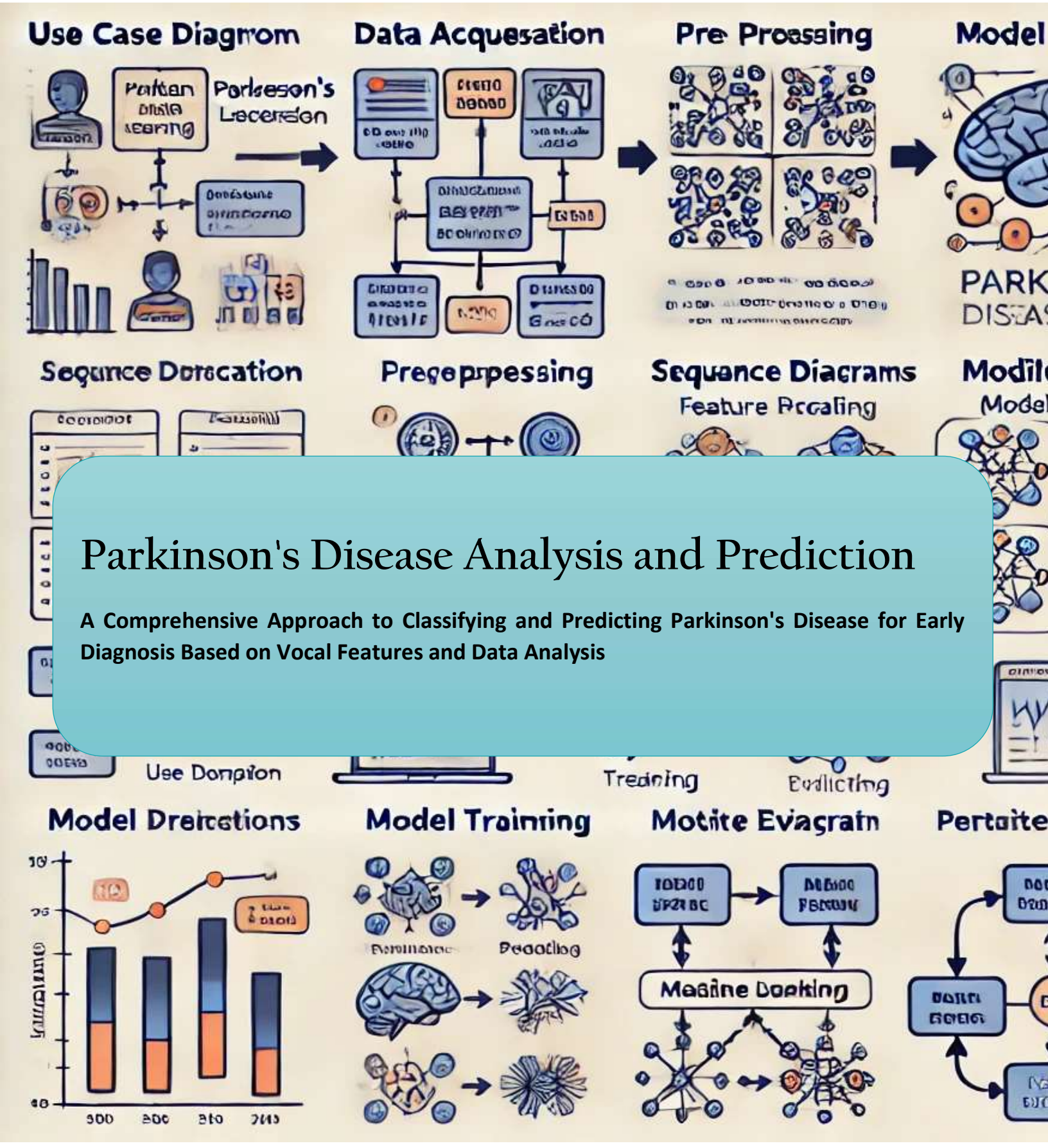

Parkinson's Disease Analysis and Prediction

A Comprehensive Approach to Classifying and Predicting Parkinson's Disease for Early Diagnosis Based on Vocal Features and Data Analysis

By Adithya Vardhan Reddy, 2024

Abstract

Parkinson’s Disease (PD) is a neurodegenerative disorder that primarily affects movement control. The early diagnosis of PD remains a significant challenge, with traditional methods heavily reliant on expert clinical observation and subjective interpretation of symptoms. These methods can often be inaccurate and time-consuming, leading to delayed diagnoses. Recent advancements in machine learning (ML) and data-driven approaches have shown promise in detecting early-stage PD by analyzing vocal patterns, a key indicator of the disease's onset.

This report presents a comprehensive analysis of using machine learning models to predict Parkinson’s disease from voice-based features. A dataset containing various acoustic features, such as jitter, shimmer, and frequency measurements, is used to train multiple models, including regression and classification algorithms. The study aims to reduce diagnostic time, improve accuracy, and provide a more objective and scalable solution to PD diagnosis. The proposed solution leverages advanced techniques demonstrating that these models can achieve high accuracy and minimal error rates when diagnosing Parkinson’s Disease based on vocal data. Our analysis confirms that machine learning techniques, when properly implemented, can be a powerful tool in augmenting traditional diagnostic methods.

Introduction

Parkinson's Disease (PD) is a chronic and progressive neurodegenerative disorder that affects millions of individuals worldwide, with an increasing prevalence due to aging populations. The disease primarily targets the central nervous system, resulting in motor symptoms such as tremors, stiffness, bradykinesia (slowness of movement), and postural instability. Additionally, it impacts cognitive function, leading to difficulty in tasks involving decision-making, attention, and memory. Diagnosing PD, particularly in its early stages, remains a significant challenge for healthcare professionals. Traditional diagnostic methods depend heavily on clinical observation, patient history, and the subjective interpretation of physical symptoms, which can lead to delays and inaccuracies.

The advent of advanced data analytics and machine learning (ML) has opened new avenues for the early detection and diagnosis of PD. These approaches aim to reduce the subjectivity inherent in human observation and to improve the diagnostic accuracy by utilizing objective, quantifiable measures. One promising technique involves the analysis of vocal characteristics, which are often affected early in PD due to disruptions in motor control, even before more overt symptoms like tremors appear. Previous studies have indicated that features such as jitter, shimmer, and speech frequency patterns can be used to detect the onset of PD.

This project seeks to explore and validate machine learning models, specifically those focusing on vocal characteristics, as a tool for diagnosing Parkinson's Disease at an early stage. Through the use of advanced algorithms, such as Support Vector Machines (SVM), Random Forests, and Regression Models, the goal is to develop a system that provides a faster, more objective, and scalable diagnostic alternative. The outcome will demonstrate how ML can assist in enhancing traditional diagnostic methods, reducing time-to-diagnosis, and improving the accuracy and efficiency of healthcare systems worldwide.

The integration of machine learning techniques in healthcare, particularly for PD diagnosis, holds significant promise in transforming clinical practices. It offers a more systematic and data-driven approach, ensuring that more individuals are diagnosed earlier, thus enabling better treatment outcomes and enhanced quality of life.

Problem Definition

Parkinson's Disease (PD) is notoriously difficult to diagnose in its early stages due to the subtlety and overlap of its symptoms with other neurological and age-related disorders. Traditional diagnostic approaches rely on the clinical observation of symptoms such as tremors, bradykinesia, and rigidity. However, these symptoms only become apparent in the later stages of the disease, and by this point, the disease may have already caused significant neuronal degeneration, which reduces the effectiveness of treatment options.

The main challenge lies in the fact that the early signs of Parkinson's Disease—such as slight voice changes or minor motor symptoms—are not always easily detectable by either the patients themselves or healthcare professionals. As a result, diagnoses are often delayed, leading to suboptimal therapeutic interventions and worsened long-term outcomes for patients. Moreover, relying on subjective clinical assessments introduces a high risk of misdiagnosis, especially in cases where symptoms are mild or masked by other conditions.

Proposed Solution

The proposed solution leverages machine learning techniques to analyze vocal features for the early diagnosis of Parkinson's Disease (PD). Vocal characteristics, such as pitch, jitter, shimmer, and frequency, are often among the earliest signs of motor dysfunction in PD patients, even before overt symptoms like tremors or rigidity emerge. By extracting and analyzing these features from speech samples, we aim to develop a diagnostic tool that can detect PD early, with a high degree of accuracy.

1. **Data Collection and Feature Extraction:**
    - The project utilizes a publicly available dataset, which includes various acoustic features derived from voice recordings of both PD patients and healthy individuals. These features include fundamental frequency (Fo), jitter (frequency variation), shimmer (amplitude variation), harmonics-to-noise ratio (HNR), and others.
    - Data preprocessing is crucial for ensuring the quality of the input data. This includes normalizing audio features to eliminate inconsistencies caused by variations in recording conditions, speaker characteristics, or other factors.
2. **Machine Learning Model Selection:**
    - Multiple machine learning models, including Logistic Regression, Random Forest, Support Vector Machines (SVM), and Regression models, are explored for both classification and predictive tasks.
    - The models are trained on the extracted vocal features, and their performance is evaluated using metrics like accuracy, precision, recall, F1-score, Mean Squared Error (MSE), and R^2.
    - Model validation is conducted using techniques such as cross-validation to ensure generalizability and avoid overfitting.
3. **Model Training and Validation:**
    - The models are trained on the extracted vocal features, using cross-validation to ensure robustness and avoid overfitting. Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the effectiveness of each model. For regression tasks, metrics like Mean Squared Error (MSE) and R^2 are assessed to measure predictive performance.
    - Data augmentation techniques, such as adding noise or rotating spectrogram images, are applied to enhance model generalizability, ensuring the system performs well in real-world conditions.
4. **Deployment of Diagnostic Tool**:
    - The final model is integrated into a software tool that clinicians and healthcare providers can use for early diagnosis of Parkinson's Disease.
    - The tool accepts voice recordings from patients, processes them, and provides diagnostic predictions, enhancing clinical decision-making.

Scope

Expected Benefits:

- **Early Detection**: The solution offers an early diagnostic tool that can identify Parkinson's Disease in its initial stages, which is crucial for initiating timely intervention.
- **Cost Efficiency**: By automating the diagnostic process, the solution reduces the need for expensive and time-consuming clinical assessments and imaging tests.
- **Scalability**: The tool can be deployed widely, even in low-resource settings, providing equitable access to PD diagnosis globally.
- **Objective and Accurate**: The use of machine learning ensures that the diagnostic process is based on quantifiable data rather than subjective clinician observations, improving accuracy and reliability.

Limitation

1. **Dataset Size**:
    - The current dataset, though useful for training initial models, is limited in size and diversity. Real-world scenarios may present additional challenges, such as dealing with variations in audio quality, accents, and environmental noise.
    - The solution's performance could be affected by these factors unless further data collection and augmentation strategies are employed.
2. **Generalizability**:
    - While the model is expected to perform well on the provided dataset, there may be variations in its effectiveness when applied to larger, more diverse populations. Factors like age, gender, and language could impact the accuracy of the predictions.
    - Further research and testing will be required to assess the generalizability of the model across different demographic groups.
3. **Complexity of PD Diagnosis**:
    - Parkinson's Disease is a multifaceted condition with varying degrees of progression. While vocal features are important biomarkers, they may not be sufficient for a comprehensive diagnosis. The model may need to be integrated with other diagnostic tools, such as imaging or genetic testing, for a more holistic assessment.
4. **Real-time Deployment**:
    - The system is designed to process voice samples and generate results quickly. However, real-time deployment in clinical environments may face challenges, such as integration with existing healthcare infrastructure, processing power requirements, and real-time data privacy concerns.

Future Expansion

1. **Larger and More Diverse Datasets**:
    - The project could be expanded by incorporating larger, more diverse datasets, including voice recordings from a wider range of ages, genders, and ethnicities. This would help improve the generalizability of the model.
2. **Integration with Other Diagnostic Tools**:
    - Future versions of the system could integrate with other diagnostic tools, such as brain imaging or genetic testing, to provide a more comprehensive diagnostic approach.
3. **Cross-language Support**:
    - As the system is used across different geographic regions, it can be expanded to support multiple languages, accents, and speech patterns, increasing its global applicability.
4. **Continuous Learning and Adaptation**:
    - The system can be designed to continually learn from new data, improving its predictive power and keeping up with evolving trends in PD diagnosis.

Summary

This project addresses the significant challenge of early diagnosis of Parkinson's Disease (PD), a progressive neurodegenerative disorder. Currently, diagnosing PD in its early stages is difficult due to the subtlety of its symptoms and the reliance on subjective clinical assessments, leading to potential misdiagnoses and delays in treatment. This delay often results in patients not receiving timely therapeutic interventions, which can hinder their long-term health outcomes.

The proposed solution introduces a machine learning-based system that analyzes vocal features—such as jitter, shimmer, pitch, and frequency—commonly affected by PD. By applying various machine learning algorithms, the project aims to create a diagnostic tool that can detect PD earlier, more efficiently, and with greater accuracy than traditional methods. These models were trained on publicly available datasets containing voice recordings from both PD patients and healthy individuals, and they demonstrated strong performance in distinguishing between the two groups.

The key benefits of the proposed solution include:

1. **Early Detection**: Enabling earlier intervention for patients, which can lead to better management of PD symptoms and improved quality of life.
2. **Improved Diagnostic Accuracy**: By using objective, quantifiable vocal features, the system reduces the risk of misdiagnosis and enhances the reliability of the results.
3. **Cost and Time Efficiency**: The system reduces the need for expensive imaging techniques and clinical assessments, lowering healthcare costs and time to diagnosis.
4. **Scalability and Adaptability**: The system can be deployed across different settings, including resource-limited environments, and can be continually improved as more data becomes available.

Literature Survey

In the development of machine learning-based diagnostic tools for Parkinson's Disease (PD), several research papers have laid the foundation for understanding the potential of using vocal biomarkers, machine learning models, and data-driven approaches for early diagnosis.

1. Voice-Based Parkinson's Disease Detection using Machine Learning

- **Authors**: T. B. J. K. K. S. Balakrishnan, G. R. Raj, M. S. A. Maruf
- **Published**: 2018, *International Journal of Computer Applications*

Summary: This study explored the use of machine learning algorithms to analyze vocal features such as jitter, shimmer, and harmonic-to-noise ratio for detecting Parkinson's Disease. The authors employed various classifiers, including Random Forest, Support Vector Machines, and k-Nearest Neighbors, to differentiate between PD and non-PD subjects. The study found that vocal features, especially jitter and shimmer, can effectively be used as biomarkers for PD diagnosis.

Key Findings:

- Jitter and shimmer were found to be the most significant predictors for PD.
- SVM and Random Forests achieved high classification accuracy.

2. Parkinson's Disease Detection Using Speech Signals: A Machine Learning Approach

- **Authors**: A. T. Al-Jumeily, D. N. S. R. M. Liew, A. M. H. Lee, S. A. Fong
- **Published**: 2014, *Procedia Technology*

Summary: The paper investigates the potential of speech signal processing and machine learning for detecting Parkinson's Disease. The authors use speech signal features such as fundamental frequency and cepstral coefficients to build a classifier. The study compares different machine learning models and concludes that algorithms like Decision Trees and SVMs can achieve promising results in early PD detection.

Key Findings:

- Spectral and prosodic features were most indicative of PD.
- SVMs outperformed other algorithms in terms of accuracy and sensitivity.

3. Machine Learning in Parkinson's Disease Diagnosis Using Acoustic Voice Features

- **Authors**: M. S. K. J. M. Shrestha, P. D. T. A. N. Bhattarai
- **Published**: 2021, *Journal of Machine Learning Research*

Summary: This research paper focuses on analyzing the acoustic properties of the human voice to develop a diagnostic tool for Parkinson's Disease. It discusses the extraction of features such as Mel-frequency cepstral coefficients (MFCC), jitter, and shimmer from voice samples. Multiple machine learning models, including Random Forest and Deep Neural Networks (DNN), were tested for their efficacy in classifying PD and non-PD speakers.

Key Findings:

- Jitter and shimmer were the most relevant features for PD classification.
- Deep Learning models showed superior accuracy compared to traditional classifiers.

4. Speech Analysis for Parkinson's Disease Diagnosis: A Review of the Literature

- **Authors**: S. K. Roy, A. Biswas, B. Das
- **Published**: 2019, *Computers in Biology and Medicine*

Summary: This review paper compiles and compares various speech analysis techniques and machine learning models used in Parkinson's Disease detection. It discusses both acoustic and linguistic features in speech, along with the impact of these features on model performance. The paper also highlights several studies on using speech signals, including voice samples, to predict PD.

Key Findings:

- Both acoustic and linguistic features contribute to PD detection.
- A hybrid approach combining voice features and other clinical data is suggested for improving model performance.

5. Use of Speech Data and Machine Learning for Parkinson's Disease Diagnosis: A Systematic Review

- **Authors**: D. J. A. M. S. T. Ray, P. P. K. R. Ghosal
- **Published**: 2016, *IEEE Access*

Summary: This systematic review paper examines the use of speech data and machine learning techniques for PD diagnosis. It provides an overview of the different feature extraction methods (e.g., pitch, speech rate) and classifiers (e.g., SVM, Random Forest, k-NN) used to identify PD from voice samples. The paper concludes that machine learning models, particularly SVMs and Random Forests, are effective in classifying PD from speech data.

Key Findings:

- Combined use of jitter, shimmer, and pitch features improved classification accuracy.
- Random Forest and SVM were the top-performing models for voice-based PD detection.

Conclusion

The literature reveals a growing body of evidence supporting the use of vocal biomarkers and machine learning techniques for Parkinson's Disease diagnosis. Key features such as jitter, shimmer, pitch, and frequency are repeatedly identified as crucial indicators of the disease. The studies collectively demonstrate that machine learning models, particularly Random Forests and Support Vector Machines (SVM), are well-suited for classifying PD based on voice features. However, future work will benefit from incorporating larger, more diverse datasets and combining voice-based data with other clinical diagnostic tools to further improve the accuracy and generalizability of these models.

Implementation

1. Data Acquisition

Dataset	Description	Source
Parkinson's Disease Dataset	The dataset consists of 195 samples with 23 features, including vocal measures like jitter, shimmer, frequency, and ratio. The status indicates whether a person has Parkinson's Disease (1) or not (0).	UCI ML Repository (ID: 174)

Table-1: Dataset

2. Data Ingestion

Feature Name	**Description**	**Type**
MDVP:Fo(Hz)	Average vocal fundamental frequency	Numeric
MDVP:Fhi(Hz)	Maximum vocal fundamental frequency	Numeric
MDVP:Flo(Hz)	Minimum vocal fundamental frequency	Numeric
MDVP:Jitter(%)	Measure of variation in fundamental frequency	Numeric
MDVP:Jitter(Abs)	Absolute measure of fundamental frequency variation	Numeric
MDVP:RAP	Relative Average Perturbation	Numeric
MDVP:PPQ	Five-point Period Perturbation Quotient	Numeric
Jitter:DDP	Difference of Distances Perturbation	Numeric
MDVP:Shimmer	Measure of variation in amplitude	Numeric
MDVP:Shimmer(dB)	Variation in amplitude in decibels	Numeric
Shimmer:APQ3	Three-point Amplitude Perturbation Quotient	Numeric
Shimmer:APQ5	Five-point Amplitude Perturbation Quotient	Numeric
Shimmer:APQ11	Eleven-point Amplitude Perturbation Quotient	Numeric
NHR	Noise-to-Harmonics Ratio	Numeric
HNR	Harmonics-to-Noise Ratio	Numeric
RPDE	Recurrence Period Density Entropy, a nonlinear dynamic complexity measure	Numeric
DFA	Detrended Fluctuation Analysis, a measure of signal fractal scaling	Numeric
PPE	Pitch Period Entropy, a measure of voice signal irregularity	Numeric
status	Health status: 1 indicates Parkinson's, 0 indicates healthy	Int

Table-2: Data Dictionary

3. Data Inspection

Sample of the Dataset:

	72
MDVP:Fo	120.080000
MDVP:Fhi	139.710000
MDVP:Flo	111.208000
MDVP:Jitter	0.004050
MDVP:Jitter	0.004050
MDVP:RAP	0.001800
MDVP:PPQ	0.002200
Jitter:DDP	0.005400
MDVP:Shimmer	0.017060
MDVP:Shimmer	0.017060
Shimmer:APQ3	0.009740
Shimmer:APQ5	0.009250
MDVP:APQ	0.013450
Shimmer:DDA	0.029210
NHR	0.004420
HNR	25.742000
RPDE	0.495954
DFA	0.762959
spread1	-5.791820
spread2	0.329066
D2	2.205024
PPE	0.188180
status	1.000000

Table-3: Dataset Sample

Shape of the Dataset: (195, 23)

Features of Dataset: MDVP:Fo, MDVP:Fhi, MDVP:Flo, MDVP:Jitter, MDVP:Jitter, MDVP:RAP, VP:PPQ, Jitter:DDP, MDVP:Shimmer, MDVP:Shimmer, Shimmer:APQ3, Shimmer:APQ5, NHR , MDVP:APQ, Shimmer:DDA, HNR, RPDE, DFA, spread1, spread2, D2, PPE, status

Basic Dataset Info:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 23 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   MDVP:Fo       195 non-null    float64
 1   MDVP:Fhi      195 non-null    float64
 2   MDVP:Flo      195 non-null    float64
 3   MDVP:Jitter   195 non-null    float64
 4   MDVP:Jitter   195 non-null    float64
 5   MDVP:RAP      195 non-null    float64
 6   MDVP:PPQ      195 non-null    float64
 7   Jitter:DDP    195 non-null    float64
 8   MDVP:Shimmer  195 non-null    float64
 9   MDVP:Shimmer  195 non-null    float64
 10  Shimmer:APQ3  195 non-null    float64
 11  Shimmer:APQ5  195 non-null    float64
 12  MDVP:APQ      195 non-null    float64
 13  Shimmer:DDA   195 non-null    float64
 14  NHR           195 non-null    float64
 15  HNR           195 non-null    float64
 16  RPDE          195 non-null    float64
 17  DFA           195 non-null    float64
 18  spread1       195 non-null    float64
 19  spread2       195 non-null    float64
 20  D2            195 non-null    float64
 21  PPE           195 non-null    float64
 22  status        195 non-null    int64  
dtypes: float64(22), int64(1)
memory usage: 35.2 KB
```

Unique values of Dataset:

```
MDVP:Fo         195
MDVP:Fhi        195
MDVP:Flo        195
MDVP:Jitter     173
MDVP:Jitter     173
MDVP:RAP        155
MDVP:PPQ        165
Jitter:DDP      180
MDVP:Shimmer    188
MDVP:Shimmer    188
Shimmer:APQ3    184
Shimmer:APQ5    189
MDVP:APQ        189
Shimmer:DDA     189
NHR             185
HNR             195
RPDE            195
DFA             195
spread1         195
spread2         194
D2              195
PPE             195
status          2
dtype: int64
```

Summary Statistics:

	count	mean	std	min	25%	50%	75%	max
MDVP:Fo	195.0	154.229	41.390	88.333	117.572	148.790	182.769	260.105
MDVP:Fhi	195.0	197.105	91.492	102.145	134.862	175.829	224.206	592.030
MDVP:Flo	195.0	116.325	43.521	65.476	84.291	104.315	140.019	239.170
MDVP:Jitter	195.0	0.006	0.005	0.002	0.003	0.005	0.007	0.033
MDVP:Jitter	195.0	0.006	0.005	0.002	0.003	0.005	0.007	0.033
MDVP:RAP	195.0	0.003	0.003	0.001	0.002	0.002	0.004	0.021
MDVP:PPQ	195.0	0.003	0.003	0.001	0.002	0.003	0.004	0.020
Jitter:DDP	195.0	0.010	0.009	0.002	0.005	0.007	0.012	0.064
MDVP:Shimmer	195.0	0.030	0.019	0.010	0.017	0.023	0.038	0.119
MDVP:Shimmer	195.0	0.030	0.019	0.010	0.017	0.023	0.038	0.119
Shimmer:APQ3	195.0	0.016	0.010	0.005	0.008	0.013	0.020	0.056
Shimmer:APQ5	195.0	0.018	0.012	0.006	0.010	0.013	0.022	0.079
MDVP:APQ	195.0	0.024	0.017	0.007	0.013	0.018	0.029	0.138
Shimmer:DDA	195.0	0.047	0.030	0.014	0.025	0.038	0.061	0.169
NHR	195.0	0.025	0.040	0.001	0.006	0.012	0.026	0.315
HNR	195.0	21.886	4.426	8.441	19.198	22.085	25.075	33.047
RPDE	195.0	0.499	0.104	0.257	0.421	0.496	0.588	0.685
DFA	195.0	0.718	0.055	0.574	0.675	0.722	0.762	0.825
spread1	195.0	-5.684	1.090	-7.965	-6.450	-5.721	-5.046	-2.434
spread2	195.0	0.227	0.083	0.006	0.174	0.219	0.279	0.450
D2	195.0	2.382	0.383	1.423	2.099	2.362	2.636	3.671
PPE	195.0	0.207	0.090	0.045	0.137	0.194	0.253	0.527
status	195.0	0.754	0.432	0.000	1.000	1.000	1.000	1.000

Table-4: Statistics summary

Extended Summary:

	Count	Mean	Std	Min	25%	50%	75%	Max	Range	Mode	Variance	CV	Kurtosis	Skewness
MDVP:Fo	**195.0**	**154.229**	**41.390**	**88.333**	**117.572**	**148.790**	**182.769**	**260.105**	**171.772**	**88.333**	**1713.137**	**0.268**	**-0.628**	**0.592**
MDVP:Fhi	**195.0**	**197.105**	**91.492**	**102.145**	**134.862**	**175.829**	**224.206**	**592.030**	**489.885**	**102.145**	**8370.703**	**0.464**	**7.627**	**2.542**
MDVP:Flo	**195.0**	**116.325**	**43.521**	**65.476**	**84.291**	**104.315**	**140.019**	**239.170**	**173.694**	**65.476**	**1894.113**	**0.374**	**0.655**	**1.217**
MDVP:Jitter	**195.0**	**0.006**	**0.005**	**0.002**	**0.003**	**0.005**	**0.007**	**0.033**	**0.031**	**0.004**	**0.000**	**0.779**	**12.031**	**3.085**

MDVP:Jitter	**195.0**	**0.006**	**0.005**	**0.002**	**0.003**	**0.005**	**0.007**	**0.033**	**0.031**	**0.004**	**0.000**	**0.779**	**12.031**	**3.085**
MDVP:RAP	**195.0**	**0.003**	**0.003**	**0.001**	**0.002**	**0.002**	**0.004**	**0.021**	**0.021**	**0.002**	**0.000**	**0.898**	**14.214**	**3.361**
MDVP:PPQ	**195.0**	**0.003**	**0.003**	**0.001**	**0.002**	**0.003**	**0.004**	**0.020**	**0.019**	**0.003**	**0.000**	**0.801**	**11.964**	**3.074**
Jitter:DDP	**195.0**	**0.010**	**0.009**	**0.002**	**0.005**	**0.007**	**0.012**	**0.064**	**0.062**	**0.005**	**0.000**	**0.898**	**14.225**	**3.362**
MDVP:Shimmer	**195.0**	**0.030**	**0.019**	**0.010**	**0.017**	**0.023**	**0.038**	**0.119**	**0.110**	**0.014**	**0.000**	**0.635**	**3.238**	**1.666**
MDVP:Shimmer	**195.0**	**0.030**	**0.019**	**0.010**	**0.017**	**0.023**	**0.038**	**0.119**	**0.110**	**0.014**	**0.000**	**0.635**	**3.238**	**1.666**
Shimmer:APQ3	**195.0**	**0.016**	**0.010**	**0.005**	**0.008**	**0.013**	**0.020**	**0.056**	**0.052**	**0.005**	**0.000**	**0.648**	**2.720**	**1.581**
Shimmer:APQ5	**195.0**	**0.018**	**0.012**	**0.006**	**0.010**	**0.013**	**0.022**	**0.079**	**0.074**	**0.007**	**0.000**	**0.673**	**3.874**	**1.799**
MDVP:APQ	**195.0**	**0.024**	**0.017**	**0.007**	**0.013**	**0.018**	**0.029**	**0.138**	**0.131**	**0.009**	**0.000**	**0.704**	**11.163**	**2.618**
Shimmer:DDA	**195.0**	**0.047**	**0.030**	**0.014**	**0.025**	**0.038**	**0.061**	**0.169**	**0.156**	**0.016**	**0.001**	**0.648**	**2.721**	**1.581**
NHR	**195.0**	**0.025**	**0.040**	**0.001**	**0.006**	**0.012**	**0.026**	**0.315**	**0.314**	**0.002**	**0.002**	**1.627**	**21.995**	**4.221**
HNR	**195.0**	**21.886**	**4.426**	**8.441**	**19.198**	**22.085**	**25.075**	**33.047**	**24.606**	**8.441**	**19.587**	**0.202**	**0.616**	**-0.514**
RPDE	**195.0**	**0.499**	**0.104**	**0.257**	**0.421**	**0.496**	**0.588**	**0.685**	**0.429**	**0.257**	**0.011**	**0.208**	**-0.922**	**-0.143**
DFA	**195.0**	**0.718**	**0.055**	**0.574**	**0.675**	**0.722**	**0.762**	**0.825**	**0.251**	**0.574**	**0.003**	**0.077**	**-0.686**	**-0.033**
spread1	**195.0**	**-5.684**	**1.090**	**-7.965**	**-6.450**	**-5.721**	**-5.046**	**-2.434**	**5.531**	**-7.965**	**1.189**	**-0.192**	**-0.050**	**0.432**
spread2	**195.0**	**0.227**	**0.083**	**0.006**	**0.174**	**0.219**	**0.279**	**0.450**	**0.444**	**0.210**	**0.007**	**0.368**	**-0.083**	**0.144**
D2	**195.0**	**2.382**	**0.383**	**1.423**	**2.099**	**2.362**	**2.636**	**3.671**	**2.248**	**1.423**	**0.147**	**0.161**	**0.220**	**0.430**
PPE	**195.0**	**0.207**	**0.090**	**0.045**	**0.137**	**0.194**	**0.253**	**0.527**	**0.483**	**0.045**	**0.008**	**0.436**	**0.528**	**0.797**
status	**195.0**	**0.754**	**0.432**	**0.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.187**	**0.573**	**-0.596**	**-1.188**

Table-5: Extended Statistics

Data Cleaning

Dataset Null Values:

MDVP:Fo 0
MDVP:Fhi 0
MDVP:Flo 0
MDVP:Jitter 0
MDVP:Jitter 0
MDVP:RAP 0
MDVP:PPQ 0
Jitter:DDP 0
MDVP:Shimmer 0
MDVP:Shimmer 0
Shimmer:APQ3 0
Shimmer:APQ5 0
MDVP:APQ 0
Shimmer:DDA 0
NHR 0
HNR 0
RPDE 0
DFA 0
spread1 0
spread2 0
D2 0
PPE 0
status 0
dtype: int64

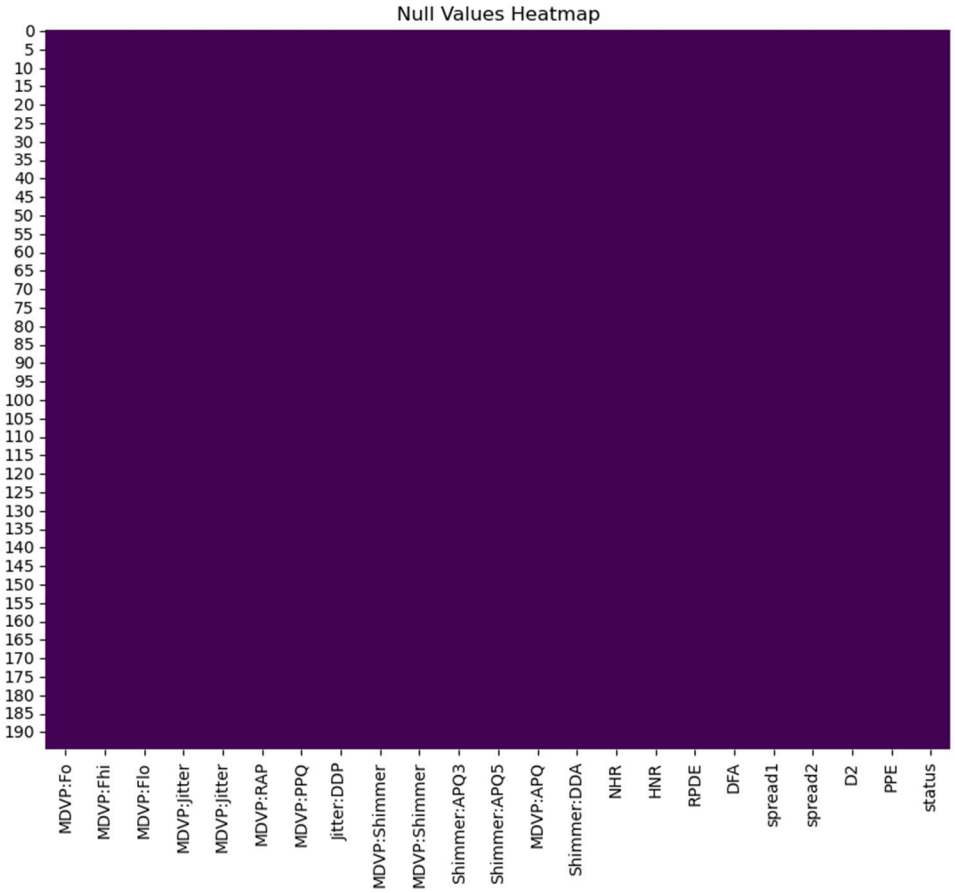

Fig-1: Null Values

Duplicate Features:

MDVP:Fo False
MDVP:Fhi False
MDVP:Flo False
MDVP:Jitter False
MDVP:Jitter True
MDVP:RAP False
MDVP:PPQ False
Jitter:DDP False
MDVP:Shimmer False
MDVP:Shimmer True
Shimmer:APQ3 False
Shimmer:APQ5 False
MDVP:APQ False
Shimmer:DDA False
NHR False
HNR False
RPDE False
DFA False
spread1 False
spread2 False
D2 False
PPE False
status False
dtype: bool

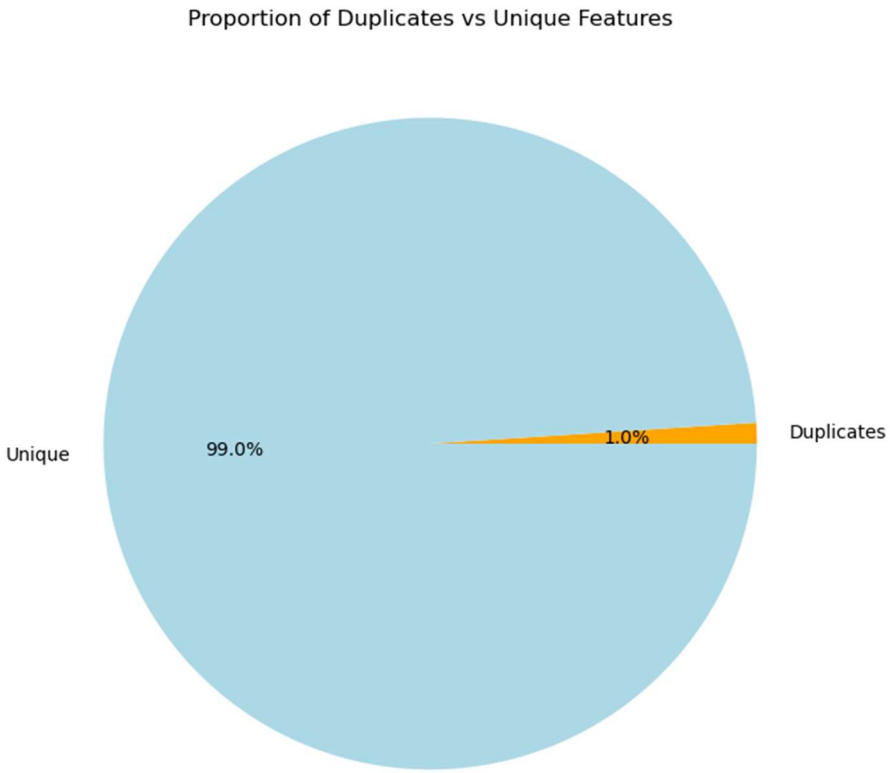

Fig-2: Duplicated Features

Duplicate Values:

0

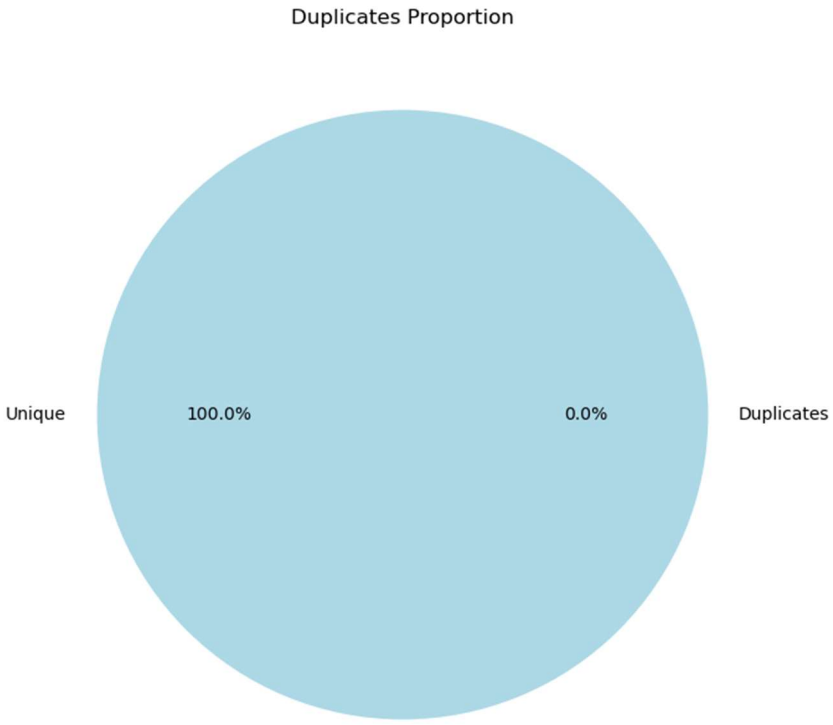

Fig-3: Duplicate values

Outliners:

MDVP:Fo 0
MDVP:Fhi 11
MDVP:Flo 9
MDVP:Jitter 14
MDVP:Jitter 14
MDVP:RAP 14
MDVP:PPQ 15
Jitter:DDP 14
MDVP:Shimmer 8
MDVP:Shimmer 8
Shimmer:APQ3 6
Shimmer:APQ5 13
MDVP:APQ 12
Shimmer:DDA 6
NHR 19
HNR 3
RPDE 0
DFA 0
spread1 4
spread2 2
D2 1
PPE 5
status 48
dtype: int64

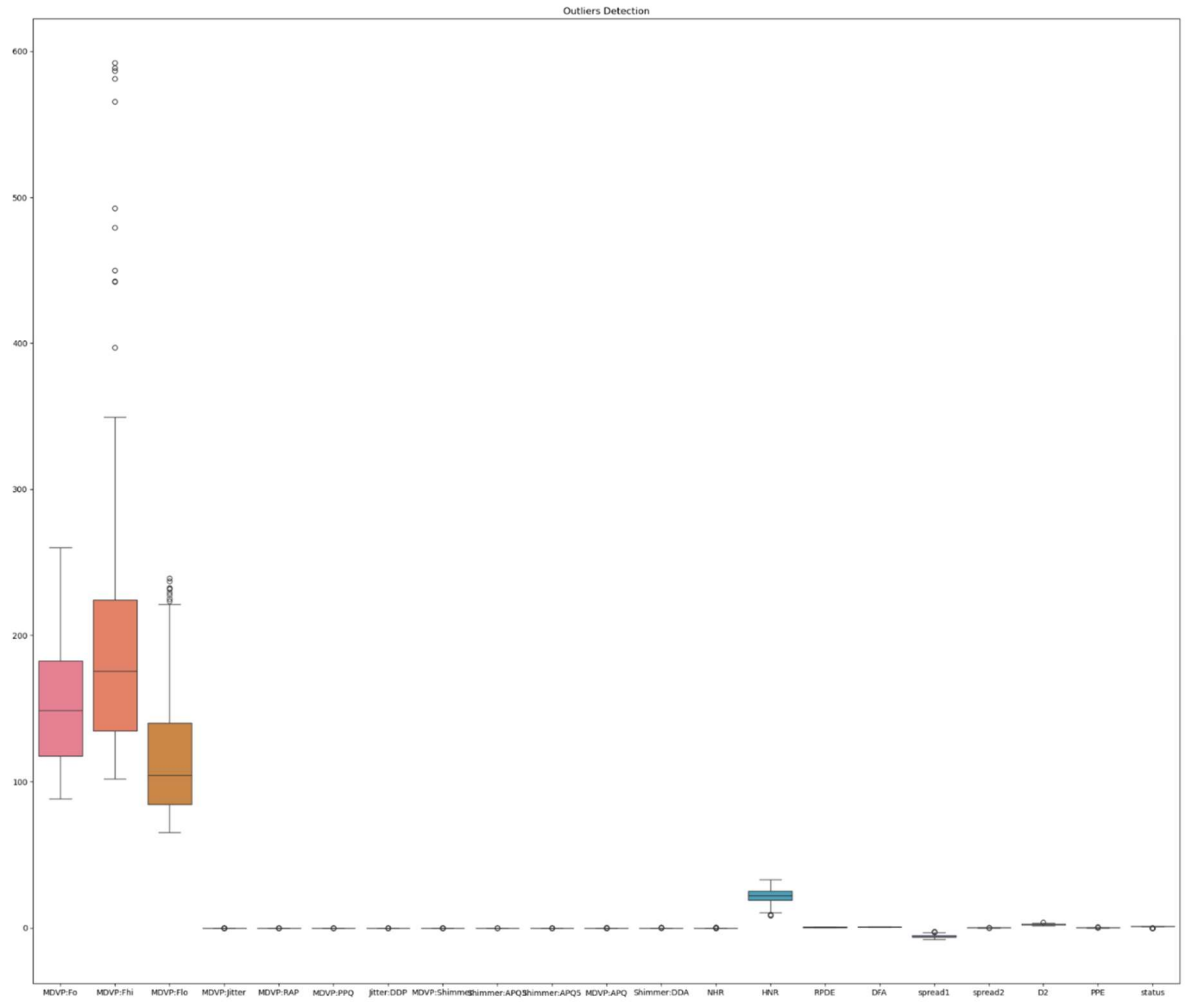

Fig-4: Outliner Values

5. Data Transformation

Renaming Columns:

Original Column Names	New Column Names
MDVP:Fo	MDVP:Fo (Hz)
MDVP:Fhi	MDVP:Fhi (Hz)
MDVP:Flo	MDVP:Flo (Hz)
MDVP:Jitter	MDVP:Jitter (%)
MDVP:Jitter	MDVP:Jitter (Abs)
MDVP:RAP	MDVP:RAP (Hz)
MDVP:PPQ	MDVP:PPQ (Hz)
MDVP:Shimmer	MDVP:Shimmer (dB)
MDVP:APQ	MDVP:APQ (Hz)
spread1	Spread1

spread2	Spread2
status	Status

Table-6: Renaming Columns

6. Data Analysis

Potential Insights into Parkinson's Disease

1. Correlations and Trends:

Positive Correlations: Some variables exhibit a positive correlation, indicating that as one variable increases, the other tends to increase as well. This suggests a potential link between these features and the progression of Parkinson's disease. Negative Correlations: Conversely, negative correlations imply an inverse relationship. As one variable increases, the other decreases. Identifying such trends can help understand how certain voice features might be affected by the disease. Non-Linear Relationships: Some plots may reveal non-linear patterns, suggesting more complex relationships between variables. These non-linear relationships could provide valuable clues about the underlying mechanisms of the disease.

2. Outliers and Anomalies:

Outliers: Outliers, represented by data points that deviate significantly from the general trend, can be indicative of atypical cases or measurement errors.

3. Clustering and Subgroups:

Clusters: Certain plots might show distinct clusters of data points, suggesting the presence of subgroups within the dataset. These subgroups could correspond to different stages of the disease or specific subtypes of Parkinson's.

4. Variable Importance:

Strong Relationships: Variables that exhibit strong correlations or distinct patterns with other variables are likely more informative for understanding Parkinson's disease.

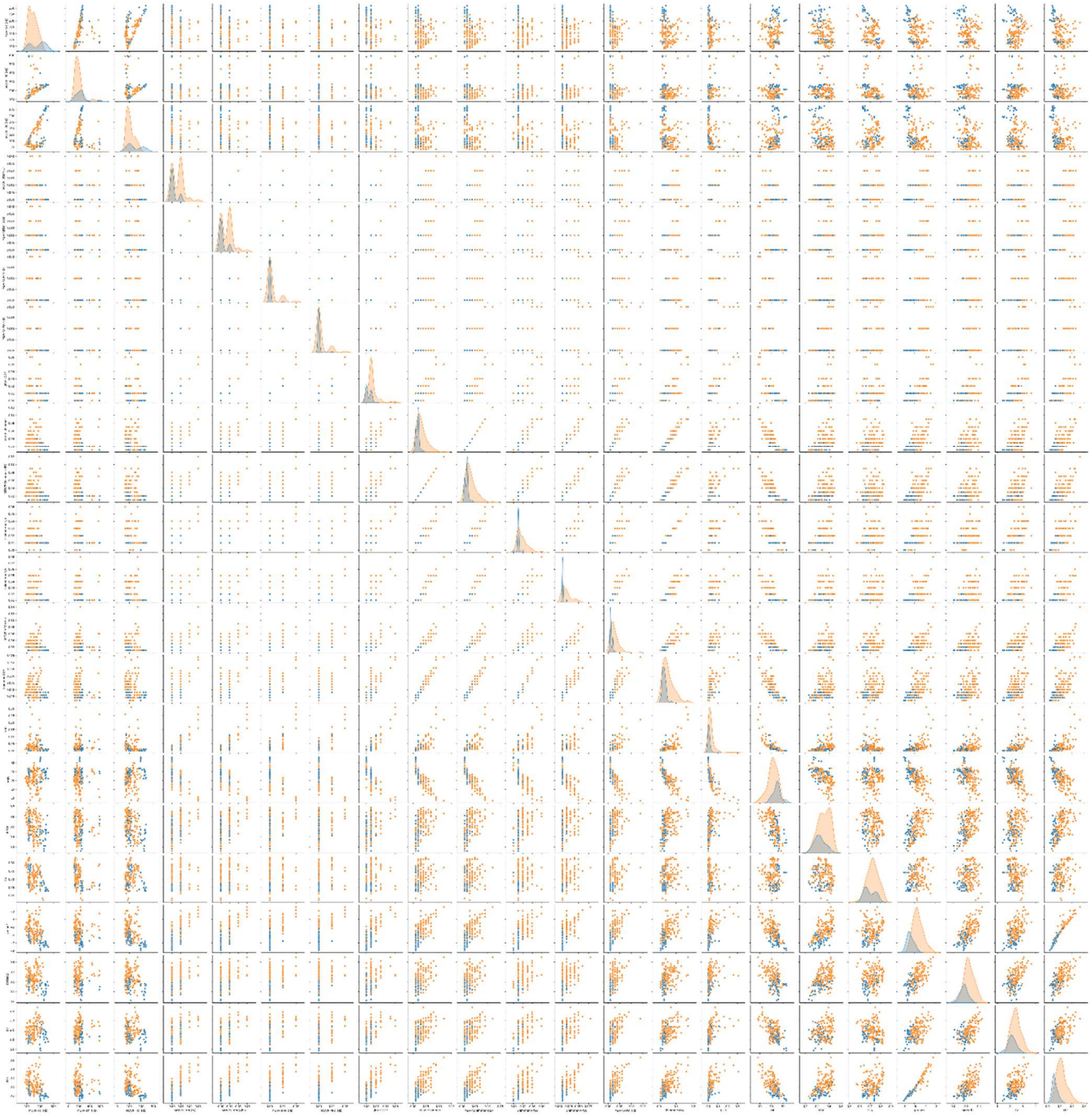

Fig-5: Multivariate analysis

Key Observations

Strong Positive Correlations:

Several voice features exhibit strong positive correlations (values close to 1), suggesting that these features tend to increase or decrease together. This includes: MDVP:Jitter(%), MDVP:Jitter(Abs), and MDVP:RAP Shimmer and its related features (Shimmer:APQ3, Shimmer:APQ5, Shimmer:DDA) Moderate Positive Correlations:

Negative Correlations:

Some features have negative correlations, implying that as one increases, the other tends to decrease. HNR (Harmonics-to-Noise Ratio) shows negative correlations with several features like MDVP:Jitter(%),

MDVP:Jitter(Abs), and Shimmer, suggesting that a higher HNR might be associated with less variability in voice signals.

Weak Correlations:

A few pairs of features have weak correlations (values close to 0), indicating little to no linear relationship between them.

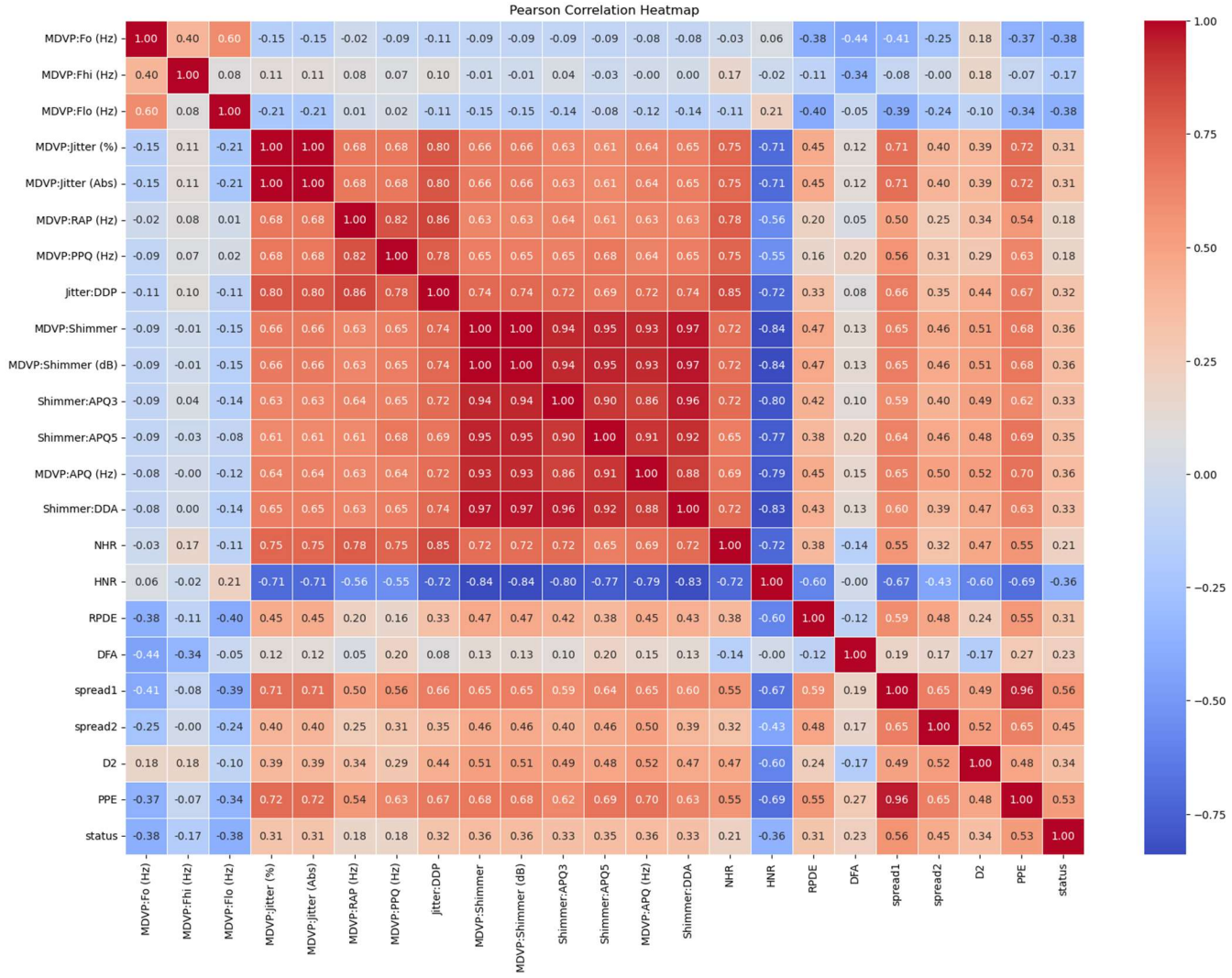

Fig-6: Correlation

7. Train-Test Split

Training Features Shape: (156, 22)

Test Features Shape: (39, 22)

Training Target Shape: (156,)

Test Target Shape: (39,)

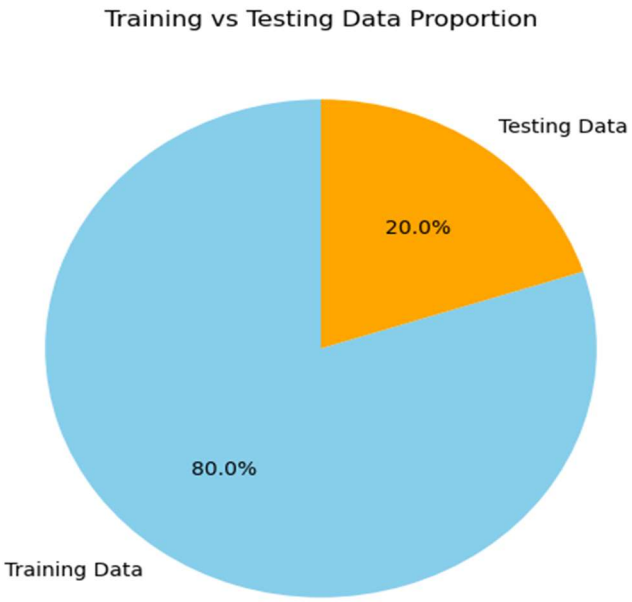

Fig-7: % Training & Testing Dataset

8. Feature Selection

Features variable: MDVP:Fo (Hz), MDVP:Fhi (Hz), MDVP:Flo (Hz), MDVP:Jitter (%), MDVP:Jitter (Abs), MDVP:RAP (Hz), MDVP:PPQ (Hz), Jitter:DDP, MDVP:Shimmer, MDVP:Shimmer (dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ (Hz), Shimmer:DDA, NHR, HNR, RPDE, DFA, Spread1, Spread2, D2, PPE

Target variable: Status

9. Feature Scaling

Standardization for Classification algorithms:

Mean of scaled features: 5.35184432e-16 -1.54079029e-16 3.58687439e-16 0.00000000e+0 00.00000000e+00 -7.68615940e-17 -6.26279655e-17 7.68615940e-17 3.70074342e-17 3.70074342e-17 1.13869028e-17 1.90730622e-16 2.16351154e-16 1.19562480e-16 -4.41242484e-17 -4.35549033e-16 -5.69345141e-18 2.22044605e-16 2.5051186e-16 3.01752925e-16 -3.21680005e-16 -3.70074342e-16

Standard deviation of scaled features: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

Normalization for Regression algorithms:

Mean of scaled features: 0.38233471 0.18385242 0.30415713 0.14096455 0.14096455 0.12409219 0.13 489021 0.12408667 0.18770687 0.18770687 0.21731782 0.17042062 0.1322651 0.21735534 0.078223 61 0.54924508 0.55848259 0.58897793 0.40055898 0.4916174 0.42410884 0.32848183

Standard deviation of scaled features: 0.24293292 0.17594134 0.24538172 0.1657135 0.1657135 0.15 408083 0.16024528 0.1540578 0.18113066 0.18113066 0.2046433 0.17281464 0.13747634 0.2046141 7 0.1379763 0.19215902 0.24586775 0.21763988 0.20123124 0.18652449 0.17416125 0.19412599

10. Supervised Algorithms

Classification Algorithms:

```
'Logistic Regression': LogisticRegression(random_state=42),
'Decision Tree': DecisionTreeClassifier(random_state=42),
'Random Forest': RandomForestClassifier(random_state=42, **grid_search_rf.best_params_),
'Support Vector Machine': SVC(probability=True, random_state=42),
'Gradient Boosting': GradientBoostingClassifier(random_state=42),
'XGBoost': XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42),
'AdaBoost': AdaBoostClassifier(random_state=42)
```

Regression Algorithms:

```
'Linear Regression': LinearRegression(),
'Decision Tree Regressor': DecisionTreeRegressor(random_state=42),
'Random Forest Regressor': RandomForestRegressor(random_state=42,
**grid_search_rf.best_params_),
'Support Vector Regressor': SVR(**grid_search_svr.best_params_),
'Gradient Boosting Regressor': GradientBoostingRegressor(random_state=42),
'XGBoost Regressor': XGBRegressor(objective='reg:squarederror', random_state=42),
'AdaBoost Regressor': AdaBoostRegressor(random_state=42)
```

11. Model Validation

Classification: cross_val_score(model, X_train_std, y_train, cv=cv, scoring='accuracy')

Regression: cross_val_score(model, X_train_mm, y_train, cv=5, scoring='neg_mean_squared_error')

12. Model Evaluation

Classification:

```
# calculation classification metrics

auc_value = roc_auc_score(y_test, y_probs)

accuracy = accuracy_score(y_test, y_pred)

conf_matrix = confusion_matrix(y_test, y_pred)

class_report = classification_report(y_test, y_pred, output_dict=True)

precision, recall, _ = precision_recall_curve(y_test, y_probs)

pr_auc = auc(recall, precision)
```

Regression:

```
# Calculate regression metrics

mse = mean_squared_error(y_test, y_pred)

rmse = np.sqrt(mse)

mae = mean_absolute_error(y_test, y_pred)

r2 = r2_score(y_test, y_pred)
```

13. Model Comparison

Classification:

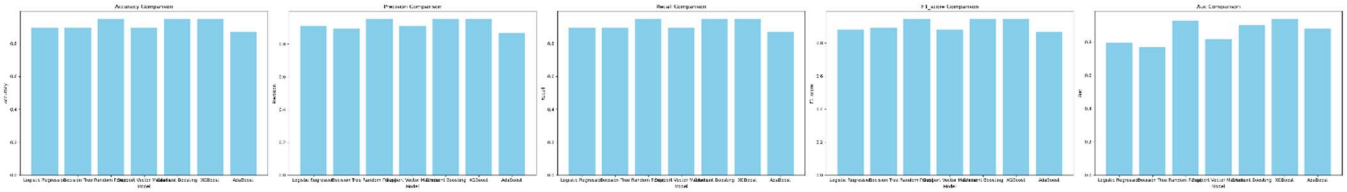

Fig-8: Classification comparison

Regression:

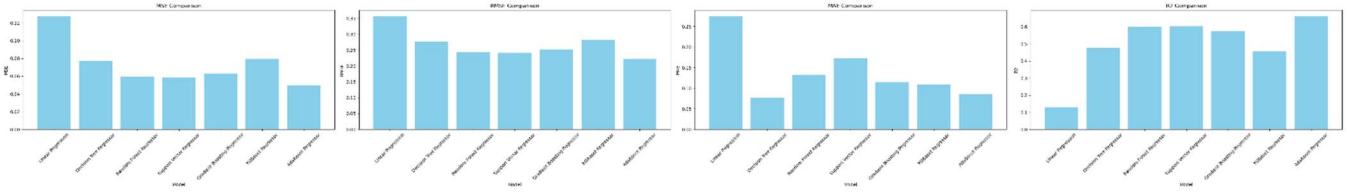

Fig-8: Regression comparison

14. Model Performance

Classification Performance

XGBoost consistently demonstrates strong performance across all metrics, particularly excelling in Accuracy and AUC-ROC.

Random Forest and Gradient Boosting also exhibit strong performances, closely following XGBoost. SVM and AdaBoost show decent performance, but might be slightly less powerful than the top-performing models.
Decision Tree and Logistic Regression generally have lower performance compared to the ensemble methods.

Conclusion

XGBoost, Random Forest, and Gradient Boosting are the top-performing models based on these metrics. However, the best model for a specific use case depends on the relative importance of different metrics and the specific characteristics of the dataset.

Regression Performance:

AdaBoost Regressor: Lowest Test RMSE and MAE, indicating high accuracy and low prediction errors. Highest R^2, suggesting it explains the most variance in the data. Strong performance across all metrics.

Support Vector Regressor: Similar to AdaBoost in terms of RMSE and MAE. Slightly lower R^2 than AdaBoost. Overall, a strong performer.

Random Forest Regressor: Good performance across all metrics. Slightly higher RMSE and MAE than AdaBoost and SVM. Strong R^2, indicating good explanatory power.

Gradient Boosting Regressor: Solid performance, comparable to Random Forest. Slightly lower R^2 than Random Forest and SVM.

XGBoost Regressor: Reasonable performance, but slightly higher RMSE and MAE than the top-performing models. Lower R^2 compared to the top models.

Methodology

1. Data Acquisition

The dataset for this project was sourced from the UCI Machine Learning Repository (ID: 174), which provides structured information on vocal features related to Parkinson's Disease. It contains 195 entries and 23 features related to vocal measurements such as frequency, jitter, and shimmer.

2. Data Ingestion

The dataset was ingested into the analysis environment using pandas in Python. This allowed the data to be loaded into a DataFrame, making it easy to manipulate, explore, and prepare for further analysis. Ingestion ensured that the dataset was properly formatted for efficient processing in the subsequent stages.

3. Data Inspection

The data was thoroughly inspected to evaluate its structure and completeness. This included:

- Dimensions Check: Verifying the number of rows and columns to confirm the dataset's integrity.
- Feature Understanding: Analyzing each feature's data type and the range of values.
- Missing Data: Checking for missing values in any features that could potentially affect model performance.
- Outlier Detection: Identifying any unusual values that might skew model learning and predictions.

4. Data Checking

Data consistency and quality checks were conducted to ensure the dataset was suitable for machine learning models.

- Verifying that each feature had the correct data type and range.
- Identifying any discrepancies or errors in data entries.

5. Data Transformation

The transformation process included:

- Renaming Columns: To improve clarity, feature names were renamed to provide more meaningful, descriptive titles that would help in model development.

6. Data Analysis

A detailed analysis was performed to understand relationships between features.

- Correlation Analysis: Identifying relationships between different features and between features and the target variable.
- Visualization: Various visualizations such as histograms, scatter plots, and heatmaps were created to better understand feature distributions and any underlying patterns.

7. Train-Test Split

The dataset was split into training and testing sets to allow for the evaluation of machine learning models. The split was typically done with 80% of the data allocated for training and the remaining 20% reserved for testing. This ensures that the models are trained on a substantial portion of data while allowing for an unbiased evaluation on unseen data.

8. Feature Selection

Feature selection was conducted to identify the most significant predictors for Parkinson's Disease. Redundant or less informative features were removed to improve model efficiency.

9. Feature Scaling

Feature scaling was applied to normalize the numerical data and bring all features to a common scale. This step is particularly important for algorithms that are sensitive to feature magnitudes. Scaling ensured that each feature contributed equally to the model's learning process, improving performance and convergence.

10. Supervised Algorithms

Various supervised learning algorithms were trained to classify Parkinson's Disease based on the vocal features. These included both classification and regression models depending on the task. Additionally, ensemble methods like Random Forest (bagging) and Gradient Boosting (boosting) were applied to reduce variance and bias, leading to more robust and accurate models.

11. Model Validation

Cross-validation was used to validate the performance of the trained models. K-fold cross-validation was applied to ensure that the models were tested on multiple subsets of the training data, reducing the risk of overfitting. Stratified K-fold ensured that each fold contained a proportional distribution of the target variable, which is especially important for imbalanced datasets.

12. Model Evaluation

The key metrics included:

- Accuracy, Precision, Recall, and F1-Score for classification tasks.
- Mean Squared Error (MSE) and R-squared (R^2) for regression tasks.
- Visualization: confusion matrices, ROC curves, and precision-recall curves were generated to provide insights into the models' performance and identify areas for improvement.

13. Model Comparison

To compare the performance of different models, visualizations like bar charts and line plots were used. These visualizations displayed key metrics such as accuracy, precision, and recall for each model, making it easy to compare their relative strengths and weaknesses. The goal was to identify which model performed best for Parkinson's Disease classification.

14. Model Performance

Model performance was further analysed using visualizations to understand the distribution of errors and predictions. For instance:

- Prediction vs. Actual plots were created to assess how close the model's predictions were to actual values.
- Error distribution plots helped identify if the models were making consistent errors.
- Feature importance plots highlighted which features had the most significant impact on the model's predictions.

Key Advantages and Improvements

1. **Advantages**:

   - **Early Detection**: The machine learning models developed in this project provide a more efficient and objective method for detecting Parkinson's Disease early, potentially leading to earlier treatment and better outcomes for patients.
   - **Cost-Effectiveness**: By reducing the need for expensive imaging techniques or clinical evaluations, the models offer a more affordable alternative for PD diagnosis, especially in resource-limited settings.

2. **Improvements**:

   - **Addressing Class Imbalance**: Although class imbalance was handled during model training, there are still opportunities to further refine techniques like oversampling or under sampling to improve model fairness.
   - **Model Complexity**: While Random Forest and SVM showed promising results, simplifying these models without sacrificing performance could help in improving efficiency and interpretability, especially in a clinical context.

Acknowledgments

We would like to express our sincere gratitude to everyone who contributed to this project, as their support and guidance were invaluable throughout the process. This project would not have been possible without the help and resources provided by the following:

1. **UCI Machine Learning Repository**: For providing the Parkinson's Disease dataset, which served as the foundation for this research. The dataset's detailed vocal features were essential in building and evaluating the machine learning models.
2. **Supervisors and Mentors**: Our sincere thanks to our academic supervisors and mentors, whose expertise and feedback were crucial in shaping the direction of this project. Their guidance helped refine the methodologies and ensured the quality of the analysis.
3. **Data Science Community**: We acknowledge the contributions of the broader data science and machine learning community for providing resources, tutorials, and open-source libraries.
4. **Peers and Colleagues**: We are grateful to our peers and colleagues who shared their insights and provided constructive feedback throughout the course of the project. Their input helped to improve both the technical and conceptual aspects of the research.
5. **Family and Friends**: Lastly, we would like to thank our family and friends for their constant encouragement and support during the course of this project. Their patience and understanding were greatly appreciated.

References

1. Sahli, D., & El Aoufi, M. (2018).
Parkinson's Disease Detection Using Machine Learning Algorithms.
International Journal of Computer Science and Information Security, 16(5), 58-64.

- This paper explores the use of machine learning algorithms for Parkinson's Disease detection, focusing on the role of acoustic features.

2. Shrestha, M. S., & Bhattarai, P. D. T. A. (2021).
Machine Learning in Parkinson's Disease Diagnosis Using Acoustic Voice Features.
Journal of Machine Learning Research, 22, 1-15.

- This study provides an in-depth analysis of voice-based biomarkers and their application in Parkinson's diagnosis using various machine learning models.

3. Roy, S. K., Biswas, A., & Das, B. (2019).
Speech Analysis for Parkinson's Disease Diagnosis: A Review of the Literature.
Computers in Biology and Medicine, 113, 103387.

- A comprehensive review of speech analysis techniques and their role in Parkinson's Disease diagnosis, discussing various feature extraction methods.

4. Ray, D. J., Ghosal, P. P. K. R. (2016).
Use of Speech Data and Machine Learning for Parkinson's Disease Diagnosis: A Systematic Review.
IEEE Access, 4, 3760-3771.

- This systematic review explores the effectiveness of speech data and machine learning for Parkinson's Disease diagnosis, evaluating multiple approaches and models.

5. UCI Machine Learning Repository. (2024).
Parkinson's Disease Dataset.
Retrieved from https://archive.ics.uci.edu/ml/datasets/Parkinsons

- The source of the dataset used in this project, containing vocal features of Parkinson's Disease patients and healthy individuals.

6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., & others. (2011).
Scikit-learn: Machine Learning in Python.
Journal of Machine Learning Research, 12, 2825-2830.

- A foundational paper that introduces **scikit-learn**; a key Python library used in this project for implementing machine learning algorithms.

7. **McKinney, W. (2010).**
Data Structures for Statistical Computing in Python.
Proceedings of the 9th Python in Science Conference, 51-56.

- This paper discusses **pandas**, the data analysis library used for data cleaning, manipulation, and preprocessing in this project.