost	91.3	92.6	91.33
2	Bagging	81.36	77.2	88.1
3.	Neural Network	77.3	66.4	75.8
4.	Decision Tree	80.9	75.3	82.2
5.	**Random Forest**	**96.6**	**93.6**	**88.7**
6.	Support vector machine	83.05	82.7	81.1
7.	Linear Regression	81.36	82	87.2

Table 7: 15 Features selected by Boruta

S.No.	CLASSIFIERS	ACCURACY	ROC	PRECISION
1	AdaBoost	93.9	90.2	91
2	Bagging	89	85.3	85
3.	Neural Network	81.3	75	80.1
4.	Decision Tree	77.3	75.3	80.1
5.	**Random Forest**	**94.9**	**94.9**	**94.2**
6.	Support vector machine	83.5	85.3	82
7.	Linear Regression	84.7	94.7	91.5

Table 8: 10 Features selected by Boruta

S.No	CLASSIFIERS	ACCURACY	ROC	PRECISION
1	AdaBoost	93.2	92	91.6
2	Bagging	86.44	87.1	92.7
3.	Neural Network	60.4	89	85.3
4.	Decision Tree	58.3	77.8	78
5.	**Random Forest**	**91.53**	**97**	**91.1**
6.	Support vector machine	86.4	93.3	88.2
7.	Linear Regression	89.8	93.8	90

Table 9: 5 Features selected by Boruta

S. No	CLASSIFIERS	ACCURACY	ROC	PRECISION
1	AdaBoost	91.2	82	89
2	Bagging	86.44	76.2	89.4
3.	Neural Network	77.9	78	75.5
4.	Decision Tree	60.4	72	63.2
5.	**Random Forest**	**88.14**	**94.6**	**86.4**
6.	Support vector machine	87	92.8	88.2
7.	Linear Regression	81.36	90.1	80.9

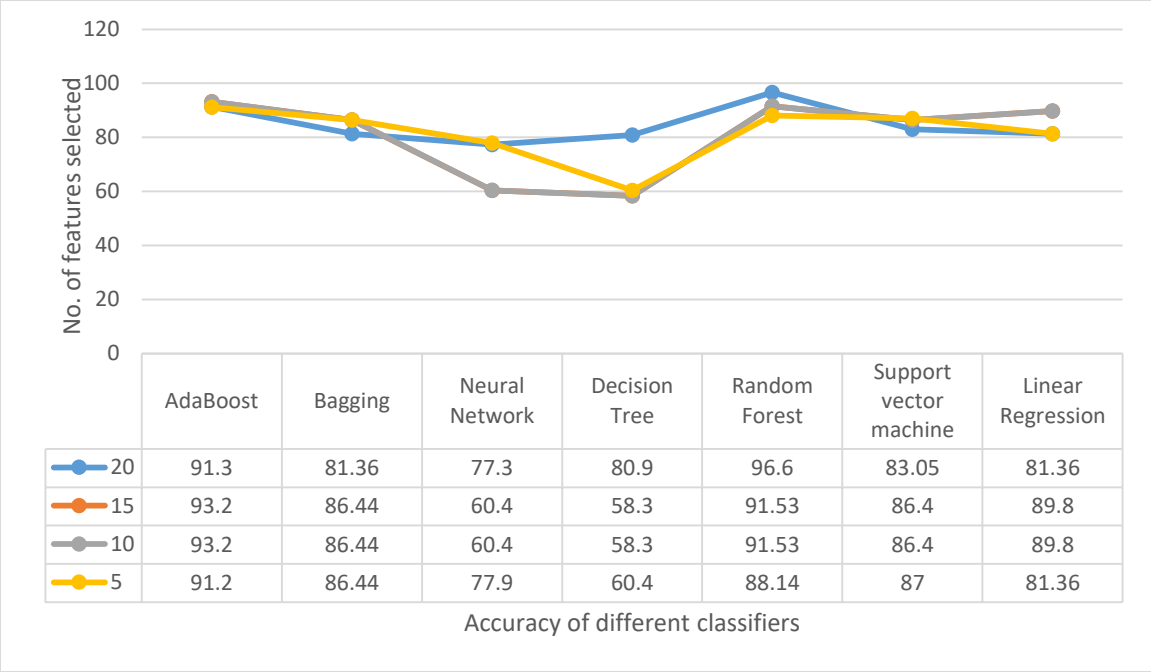

Fig 15: Accuracy versus No. of selected features.

It is observed that the random forest with 20 numbers of features selected by boruta feature selection algorithm provided the overall accuracy 96.6%, ROC value 93.6 and precision of 88.7 techniques has shows best results when compared with 5,10 and 15 feature's performance metrics as shown in fig 15.The graph shows the accuracy versus feature selection plot, where y axis represents the number of features selected by boruta feature selection algorithm and x axis represents the hit value of accuracy.

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

In this work, various prediction models for Parkinson's disease detection. For this purpose seven machine learning techniques i.e. are used such as adaptive boosting, bagging, neural networks, random forest, decision tree, SVM and linear regression. To obtain the desired results, error rates are calculated i.e. AAE and ARE as well as four performance metrics are evaluated. These four metrics are accuracy, sensitivity, ROC, specificity.

From the results, Random forest outstands from all the other ML techniques with the accuracy of 87%, Precision 85.0%, ROC 96.4%. After that , we tried to selected the most important and minimum number of features from the speech articulation data of 31 people where we have 23 features as explained in chapter 4 in dataset description .For that we have used Boruta feature selection whose working is shown in fig 12 by changing the number of features selected in multiples of 5 ie firstly we check over 20 features than 15 features, 10 features and lastly 5 features. From all the experiments random forest with 20 features selection outstands from all the other ML techniques as it is giving the overall accuracy 96.6%, ROC value 93.6 and precision of 88.7 which is better in comparison to all other machine learning techniques when compared with 5,10 and 15 feature's performance metrics.

7.2 Future scope

In this study we have used machine learning techniques, however very few researches have been done on deep learning methods. In future, the work can be extended by using autoencoders to reduce the number of feature and to extract the most important from them. Also the dataset used in this work is not so complex , so autoencoder did not learn well from that but with complex dataset it would definitely give better results.

REFERENCES

[1] Kamal Nayan Reddy, Challa, Venkata Sasank Pagolu and Ganapati Panda, "An Improved Approach for Prediction of Parkinson's Disease using Machine Learning Techniques", in *Procedings of the International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016,* pp. 1446-145, 2016.

[2] Geeta Yadav, Yugal Kumar and G. Sahoo, "Predication of Parkinson's disease using Data Mining Methods: a comparative analysis of tree, statistical and support vector machine classifiers", in *Procedings of the National Conference on Computing and Communication Systems (NCCCS),* pp. 1-4, 2012.

[3] Paolo Bonato, Delsey M. Sherrill, David G. Standaert, Sara S. Salles and Metin Akay, "Data Mining Techniques to Detect Motor Fluctuations in Parkinson's Disease", in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4766-4769, 2004.

[4] Sonu S. R., Vivek Prakash and Ravi Ranjan, "Prediction of Parkinson's Disease using Data Mining", in *Proceedings of the International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS),* pp. 1082-1085, 2017.

[5] Aarushi Agarwal, Spriha Chandrayan and Sitanshu S Sahu, "Prediction of Parkinson's Disease using Speech Signal with Extreme Learning Machine", in *Proceedings of the International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp. 1-4, 2016.

[6] Akshaya Dinesh and Jennifer He, "Using Machine Learning to Diagnose Parkinson's Disease from Voice Recording", in *Proceedings of the IEEE MIT Undergraduate Research Technology Conference (URTC),* pp. 1-4, 2017.

[7] Giulia Fiscon, Emanuel Weitschek, Giovanni Felici and Paola Bertolazzi, "Alzheimer's disease patients classification through EEG signals processing", in *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. pp 1-4, 2014.

[8] Pedro Miguel Rodrigues, Diamantino Freitas and Joao Paulo Teixeirab, "Alzheimer electroencephalogram temporal events detection by K-means", in *Proceedings of the International Conference on Health and Social Care Information Systems and Technologies HCIST.* pp. 859 – 864, 2012.

[9] Elva Maria Novoa-del-Toro, Juan Fernandez-Ruiz, Hector Gabriel Acosta-Mesa and Nicandro Cruz-Ramirez, "Applied Macine Learning to Identify Alzheimer's Disease through the Analysis of Magnetic Resonance Imaging", in *Proceedings of the International Conference on Computational Science and Computational Intelligence,* pp. 577-582, 2015.

[10] Daniel Johnstone1, Elizabeth A. Milward1, Regina Berretta1 and Pablo Moscato1, "Multivariate Protein Signatures of Pre-Clinical Alzheimer's Disease in the Alzheimer's Disease Neuroimaging Initiative (ADNI) Plasma Proteome Dataset", in *Proceedings of the Disease Neuroimaging Initiative*, vol-7, pp. 1-17, 2017.

[11] Jason Orlosky, Yuta Itoh, Maud Ranchet, Kiyoshi Kiyokawa, John Morgan, and Hannes Devos, "Emulation of Physician Tasks in Eye-tracked Virtual Reality for Remote Diagnosis of Neurodegenerative Disease", in *Proceedings of the IEEE Transactions on Visualization and Computer Graphics*, vol. 23, pp. 1302 – 1311, 2017.

[12] Mathew J. Summers, Vienna, Austria, Alessandro E. Vercelli, Georg Aumayr, Doris M. Bleier and Ludovico Ciferri,"Deep Machine Learning Application to the Detection of Preclinical Neurodegenerative Diseases of Aging", in *Proceedings of the Scientific Journal on Digital Cultures*, vol. 2, pp. 9-24, 2017.

[13] Bianca Torres, Raquel Luiza Santos, Maria Fernanda Barroso de Sousa, Jose Pedro Simoes Neto, Marcela Moreira Lima Nogueira, Tatiana T. Belfort1, Rachel Dias1, Marcia and Cristina Nascimento Dourado, "Facial expression recognition in Alzheimer's disease: a longitudinal study", pp. 383-389, 2014.

[14] Smitha Sunil and Kumaran Nair, "An exploratory study on Big data processing: a case study from a biomedical informatics", 3rd MEC International Conference on Big Data and Smart City, pp. 1-4, 2016.

[15] C. Kotsavasilogloua, N. Kostikis, D. Hristu-Varsakelis and M. Arnaoutoglouc, "Machine learning-based classification of simple drawing movements in Parkinson's disease", in *Proceedings of the Biomedical Signal Processing and Control*, pp. 174–180, 2017.

[16] Santosh S. Rathore and Sandeep Kumar, "An empirical study of some software fault prediction techniques for the number of faults prediction", in *Proceedings of the Soft Computing*, vol. 21, pp 7417–7434, 2017.

[17] Arvind Kumar Tiwari, "Machine Learning Based Approaches for Prediction of Parkinson's Disease", in *Proceedings of the Machine Learning and Applications: An International Journal (MLAIJ)*, vol.3, pp. 33-39, 2016.

[18] Polina Mamoshina, Armando Vieira, Evgeny Putin and Alex Zhavoronkov, "Applications of Deep Learning in Biomedicine", *in Proceedings of the American Chemical Sociecty Mol. Pharmaceutics*, pp. 1445−1454, 2016.

[19] Alexis Elbaz,James H. Bower, Brett J. Peterson, Demetrius M. Maraganore, Shannon K. McDonnell, J. Eric Ahlskog, Daniel J. Schaid, Walter A. Rocca, "Survival Study of Parkinson Disease in Olmsted County, Minnesota ", Arch Neurol. Vol. 60 pp. 91-96, 2003.

[20] Tanner CM, Ross GW, Jewell SA ," Occupation and risk of Parkinsonism: a multicenter case- control study" Arch Neurol,66(9):1106–1113,2009.

[21] V. A. Sukhanov, I. D. Ionov, and L. A. Piruzyan, "Neurodegenerative Disorders: The Role of Genetic Factors in Their Origin and the Efficiency of Treatment" in *Proceedings of the Human Physiology US National Library of Medicine National Institutes of Health*, vol. 31, pp. 472–482, 2005.

[22]Marras C, Tanner C."Epidemiology of Parkinson's Disease", Movement Disorders: Neurologic Principles and Practice, 2nd ed.2004, Watts, RL, Koller, WC (Eds). The McGraw-Hill Companies:New York, pp. 177.

[23]http://www.orionpharma.co.uk/Products-and-Services-Orion/Parkinsons-disease/10-facts-about-Parkinsonsdisease/

[24] Cnockaert, L., Schoentgen, J., Auzou, P., Ozsancak, C.,Defebvre, L., & Grenez, F., "Low frequency vocal modulations in vowels produced by Parkinsonian subjects", Speech Communications, vol 50, pp. 288-300, 2008.

[25] Kenneth Revett, Florin Gorunescu and Abdel-Badeeh Mohamed Salem, "Feature Selection in Parkinson's disease: A Rough Sets Approach", Proceedings of the International Multi onference on Computer Science and Information Technology, pp. 425 – 428,2004, ISBN 978-83-60810- 22-4.

LIST OF PUBLICATION

[1] Kirti Sharma and Ashutosh Mishra," Prediction of Parkinson Disorder using machine learning techniques", *8th IACC International Advance Computing Conference 2018.* [Communicated]