Ensemble Classifier to Enhance Computer Aided Diagnosis of Parkinson Disease

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

Master of Engineering
in
Computer Science and Engineering

Submitted By
Harkawalpreet Kaur
(Roll No. 801632008)

Under the supervision of:
Dr. Avleen Kaur Malhi
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, "*Ensemble classifier to enhance computer aided diagnosis of Parkinson disease*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *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. Avleen Kaur Malhi* 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:

Harkawalpreet Kaur

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

14/6/18

Dr. Avleen Kaur Malhi

Assistant Professor
Computer Science and Engineering Department

Countersigned by

(Dr. Maninder Singh)
Head
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala

(Dr. S S Bhatia)
Dean (Academic Affairs)
Thapar Institute of Engineering
and Technology
Patiala

ACKNOWLEDGEMENT

Thanks to the almighty for showering the blessings

Nothing concrete can be achieved without a combination of inspiration. Although writing a few words on a piece of paper is not a proper way of acknowledging those people who has helped me in writing this dissertation, yet the words coming from our heart and soul need no mode of communication.

My first, and most earnest, acknowledgement must go to my advisor and mentor Dr. Avleen Kaur Malhi and my Head of Department Dr. Maninder Singh. They have been instrumental in ensuring academic and moral well being ever since. In every sense, none of this work would have been possible without them.

Far too many people to mention individually have assisted in so many ways. They all have my sincere gratitude and I thank several anonymous publishers of various journals and research papers for their help. I also owe a huge debt of gratitude to my friends who were instrumental in the success.

A penultimate thank you goes to my wonderful parents, for always being there when I needed them most.

Harkawalpreet Kaur

ABSTRACT

Parkinson's disease is a neurodegenerative disorder of the central nerve system which affects movements. Data from 42 persons having early stage of Parkinson's disease was collected with a total number of 5875 voice recordings present in dataset. The different machine learning models were used to predict the motor Unified Parkinsons disease rating score (UPDRS) score from the various voice measures. Then the actual and predicted values for various evaluation parameters (Correlation, R Square, RMSE, Accuracy) are calculated and results are compared. After comparing the results of the various models, the top 5 models are ensembled and results are calculated to give stronger overall prediction. The aim of ensembled model is to calculate the UPDRS from various voice measures with higher accuracy of 99.6%. K-Fold validation approach is used to measure the robustness of ensembled model.

Index Terms—Parkinson's disease, Machine Learning, Ensemble, UPRDS

LIST OF ABBREVIATIONS

1. PD	Parkinson's Disease
2. MS	Motor Symptoms
3. NMS	Non-Motor Symptoms
4. UPDRS	Unified Parkinson's Disease Rating Scale

TABLE OF CONTENT

Chapter No. **Page No.**

LIST OF FIGURES

Figure No. **Page No.**

LIST OF TABLES

Table No. **Page No.**

CHAPTER 1

INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disorder of central nerve system which affects the body movements. It is a progressive disorder that affects movements of body. Parkinson's person's muscles are weaker than the individual which is healthy and may assume an unusual postures. It belongs to the group of conditions called movement disorder. It describes neurological behaviour which includes abnormal body movements such conditions as Tourette syndrome and cerebral palsy.

1.1 INTRODUCTION

Parkinson's disease was first introduced by Doctor James Parkinson as shaking palsy in 1817 [1]. This disease is most common among the elders and it is the second disease after Alzheimer [2]. Approximately 60,000 adults are diagnosed out of one million adults in the annually. The real figure is much higher than that when counting the people those are undetected. Parkinson's disease causes various side effects and signs. The signs and side effects are classified into two categories Motor Symptoms (MS) and Non-Motor symptoms (NMS) as shown in figure 1.1.

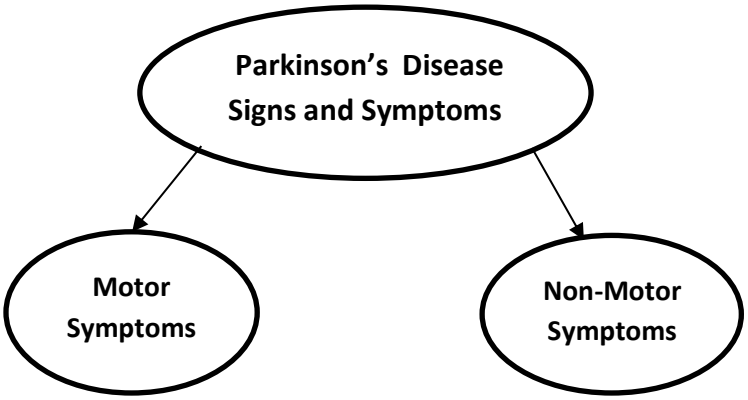

Figure 1.1: PD Symptoms Classification

Motor side effects affect movements of muscles and non-motor side effects include problems like brain problems, sleep problems and sensory problems.

After the age of 50, symptoms starts to appear. When signs and symptoms develop ranges 21 to 40 years in individual it is called as young-onset Parkinson's disease [3]. Vocal impairment is also common [12] [14].

Figure 1.2: Fixed Inexpressive Face [29]

Despite from the tremor and slow movements, fixed impressive face is also noticed in patients as shown in the figure 1.2. This is due to poor control upon the facial muscles movements and coordination. Parkinson's disease affects the voice too. Degrading performance in voice with PD progression is supported by evidence [7][17][18]. Dysphonia (hoarseness, breathiness and creakiness in the voice) and Hypohonia (reduced voice volume) are more generalized speech disorders [8] [12]. Speech disturbance is most common noticed symptom in the PD patient. It has been found from the research that 90% of PWP (people with Parkinson's Disease) are affected with motor problems. The different symptoms of PD are shown in figure 1.3.

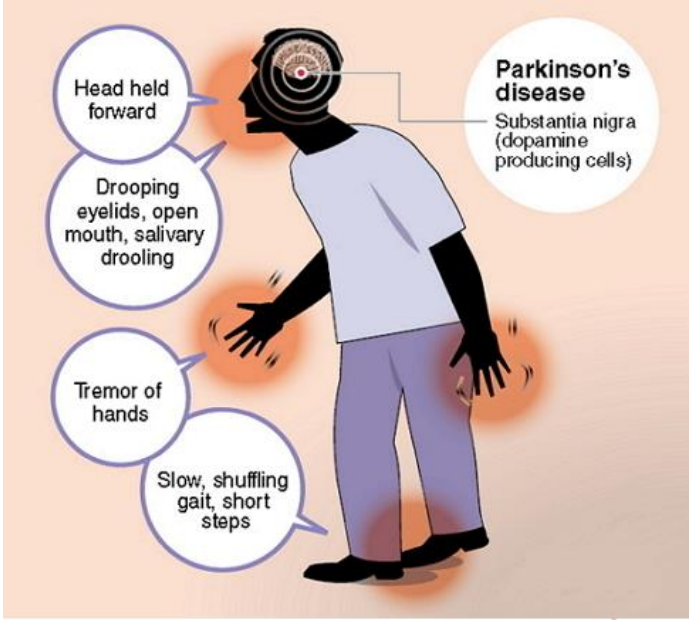

Figure 1.3: Affects of Parkinson's disease on muscles [30].

Parkinson's disease include following common symptoms:

- Slow body movement
- Trouble in speaking
- Stiff muscles
- Problems in balancing and walking
- Tremor of arms, hands or legs

1.1.1 CAUSE OF PARKINSON'S DISEASE

The root cause of PD is falling and low dopamine levels in the patients [12]. Dopamine is cerebrum which act as a connection that sends message to the part of mind that controls developments and coordination. In the brain, there are nerve cells which is called as neurons which are responsible for producing the dopamine. PD basically affects the neurons as a result of which the level of dopamine decreases because of which the unusual action of the mind prompting the indication of Parkinson's, leaves a man unfit to control movements. As dopamine level decreases, the PD progresses. Dopamine level of healthy

person is more than those who are prone to Parkinson's disease. Figure 1.4 shows the dopamine level of healthy person versus PD patient.

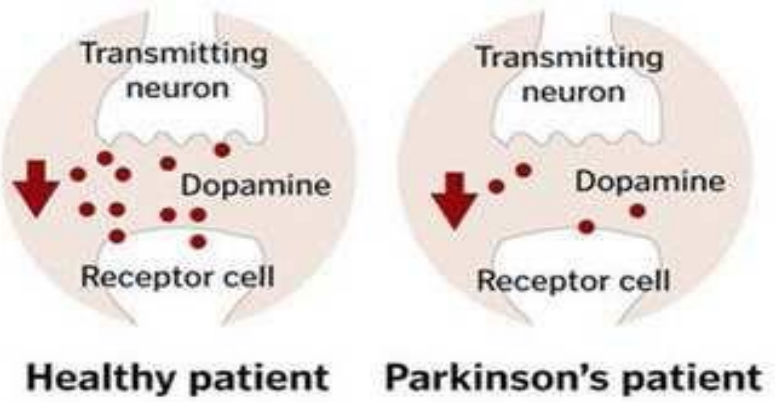

Figure 1.4: Dopamine Level in Patient [31]

To track the progression of Parkinson's disease, the UPDRS score is used [22]. Trained medical staff is required to examine the patient and presence of patient in clinic which is time consuming. [8]. The target for these medical measurement is to find UPRDS. The main motive of the work is to train the different machine learning regression models to achieve the best performance for determining the UPDRS score for analyzing the progression of PD.

The UPDRS that tells the severity and presence of PD symptoms. For untreated patients its span range 0 to 176 with 0 reflects healthy status and 176 reflects the complete disabilities and consist of 3 sections:

- Mentation
- Behaviour and temper
- Motor

1.2 AIMS AND OBJECTIVES

The objectives of the work are summarized as:

- To study, analyze and explore the already existing diagnosis methods for Parkinson's disease and identifying gaps and limitations in the manual procedures and existing diagnosis systems proposed so far.

- To execute the system using 25 Machine learning regression models on publicly available datasets of Voice measures of PD patients so as to predict unified Parkinson's disease rating scale(UPDRS) and evaluate the results using correlation, RSquare, RMSE, Accuracy and Time taken.
- To propose an Ensemble Classifier to Enhance Computer Aided Diagnosis of Parkinson Disease by ensembling the top 5 models.
- To test and cross validate all the results using validation approach.

1.3 IMPORTANCE

Achieving higher accuracy in prediction of UPDRS for PD is very crucial task. PD has several motor symptoms. It is also important to identify PD as soon as possible so that patient can start their treatment early. Detecting Parkinson's disease at early stage is one of the major tasks. Therefore, if that technique used for UPDRS score prediction gives high accuracy then it will be good for all PD patients and also helpful for doctors. There are very few techniques of machine learning used in this area to overcome limitation of detection of PD at early stage so by comparing all techniques will help in estimating the best technique among them for UPDRS score prediction for PD.

1.4 LAYOUT OF THE THESIS

The contents of the thesis are organized as follows:

Chapter 1 Introduces the Parkinson's disease and its symptoms and progression analysis along with the objectives. It outlines the aim and objective of the thesis along with the proposed approach and importance .It concludes with the layout of the thesis.

Chapter 2 provides the review about the existing work in a brief literature survey. Chapter 3 presents the problem statement.

Chapter 4 provides the details about the machine learning models to be used for Parkinson's disease prediction. It reviews the concept of Machine Learning Regression Models along with their types. It also explains the concept of Feature Selection techniques and Ensembling along with the cross-validation concept.

Chapter 5 investigates the performance of Regression Models through evaluation parameters RMSE, Correlation, RSquare and Accuracy. Results are presented in the form

of graphs and tables to evaluate the performance of Machine Learning Models and ensemble Model. Results are cross-validated by K-fold validation Technique.

Chapter 6 discusses together the summary of thesis and the conclusions drawn from the results obtained through model evaluation. It also gives some suggestions for further research that can be carried out in this field.

CHAPTER 2

LITERATURE SURVEY

Keeping in view the importance of Parkinson's disease diagnosis and to study different machine learning models to predict the UPDRS score and its performance, literature review has been done as follows:

2.1 PARKINSON'S DISEASE

Parkinson's disease was first introduced by Doctor James Parkinson as shaking palsy in 1817 [1]. J. William defined the Parkinson's disease as an ailment that influences the piece of your mind that controls how you move your body. It can go ahead so gradually that you don't see it at first. Be that as it may, after some time, what begins as a little instability in your grasp can affect how you walk, talk, rest, and think. L.D. Lau *et al.* introduced that among the elders, Parkinson's disease is usual and it is the 2nd common neurological disease after Alzheimer [2][3].

Nerve cell harm in the cerebrum causes dopamine levels to drop, prompting the manifestations of Parkinson's. Parkinson's regularly begins with a tremor in one hand. Different side effects are moderate development, firmness and loss of adjust. Pharmaceutical can help control the side effects of Parkinson's. Will probably get it when you're 60 and older. It's likewise feasible for it to begin when you're more youthful, yet that doesn't occur so regularly. There's no cure for Parkinson's illness, yet you can get treatment and support to help deal with the side effects. Where it counts in your cerebrum, there's a zone called the substantia nigra. A portion of its cells make dopamine, a connection that carries messages around your brain. For example, when you have kick a ball, dopamine rapidly conveys a message to the nerve cell that controls that movement.

MCD. Rijk *et al.* [3] studied on the Prevalence of Parkinson's disease in Europe based on population and found 6000 approximately in the USA are thought to line with Parkinson's Disease annually diagnosed. A clinical examination were utilized to identify potential PD cases. The general ordinariness (per 100 masses) in individuals 65 years of age and more

settled was 1.8, with a headway from 0.6 for those age 65 to 69 years to 2.6 for those 85 to 89 years. There were no sex separates in commonness of PD.

At the point when that framework is functioning admirably, your body moves easily and equitably. However, the substantia nigral cells begin to die when person has PD. There's no supplanting them, so your dopamine levels reduced and you can't receive the same number of messages to control your body. At an early stage, you won't see anything unique. Be that as it may, as an ever-increasing number of cells pass on, you achieve a tipping point where you begin to have manifestations.

2.2 SYMPTOMS AND CLINICAL FEATURES OF PARKINSON'S DISEASE

J. Jankovic [4] defined the PD is progressive neurological disorder. It is classified by two categories non-motor and motor features. The nearness and particular introduction of these highlights are utilized to separate PD from related parkinsonian issue. A standout amongst the most widely recognized motor issues of Parkinson's sickness is discourse aggravation. M.Politis *et al.* [5] defined the symptoms. The motor symptoms (MS) comprise of the cardinal ternion of bradykinesia, inflexibility. Propelled PD is regularly headstrong to traditional treatments furthermore, muddled by motor and non-motor issues. Non-motor symptoms (NMS) have developed with more prominent criticalness as of late as they have an expansive effect on personal satisfaction measures and societal expenses of PD. These symptoms comprise of state of mind changes, psychological, decay, torment, rest unsettling influence, and autonomic dysfunction. Moreover, NMS have been accounted for to relate with propelling age and seriousness of PD, albeit a portion of these side effects, for example, discouragement, olfactory issues, blockage, and rest issue can happen ahead of schedule throughout the illness, even at a premotor stage. The expanding scope of MS and NMS found in PD has been accounted for as a quantitative measure of commonness, featuring the clinical expansiveness of the illness.

2.3 CAUSE OF PARKINSON'S DISEASE PROGRESSION

J. Lotharius *et al.* [6] provided potential pathogenic mechanism which underlies the loss of dopaminegic neurons in PD which produces chemical named as Dopamine. Dopamine produced by nerve cells present in brain cause the motor symptoms when the dopamine level decreases. Person with Parkinson's disease has less amount of dopamine (which act as a connector that sends message to the brain regarding the muscular activities) is not able to control his body movements.

S. Skodda *et al.* [7] discovered that dysprosody in PD demonstrates trademark changes after some time, yet demonstrate no reasonable relationship with general engine weakness as evaluated by UPDRS engine score. Along these lines, he speculated that the fundamental system could be autonomous from dopaminergic shortages. G.Boka *et al.* [8] suggested that actuated glial cells showed in Parkinson's illness in the substantia nigra may take part in the section of nerve cell passing by giving perilous matter, for example, cytokines. Among these mixes, tumor degradation factor-alpha (TNF) is of enthusiasm since it can initiate cell end. TNF-immunoreactive glial cells which are present in substantia nigra of PD patients were recognized by them. Cornelia Hampe *et al.* [9] depicted the biochemical portrayal of variations conveying amino-corrosive substitutions all through the succession. Transformations in the RING fingers area diminished the solvency of the protein in cleanser and expanded its propensity to shape obvious totals. No transformations considered bargained the authoritative of Parkin via progression of protein accomplices/substrates.

2.4 TELEMONITORING OF PARKINSON'S DISEASE

Neurological disorders, counting (PD), epilepsy, Alzheimer's significantly influence the families and lives of patients. PD influences more than 1 million individuals which live in North Americas. Besides, a maturing populace implies this number is relied upon to ascend as studies recommend quickly expanding commonness rates after the age of 60. Not with standing expanded social disconnection, the budgetary weight of PD is noteworthy and is evaluated to ascend later on. Right now, there is no cure, despite the fact that pharmaceutical is accessible offering huge easing of side effects, particularly at the

beginning periods of the ailment. The vast majority with Parkinson's (PWP) malady will along these lines be significantly subject to clinical mediation.

It demonstrated that around 90 percent of PWP display various type of vocal impairment. Vocal indications that incorporate weakness in the typical generation of vocal sounds (dysphonia) and issues with the ordinary enunciation of discourse (dysarthria). Dysphonic indications regularly incorporate reduced loudness, breathiness, roughness. There are countless and novel estimation techniques for the appraisal of voice issue furthermore, the behaviour of PD-particular dysphonia is genuinely settled, none of the techniques for effectively portraying such dysphonia within the sight of relevant puzzling variables for example, subject sex and very factor acoustic situations. Hence, M.A. Little *et al.* [10] presented another measure of dysphonia named as pitch period entropy, a strong measure delicate to watched changes in discourse particular to PD. UPDRS that shows the nearness and seriousness of side effects. For unattended patients, it traverses the range 0–176, with 0 reflects healthy status and 176 reflects complete disabilities.

Previous research have concentrated on separating PWP from healthy persons. A.Tsanas *et al.* [11] broadened this idea of mapping the seriousness of voice side effects to UPDRS. The promise for telemonitoring of PD depends energetically on the diagram of fundamental tests that can demonstration normally directed remotely. Since the record of talk signals are non-obtrusive and can be promptly incorporated in to telemedicine uses, such tests are extraordinary rivals in such way.

2.5 SPEECH IMPAIRMENT AND VOICE DISORDER

D.Hanson *et al.* [12] proposed relationship of vocal variation from the norm and general neurologic side effects with the laryngoscopic examination which prompts the conclusion that the phonatory irregularities noted in Parkinson's illness are identified with unbending nature in the phonatory stance of the larynx. A.Ho *et al.* [13] categorized speech impairment in two hundred PWP into 5 levels of general seriousness and portrayed the comparing compose (voice, verbalization, familiarity) and degree (appraised on a 5 pt. scale) of impedance for each level.

From 2-min conversational discourse tests, features of voice, familiarity and enunciation were surveyed by 2 prepared raters. Voice was observed to be the main deficiency, as often

as possible influenced and impeded to a more prominent degree than different highlights in these underlying levels. Familiarity deficiencies showed after, articulatory hindrance coordinating voice debilitation in recurrence and degree at the 'Extreme' level. In the last phase of `Profound' impedance, explanation was the most as often as possible hindered include at the least level of execution. A.Ho *et al.* [14] represented the unmistakable quality of voice discourse motor handles shortfalls, and making sync with deficiencies of engine set and engine set unsteadiness in skeletal handles stride and penmanship. Displaying and surrogate information thinks about have indicated noteworthy nonlinear and non-Gaussian irregular attributes in these sounds. M.A.Little *et al.* [15] found that existing apparatuses are restricted to dissecting voices showing close periodicity, and don't represent this inalienable biophysical nonlinearity and non-Gaussian haphazardness, frequently utilizing direct flag preparing techniques harsh to these properties. They don't straightforwardly quantify the two primary biophysical side effects.

Voice issue emerge because of physiological disease or mental issue, mischance, abuse of the voice, or medical procedure influencing the vocal overlays and profoundly affect the patient's life. This impact is considerably more outrageous when the people are proficient voice clients, for example, vocalists, performing artists, radio and TV moderators, for instance. Ordinarily utilized by discourse clinicians. J. A. Logemann *et. al.* [16] noted the frequency of occurrence of speech and voice side effects in PD patients and divide the symptoms into 5 groups.

2.5.1 VOICE CHARACTERISTICS IN PD PATIENTS

R.J. Holmes *et al.* [17] analyzed voice attributes of patients with Parkinson's infection as indicated by malady seriousness. The voice attributes of thirty patients with beginning period PD and thirty patients with later stage PD were contrasted and information from 30 typical control subjects was also collected. In correlation with controls and beforehand distributed standardizing information, both later and early stage voices of PD patients were portrayed perceptually by restricted pitch and din changeability, hoarseness, cruelty and decreased commotion. High modular pitch levels additionally described the voices of guys in both early and later phases of PD. Albeit less understandable, the present information likewise proposed that the voices were described by abundance jitter, a high-

talking essential recurrence for guys and a diminished principal recurrence fluctuation for females. While a few of these voice highlights did not seem to weaken with sickness movement (i.e. brutality, high modular contribute and talking key recurrence guys, essential recurrence inconstancy in females, low force), rasp, monoloudness, monopitch, low din and diminished greatest phonational recurrence run were all more regrettable in the later phases of PD.

2.5.2 VARIABILTY IN FREQUENCY DURING SPEECH

B. Harel *et al.* [18] presented the diagnostics and recovery of Parkinson's disease (PD) that showed the present data relating to novel strategies to assess side effects, restoration, new uses of cerebrum imaging and obtrusive techniques to the investigation of PD. Analysts have just as of late centered around the non-motor side effects of PD, which are ineffectively perceived. The non-motor manifestations of PD significantly affect quiet personal satisfaction and mortality, and incorporate psychological disabilities, autonomic, gastrointestinal, and tactile side effects. In depth dialog of the utilization of imaging devices to consider ailment systems is likewise given, with accentuation on the irregular system association in parkinsonism. Profound mind incitement administration is an outlook changing treatment for PD, fundamental tremor. Ongoing years, new methodologies of early diagnostics, preparing projects and medicines have boundlessly enhanced the lives of individuals with PD, generously diminishing indications and fundamentally postponing incapacity. PD comes about basically from the demise of neurons which is called dopaminergic neurons. Present PD meds treat indications; none stop or retard dopaminergic neuron degeneration. The principle hindrance to creating neuroprotective treatments is a restricted comprehension of the key sub-atomic instruments that incite neurodegeneration. Beforehand involved offenders in PD neurodegeneration, mitochondrial brokenness and oxidative pressure, may likewise act to a limited extent by causing the collection of misfolded proteins, notwithstanding creating different injurious occasions in dopaminergic neurons. Neurotoxin-based models have been vital in explaining the sub-atomic cascade of cell passing in dopaminergic neurons. PD models in view of the control of PD qualities ought to demonstrate profitable in

clarifying critical parts of the illness, for example, particular powerlessness of dopaminergic neurons to the degenerative procedure.

2.6 RATING SCALE FOR DISABILITY IN PARKINSON'S DISEASE

C.Ramaker *et al.* [19] reviewed the clinometric properties of rating scales used for the assessment of PD. He conducted the systematic review of different scales used for the assessment of PD. It is particularly used for motor impairment.He described eleven scales for identifying the PD. It outcomes reliability , responsiveness and validity.Out of these 11 sclaes he evaluated 3 scales named as NUDS(Northwestern University Disability Scale) , UPDRS (Unified Parkinsosn's Disease Rating Scale) and CURS (Columbia University Ratinf Sclae).All these sclaes were used in contrast with the clinical system used for detection of PD , it was noticed that these three scales gave high reliability, validity and accuracy in prediction.From the evidences it was proved trhat all these 3 scales have medium to good validity.

2.7 DIFFERENT APPROACH FOR DIAGNOSIS OF PARKINSON'S DISEASE

Parkinson's ailment is 2nd most general neurodegenerative issue, after Alzheimer's. D.Bazazeh et al. [20] proposed the approach in light of machine learning frameworks. The purpose of machine learning (ML) frameworks has been seen over a wide group of employments in bioinformatics. Biomarkers are described as an objective measure of natural parameters that can break down an ailment, screen its development, or envision medicinal pathologies. Biomarkers keep running from genetic. Biomarker recognizing evidence is a back to back and dreary process that involves various essential advances, consisting data preprocessing, show decision, biomarker endorsement and feature extraction. It contains numerous basic advances, including highlight extraction, information preprocessing, demonstrate choice approval.

2.7.1 CLINICAL DECISION SUPPORT SYSTEM

A.R. Muhammed *et al.* [21] composed the equipment to obtain precise displacement from tri-axial gyroscope and apply a progression of procedures to separate diverse highlights in time and recurrence spaces. A total of one hundred four people presented in our study, Clinical Decision Support System (CDSS) with overall accuracy of 82.43% is created by using this dataset. Moreover, CDSS was likewise utilized as a first demonstrative device in a genuine healing facility setting with a precision of 77.78%.

2.7.2 FEATURE SELECTION

For feature selection A. B. Soliman *et al.* [22] compared the Filter and Wrapper methods. Reducing the number of features lead to more efficient machine learning algorithms. In filter method he applied some statistical approach to rank the features according to its importance and then sort based upon the rank. The features having lowest rank removed from the dataset. In wrapper method, he chose an arrangement of various features and assessed and in addition contrasted every combination with different combinations and afterward utilized prescient model to assess a group of features and assign score in light of model performance. K. Revet *et al.* [23] proposed rough set theory for feature selection. It is a new technique in data mining used to extract the pattern from data. Its basic concept to reduce the data elements from the Decision tree based on the information associated to the particular attribute or feature.

2.7.3 ENSEMBLE

T. G. Dietterich *et al.* [24] proposed the different ensembling methods like error-correcting output coding, Bagging, and boosting. He compared these three methods and gave the conclusion that ensemble methods are better in performance than the individual models. P.Shrivastava et al. [25] proposed neural network model for prediction of PD with feature selection technique Genetic Algorithm and achieve 79.93% accuracy and 93.60 % accuracy by using Neural Network with Binary Bat feature selection technique. H. L. Chen [26] proposed a concept of using the KELM classifier which give accuracy 94.19%. R.Prashanth et al. [27] used Boosted Tree with multimodel feature selection technique and achieve 95.08% accuracy. N. Fayyazifar *et al.* [28] proposed adaBoost and Bagging

algorithms as models to detect PD and obtained 96.55 percent and 98.28 percent accuracy by using AdaBoost and Bagging algorithms.

2.8 COMPARATIVE STUDY

The comparative study of the related literature has been done in this section. All the related techniques applied for PD have been analysed and compared. The summary of previous methods used in literature review is shown in the table 2.1.

Table 2.1: Summary of state-of-the-art literature

Author	Description	Purpose	Results
M. Politis et al. [5]	Used clinical method to identify the symptoms of PD.	To classify the symptoms of PD and common symptoms.	90% people with Parkinson's affected with Motor symptoms.
J. Lotharius, P. Brundin [6]	Used Pathogenic Mechanism.	To identify the level of Dopamine.	PD is progressive, Dopamine level decreases as it progresses.
R. Claudia [19].	Used Scales like CURS, NUDS and UPDRS.	For assesment of motor impairment in PD patients.	UPDRS provides good degree of assessment to identify the progression of PD.
M. A. Raza et al.[21].	Resting and tremour is studied in PD patients.	Provide the diagnostic tool for PD.	Clinical Decision Support system gave 82.43 % accuracy.

A. B. Soliman, M. Fares, M. M. Elhefnawi, M. Al-Hefnawy [22].	Used filter and Wrapper Method.	Feature Selection.	Reduce the dataset.
K. Revett, F. Gorunescu, A. B. Salem [23].	Used Rough Set Theory	Feature Selection	Find out best combination of features.
T. G. Dietterich [24].	Used Bagging, Boosting and Coding.	Ensemble the Models.	Ensembling of models provided better performance than individual learners.
P. Shrivastava, A. Shukla, P. Vepakomma, N. Bhansali, K. Verma [25].	Used Neural network model with genetic algorithm and Binary Bat algorithm.	To diagnose the PD.	Give accuracy 79.93% and 93.60% respectively.
R. Prashanth et al. [26]	Used Boosted Tree with Multimodel Feature selection technique.	To give good accuracy detection of PD.	Accuracy was 95.08.
H. L. Chen [27]	Used KLEM classifier with mRMR filter.	Early Diagnosis of PD.	Accuracy given by this model was 94.19.
Fayyazifar N, Samadiani N [28].	Used Adaboost and Bagging Algorithm.	To reduce the maximum no. of features and give better results by using ensembled model.	Bagging performed well with 98.28% accuracy and Adaboost with 96.55% in PD

			detection from voice measures.

Table 2.2: Comparison of existing methods [28]

Methodology	**Feature Selection Technique**	**Number of selected Features**	**Accuracy (%)**
Neural network	Genetic algorithm	8	79.93%
Neural network	Binary Bat algorithm	6	93.60%
Boosted Tree	multimodal	22	95.08%
KELM classifier	mRMR filter	15	94.19%
AdaBoost	Genetic algorithm	6	96.55
Bagging	Genetic algorithm	7	98.28

The table 2.2 gives the comparison of the exiting methods proposed so far for the detection of PD with comparison of features selected and accuracy. Existing approaches used ensemble model with 1 base learner which did not give stronger results. Proposed approach provide the ensembled model consist of top 5 best models as base learners which gives single stronger overall prediction calculated by the different 5 best models by using minimum no. of features and cross validate the results.

CHAPTER 3

PROBLEM STATEMENT

3.1 PROBLEM DEFINITION

Parkinson's disease is a neurological progressive disorder. It is classified by two types of symptoms:

- Motor symptoms
- Non- Motor Symptoms

Motor symptoms includes the movements of body or muscles like shaking of hands, difficulty in walking, thinking, speech disorder like voice fluctuations or impairment etc. These disorders of PD does not lead to the death of patient directly but its symptoms gets worse the patient's condition as it progresses by time. Unfortunately, no definitive diagnostic test is available. Therefore, creating an automatic system is very important which can detect the disorder in early stages. PWP are having some unpretentious disorders which are not identifiable by listeners directly but useful for detection of PD by utilizing acoustical analysis. Voice is the only feature which gets affected in the initial stage of PD.

Voice recorded after that various important attributes have been extracted to differentiate patients of PD from healthier person. Voice disorder progresses by time, the progression of disease is evaluated by UPDRS scale. UPDRS scale is needed to measure the progression of PD because it quantifies the motor symptoms. UPDRS scale is made bases on the medical observations by the specialist. There is no analytic test for detecting the PD at early stage and to know its progression score. Therefore, it is required to develop a framework which can better predict the UPDRS score so as to know the progression of disease in terms of score at early stage.

3.2 GAP ANALYSIS

- The presently Parkinson's disease detection techniques used less samples.
- Accuracy achieved by these methods is less.

- Cross validation of results was not done in previous works in order to validate the results so as to ensure the reliability of the system.

CHAPTER 4

METHODOLOGY

We propose a method to diagnose the Parkinson's disease by detecting the UPDRS score that only uses dataset of voices of PD patients which are captured at patients home. Total 5875 voices are collected. Then, feature extraction is done to extract the necessary parameters. Various machine learning regression models are applied to these extracted features with 70% of the data from the dataset is used for train the system and 30% of the data from the dataset is used to test the system and then ensembling is done to get better results. Diagrammatic representation of methodology is shown in figure 4.1.

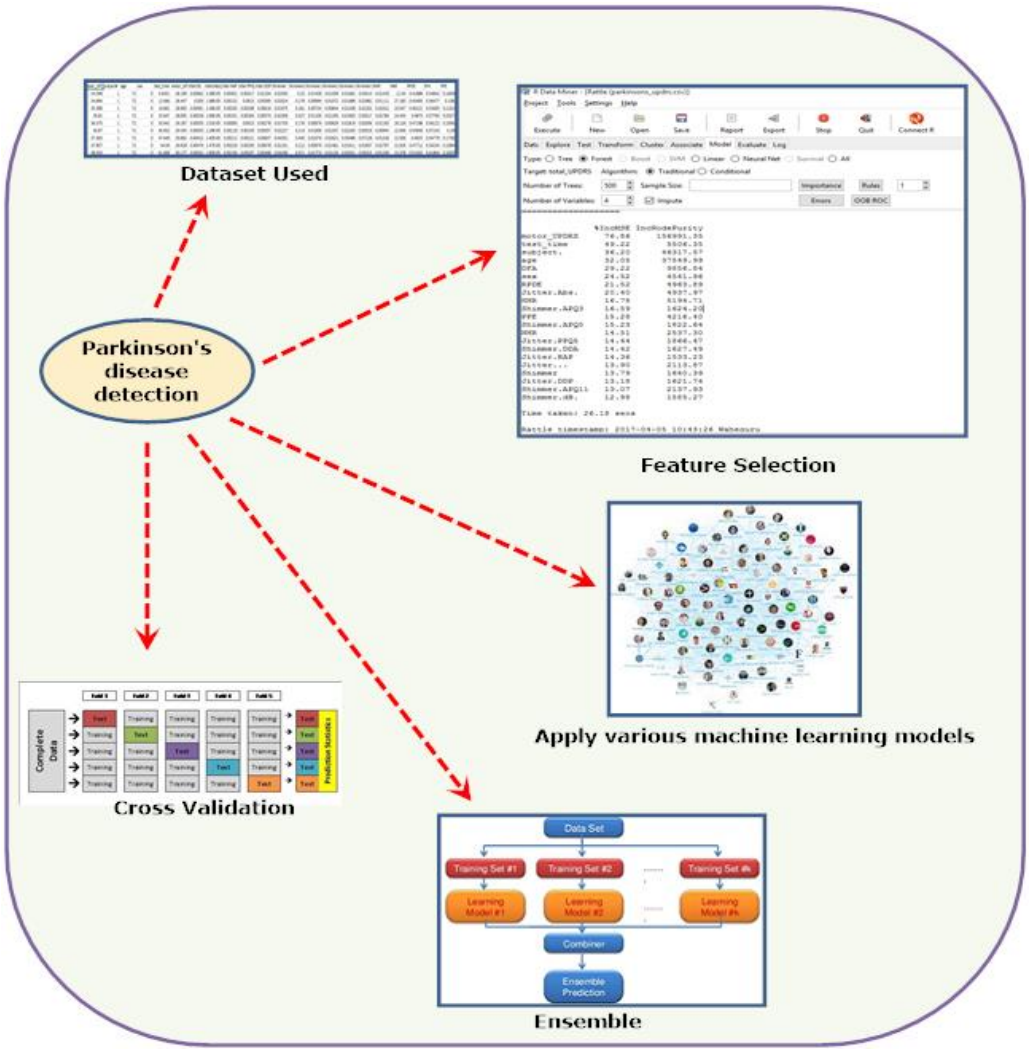

Figure 4.1: Diagrammatic representation of the proposed approach

4.1 PROPOSED METHODOLOGY

The machine learning approach has been used for the prediction of PD. The detailed methodology is described below (as shown in figure 4.2):

- Different 25 regression models of machine learning are applied on the training dataset to predict the results using R Studio. 70% of the data from the dataset is used to train the system and results are predicted by using 30% of the test data.
- Features are selected from the dataset to improve the results using Rattle.
- Executing all the 25 models, top five models with best performance are chosen.
- Ensembling of the top best 5 models is done and the results are evaluated to enhance the results of the resultant prediction.
- Cross validation is then taken into consideration. K-fold validation is a category of cross validation which measures the robustness of the model.
- The results of machine learning models are analyzed using graphs in which the ensembling of the top best 5 models is done and the results are evaluated to enhance the results of the resultant prediction.
- Cross validation is then taken into consideration. K-fold validation is a category of cross validation which measures the robustness of the model.
- The results of machine learning models are analyzed using graphs and comparison of techniques.

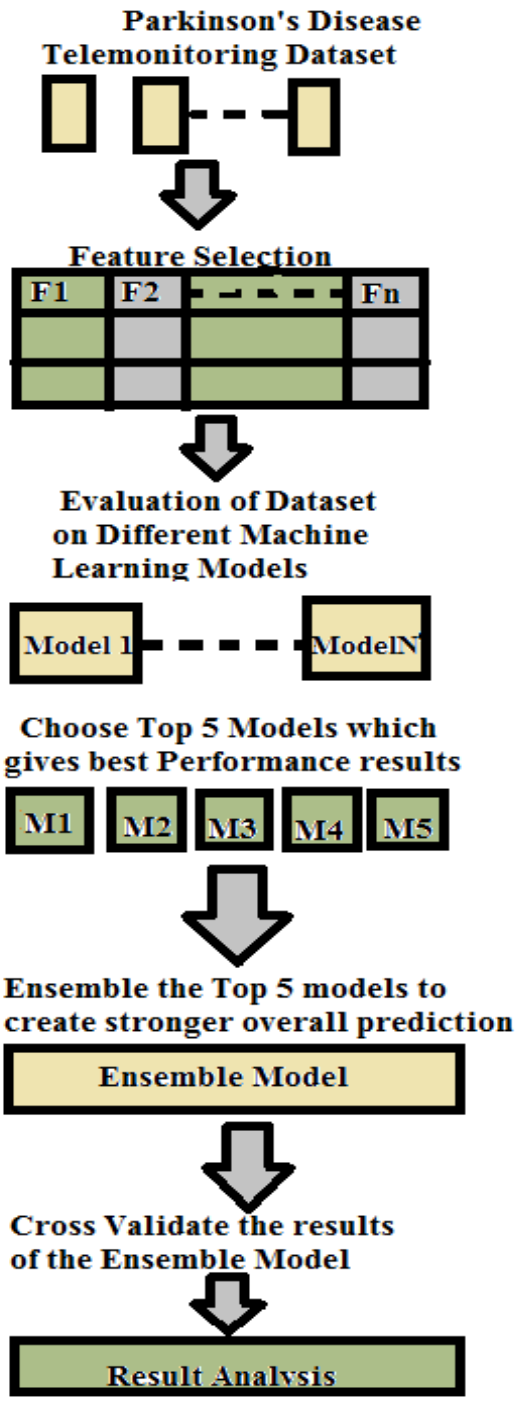

Figure 4.2: Proposed Approach for Parkinson's disease detection

4.2 DATASET DESCRIPTION

This dataset contains the number of biomedical voice measurements. The dataset is collected from 42 persons having early stage Parkinson's disease [8]. The records were captured at patient's home. The dataset was made by Max Little and Athanasios Tsanas of the University of Oxford, as a team with ten medicinal focuses in the US and Intel Corporation who built up the telemonitoring gadget to capture the signals of speech [10]. The dataset comprises of number of traits those are subject gender, subject, age, total UPDRS, time interim from motor UPDRS, basic recruitment date, subject number and 10 measures for biomedical voice. Jitter, Jitter (Abs), Jitter: PPQ5 are various parameters of variation in base frequency [7] [8].

Table 4.1: Dataset Description

Feature Set	Description
Subject	Integer that uniquely identifies each individual
Age	Age of an individual
Test_time	Individual gender '0' shows Male, '1' shows Female
Motor_UPDRS	Time since recruitment into trail. The integer parts the number of days since recruitment.
Total_UPDRS	Clinical's motor UPDRS_score
Jitter(%)	Random variability of vocal vibration, contributes to harsh voice quality.
Jitter(Abs)	Cycle to cycle variation of fundamental frequency i.e. the average absolute difference between two consecutive periods.
Jitter:PPQ5	Five point period perturbation quotient. It is calculated as the average absolute difference between a period and average of it and the four closest neighbours, divided by the average period.
Shimmer:APQ3	Three point Amplitude Perturbation Quotient. It is defined as the average absolute difference between the amplitude of a period and avaerage of the amplitudes of its neighbours, divided by the average amplitude.

Shimmer:APQ5	Five-point Amplitude Perturbation Quotient. It is defined as the average absolute difference between the amplitude of a period and average of the amplitudes of it and its four closest neighbours, divided by the average amplitude.
NHR , HNR	Two measures of ratio of noise to total components in the voice.
RPDE	A non linear dynamical complexity measure.
DFA	Signal fractal scaling component.
PPE	A non linear measure of fundamental frequency variation.

Different parameters of variation in amplitude are: Noise to Harmonic Ratio (NHR), Harmonic-Noise Ratio (HNR) and Personal Protective Equipment (PPE). Total numbers of 5875 voice recordings from individuals were present. The main objective of the dataset is to predict the motor UPDRS score from various voice measures. Features used in this methodology are shown in above table 4.1.

4.3 FEATURE SELECTION

The primary thought of feature selection is, to find out the most reliable features, as they act as an important factor in the whole prediction process. The main purpose of feature selection is to find out the most reliable features, as they act as an important factor in the whole prediction process. The original dataset is shown in below figure 4.3.

total_UPD	subject#	age	sex	test_time	motor_UP	Jitter(%)	Jitter(Abs)	Jitter:RAP	Jitter:PPQ	Jitter:DDP	Shimmer	Shimmer(	Shimmer:	Shimmer:	Shimmer:	Shimmer:	NHR	HNR	RPDE	DFA	PPE
34.398	1	72	0	5.6431	28.199	0.00662	3.38E-05	0.00401	0.00317	0.01204	0.02565	0.23	0.01438	0.01309	0.01662	0.04314	0.01429	21.64	0.41888	0.54842	0.16006
34.894	1	72	0	12.666	28.447	0.003	1.68E-05	0.00132	0.0015	0.00395	0.02024	0.179	0.00994	0.01072	0.01689	0.02982	0.01111	27.183	0.43493	0.56477	0.1081
35.389	1	72	0	19.681	28.695	0.00481	2.46E-05	0.00205	0.00208	0.00616	0.01675	0.181	0.00734	0.00844	0.01458	0.02202	0.02022	23.047	0.46222	0.54405	0.21014
35.81	1	72	0	25.647	28.905	0.00528	2.66E-05	0.00191	0.00264	0.00573	0.02309	0.327	0.01106	0.01265	0.01963	0.03317	0.02784	24.445	0.4873	0.57794	0.33277
36.375	1	72	0	33.642	29.187	0.00335	2.01E-05	0.00093	0.0013	0.00278	0.01703	0.176	0.00679	0.00929	0.01819	0.02036	0.01163	26.126	0.47188	0.56122	0.19361
36.87	1	72	0	40.652	29.435	0.00353	2.29E-05	0.00119	0.00159	0.00357	0.02227	0.214	0.01006	0.01337	0.02263	0.03019	0.00944	22.946	0.53949	0.57243	0.195
37.363	1	72	0	47.649	29.682	0.00422	2.40E-05	0.00212	0.00221	0.00637	0.04352	0.445	0.02376	0.02621	0.03488	0.07128	0.01326	22.506	0.4925	0.54779	0.17563
37.857	1	72	0	54.64	29.928	0.00476	2.47E-05	0.00226	0.00259	0.00678	0.02191	0.212	0.00979	0.01462	0.01911	0.02937	0.02797	22.929	0.47712	0.54234	0.23844
38.353	1	72	0	61.669	30.177	0.00432	2.85E-05	0.00156	0.00207	0.00468	0.04296	0.371	0.01774	0.02134	0.03451	0.05323	0.01338	22.078	0.51563	0.61864	0.20037

Figure 4.3: Original Dataset [11]

Effective feature selection eliminates the redundant variables and keep the best variables which will predict better in the model. Feature selection is important so as to reduce the extra computation stress from the model. Because the less no. of features which are relevant to the target results can give better results in less amount of time. So as the

performance of model will also increase. When presented data is of high dimension, model usually choke because:

- Training time increases exponentially with number of features.
- Models have increasing risk of overfitting with increasing number of features.

Feature selection methods help with these problems by reducing the dimension of data without losing the total information. It also helps to make sense of the features and its importance of the variables that are described as below:

4.3.1 %INCMSE - It is computed from permuting test data: For each tree, the prediction error on test is recorded which is Mean Squared Error (MSE). Then after permuting each predictor variable, the same procedure is done. It is the most informative and robust measure. It is defined as the increase in MSE of prediction as a result of any variable i being permuted. The higher the value of %IncMSE the more important it is. %IncMSE of jth is calculated by using the following equation:

$$\%\text{IncMSE} = ((\text{mse}(j)-\text{mse}(0))\div\text{mse}(0)*100 \quad (1)$$

Table 4.2: Feature selection using %IncMSE

Features	%IncMSE
Motor_UPDRS	76.86
Test_time	49.22
Subject	36.20
Age	32.50
DFA	29.22
Sex	24.52
RPDE	21.52
Jitter.Abs.	20.40
HNR	16.78
Shimmer.APQ3	16.59
PPE	15.28
Shimmer.APQ5	15.23

NHR	14.51
Jitter.PPQ5	14.44
Shimmer.DDA	14.42
Jitter.RAP	14.36
Jitter…	13.90
Shimmer	13.79
Jitter.DDP	13.18
Shimmer.APQ11	13.07
Shimmer.Db.	12.98

4.3.2 INCNODEPURITY - It is the loss function which is chosen by using splits. It is the mse value for regression. More important variables has highest value of node purities. This means that find the split which has small intra node variance and higher inter node variance. Feature selection using IncNodePurity shown in table 4.3.

Table 4.3: Feature selection using IncNodePurity

Features	**IncNodePurity**
Motor_UPDRS	156991.35
Test_time	5506.35
Subject	46317.57
Age	37549.99
DFA	9856.84
Sex	4541.96
RPDE	4963.89
Jitter.Abs.	4937.97
HNR	5194.71
Shimmer.APQ3	1624.20
PPE	4216.40
Shimmer.APQ5	1822.64
NHR	2537.30
Jitter.PPQ5	1866.47

Shimmer.DDA	1627.49
Jitter.RAP	1533.23
Jitter…	2113.87
Shimmer	1640.39
Jitter.DDP	1621.74
Shimmer.APQ11	2137.93
Shimmer.Db.	1585.27

The tables 4.2 and 4.3 of feature selection shows the values for %IncMSE and IncNodePurity for 21 attributes of PD person's voice and sex. Based on these values, the features get reduced by 5 attributes named as Jitter, Shimmer, Jitter.DDP, Shimmer.APQ11, Shimmer.Db. The new dataset after feature selection is shown below in figure 4.4.

total_UPD	subject#	age	sex	test_time	motor_UP	Jitter(Abs	Jitter:RAP	Jitter:PPQ	Shimmer:	Shimmer:	Shimmer:	NHR	HNR	RPDE	DFA	PPE
34.398	1	72	0	5.6431	28.199	3.38E-05	0.00401	0.00317	0.01438	0.01309	0.04314	0.01429	21.64	0.41888	0.54842	0.16006
34.894	1	72	0	12.666	28.447	1.68E-05	0.00132	0.0015	0.00994	0.01072	0.02982	0.011112	27.183	0.43493	0.56477	0.1081
35.389	1	72	0	19.681	28.695	2.46E-05	0.00205	0.00208	0.00734	0.00844	0.02202	0.02022	23.047	0.46222	0.54405	0.21014
35.81	1	72	0	25.647	28.905	2.66E-05	0.00191	0.00264	0.01106	0.01265	0.03317	0.027837	24.445	0.4873	0.57794	0.33277
36.375	1	72	0	33.642	29.187	2.01E-05	0.00093	0.0013	0.00679	0.00929	0.02036	0.011625	26.126	0.47188	0.56122	0.19361
36.87	1	72	0	40.652	29.435	2.29E-05	0.00119	0.00159	0.01006	0.01337	0.03019	0.009438	22.946	0.53949	0.57243	0.195
37.363	1	72	0	47.649	29.682	2.40E-05	0.00212	0.00221	0.02376	0.02621	0.07128	0.01326	22.506	0.4925	0.54779	0.17563
37.857	1	72	0	54.64	29.928	2.47E-05	0.00226	0.00259	0.00979	0.01462	0.02937	0.027969	22.929	0.47712	0.54234	0.23844
38.353	1	72	0	61.669	30.177	2.85E-05	0.00156	0.00207	0.01774	0.02134	0.05323	0.013381	22.078	0.51563	0.61864	0.20037
38.849	1	72	0	68.688	30.424	2.70E-05	0.00258	0.00253	0.0203	0.0197	0.06089	0.018021	22.606	0.50032	0.58673	0.20117

Figure 4.4: New Dataset after Feature Selection

4.4 EVALUATION OF DATASET ON DIFFERENT MACHINE LEARNING MODELS

The datasets are evaluated on various machine learning models and their results are compared based on various parameters.

4.4.1 MACHINE LEARNING REGRESSION MODELS

Regression models falls under the class of supervised machine learning which the subset of machine learning algorithms is. One of the principle essential element in the supervised learning is that the connections between target output variable and input features to predict

the incentive for new information and the model conditions. Regression is the parametric strategy. It is utilized to anticipate consistent (subordinate) variable given an arrangement of autonomous factors. It is of parametric in nature since it takes some specific suspicions in light of the dataset. Regression algorithms predicts the output values in light of the input features from the information fed in the framework to prepare it. There are two types of analysis techniques:

- Single variable: It is used to model the relationship between single input independent variable and an output dependent variable using a linear model i.e. Line.
- Multi variable: It is used to model the relationship between multiple independent variables and an output dependent variable using linear model.

Regression problem requires the prediction of a quantity which holds real valued and discrete input variables. Regression is the method of predicting continuous quantity. Here, the target or output variable in the dataset is total_UPDRS which holds the continuous values act as a dependent variable in this regression analysis. It is multi variable regression problem so the multiple independent input variables in this problem are described in table 4.1. Different machine learning regression models applied on the dataset and the methods as well as packages used by them to predict the total_UPDRS are shown in table 4.4.

Table 4.4: Methods and Packages used by Different Model

Model	**Method**	**Package**
Bagged MARS	bagEarth	Earth
Kknn	Kknn	kknn, caret
randomForest	randomForest	randomForest
projection Pursuit Regression	Ppr	NA
Boosted Generalized Linear	Glmboost	mboost, plyr
Bagged CART	Treebag	caret, ipred, plyr
linearModel	Glm	NA
CART2	rpart1SE	rpart, caret
Least Angle Regression1	Lars	lars, caret
Elasticnet	Enet	elasticnet, caret
Least Angle Regression2	Lars	lars, caret
Relaxed Lasso	Relaxo	relaxo, plyr,caret
neuralNetwork	Nnet	Nnet

Lasso	Lasso	elasticnet, caret
Ridge Regression	Ridge	Elasticnet
decisionTree	Rpart	Rpart
CART3	rpart2	rpart, caret
partial Least Squares1	Kernelpls	pls, caret
partial Least Squares3	Simpls	pls, caret
partial Least Squares2	Pls	pls, caret
CART1	Rpart	rpart, caret
Independent component Regression	Icr	FastICA
BoostedLM	BstLm	bst, plyr, caret
PCA	Pcr	pls, caret
Supervised PCA	Superpc	Superpc

4.4.2 MODEL EVALUATION PARAMETERS: The dataset is evaluated by regression models by calculating the following evaluation parameters of regression:

• **CORRELATION(r):** Linear association between the predicted numeric target value and the actual numeric value is measured by the correlation coefficient. Estimation of the correlation coefficient reliably lie between - 1 and +1. A correlation coefficient of +1 suggests that 2 variables are perfectly related in a positive straight manner, a correlation coefficient of - 1 infers that two components are faultlessly related in a negative direct manner, and an association coefficient of 0 infers that there is no immediate relationship show between the two elements.

The relationship between's two x and y factors are calculated as:

$$Corr(r) = \frac{\sum(x-mean(x))(y-mean(y))}{\sqrt{\sum(x-mean(x))\wedge 2(y-mean(y))\wedge 2}} \quad (2)$$

•**R SQUARE (R):** The Square of the Correlation(r) value can be interpreted as the proportion of the information in the data that is explained by the model.

$$R = (r)^2 \quad (3)$$

• **RMSE:** The Root Mean Square Error (RMSE) metric is defined as a distance measure between the predicted value and the actual value. The smaller the value of the RMSE, the

better is the predictive accuracy of the model. RMSE value 0 means a model has perfect and correct predictions. RMSE is calculated by using equation 4

$$RMSE = \sqrt{\frac{1}{N\sum_{n=1}^{N}(actual-predicted)\wedge 2}} \quad (4)$$

• **ACCURACY:** To calculate the overall match between predicted and actual values, accuracy is used by machine learning models.

• **TOTAL TIME:** The time between the starting of the model and the completion of the model that is, the total time taken by the model in seconds to run successfully.

4.5 ENSEMBLE

Ensemble learning includes consolidating numerous predictions determined by various methods with a specific end goal to create a stronger overall prediction. In this methodology, top five models with highest accuracy are ensembled as shown in figure 4.5.

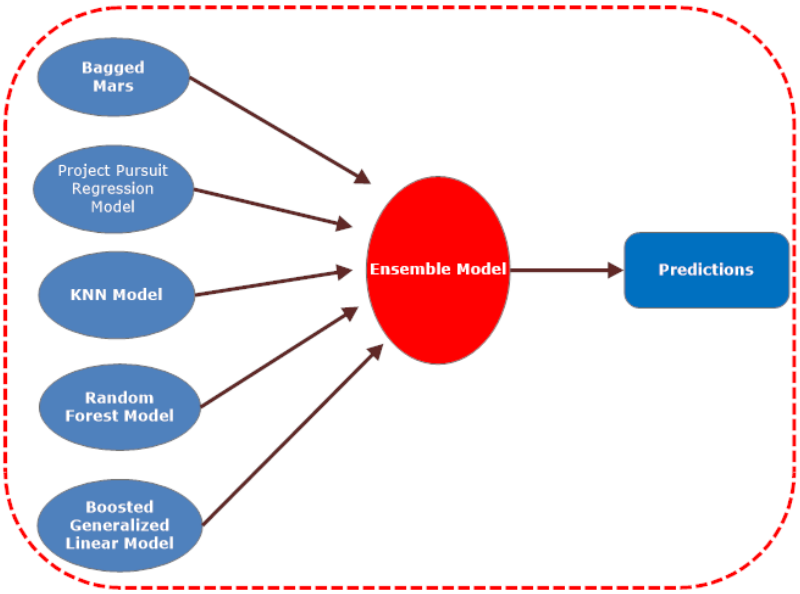

Figure 4.5: Ensembling of Top Models

The prediction of the top models is combined and then the average of the combined predictions is found out. Then evaluation parameters (Correlation, R Square, RMSE, Accuracy) between the actual and ensemble prediction are evaluated. The accuracy of the ensembled model becomes more than the individuals top model's accuracy. In such a way the ensembled model improves the performance and gives the stronger overall prediction results.

4.5.1 DESCRIPTION OF TOP MODELS

The top five models selected based on the performance can be described as below:

- **BAGGED MARS:** Bagged MARS is a type of regression analysis. This analysis given by Jerome H. Friedman in 1991. It is a non-parametric regression method. It can be viewed as an augmentation of linear models that automatically models connections between factors and nonlinearities. The expression "MARS" is trademarked and authorized to Salford Systems. With a specific end goal to maintain a strategic distance from trademark encroachments, numerous open source executions of MARS are called "Earth".

- **k-NEAREST NEIGHBORS MODEL (KNN):** k-nearest neighbors can be utilized for both classification and regression predictive issues. KNN calculation fairs over all parameters of considerations (those are straightforwardness to translate output, calculation time and predictive power). It is usually utilized for its ease of interpretation and low calculation time. KNN algorithm is one of the easiest classification algorithm. KNN algorithm can likewise be utilized for regression issues. The main contrast from the talked about system will utilize averages of nearest neighbors instead of voting from nearest neighbors. K-Nearest Neighbors, or KNN, is a group of classification and regression algorithms in light of similarity (Distance) figuring between occasions. Nearest Neighbor actualizes repetition learning and it depends on a nearby normal computation.

- **RANDOM FOREST MODEL:** Tin Kam Ho introduced the first algorithm for the Random Forests. It is an ensemble learning method in which the subtrees are learned so that the resulting prediction from all sub-trees have less correlation so as to solve the regression, classification and other problems. Random Forests are an improvement over

bagged decision trees. The learning algorithm is permitted to look through all factors and every single variable incentive keeping in mind the end goal to choose the most ideal split point, in CART while choosing the split point. This procedure changed by the Random Forest so that learning algorithms are restricted to an arbitrary example of highlights of which to search. The number of highlights that can be sought at each split point (m) must be indicated as a parameter to the algorithm. One can try different values and tune it using cross validation.

- **PROJECT PURSUIT REGRESSION MODEL:** PPR is a measurable model which is an expansion of added substance models which is a nonparametric relapse technique and utilizations a one-dimensional smoother to fabricate a limited class of nonparametric relapse strategies.

- **BOOSTED GENERALIZED LINEAR MODEL:** The Boosted generalised linear model is an adaptable speculation of customary slightest squares relapse. It sums up straight relapse. By enabling the direct model to be identified with the response variable by means of a connection work it sums up linear regression.

4.6 CROSS VALIDATION

It is the way to estimate, how better the results given by system on new unseen dataset, after it get learned through a given training dataset. It divides the data into k no. of equal sized subsets, out of which union of k-1 subsets used for training while the rest subsets used for evaluation of performance.

A way to estimate how well the results learned from a given training data set is going to generalize on unseen new data. It partitions the data into k number of subsets of equal size and then use the union of k-1 subsets for training while remaining subsets for performance evaluation. The performance of each subset is calculated first then results are averaged to get final evaluation. A mainstream setting of k and for this situation is called as K-fold validation where k is number of training samples. It is also called LOO (Leave-one-out). 8-fold validation is shown in figure 4.6.

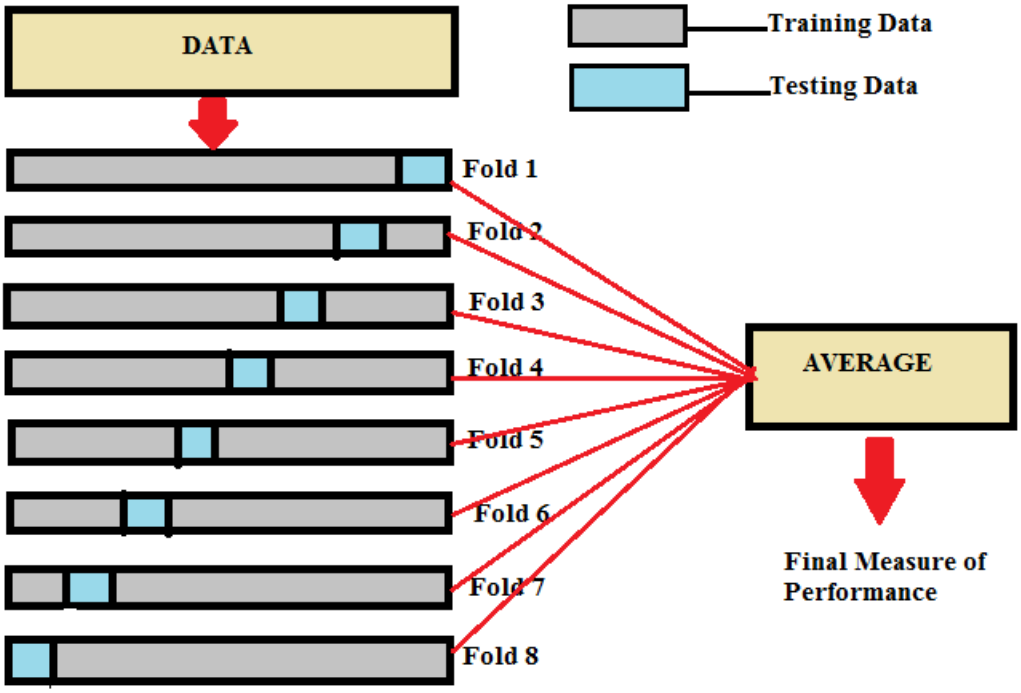

Figure 4.6: K-Fold Cross Validation (Here K=8)

Cross validation technique is used to validate the predictive models and analyse statistical results. It estimates how accurately any predictive model will perform. In this technique the original sample is partitioned into a training set to train the model, and a test set which is used for system evaluation. In this procedure cross validation is utilized to validate the predicted results, in which data get rearranged or shuffled on irregular premise. The objective of the cross validation is to characterize a test dataset which is utilized for testing the framework and it likewise diminishes the issue of overfitting. The dataset is rearranged eight times and the outcomes are cross validated. Cross validation comes about regarding different assessment parameters such as Accuracy and Correlation, R Square and RMSE.

CHAPTER 5

RESULT ANALYSIS

5.1 INTRODUCTION

This chapter deals with various machine learning regression models to predict Parkinson's disease in various individuals. Parkinson's disease is the turmoil of the sensory system which influences the body developments in any person. Parkinson's incorporates the glitch and going of key nerve cells in the brain, called neurons. Parkinson's basically impacts neurons in a zone of the cerebrum. A segment of these lessening neurons convey dopamine, an engineered that transmit messages to the bit of the cerebrum that handles advancement and movements. As PD progresses, the measure of dopamine conveyed in the brain diminishes, left a man unfit to handle body movements.

It causes different indications and signs. These signs and manifestations can be characterized into two classifications: motor and non-motor side effects. Motor side effects influence development of muscles and non-motor side effects incorporate issues like neurobehavioral issues, rest issues, tangible issues. A standout amongst the most widely recognized engine issues of Parkinson's infection is discourse unsettling influence. The dataset is gathered from 42 people having beginning time Parkinson's sickness. The records were caught at patients home. The dataset comprises of number of traits those are subject sexual orientation, date, motor UPDRS, subject number, subject, age, add up to UPDRS, and ten biomedical voice measures. Jitter, Jitter: PPQ5, Jitter (Abs) are different measures of variety in central recurrence. A few measures of variety in adequacy. Add up to quantities of 5875 voice chronicles from patients are taken. The primary target of the dataset is to anticipate the engine UPDRS score from different voice measures.Different machine learning regression models applied on the dataset to evaluate the performance of the models to predict the UPDRS score. The evaluation parameter calculated by the models are correlation, RSquare, RMSE, Accuracy and Time taken. The models are trained by the 70% of the data available and 30% of data used for testing the data. When you run the algorithm over your training data, what you get and what you use to make predictions on new data is called model.

5.2 PERFORMANCE COMPARISON OF DIFFERENT MACHINE LEARNING MODEL

This section covers the performance comparison of various machine learning models used.

5.2.1 TOOLS USED

- Rattle
- Weka
- R Studio

The models are trained on the 70% of the dataset using R Studio. Then after training the models 30% data is passed to them to evaluate the prediction results. The comparison of evaluation parameters of different machine learning regression models is described as below mentioned.

5.2.2. COMPARISON W.R.T CORRELATION (r)

It is also called the coefficient of correlation. It quantifies the degree to which the two variables are related. It ranges between -1 and +1.

Table 5.1: Comparison of Evaluation Parameters of Different Models

Model	R	RMSE	Correlation	Accuracy
Bagged MARS	0.97	1.32	0.98	99.38
Kknn	0.98	0.79	0.99	98.47
randomForest	0.98	1.42	0.99	97.62
projection Pursuit Regression	0.94	1.93	0.97	95.01
Boosted Generalized Linear	0.9	2.33	0.95	88.43
Bagged CART	0.92	2.42	0.96	88.2
linearModel	0.9	2.35	0.95	87.86

CART2	0.9	2.6	0.95	87.52
Least Angle Regression1	0.9	2.37	0.95	87.12
Elasticnet	0.9	2.41	0.95	87.07
Least Angle Regression2	0.9	2.39	0.95	87.07
Relaxed Lasso	0.9	2.45	0.95	87.01
neuralNetwork	0.9	2.36	0.95	86.95
Lasso	0.9	2.39	0.95	86.9
Ridge Regression	0.9	2.42	0.95	86.84
decisionTree	0.88	2.65	0.94	85.08
CART3	0.88	2.87	0.94	84.12
partial Least Squares1	0.88	2.94	0.94	82.42
partial Least Squares3	0.88	2.92	0.94	82.36
partial Least Squares2	0.88	2.96	0.94	81.4
CART1	0.72	4.49	0.85	61.88
Independent component Regression	0.66	4.89	0.81	59.27
BoostedLM	0.88	5.48	0.94	52.64
PCA	0.52	6.04	0.72	49.86
Supervised PCA	0.9	29.18	0.95	0

In the above table 5.1, the values of correlation like 0.98, 0.99 are more closer to the 1 which shows the model predicted values are closely related to the actual observed values of data.

5.2.3 COMPARISON W.R.T R

It is also called Coefficient of Determination. It gives the measure of how well the regression represents the data. Its value lies between 0 and 1 and denotes the strength of the linear relationship between the actual and predicted total_UPDRS values. For example, in table 5.1 comparison of R values 0.97 tells that 97% of total variation in actual can be explained by the relationship between actual and predicted. It shows the strength of the regression equation which is used to predict the total_UPDRS.

5.2.4 COMPARISON W.R.T RMSE

It gives the result of difference between predicted and actually observed values of the model. It defines the error between the data's actual values and predicted values (shown in table 5.1). It tells how close the actual data points are to the predicted data values. From the above table 5.1 values of RMSE, it is observed lesser the value of the RMSE value more will be the accuracy of model.

5.2.5 COMPARISON W.R.T ACCURACY

Accuracy is used to calculate the overall match between actual and predicted total_UPDRS values (shown in table 5.1) given by the model. More the accuracy better the performance of the model to predict the total_UPDRS.

5.2.6 GRAPHICAL REPRESENTATION USING SCATTER PLOTS

It contains the set of points plotted on horizontal and vertical axes. It shows the relationship between the two set of values and find out the correlation between them. The Y-axis shows the actual total_UPDRS and the X-axis shows the predicted value of the total_UPDRS by the models. Each dot in these plots represents the person's actual total_UPDRS value versus their predicted total_UPDRS value. Data points are grouped very close to each other in these scatter plots that indicates the strong +ve correlation such that it represents the linear relationship. The scatter plots of the top 5 models are shown in figure 5.1.

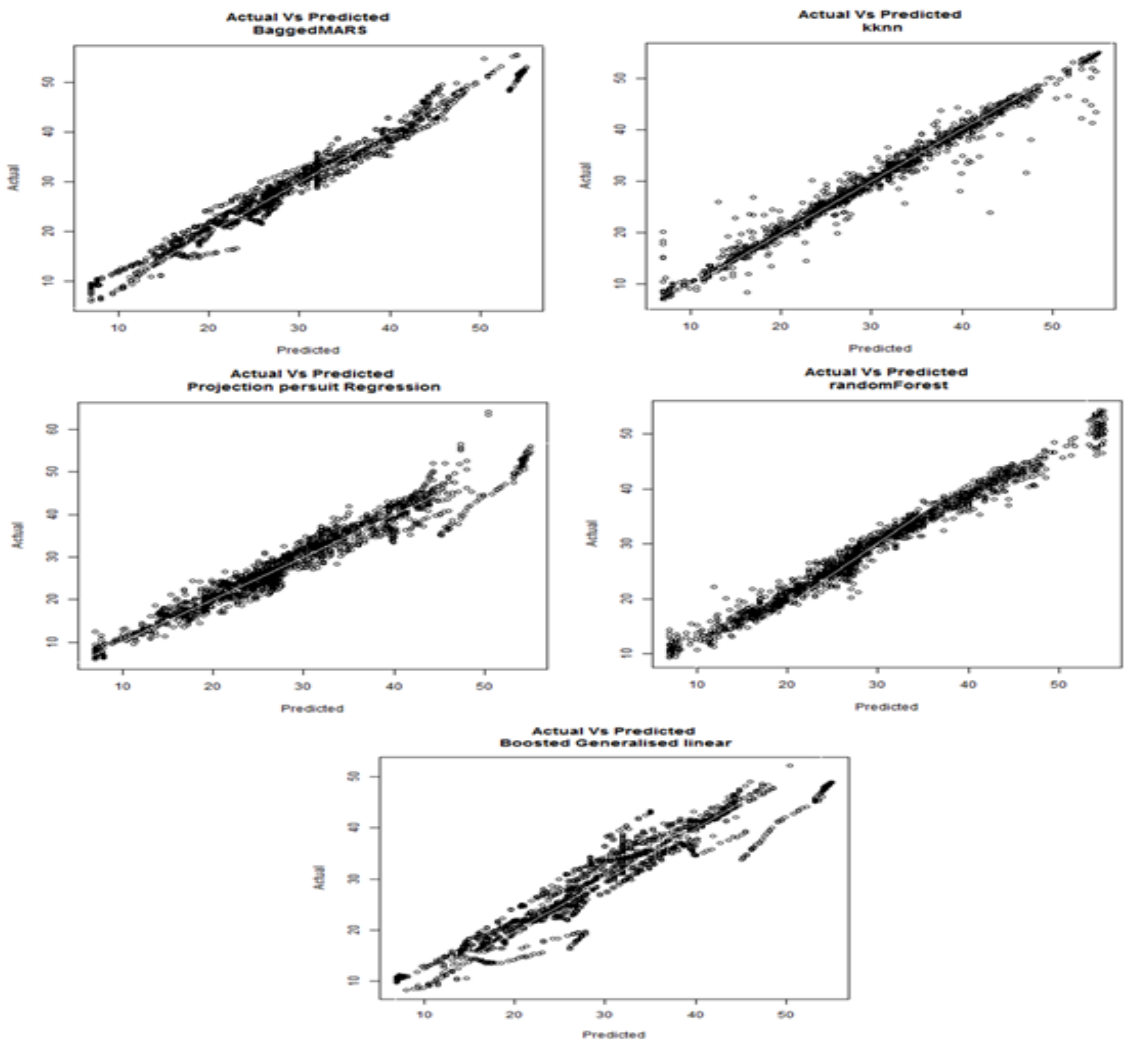

Figure 5.1: Scatter plots of top five models

5.3 ENSEMBLED RESULTS

Ensemble learning involves combining multiple model predictions. It gives better performance than an individual model. In this methodology, top five models with highest accuracy are ensembled and the evaluation parameters are calculated for the ensembled model as shown in table 5.2:

- Bagged MARS
- Kknn
- randomForest

- projection Pursuit Regression
- Boosted Generalized Linear

Table 5.2: Ensembled Model Results

Correlation(r)	**R**	**RMSE**	**Accuracy**
0.99	0.98	1.18	99.6

The correlation signifies the degree of relation, 0.99 is closer to 1 which indicates the models predicted value is in strong relation with the observed actual value. The R value 0.98 shows the 98% of data is closest to the line of best fit. The RMSE shows the error of 1.18 indicates the difference between the actual observed values and the models prediction. The accuracy defines the performance of model to predict the new data point after training and testing which is 99.6%.

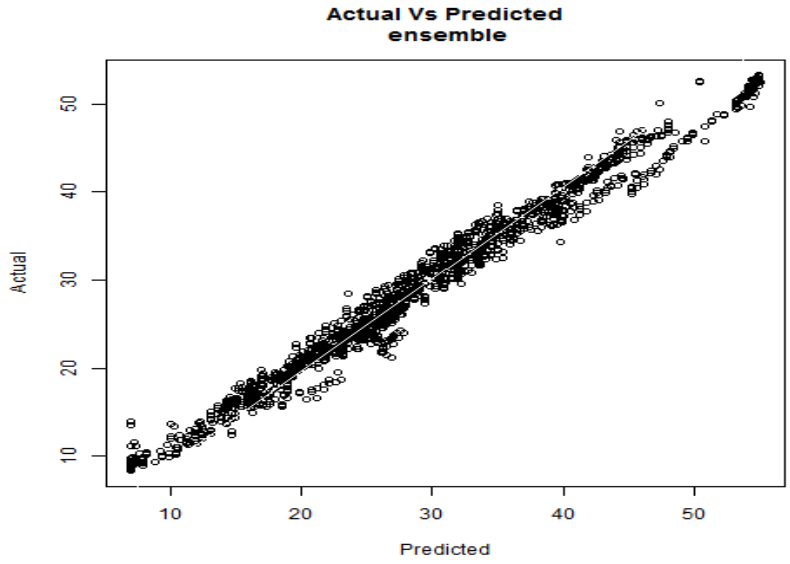

Figure 5.2: Scatter Plot of Ensemble Model

The comparison between the actual values and predicted values of total_UPDRS calculated by the ensemble model is shown in the above scatter plot. Each dot in this plots represents the person's actual total_UPDRS value versus their predicted total_UPDRS value.

5.4 CROSS VALIDATION RESULTS

It is a technique in which original dataset is partitioned into training set to train the model and the test data to evaluate it by the predictive models. In this work the original data set is partitioned into 70% to train the model and 30% to validate the model. Original sample is divided into 8 subset randomly. Out of 8 subset 1 subset is used for testing the data and rest 7 subsets used as tarining data. The following are the tables 5.3 and 5.4 which shows the 8-fold cross validation results for Accuracy, R and Correlation, RMSE respectively.

Table 5.3: 8-Fold Cross-Validation w.r.t Accuracy and R Values

Runs	**R**	**Accuracy**
1	0.98	99.09
2	0.98	99.32
3	0.98	99.48
4	0.98	99.15
5	0.98	99.32
6	0.98	99.90
7	0.98	99.55
8	0.98	99.66

Table 5.4: 8-FoldCross-Validation w.r.t Correlation and RMSE Values

Runs	**Correlation**	**RMSE**
1	0.99	1.32
2	0.99	1.29
3	0.99	1.3
4	0.99	1.27
5	0.99	1.29
6	0.99	1.32
7	0.99	1.29
8	0.99	1.23

The 8 results are then combined to get single estimation by averaging them as shown in Table 5.5.

Table 5.5: Average Estimated Result of 8-fold Cross-Validation

Correlation(r)	**R**	**RMSE**	**Accuracy**
0.99	0.98	1.28	99.43

The advantage of this method is that all observations are used for both training and validation. It helps improve machine learning results by combining multiple models.

5.5 COMPARISON ANALYSIS

The results of research work is compared with Neural Network, Boosted Tree, KELM classifier, Adaboost, Bagging algorithms, on the basis of accuracy and from our results it is seen that proposed method gives better results than all these models. Table 5.6 and Figure 5.3 shows comparison of different models.

Table 5.6: Comparison of Different Models Based on Accuracy

Author Names	Methodology(Model+Feature Selection Technique)	Number of selected Features	Accuracy(%)
P. Shrivastava et al. 1st Method	Neural network + Genetic algorithm[25]	8	79.93%
P. Shrivastava et al. 2nd Method	network + Binary Bat algorithm[25]	6	93.60%
H. L. Chen et al.	Boosted Tree + multimodal[26]	22	95.08%
R. Prashanth et al.	KELM classifier + mRMR filter[27]	15	94.19%
N. Fayyazifar et al. 1st Method	AdaBoost + Genetic algorithm[28]	6	96.55%

N. Fayyazifar et al. 2nd Method	Bagging + Genetic algorithm[28]	7	98.28%
Proposed Method	The Proposed Method (Ensembled Model)	17	99.6%

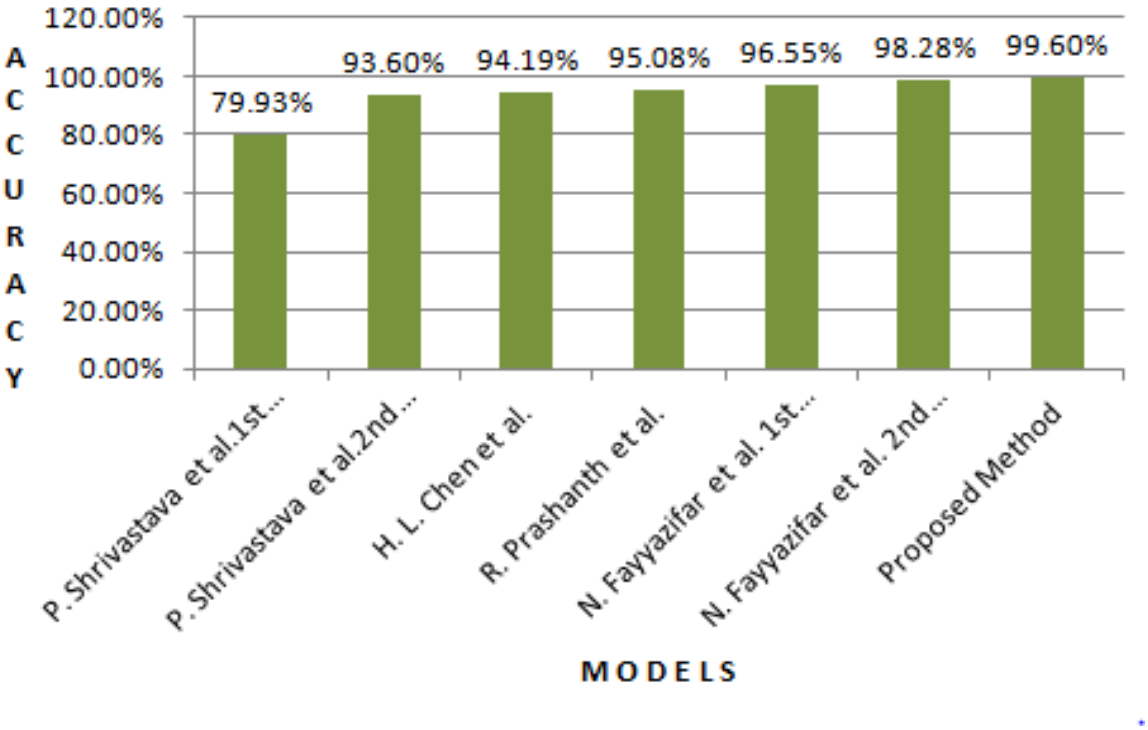

Figure 5.3: Graphical Representaion of Different Models Based on Accuracy

From the table 5.6 and figure 5.3, it can be analysed that the proposed ensemble model outperforms the state-of-the-art techniques.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

This work aims at studying and analyzing the performance of the machine learning regression models for diagnosis of Parkinson's disease by predicting the total_UPDRS. This chapter provides the conclusions drawn from the analysis of the voice dataset of Parkinson's patients and future scope.

6.1 CONCLUSION

Parkinson's disease is a dynamic issue that influences the nerve cells in the mind which produces dopamine. Dopamine level gets reduced as the disease progresses, as a result of which person becomes unable to control its body movements. The voice is most regularly influenced and weakened to more noteworthy degree than some other element in the underlying phase of the Parkinson's illness. The UPDRS scale is utilized for the evaluation of the seriousness of Parkinson's disease side effects. As there are number of features present in the dataset, the feature selection techniques are applied on the dataset to get the important features which are only required for the evaluation. It reduces the overfitting problem.

To avoid the overfitting problem, to enhance the model performance, to reduce the volume of data feature selection techniques are used. %IncMSE and IncNodePurity are used in this approach to select the relevant features from dataset. The system is then trained with 70% of the data set and 30% of the dataset is used for testing. In this the system is executed by using 25 machine learning regression models to evaluate the performance parameters like RMSE, Correlation, R and Accuracy. The results are sorted on the basis of the accuracy of the models. Out of the 25 machine learning models Bagged Mars model results in highest accuracy of 99.38%. The performance of the models Bagged MARS> kknn> randomForest> projection Pursuit Regression> Boosted Generalized Linear as in terms of the accuracy 99.38> 98.47>97.62>95.01>88.43>88.2 respectively.

The top five models with best performance are selected and ensembled together. Bagged MARS, KNN Model, Project Pursuits Projection Model, Random Forest and Boosted Generalized Linear models gives the best results and are ensembled. The ensembled accuracy obtained is 99.6%. After this, all the results of 8-fold cross-validation is then averaged to give single estimation value of 99.4% accuracy.

6.2 FUTURE WORK

As a future work a laboratory is planned to collect data from the individuals affected with Parkinson's disease and healthy persons. The dataset can be collected by using vocal tests from other languages and tested. Progression of dysprosody in Parkinson's disease with overtime can also be predicted from the voice dataset by machine learning methods.

REFERENCES

[1] J. William Langston, "Parkinsons disease: current and future challenges", *Neurotoxicology*, vol. 23, no. 4, pp. 443–450, 2002.

[2] L. ML. De Lau and M. MB. Breteler, "Epidemiology of Parkinson's Disease" , *The Lancet Neurology*, vol. 5, no. 6, pp. 525–535, 2006.

[3] MC. De Rijk, L.J. Launer, K. Berger, MMB. Breteler, JF. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, "Prevalence of parkinson's disease in europe.", *A collaborative study of. Neurology*, vol. 54, no. 5, pp. S21–S23, 2000.

[4] J. Jankovic, "Parkinsons disease: clinical features and diagnosis.", *Journal of Neurology, Neurosurgery & Psychiatry, vol.* 79, no. 4, pp. 368–376, 2008.

[5] M. Politis, K. Wu, S. Molloy, P. G. Bain ,K. Chaudhuri, P. Piccini, "Parkinson's disease symptoms: the patient's perspective." *Movement Disorder*, vol. 15, no. 11, pp. 1646-51, 2010.

[6] J. Lotharius, P. Brundin, "Pathogenesis of Parkinson's disease: dopamine, vesicles and α-synuclein." *Nature Reviews Neuroscience*, vol. 3, no. 12, pp. 932, 2002.

[7] S. Skodda, H. Rinsche, and U. Schlegel, "Progression of dysprosody in Parkinson's disease overtime—A longitudinal study," *Movement Disordor,* vol. 24, no. 5, pp. 716–722, 2009.

[8] G. Boka, P. Anglade, D. Wallach, F. Javoy-Agid, Y. Agid, E.C. Hirsch, "Immunocytochemical analysis of tumor necrosis factor and its receptors in parkinson's disease.",*Neuroscience letters*, vol. 172, no. 1, pp. 151–154, 1994.

[9] C. Hampe, H. Ardila-Osorio, M. Fournier, A. Brice, and O. Corti, "Biochemical analysis of parkinson's disease-causing variants of parkin, an e3 ubiquitin–protein ligase with monoubiquitylation capacity." *Human molecular genetics*, vol. 15, no.13, pp. 2059–2075, 2006.

[10] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, L. O. Ramig, " Suitability of dysphonia measurements for telemonitoring of parkinson's disease." *IEEE transactions on biomedical engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[11] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. "*IEEE transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.

[12] D. Hanson, B. Gerratt, and P. Ward, "Cine graphic observationsof laryngeal function in Parkinson's disease," *Laryngoscope*, vol. 94, pp. 348–353, 1984.

[13] A. Ho, R. Iansek, C. Marigliani, J. Bradshaw, S. Gates, "Speech impairment in a large sample of patients with parkinsons disease." *Behavioural neurology*, vol. 11, no. 3, pp. 131–137, 1999.

[14] A. Ho, R. Iansek, C. Marigliani, J. Bradshaw, S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," Behav. Neurol., vol. 11, pp. 131–37, 1998.

[15] M. A. Little, P. E. McSharry, S. Roberts, D. Costell, I. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection." *BioMedical Engineering OnLine*, vol. 6, no. 1, pp. 23, 2007.

[16] J. A. Logemann, H. B. Fisher, B. Boshes, E. R. Blonsky, "Frequency and coocurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *Speech Hear. Disordor*, vol. 43, pp.47–57, 1978.

[17] R. J. Holmes, J. M. Oates, D. J. Phyland, A. J. Hughes, "Voice characteristics in the progression of Parkinson's disease," *Int. J. Lang. Commun. Disord.*, vol. 35, pp. 407–418, 2000.

[18] B. Harel, M. Cannizzaro, P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain Cogn.*, vol. 56, pp. 24–29, 2004.

[19] R. Claudia, "Systematic evaluation of rating scales for impairment and disability in Parkinson's disease." *Movement Disorders* , vol. 17.5, pp. 867-876, 2002.

[20] D. Bazazeh, R. M. Shubair, W. Q. Malik, " Biomarker discovery and validation for Parkinson's Disease: A machine learning approach.", *Bio-engineering for Smart Technologies (BioSMART), 2016 International Conference*, pp. 1-6, IEEE, 2016.

[21] M. A. Raza, Q. Chaudry , S. M. Zaidi , M. B. Khan, "Clinical decision support system for Parkinson's disease and related movement disorders." *Acoustics, Speech*

and Signal Processing (ICASSP), 2017 IEEE International Conference pp. 1108-1112, IEEE, 2017

[22] A. B. Soliman, M. Fares, M. M. Elhefnawi, M. Al-Hefnawy, " Features selection for building an early diagnosis machine learning model for Parkinson's disease. InArtificial Intelligence and Pattern Recognition (AIPR)." *International Conference on 2016* , pp. 1-4, IEEE, 2016.

[23] K. Revett, F. Gorunescu, A. B. Salem, "Feature selection in Parkinson's disease: A rough sets approach."*Computer Science and Information Technology*, pp. 425-428, IEEE, 2009.

[24] T. G. Dietterich, "Ensemble methods in machine learning", *International workshop on multiple classifier systems,* vol. 33, pp. 1-15, 2000.

[25] P. Shrivastava, A. Shukla, P. Vepakomma, N. Bhansali, K. Verma,"A survey of nature-inspired algorithms for feature selection to identify Parkinson's disease", *Comput Meth Prog Bio*, vol. 139, pp. 171-179, 2017.

[26] H. L. Chen, G. Wang, C. Ma, Z.N. Cai, W. B. Liu, S. J. Wang, "An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson′ s disease", *Neurocomputing*, vol. 184, pp. 131-144, 2016.

[27] R. Prashanth, S.D. Roy, P.K. Mandal, S. Ghosh, "High-Accuracy Detection of Early Parkinson's Disease through Multimodal Features and Machine Learning", *Int J Med Inform,* vol. 90, pp.13-21, 2016.

[28] N. Fayyazifar, N. Samadiani, "Parkinson's disease detection using ensemble techniques and genetic algorithm.", *Artificial Intelligence and Signal Processing Conference (AISP)*, vol. ,pp. 162-165, IEEE, 2017.

[29] http://www.braintrainuk.com/other-conditions-that-neurofeedback-supports/neurofeedback-for-parkinsons/. [accessed on 20 May 2018]

[30] https://www.joinusworld.org/community/3649-parkinson%E2%80%99s-disease/ [accessed on 20 May 2018]

[31] http://www.newhealthadvisor.com/Early-Onset-Parkinson's-Disease.html [accessed on 20 May 2018]

PUBLICATIONS

[1] H. Kaur , A. Malhi , "Ensemble Classifier to Enhance Computer Aided Diagnosis of Parkinson's Disease" , *The 9th International Conference on Computing, Communication and Networking Technologies(ICCCNT), 2018 (Presented).*