PARKINSONS DISEASE DETECTION AND MODELING ANALYSIS

Abstract:

Parkinson's disease (PD) is a neurodegenerative disease that affects the neural, behavioural, and physiological systems of the brain. The common symptoms of this disease are a slowness of movement known as 'bradykinesia', loss of automatic movements, speech/writing changes, and difficulty with walking at early stages.

Parkinson's disease (PD) has affected millions of people worldwide and is more prevalent in people, over the age of 50. Even today, with many technologies and advancements, early detection of this disease remains a challenge. This necessitates a need for the machine learning-based automatic approaches that help clinicians to detect this disease accurately in its early stage. Thus, the focus of this research paper is to provide an insightful survey and compare the existing computational intelligence techniques used for PD detection. To save time and increase treatment efficiency, classification has found its place in PD detection.

The dataset for the disease is acquired from UCI, an online repository of large data sets. A comparative study on different classification methods is carried out to this dataset by applying the feature relevance analysis and the accuracy analysis to come up with the best classification rule, also the intention is to sieve the data such that the healthy and people with Parkinson will be correctly classified.

Keywords

Knowledge Data Discovery (KDD), Data Mining, Error Rate, Classification, Classification metrics, Parkinson Disease.

Introduction

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

After decades of exhaustive study, the causes of PD are still unknown. Many of the researchers think that a combination of genetic and environmental factors, such as exposure to the environmental toxin, head injury, rural living, drinking water, manganese and exposure to pesticides, are responsible for PD. These factors may vary from person to person.

In this study, we will analyse the patients' data who are diagnosed with the disease. Using speech data from subjects is expected to help the development of a non-invasive diagnostic. People with Parkinsonism (PWP) suffer from speech impairments like dysphonia (defective use of the voice), hypophonia (reduced volume), monotone (reduced pitch range), and dysarthria (difficulty with articulation of sounds or syllables). Therefore, our analysis in this project will be based on voice parameters of the affected.

Literature Review:

S.No	Title	Author	Website	Year of Publication	Dataset	Classification Algorithms	Classification Metrics
1	Parkinson Disease Classification Using Data Mining Algorithms	-Dr.R.Geetha Ramani (Professor & Head Department of Computer Science Engineering) -G. Sivagami (Master of Engineering Department of Computer Science and Engineering)	Science Direct	9 Oct 2011	UCI, an online repository of large data sets	Binary Logistic Regression C4.5 ID3 C-RT (Classification and regression tree) K-NN (K-nearest neighbour) LDA (Linear discriminant analysis) Random Tree (Rnd Tree) PLS (Partial Least Square Regression) SVM (Support Vector Machine)	**Accuracy:** LDA, C4.5, K-NN:>90% RndTree: 100% **Error Rate:** Binary Logistic Regression-0.1385 C4.5-0.0410 ID3-0.2462 C-RT-0.0462 K-NN-0.0256 LDA-0.0821 Rnd Tree-0 PLS-0.2923 SVM-0.1128
2	Imperative Role of Machine Learning Algorithm for Detection of Parkinson's Disease	-Arti Rana -Ankur Dumka -Rajesh Singh -Manoj Kumar Panda -Neeraj -Priyadarshi -Bhekisipho Twala	MDPI	19 Aug 2022	-Web of Science database -Science Direct -IEEE -Scopus Database	**Supervised ML:** Regression Decision Tree Classification K-NN Naïve bayes SVM Logistic regression **Unsupervised ML:** Clustering PCA K-Means Neural networks Hierarchical Hidden Markov models	**Accuracy:** K-NN-90% SVM-91.4% CNN-88.89%

3.	Comparative Study of Existing Machine Learning Approaches for Parkinson's Disease Detection	-Gunjan Pahuja -T. N. Nagabhushan	IEEE	2018		Artificial Neural Network (ANN) Support Vector Machine (SVM) K-Nearest Neighbour (K-NN) Naïve Bayes (NB) Principal Component Analysis (PCA) Fuzzy K-Nearest Neighbour (FKNN)	**Accuracy:** ANN-92.9% 9 parallel neural networks -91.2% Information gain+ANN -83.33% PCA+FKNN -96.07% PCA+SVM -87.21%

Methodology of the study :

Dataset:

The dataset for the disease is acquired from UCI, an online repository of large data sets. The dataset was created by Athanasios Tsanas and Max Little of the University of Oxford, in collaboration with 10 medical centres in the US and Intel Corporation who developed the tele-monitoring device to record the speech signals. The original study used a range of linear and nonlinear regression methods to predict the clinician's Parkinson's disease symptom score on the UPDRS scale.

This dataset is composed of a range of biomedical voice measurements from various people with early-stage of Parkinson's disease recruited to a six-month trial of a tele-monitoring device for remote symptom progression monitoring. The recordings were automatically captured in the patient's homes.[1]

Attribute Information of Parkinson's Dataset:

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

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

Table1 : Information about attributes in Parkinson's Dataset

DATA VISUALIZATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Different types of visualizations (used in our project):

Box Plot: It is a type of chart that depicts a group of numerical data through their quartiles. It is a simple way to visualize the shape of our data. It makes comparing characteristics of data between categories very easy.

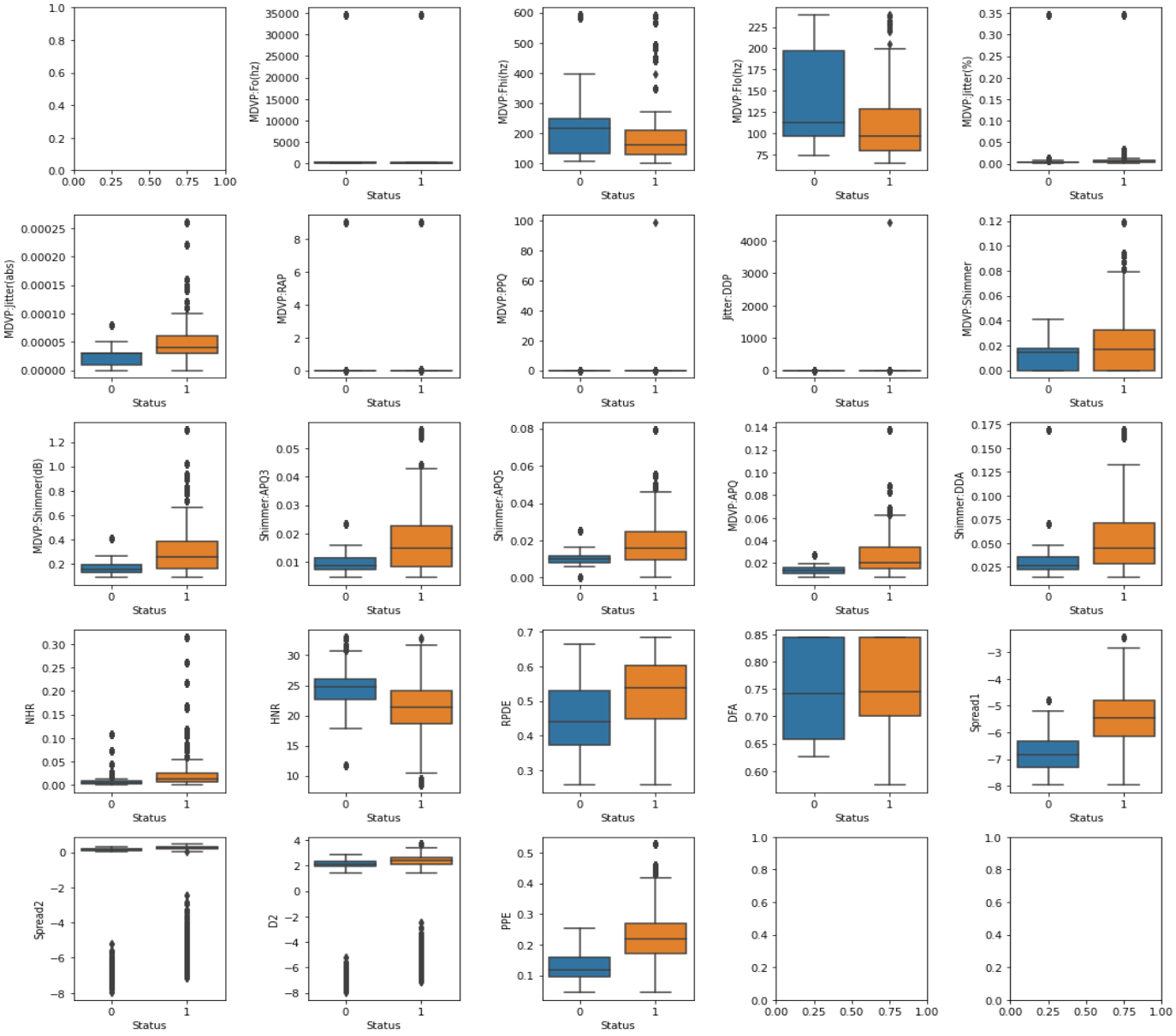
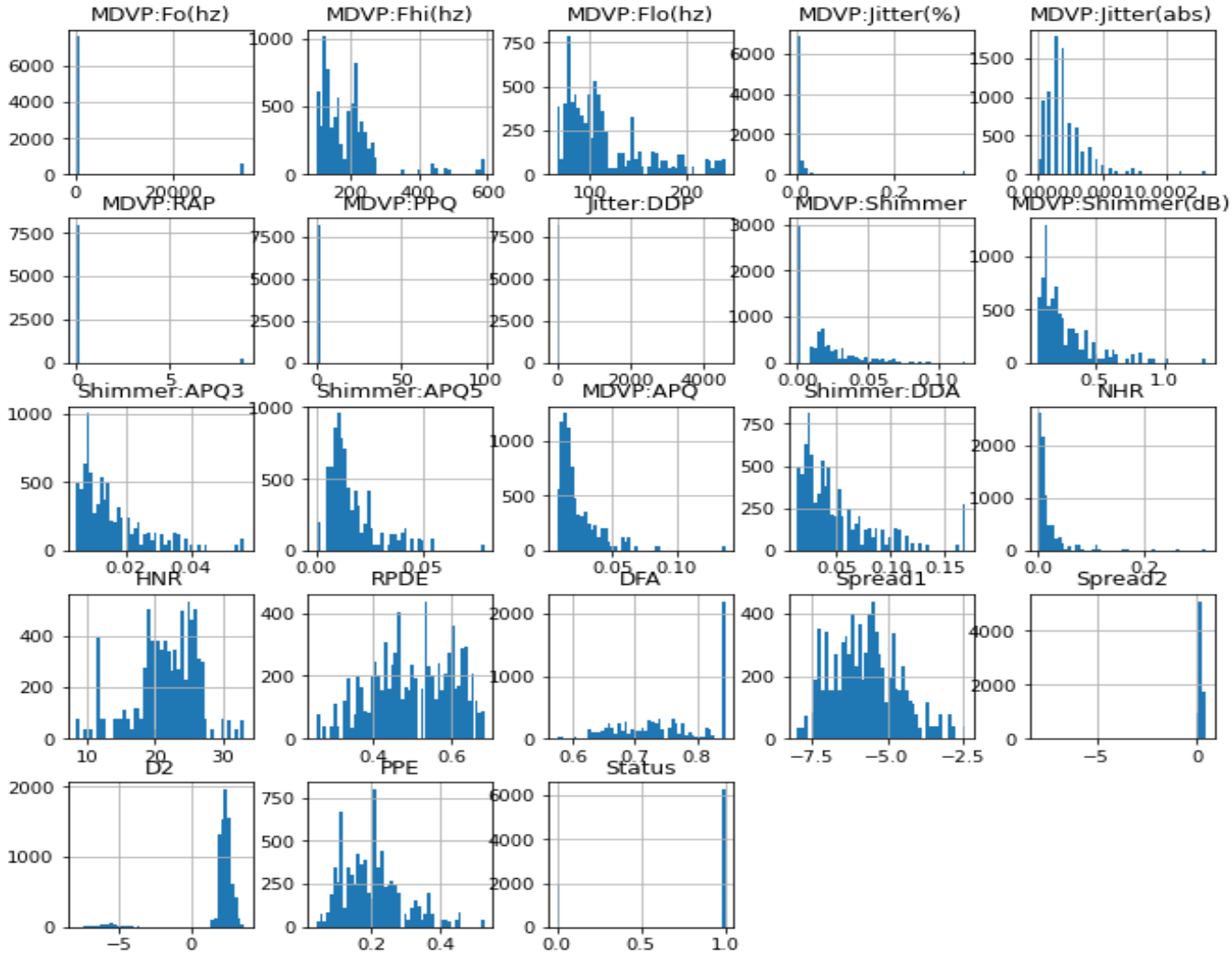
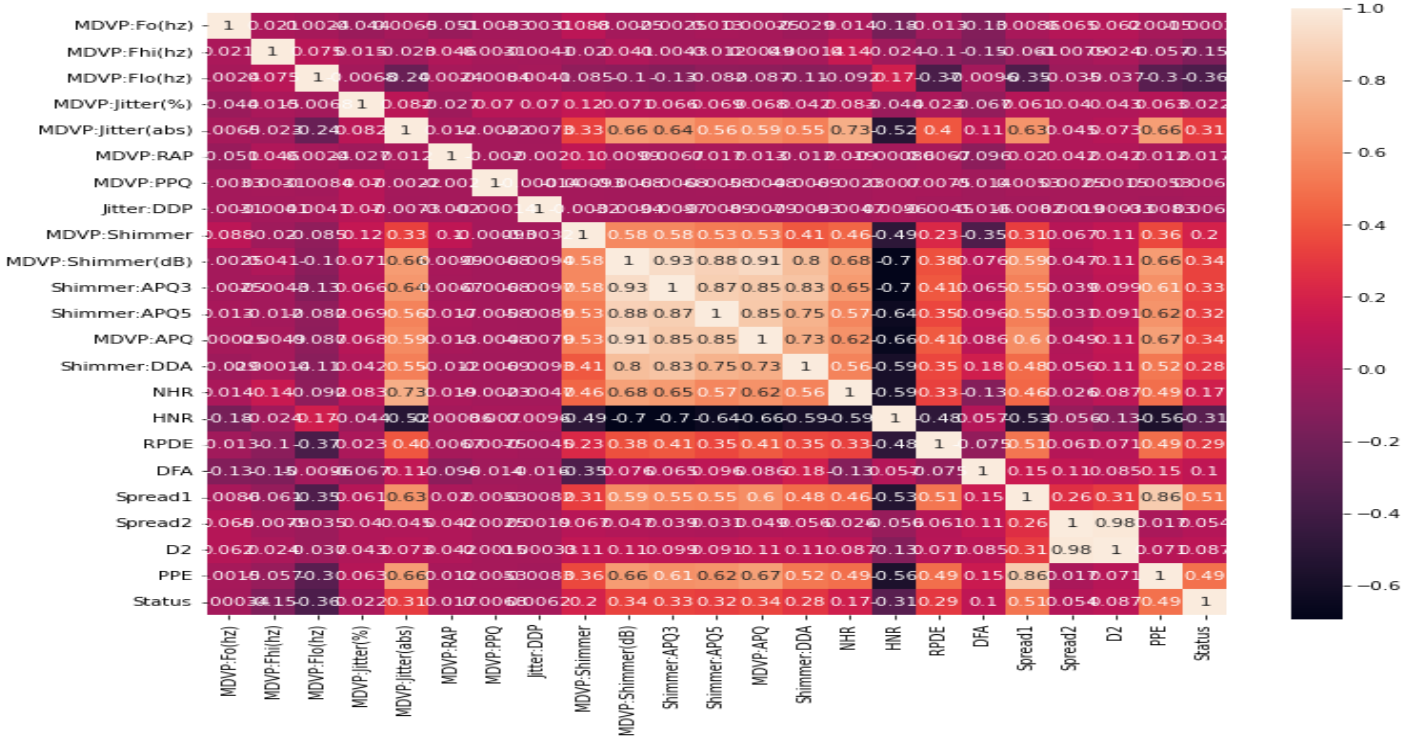


Figure 1: Comparison of Characteristics of data using Box Plot.

HISTOGRAM: A histogram is a chart that displays numeric data in ranges, where each bar represents how frequently numbers fall into a particular range. Like a bar chart, histograms consist of a series of vertical bars along the x-axis. Histograms are most commonly used to depict what a set of data looks like in aggregate. At a quick glance, histograms tell whether a dataset has values that are clustered around a small number of ranges or are more spread out.



Heatmap: Heatmap visualization or heatmap data visualization is a method of graphically representing numerical data where the value of each data point is indicated using colors. The most commonly used color scheme used in heatmap visualization is the warm-to-cool color scheme, with the warm colors representing high-value data points and the cool colors representing low-value data points.



DATA PRE-PROCESSING:

- Our first step is going through the dataset and identify any missing value to take necessary measures. This step is essential
- to prepare the data for fruitful analysis. There are no missing values in our dataset.
- A glance at the data and we realized that it may have duplicate observations. We have to remove the duplicates as this will
- increase the likelihood of our model to overfit.
- Our data set contains 49920 tuples with 24 attributes. After removing duplicates our data is of (8223).
- As Data type of all attributes is Float and Status is number. So, there is no need of conversion from Categorical to Numerical.
- Our next step is dimensionality reduction. The dataset is very large with 24 variables and some of the variables have high
- correlations between them, so we are expected to reduce the number of dimensions for better interpretation of the data.

As our dataset contains 1st attribute as Name, it doesn't affect our results. So, we dropped name column.

Normalization:

Min-Max scaling was used for data reduction. Min-max normalization performs a linear

transformation on the original data. This technique gets all the scaled data in the range (0,1).

Data splitting:

Data splitting is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model.

- We have considered approximately 50k (49,920) instances or tuples in our data set, before pre-processing, And nearly 8k (8223) instances after pre-processing.
- This whole data set cannot be used for just training. Hence it is being split into two parts, one for training and the other for testing the model.
- Here we have followed the pattern of 65-35% splitting fashion where 65% (5,334 tuples) for training and other 35% (2,879 tuples) for testing.
- Python libraries were used to split the data.

DATA CLASSIFICATION:

An overview of the Algorithms used for the classification of Parkinson's dataset are discussed here.

K- Nearest Neighbour (K-NN):

K-Nearest Neighbour is one of the simplest classification algorithms we have used to analyse the model.
The k-nearest neighbour algorithm classifies objects on closest training examples in the feature space.
KNN is a lazy learning algorithm where the function is only approximated locally and all computation is delayed until classification. It does not need any training data points for model generation. All training data used in the testing phase. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. The model structure is determined from the dataset. This is very helpful for real-world datasets where mathematical theoretical assumptions are not followed.
An object is classified by a majority vote of its neighbours, with the object being assigned to the class that is most common amongst its k nearest neighbours (k is a positive integer, typically small and >= 5). If k = 1, then the object is simply assigned to the class of its nearest neighbour. The neighbours are taken from a set of objects for which the correct classification is known.

Naïve Bayes:

One of the classification algorithms we have implemented here is naïve bayes algorithm.

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for Bayes' theorem is given as:

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

Where, P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence

Decision Tree

In decision tree classification we have used 2 methods.

1.Entropy method

2. Gini Index

- **Entropy method**

To find information gain of an attribute first we have to find ENTROPY of whole dataset and the entropy of individual values

of the attribute. Entropy is the measurement of impurities or randomness in the data points.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Where pi is the proportion of a label.

- **Gini Index**

Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly

when selected randomly.

While designing the decision tree, the features possessing the least value of the Gini Index would get preferred.

The Gini of a dataset is

$$\text{Gini} = 1 - (\sum p_i^2)$$

Where pi is the proportion of a label.

Experimental analysis:

Table 1 shows the performance comparison of Gaussian NB, Decision Tree and K-NN with different variants on PD voice dataset. It is clear from Table that K-NN with K=1, K-NN with K=15 outperformed the other variants that are investigated in this study.

Table 2: Performance comparison of Gaussian NB, KNN and Decision Tree on PD voice dataset

Algorithm name	KNN				Gaussian Naïve Bayes		Decision Tree			
Metrics	K=1		K=15		0	1	Gini Index		Entropy	
	0	1	0	1			0	1	0	1
Precision	0.93	0.98	0.92	0.98	0.93	0.96	0.85	0.95	0.91	0.90
Recall	0.93	0.98	0.93	0.98	0.87	0.98	0.83	0.96	0.81	0.98
F1-Score	0.93	0.98	0.93	0.98	0.90	0.97	0.84	0.95	0.84	0.94
Support	675	2204	675	2204	675	2204	675	2204	2879	2204
Accuracy	96.5266				76.1723		92.7058		90.0312	

For comparison purpose, classification accuracies of the previous methods which were investigated on for PD diagnosis using voice data are listed in Table 2.

Table 3: Classifier performance comparison with studies available in the literature on vocal dataset

Method	Accuracy
ANN	92.9
9 parallel neural networks	91.2
Information Gain+ ANN	83.33
PCA+FKNN	96.07
PCA+SVM	87.21

Confusion matrix:

K-NN:

K=1: [[629 46]
 [52 2152]]

K=15: [[587 88]
 [46 2158]]

Naïve Bayes:

[[368 307]
 [490 1714]]

Decision Tree:

➔ Gini index:

[[560 115]
 [95 2109]]

➔ Entropy:

[[431 244]
 [43 2161]]

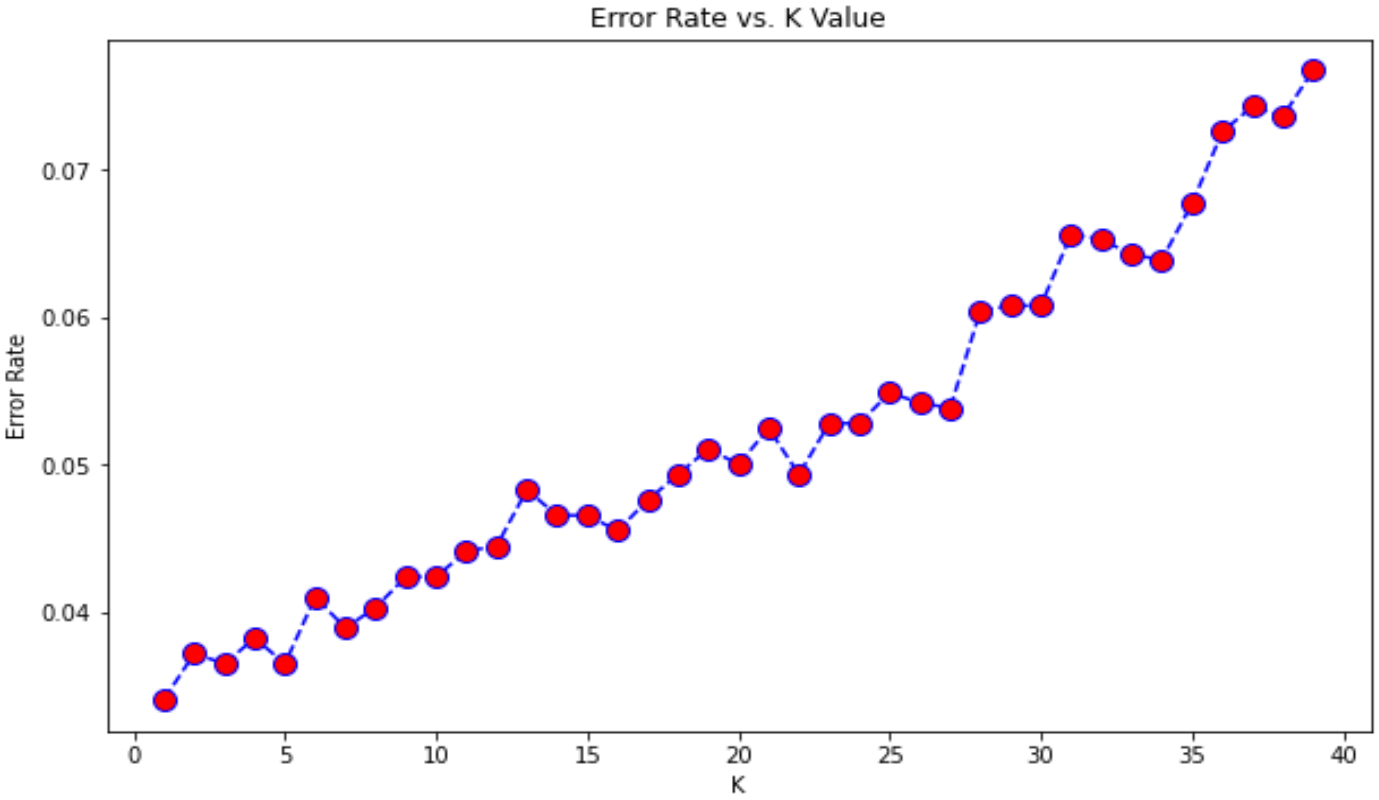


Figure 4: K-Nearest Neighbour Classification Algorithm

Conclusion:

Managing PD in day-to-day life is very challenging for an individual. Therefore, a good screening procedure will be beneficial, especially in circumstances where a physician's treatment is not necessary. Thus, for the diagnosis of PD, ML algorithms were evaluated. The main aim of this review was to identify existing ML-based research to diagnose PD in terms of voice attributes and to determine the most appropriate technique to diagnose the PD with an accuracy, precision and recall rate. The Dataset considered here is Un-balanced, so based on precision and recall we decided the best model. From the review, it was observed that the best recall and precision for voice features to diagnose PD was obtained by K-NN classification algorithm with a precision of 0.925 and recall of 0.98.

sss

ACKNOWLEDGEMENT

We would like to express our deep gratitude towards our project guide **Ms**. **J**. **Sreedevi** Associate Professor, Department of Computer Science and Engineering, MGIT, for her guidance with unsurpassed knowledge and immense encouragement. We are grateful for providing us with the required facilities for the completion of the project work.

We are very much thankful to the Principal **G. Chandra Mohan Reddy** and Management, MGIT, for his inspiration and cooperation to carry out this work.

We express our thanks to Project Coordinator **Ms. J. Sreedevi**, for her Continuous Aid and Inducement.

We would like to thank our parents, friends, and classmates for their support throughout our project period. In the end, we thank everyone who provided us help directly or indirectly in completing this project successfully.

PROJECT STUDENTS

P. AYESHA KHANAM (20261A6741)

P. VENKAT CHARAN (20261A6742)

P. DEEKSHITHA (20261A6743)

P. SRIKAMAL (20261A6744)

P. LAXMIKANTH (20261A6745)

P. HARSHIKA (20261A6746)

R. MADHU ARYAN (20261A6747)

R. RISHI (20261A6748)

S. MAHESHWARI (20261A6749)

S. LIKITHA (20261A6750)

REFERENCES:

1. https://www.kaggle.com/datasets/margot234/parkinsons-disease-dataset
2. https://ieeexplore.ieee.org/document/7707401/
3. https://www.mdpi.com/1784636
4. https://www.tandfonline.com/doi/abs/10.1080/03772063.2018.1531730
5. https://www.kaggle.com/code/aniruddhadeswandikar/parkinsons-disease-classification
6. https://github.com/pqrst/ParkinsonsDiseaseDataAnalysis
7. https://www.analyticsvidhya.com/blog/2021/01/a-guide-to-the-naive-bayes-algorithm/
8. https://www.kaggle.com/code/parhamzm/parkinson-s-disease-pd-classification/data
9. https://towardsdatascience.com/a-comprehensive-guide-to-a-classification-project-data-cleaning-and-exploration-88edd5617ce2
10. https://www.geeksforgeeks.org/k-nearest-neighbors-with-python-ml/
11. https://www.kaggle.com/code/aniruddhadeswandikar/parkinsons-disease-classification

2. J. Rusz, M. Novotný, J. Hlavnička, T. Tykalová and E. Růžička, "High-Accuracy Voice-Based Classification Between Patients With Parkinson's Disease and Other Neurological Diseases May Be an Easy Task With Inappropriate Experimental Design," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 8, pp. 1319-1321, Aug. 2017, doi: 10.1109/TNSRE.2016.2621885.

4. Gunjan Pahuja & T. N. Nagabhushan (2021) A Comparative Study of Existing Machine Learning Approaches for Parkinson's Disease Detection, IETE Journal of Research, 67:1, 4-14, DOI: 10.1080/03772063.2018.1531730

4. Gunjan Pahuja & T. N. Nagabhushan (2021) A Comparative Study of Existing Machine Learning Approaches for Parkinson's Disease Detection, IETE Journal of Research, 67:1, 4-14, DOI: 10.1080/03772063.2018.1531730