Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 9 August, 2021: 6753 – 6771

Research Article

# Parkinson's disease Prediction using Quasi Optimal Optimization Algorithm over Big Data

**Siva Sankara Reddy Donthi Reddy [1]*, Udaya Kumar Ramanadham [2]**
Research Scholar, Department of Computer Science & Engineering, BIHER, Chennai, Tamilanadu, India[1]
Professor, Department of Information Technology, BIHER, Chennai, Tamilanadu, India[2]

## Abstract

Parkinson's Disease (PD) is a neurodegenerative disorder of the central nervous system of people worldwide. It can affect mostly the motor functions. The PD is observed by bradykinesia, rigidity, resting tremor, postural instability, sleeping problems, speech problem, and disordering of the vocal cord at an early stage. The voice disorders the PD patients more than 90%. If the disease is predicted at an early stage, then the doctor can decide to give treatment for increasing the patient's living period. Here, we aim that to predict PD using patient voice recording data set using Big Data Analytics (BDA). In our approach, we propose a disease prediction model that uses machine learning-based classifier algorithms such as Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Neural Network (NN), and Algorithm Quasi (AQ). The result shows an average accuracy of 96.66. The recorded voices of patients are converted to voice parameters like jitter, shimmer, Harmonic to Noise Ratio (HNR), Recurrence Period Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA), Pitch Period Entropy (PPE), and Unified Parkinson's Disease Rating Scale (UPRS) by using R Programming. The status of Parkinson's Disease is found based on testing patient voice data set whether a person has Parkinson's disease or not.

**Keywords:** Parkinson's disease, Motor Function, Big Data Analytics, Logistic Regression, Random Forest, Harmonic to Noise Ratio, Pitch Period Entropy

## 1. Introduction

Parkinson's Disease(PD) is a neurodegenerative of the human brain system over time. It damages brain cells and affects the person's quality of life. The brain cells generate dopamine which is a hormone and it is neurotransmitter [1]. Dopamine is a chemical that is used to send signals to brain cells and it controls movement and coordination. When a person gets PD, the dopamine has degenerated in human brain cells and it is not able to control the movement and activity of muscles. Millions of people are living with PD throughout the world [2]. The PD gets for people who crossed age more than 60 and in 1% of the population.

The reason for PD occurrence is unknown and no cure for this disease but PD patients living period can be increased by giving treatment such as medication and surgery of damaged cells. The nerve cells of the brain may have died periodically and malfunction of nerve cells because of the occurrence of PD [3]. Initially, the PD damages neurons of substantia nigra, and these affected neurons could not produce dopamine which can be used to send information to brain cells. It can control the movement of muscles and coordination. When PD progression grows, the production of dopamine is reduced in brain cells. So, the person is unable to control the movement normally [4].

Using PD patient data analysis, the best practices of treatment, reliable outcomes, and low-cost health care delivery policies are predicted and analyzed in a better way with consideration of technological development in information technology [5]. PD damages a person’s life. So, it is necessary to predict PD in the early stage. If it is identified early, then the treatment will be given in a better way and can be avoided operation [6].

Many PD patients are suffered from speech disturbance which is the most common motor problem. Most of the PD patients are suffered from speech impairment. PD diagnosis is more widespread with speech impairments. Generally, speech disorders are associated with weakness, slowness, or incoordination of the muscles used to produce speech that result from neurologic impairments in PD patients [7]. Speech disturbance occurs in the following ways.

**Hypophonia speech**: PD patients can get soft speech because of weakness in the vocal musculature.

**Monotonic speech**: The speech quality may be soft or hoarse or monotonous.

**Festination speech**: The speech becomes excessively rapid, soft, breathy, and poorly intelligible.

To find out the severity of speech impairment sign, there are two types of best vocal tests for this purpose:

**Sustained phonation:** The patient is asked to say a single vowel while holding its pitch as constant as long as possible.

**Running speech:** The linguistic can show possible impairment signs of vocal disorder when the patient tells a sentence.

The earlier researches had two main issues. i) all the voice samples were considered as single classifiers ii) the statistical metrics were used to summarize the vocal samples of each subject irrespective of the discriminating ability of each vocal test [7]. Only one or a few types of vocal tests of gathered data sets are done in most of the previous PD research. We have concentrated on multiple sound recordings.

There are three main key folds in a research study: i) To apply a unique classifier for vocal samples of each type ii) To omit less discriminating vocal tests and iii) To present more representative vocal tests in our proposed method [7].

The rest of the paper is organized as follows. The previous studies of this domain are reviewed in the literature survey section. In the methods section, dataset description, proposed method and evaluation metrics can be found. The results are presented in the results section. In the discussion section, the demonstration of results is presented. The final section includes the conclusion and future work.

## 2. Literature Review

Marcos L. Carneiro et al. [8] proposed a model which is an automated machine learning system to predict the status of PD and diagnose to find out if a person is having a disease or not. The person is differentiated by considering several key elements such as stride gap, stand path, swing gap, double support intervals. By using a pressure sensor, the footstep calculations are measured and the double support victims are used for footstep calculations. The force-sensitive of the foot is calculated and determined the dynamic position using machine learning algorithms. The Python language was used for predicting PD by using SVM, K- Means, NB, Linear Discriminate Analysis, and Decision Tree. It can diagnose and predict the PD in less time.

Christian Herff et al. [9] introduced the Experience Sampling Method (ESM) model for predicting the beta oscillation of brain waves. By using a motor with regular intervals, the fluctuation is made. The motor fluctuations are partitioned into different states such as ON and OFF. The ON state states the motor treatment is more effective and the OFF is not effective. The deep brain wave is observed with the help of motor fluctuation in graphical representation. The dopaminergic distortion is varied by the fluctuation of the motor. This method is more useful for more researchers for predicting accurate results.

Ancy Carshia S et al. [10] proposed a model to compare patterns of brain waves of Alzheimer's disease and Parkinson's disease. In both diseases, the status of brain fluctuations is found by comparing static and dynamic stages. By comparison, the person's unhealthiness can be predicted. The motor is used to make the brain fluctuations and the ON and OFF states are recorded in databases. The brain wave stages are monitored through implanting the brain network. By using a neural topology system, graphical patterns are obtained. In both diseases, the memory region and motor region are compared. By using ON and OFF states, the two diseases' neural regions are compared. The better results are produced using motor operations such as ON and OFF. It can produce a good health condition state reliably. The motor state fluctuations are predicted using a neural network. This system can give patterns of brain waves continuously.

Igor Škrjanc et al. [11] proposed a learning method for determining the Parkinson’s disease status. This method determines a person's healthiness. Based on clinical data, the analysis is made using machine learning techniques. It can specify the data range of disease and also the brain part activations. Parkinson’s disease is attacked if the neural region is affected in the major area of the brain. Based on beta oscillation, the brain wave pattern is determined. To predict the disease deeply, the nursing data is played a vital role. Already many researchers are done research work on prediction and controlling the disease for increasing the patient’s life period. Machine learning is given more help for effective treatment and for controlling Parkinson’s disease in this research.

Jack W. Judy et al. [12] proposed the Freezing-of-Gait Detection using wearable sensor technology model to detect PD at the earliest. The PD can damage the locomotion of the patient. For observing the patient locomotion status, the gait episodic motor system is used. The algorithm can be learned easily for monitoring the locomotion of patients. The locomotion status of the patient is monitored and it will be updated in the database system for future purposes. In every movement, waveform generates. In this model, the KNN classifier is used for classifying data elements of the data set, and the self-mapping technology method is used to formulate the design of the system. This model produces appropriate results.

Yanan Zhang et al. [13] proposed a classification model which classifies persons who have PD and don’t have. To classify the affected people with PD in this model, machine learning algorithms were used. The main problem of PD is gait disorder. This sort of gait disorder can find by using data clustering technology. The clinicians will suggest PD patients control the disease based on the classification of PD data. This model has realized four types of gait disorders and the average accuracy classification is 85.7%.

Satyabrata Aich et al. [14] proposed a classification model that can help for classifying the people who are affected by Parkinson’s disease and not affected, which is based on voice data set. In this model, two algorithms are used for reduction of the data and collecting the features such as Principal Component

Analysis (PCA) and Genetic Algorithm (GA) respectively. These approaches are used to compare the performances of various metrics. The accuracy is evaluated as 97.57% using SVM with RBF and Genetic Algorithms. This analysis is more useful for clinicians to classify people who are affected with PD and not affected from PD based on voice data set.

As per related work, Parkinson’s disease can be predicted by machine learning techniques over big data. AQ, Random Forest, Decision Tree, Neural network, and SVM show good performance in improving the accuracy of prediction. Generally, there may not be any symptoms at an early stage. So, it is required some advanced prediction techniques to predict the disease at its early stage.

## 3. Methods

### 3.1 Data Set Description

The PD voice dataset is obtained from the UCI machine learning repository system. The data set consists of various attributes like MDVP: Fo(Hz), MDVP: Fhi(Hz), MDVP: Flo(Hz), MDVP: Jitter(%), MDVP: Jitter(Abs), MDVP: RAP, MDVP: PPQ, Jitter: DDP, MDVP: Shimmer, MDVP: Shimmer(dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, Shimmer: DDA, NHR, HNR, status, RPDE, DFA, spread1, spread2, D2, PPE [15]. The attribute id, attribute name, and description of each one are added to table 1.

The PD dataset consists of 195 instances of biomedical voice measurements of 31 people and 23 people are affected with PD out of 31 people. In this data set, each column represents particular voice measure and there are total 195 voice recording of individuals which corresponds in each row. The data set consists of different occurrences of one voice recording in each row and each voice is recorded 6 times. There are 5 to 6 records for 23 different parameters which are opted by every individual. The status column denotes that whether an individual is affected with PD or healthy and ‘0’ for healthy and 1 for PD affected [15]. The classification and supervised learning methods (KNN, DT, SVM, RF, NN, NB, LR, and AQ) are executed on the retrieved voice data set [16].

The various models are compared and analyzed by using the R language. We can easily plot the graphs and visualized the results with the GUI feature of R. R language provides open-source packages which are downloaded and imported as required. These packages are more supported for modeling, plotting, and predicting the results. After preprocessing of the data, 80% of instances are used for training and 20% of instances for testing. The accuracies of various models are summarized with the help of a confusion matrix for each model by executing the test set.

Table 1 List of measurements applied to acoustic signals recorded from patients.

| S.No | Attribute ID | NAME of Attribute | DESCRIPTION of Attribute |
|---|---|---|---|
| 1 | A1 | MDVP: Fo (Hz) | Kay Pentax MDVP Average Vocal Fundamental Frequency |
| 2 | A2 | MDVP: Fhi (Hz) | Kay Pentax MDVP Maximum Vocal Fundamental Frequency |
| 3 | A3 | MDVP: Flo(Hz) | Kay Pentax MDVP Minimum Vocal Fundamental Frequency |
| 4 | A4 | MDVP: Jitter (%) | Kay Pentax MDVP Jitter as Percentage |

| | | | |
|---|---|---|---|
| 5 | A5 | MDVP: Jitter(Abs) | Kay Pentax MDVP Absolute Jitter in Microseconds |
| 6 | A6 | MDVP: RAP | Kay Pentax MDVP Relative Amplitude Perturbation |
| 7 | A7 | MDVP: PPQ | Kay Pentax MDVP 5 Point Period Perturbation Quotient |
| 8 | A8 | Jitter: DDP | Average Absolute Difference of Differences between Cycles, Divided By the Average Period. |
| 9 | A9 | MDVP: Shimmer | Kay Pentax MDVP Local Shimmer |
| 10 | A10 | MDVP: Shimmer(dB) | Kay Pentax MDVP Local Shimmer in Decibels |
| 11 | A11 | Shimmer:APQ3 | 3-Point Amplitude Perturbation Quotient |
| 12 | A12 | Shimmer:APQ5 | 5-Point Amplitude Perturbation Quotient |
| 13 | A13 | MDVP:APQ | Kay Pentax MDVP 11 Point Amplitude Perturbation Quotient |
| 14 | A14 | Shimmer: DDA | The average absolute difference between consecutive differences between the amplitudes of consecutive periods. |
| 15 | A15 | NHR | Noise to Harmonic Ratio |
| 16 | A16 | HNR | Harmonic to Noise Ratio |
| 17 | A17 | RPDE | Recurrence Period Density Entropy |
| 18 | A18 | D2 | Correlation Dimension |
| 19 | A19 | DFA | Detrended Fluctuation Analysis |
| 20 | A20 | Spread1, Spread2 | Non-Linear measures of Fundamental Frequency variation |
| 21 | A21 | PPE | Pitch Period Entropy |
| 22 | A22 | Status | Health status of a subject, 1-Parkinson’s Disease, 0-Healthy |

### 3.2 Overview of Proposed Method

The main aim of this research work is to propose a classification framework using the Hadoop ecosystem for analyzing the vocal test. The ecosystem is designed to big data layered framework as shown in Figure 1. The data set of Parkinson’s disease is collected from the UCI machine learning repository. The clinical test results, oral disease-related information, and symptoms are also collected from clinician labs of Parkinson’s disease patients [17]. By analyzing these data elements, it is provided a more advanced diagnosis for predicting PD in the early stage for better treatment and increase the patient life period, and reduces the treatment cost. The collected data is not in a specific format and produces in terabytes of data. So, the big data storage ecosystem is used to store large sets of data for analysis purposes. It converts from a mixed type of big data into a structured form and provides a process for retrieving actionable insights without missing any information [18].

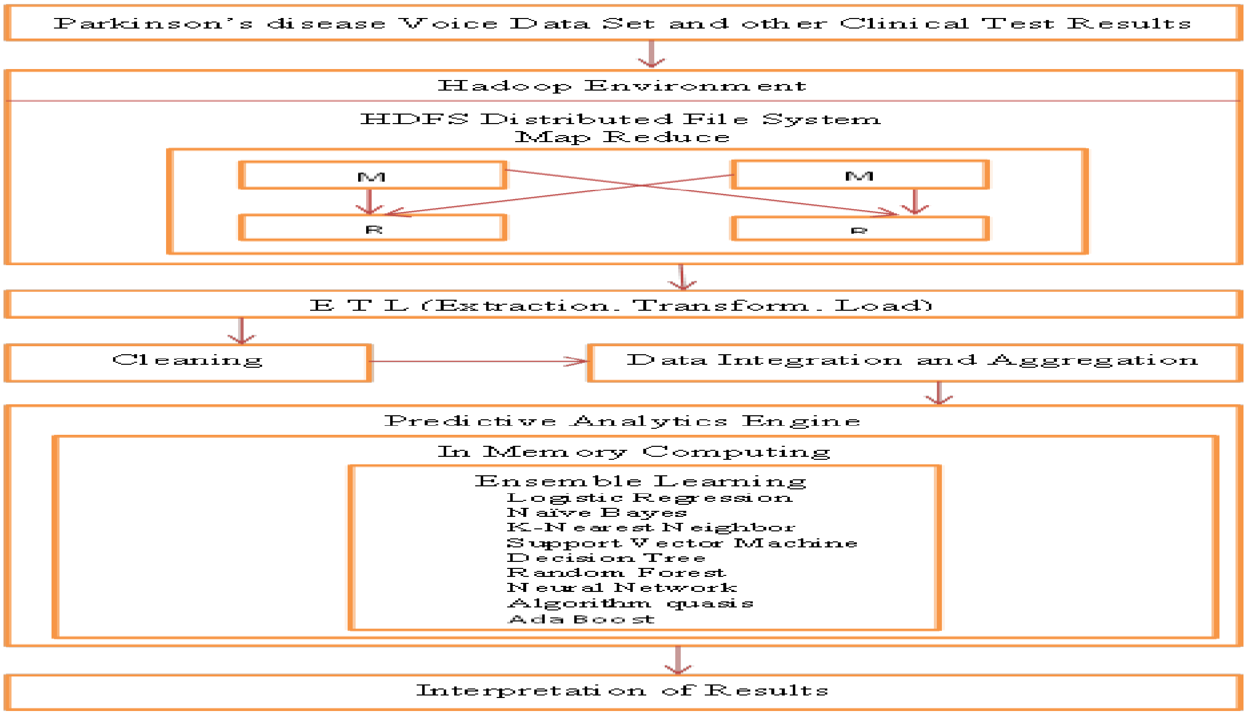
Parkinson's disease Voice Data Set and other Clinical Test Results
Hadoop Environment
HDFS Distributed File System
Map Reduce
M
M
R
R
E T L (Extraction. Transform. Load)
Cleaning
Data Integration and Aggregation
Predictive Analytics Engine
In Memory Computing
Ensemble Learning
Logistic Regression
Naïve Bayes
K-Nearest Neighbor
Support Vector Machine
Decision Tree
Random Forest
Neural Network
Algorithm quasis
Ada Boost
Interpretation of Results

Figure1. The architecture of the Proposed System.

Hadoop is a good layered framework for storing huge data and processing. The Hadoop Distributed File System (HDFS) is a part of the Hadoop platform which is used to store big data and distributed among the nodes based on the necessity of data. The HDFS can process the data reliably with good performance. HDFS is a fault-tolerant system and also a cost-effective one. In HDFS, the big data is divided into small chunks and distributed the chunks on multiple servers. The ETL (Extract, Transform, and Load) technique is accommodated for repeating operations as well as for getting data from source systems quickly [18]. HDFS is provided APIs for MapReduce applications for reading and writing data-parallel. MapReduce can solve data-parallel problems and deals with complex and large datasets with parallel programming approaches. The data is split into multiple small chunks as a map task and it is processed parallel. In this process, each map job can read a set of key, value pairs as input and produces intermediate key and value pairs as reduce task. The JOBTracker and TaskTracker mechanisms are in the MapReduce process for scheduling tasks and monitoring the operations [19].

Big Data Analytics can provide the environment to integrate various analytical techniques for providing better healthcare. To analyze the PD status, predictive analytical techniques are used. In this proposed system, the KNN, Decision Tree, SVM, Random Forest, Neural Network, Naïve Bayes, Logistic Regression, AQ are used for analyzing PD patient's health records. These classifiers experiment on the data sets for achieving the highest accuracy value for the final decision to give the treatment for the patients at the earliest [20].

NoSQL database is supported the In-Memory Computing process technology that is used for accessing data from data servers and store in primary memory for processing. This separates the data elements which are frequently referred to and passed on to the main memory. So, these data elements are accessed from RAM for processing effective and timely manner to achieve better accuracy. It provides a cost-effective treatment.

### 3.2.1 Algorithm quasi-optimal (AQ)

```
Algorithm quasi-optimal (AQ) Algorithm
Initial Requirements: $|R| > 0$ and $|S|>0$
$R' \leftarrow R$; $T=0$
while $|R'| > 1$ do
    select random r from R'
    $k \leftarrow$ STAR(r, S, LEF, maxstar)
    $R' \leftarrow R' - [R' \cap k]$
    $T \leftarrow T + k$
End while
```

**Algorithm: Star Generation Algorithm**

```
$s=0$
for all k in K do
    $s' = t + k$
    $s'' = s' \cup s$
    $s''$ = LEF ($s''$, maxstar)
    if { Mode = PD) then
        if ($q(s'') - q(s) >$ minq) then
            $s= s''$
        end if
    else
        $s= s''$
    end if
end for
```

### 3.2.2 Algorithm: C 4.5

```
Input: P- training dataset, Q - attribute
Output: decision tree
if P=null then
return error statement
end if
if Q = null
return  decision tree with single node with repeated class label in P
End if
set decision tree= {}
for a ∈ Q do
set information(a, P)=0, split information(a, P)=0
compute entropy(a)
for k ∈ values(a, P) do
endfor
endfor
set $P_{a,k}$ is subset of P with attribute a=k
Set $a_{best}$=argmaximum{gain ratio(a, P)}
attach $a_{best}$tree
for k ∈ values($a_{best}$, P)do c
call C4.5($P_{a,k}$)
endfor
return decision tree.
```

The entropy is calculated as follows.

$$\mathrm{entropy(Q)} = \sum_{i=0}^{n} \mathrm{p(Q,i)} * \mathrm{logp(Q,i)}$$

Where n is the no of classes and p(Q, i) is the proportion of instances that are assigned to the $j^{th}$ class.

The Information Gain is defined as

$$\text{gain (Q,P) = entropy (Q) - } \sum_{V \in values(Ps)} \frac{|Ps,k|}{|Ps|} entropy(Q_k)$$

Where $P_s$ - is the set of values of Q in P,

P is the subset of P induced by Q, and Q v P is the subset of P in which attribute Q has a value of k.

Gain ratio is calculated as,

gain ratio(Q,P)= gain(Q,P)/splitinf(Q,P)

Where, splitinf(Q,P) calculated as

$$\text{splitinf (Q,P)} = \sum_{V \in values(Ps)} \frac{|Ps,k|}{|Ps|} \; * \; \log \frac{|Ps,k|}{|Ps|}$$

### 3.2.3 Algorithm: MapReduce Implementation

```
procedure mapattribute (rwid(k1,k2,…))
emit($k_j$ (rwid, m))
end procedure
procedure reduseattribute($k_j$,(rwid, m))
emit($k_j$,(m, count))
end procedure
procedure reducepopulation($k_j$,(m, count))
emit($k_j$, all)
endprocedure
procedure mapcomputation($k_j$,(m, count, all))
compute entropy($k_j$)
compete info($k_j$)=count/allentropy($k_j$)
compete splitinf($k_j$)= -count/allentropy ($k_j$)
emit ($k_j$,(Info($k_j$), splitInfo($k_j$))
endprocedure
procedure reducecomputation(($k_j$,(Info($k_j$),
splitinf($k_j$))
emit($k_j$,gainratio($k_j$))
endprocedure
procedure mapupdatecount(($a_{best}$,(rwid, m))
emit($a_{best}$,(m,count'))
endprocedure
procedure maphash($a_{best}$,(m,count'))
compute nodeid=hash($a_{best}$)
emit (rwid,nodeid)
endprocedure
procedure MAP(($a_{best}$,rwid))
compute nodeid=hash($a_{best}$)
if nodeid is same with the old value then
emit(rwid,nodeid)
endif
add a new subnode
emit (rwid,nodeid,subnodeid)
endprocedure
```

### 3.2.4 Algorithm Quasis Decision Tree (AQDT)

The AQDT is modeled as shown in figure 2:

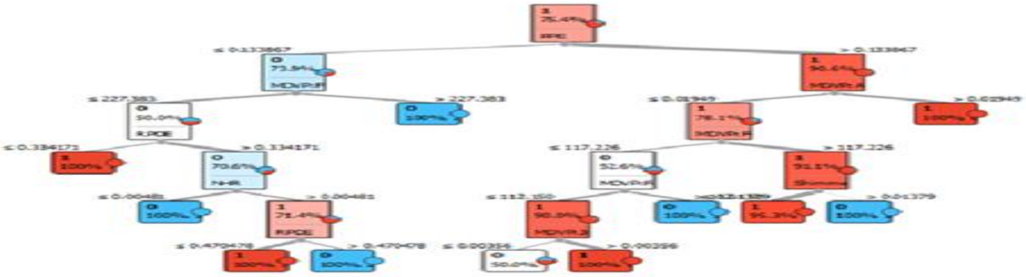

Figure 2 Algorithm Quasis Decision Tree

The AQDT is described the decision-making process. The AQ optimizes the process to make decisions at an early stage.

### 3.3 Performance Metrics

The performance and validations of classifiers are compared by using the following parameters such as Area Under Curve (AUC), Classification Accuracy (CA), F1-Score, Precision, Recall, Specificity, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE).

**Area Under Curve:** It is defined concerning the binary classification problem and is used to find the area under the ROC curve. ROC curve stands for Receiver Operating Characteristic curve which is a visual way of determining the binary classifier's performance. It is the ratio of True Positive rate (also known as recall) and False Positive rate.

$$\text{ROC curve} = \text{True positive rate}/\text{False positive rate}$$

**Classification Accuracy (CA):** Classification Accuracy is referred to as the total no of correct predictions divided by the total number of predictions, multiplied by 100**.**

$$\text{CA} = \text{no of correct predictions} / \text{total no of predictions}$$

**F1-Score:** F1-Score is referred to as the harmonic mean of recall and precision. It is used to measure the accuracy of the test dataset. It can be defined as:

$$\text{F1-score} = \frac{2*(\text{Precision}*\text{Recall})}{(\text{Precision}+\text{Recall})}$$

**Precision:** Precision is referred to as the number of correct positive results divided by the number of positive results which are predicted by the classifier.

$$\text{Precision} = \frac{(\text{True positive}))}{(\text{True Positive} + \text{False Positive})}$$

**Recall:** Recall is referred to as the number of predictions that were relevant in a dataset:

$$\text{Recall} = \frac{(\text{True positive}))}{(\text{True Positive} + \text{False Negative})}$$

**Specificity:** Specificity is referred to as the number of negatives returned by the classification model.

$$\text{Specificity} = \frac{(\text{True Negative}))}{(\text{True Negative} + \text{False Positive})}$$

**Root Mean Square Error:** Root means square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance. It is the result of dividing the square of losses and the total number of examples in the training dataset.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\Pr edicted_i - Actual_i)^2}{N}}$$

**Mean Absolute Error:** Mean Absolute Error is one of the many metrics for summarizing and assessing the quality of a machine learning model. It is used to understand the difference between the average of predicted values and actual values in the training data.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - .\hat{y}_i|$$

## 4. Results

The Hadoop framework is implemented using desktop systems for testing the experiments to measure the performance. The PD status is tested accurately using big data-based predictive analytics and less time is consumed for computation in this model. To implement the multi classifier methods, R analytic tool is used for better performance. The proposed system is used KNN, DT, SVM, RF, NN, NB, LR, and AQ classifiers.

The results are produced by using a confusion matrix. 20% of the original dataset is taken for the test set. By using a confusion matrix, the accuracy is calculated. The best result is obtained with AQ with an Area Under Curve of 97.6%. The least result is obtained with Naïve Bayes which is 89.5%. All the models were implemented using Python. The results are shown in Table 2.

Table 2. Results were obtained by applying different methods and classifiers.

| **Classification Models** | **AUC (%)** | **CA (%)** | **F1-Score (%)** | **Precision (%)** | **Recall (%)** | **Specificity (%)** | **RMSE (%)** | **MAE (%)** |
|---|---|---|---|---|---|---|---|---|
| KNN | 95.3 | 89.2 | 88.8 | 89.0 | 89.2 | 75.4 | 90.6 | 92.7 |
| DT | 96.4 | 95.5 | 98.5 | 96.5 | 94.5 | 96.7 | 89.9 | 92.4 |
| SVM | 95.7 | 90.3 | 89.4 | 91.4 | 90.3 | 70.2 | 94.9 | 89.2 |
| RF | 96.9 | 94.9 | 97.4 | 94.9 | 93.9 | 95.1 | 94.4 | 92.5 |
| NN | 94.4 | 96.4 | 96.3 | 93.5 | 96.4 | 90.4 | 95.7 | 93.4 |
| NBC | 89.5 | 75.9 | 77.5 | 83.8 | 75.9 | 83.7 | 89.5 | 96.4 |
| LR | 90.7 | 86.7 | 86.0 | 86.2 | 86.7 | 69.0 | 95.5 | 90.7 |
| AQ | 97.6 | 95.7 | 89.6 | 95.5 | 94.5 | 96.7 | 97.2 | 89.5 |

Each classifier is tested based on the trained data and predicted good and well-performed classifier on the test data. The ratio of train data is 70:30 for testing test data. So, every classifier can show good performance based on test data. We have generated graphs as shown below by using the R tool.

### A. Comparison of Area Under Curve

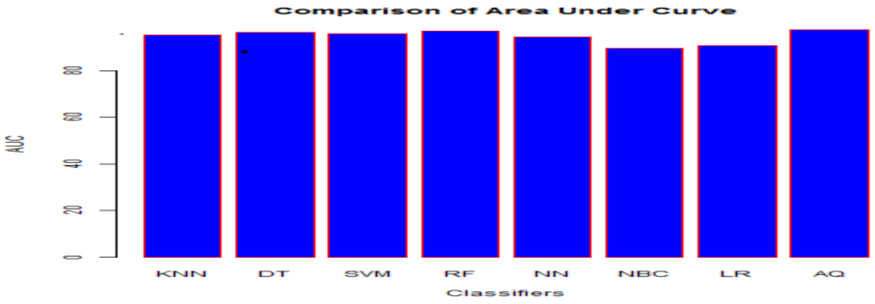

Figure 3 obtained accuracies based on the reported results in table 2.

The comparison of different Parkinson's speech prediction classification algorithms is shown in Figure 3. Various algorithms for machine learning have been used in the prediction of Parkinson's disease. The AUC of the AQ approach is shown good performance over the other ML algorithms. It is improved by 0.7 %.

### B. Comparison of Classification Accuracy

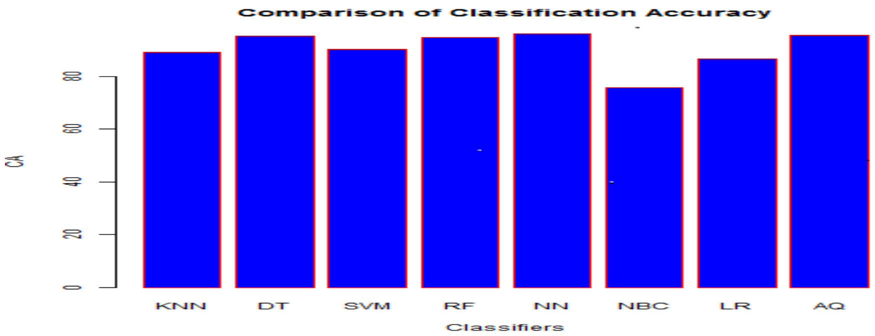

Figure 4 obtained accuracies based on the reported results in table 2.

The various algorithms for machine learning have been used in the prediction of Parkinson's disease. In figure 4, the classification accuracy of the NN approach is shown good performance over the other ML algorithms. It is improved by 0.7 %.

### C. Comparison of F1-Score

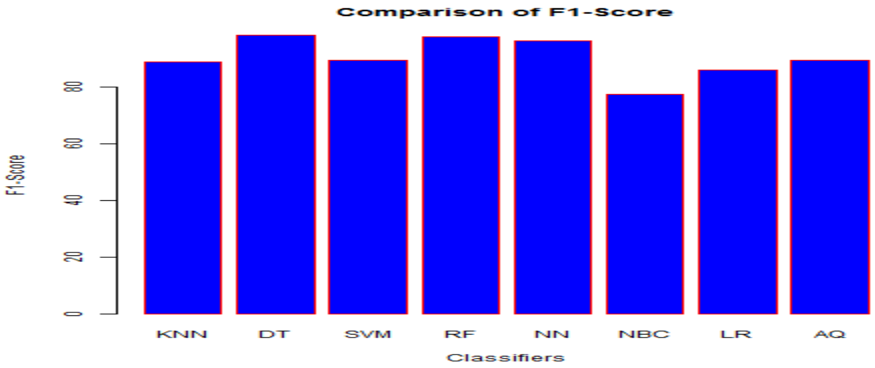

Figure 5 obtained accuracies based on the reported results in table 2.

The comparison of different Parkinson's speech prediction classification algorithms is shown in Figure 5. Figure 5 depicts the F1-Score for DT is shown as 98.5. It is shown good performance than other ML algorithms.

### D. Comparison of Precision

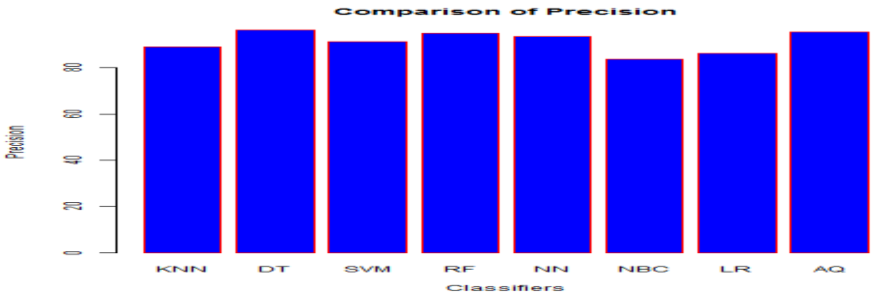

Figure 6 obtained accuracies based on the reported results in table 2.

The various algorithms for machine learning have been used in the prediction of Parkinson's disease. The comparison of different Parkinson's speech prediction classification algorithms is shown in Figure 6. It depicts the precision for DT is shown as 96.5. It is also shown good performance than other ML algorithms.

### E. Comparison of Recall

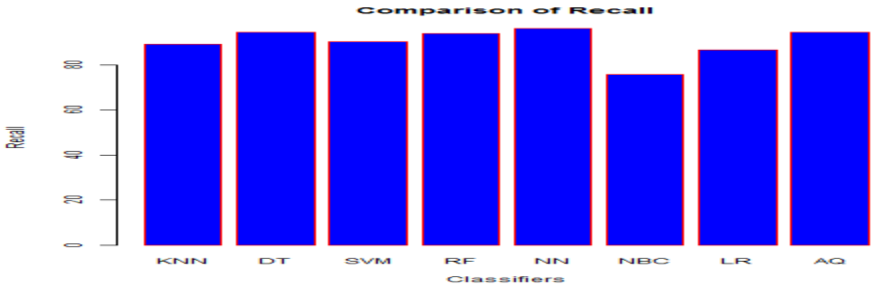

Figure 7 obtained accuracies based on the reported results in table 2.

The comparison of different Parkinson's speech prediction classification algorithms is shown in Figure 7. The recall for NN is shown as 96.4. It is shown good performance than other ML algorithms.

### F. Comparison of Specificity

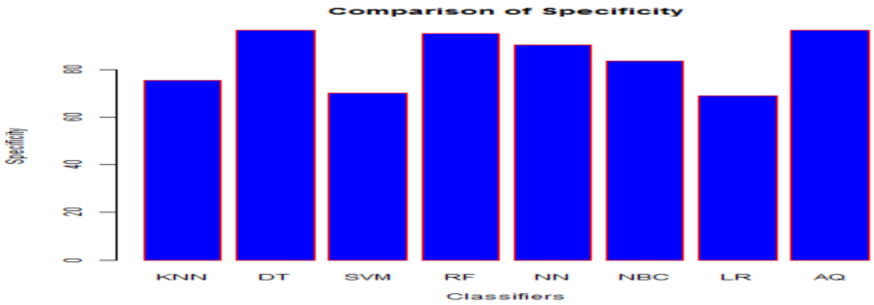

Figure 8 obtained accuracies based on the reported results in table 2.

The various algorithms for machine learning have been used in the prediction of Parkinson's disease. Figure 8 depicts the specificity for AQ is shown as 96.7. It is shown good performance than other ML algorithms.

### G. Comparison of Root Mean Square Error

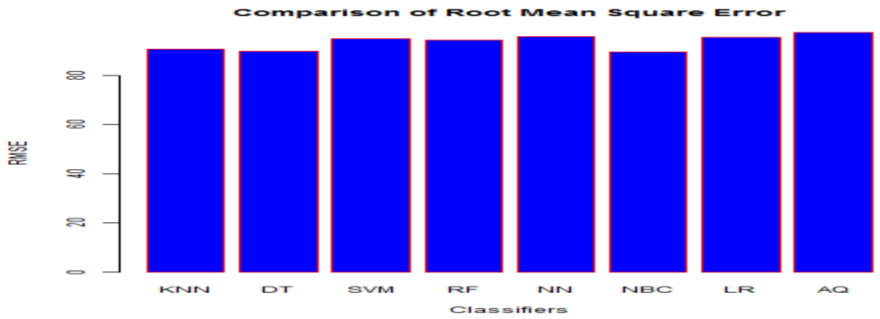

Figure 9 Obtained accuracies based on the reported results in Table 2.

The comparison of different Parkinson's speech prediction classification algorithms is shown in Figure 9. The RMSE for AQ is shown as 97.6. It is shown good performance than other ML algorithms.

## H. Comparison of Mean Absolute Error

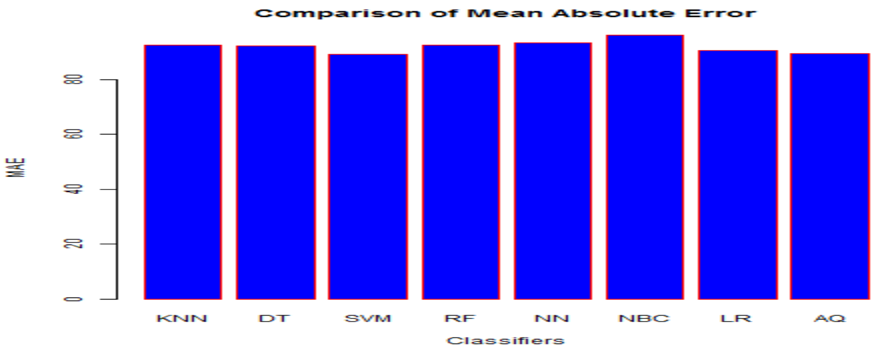

Figure 10 obtained accuracies based on the reported results in table 2.

The various algorithms for machine learning have been used in the prediction of Parkinson's disease. The comparison of different Parkinson's speech prediction classification algorithms is shown in Figure 10. The MAE for NBC is shown as 96.4. It is shown good performance than other ML algorithms.

## 5. Discussion

As shown in the above plots, most of the classifiers have shown good performance in the prediction of PD. In our case, we have seen the AQ has performed in a good manner to compare performance metrics over the voice data set. According to theoretical, we can give suggestions by considering accuracy to choose classifier, but other performance metrics may use for the practical case. In our case, we can suggest strongly AQ is the best classifier for predicting PD in the early stage for deciding to start treatment. However, it may provide a different result with the larger dataset. This result gives an idea about the performance comparison and also gives an impression to analyze more deeply for implementing in practical life.

### 5.1 Limitations:

The limitation of this study is a binary classification for early PD or healthy normal. It does not provide a comparative diagnosis but it is proving a long-term goal. There is another limitation of this study is that sensitivity is lower than the scanning process. In this research, the genetic-based features are not considered for analyzing Parkinson's disease, and also some machine learning algorithms are not applied. There are some limitations for speech which is a single biomarker for clinical diagnosis.

## 6. Conclusion and future work

In healthcare industries, big data analytics play a great role to predict accurate results at the earliest. Generally, healthcare data is complex, huge in size, and various forms. The traditional data analytical tools are not supported to store large sets of data and are not processed accurately to get results because the healthcare data may be in different forms. To achieve cost-effective treatment, greater accuracy, and transparency, powerful big data analytical tools are used to perform analytical operations on larger datasets. In this paper, multi classifiers system is proposed to work on huge PD voice data set and to find new insights, necessaries, show larger variability, improve predictive performance and perform cost-effective actions. This approach provides good opportunities and benefits in productivity, revenue, efficiency, and profitability. It helps a lot for healthcare organizations and clinicians to analyze the large data sets quickly and efficiently for diagnosing the disease. Early prediction of Parkinson's disease is a very important factor and it helps more to give treatment for PD patients in advance to increase the lifetime of PD patients. This system produces 94.31% as average accuracy. In future work, various feature selection and reduction methods will be applied to voice data sets with new features for better results in the big data ecosystem.

## Acknowledgment

I thank Prof. R. Udaya Kumar for giving esteemed guidance and support for the completion of my research work. I can also thank all the authors of various journal research articles that those are referred by me to complete my work.

## Conflict of Interest

The authors do not have any interest to declare conflict in this article.

## References

1. Olga Gavriliuc, Steffen Paschen, Alexandru Andrusca, Christian Schlenstedt, Günther Deuschl, Prediction of the effect of deep brain stimulation on gait freezing of Parkinson's disease, 2021, 87; 82-86.
2. Atiqur Rahman, Sanam Shahla Rizvi, Aurangzeb Khan, Aaqif Afzaal Abbasi, Shafqat Ullah Khan, Tae-Sun Chung, Parkinson's disease diagnosis in cepstral domain using MFCC and dimensionality reduction with SVM classifier, Mobile Information Systems, 2021; 1-10.
3. Sanjukta Krishnagopal, Rainer von Coelln, Lisa M. Shulman, Michelle Girvan, Identifying and predicting Parkinson's disease subtypes through trajectory clustering via bipartite networks, Public Library of Science, 2020; 1-15.
4. Zvezdan Pirtoˇsek,Ovidiu Bajenaru,Norbert Kov´acs, Ivan Milanov, Maja Relja, Matej Skorvanek**,** Update on the management of Parkinson's disease for general neurologists, 2020; 1-13.
5. Abhishek M. S, Chethan C. R, Aditya C. R, Divitha D, Nagaraju T. R, Diagnosis of Parkinson's disorder through speech data using machine learning algorithms, International Journal of Innovative Technology and Exploring Engineering, 2020; 9(3): 69-72.

6. Neharika D Bala, Anusuya S, Machine learning algorithms for detection of Parkinson’s disease using motor symptoms: speech and tremor, International Journal of Recent Technology and Engineering, 2020; 8(6): 47-50.
7. Mahnaz Behroozi and Ashkan Sami, A multiple-classifier framework for Parkinson’s disease detection based on various vocal tests, International Journal of Telemedicine and Applications, 2016; 1-9.
8. Juliana P. Félix, Henrique P. Corrêa, Marcos L. Carneiro, A Parkinson’s disease classification method: an approach using gait dynamics and detrended fluctuation analysis, 2019 IEEE Canadian Conference of Electrical and Computer Engineering, Edmonton, AB, Canada, 2019; 1-4.
9. M. Heijmans, J. Habets, M. Kuijf, P. Kubben, and Christian Herff, Evaluation of Parkinson’s disease at home: predicting tremor from wearable sensors, 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 2019: 584-587.
10. K. A., S. S. Prakash, S. P. and Ancy Carshia S, Investigations on the functional connectivity disruptive patterns of progressive neurodegenerative disorders, 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 2019: 800-803.
11. D. Leite, F. Gomide and Igor Škrjanc, multiobjective optimization of fully autonomous evolving fuzzy granular models, 2019 IEEE International Conference on Fuzzy Systems, New Orleans, LA, USA, 2019: 1-7.
12. P. Tahafchi and Jack W. Judy, Freezing-of-gait detection using wearable sensor technology and possibilistic k-nearest-neighbor algorithm, 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 2019: 4246-4249.
13. Y. Guo, X. Wu, L. Shen, Z. Zhang, and Yanan Zhang, Method of gait disorders in Parkinson's disease classification based on machine learning algorithms, 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, Chongqing, China, 2019: 768-772.
14. Satyabrata Aich, H. Kim, K. younga, K. L. Hui, A. A. Al-Absi and M. Sain, A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of Parkinson’s disease, 2019 21st International Conference on Advanced Communication Technology, PyeongChang Kwangwoon_Do, Korea (South), 2018; 7(3): 1116-1121.
15. N. Shamli, B. Sathiyabhama, Parkinson’s brain disease prediction using big data analytics, International Journal of Information Technology and Computer Science, 2016; 6: 73-84.
16. K.Prasanna, M.Seetha, Efficient and accurate discovery of colossal pattern sequences from biologic al datasets: doubleton pattern mining strategy(DPmine), Elsevier Procedia Computer Science**,** 2015 **;** 54: 412-421.
17. Mahalakshmi Senthilarumugam Veilukandammal, Dr. Sree Nilakanta, Dr. Baskar Ganapathysubramanian, Dr. Vellareddy Anantharam, Dr. Anumantha Kanthasamy, Dr. Auriel A Willette, Big data and Parkinson’s disease: exploration, analyses, and data challenges, proceedings of the 51st Hawaii International Conference on System Sciences, 2018; 2778-2783.

18. Vijayan A, Athithiyan G, Anandaraj K, Brain disease prediction by machine learning over big data from healthcare, International Journal of Contemporary Research in Computer Science and Technology, 2018; 4(3): 32-35.
19. Aarushi Agarwal, Spriha Chandrayan, Sitanshu S Sahu, Prediction of Parkinson's disease using speech signal with extreme learning machine, International Conference on Electrical, Electronics, and Optimization Techniques,2016.
20. K.Prasanna, M.S.P.Kumar, G.Suraya Narayana., A novel benchmark k- means clustering on continuous data, International Journal of Computer Science and Engineering, 2011; 3(8).

## AUTHORS PROFILE

**Siva Sankara Reddy Donthi Reddy** received his ME degree in Computer Science and Engineering from Sathyabama Institute of Technology and Science, Deemed to be University, Chennai in 2007 and he is presently doing his Ph.D. degree in the Department of Computer Science and Engineering from Bharath Institute of Higher Education, BIHER, Chennai. His research interests include Big Data, Cloud Computing, and Cryptography and Network Security.

**Dr. R. Udaya Kumar is a** Professor at the Department of Information Technology, Bharath Institute of Higher Education and Research, Chennai, India. He received Ph.D. from BIHER, Chennai. He has published more than 600 papers and he is supervisor for research scholars at BIHER. His research interests lie in the areas of Cloud Computing, Data Mining, Big Data Analytics, and Information Security.