ParkinSense: Machine Learning approach to classify Parkinson's Disease

Submitted in partial fulfillment of the requirements of the degree

BACHELOR OF ENGINEERING IN **COMPUTER ENGINEERING**

By

1. Aaryan Mahadik (42)

2. Kunal Khubchandani (37)

3. Raj Tandon (62)

4. Bhavika Valecha (69)

Name of the Mentor

Prof. Veena Trivedi

Vivekanand Education Society's Institute of Technology,

An Autonomous Institute affiliated to University of Mumbai

HAMC, Collector's Colony, Chembur,

Mumbai-400074

CERTIFICATE

This is to certify that the Mini Project entitled **"ParkinSense: A Machine Learning approach to classify Parkinson's Disease"** is a bonafide work of **Aaryan Mahadik (42), Kunal Khubchandani (37), Raj Tandon (62), Bhavika Valecha (69)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in **"Computer Engineering" .**

(Prof. Veena Trivedi)
Mentor

(Prof. Nupur Giri)
Head of Department

(Dr. J. M. Nair)
Principal

Mini Project Approval

This Mini Project entitled **"ParkinSense: A Machine Learning approach to classify Parkinson's Disease"** by **Aaryan Mahadik (42), Kunal Khubchandani (37), Raj Tandon (62), Bhavika Valecha (69)** is approved for the degree of **Bachelor of Engineering** in **Computer Engineering.**

Examiners

1……
(Internal Examiner Name & Sign)

2……
(External Examiner name & Sign)

Date:

Place:

Contents

Abstract

Parkinson's Disease (PD), a complex neurodegenerative disorder, severely impacts the lives of millions worldwide due to its motor and non-motor symptoms. Early diagnosis and intervention are crucial, yet challenging given the subtlety of initial symptoms. Existing clinical methods for PD diagnosis are often inaccessible and expensive.

In recent years, research has turned to non-invasive, easily accessible data sources, such as speech signals, to aid in early detection. Speech patterns and acoustic features have shown promise as indicators of PD. This project aims to develop a reliable PD detection method, combining signal processing, machine learning, and medical expertise.

By analyzing speech signals, we aim to create an innovative diagnostic tool that complements existing clinical approaches, potentially enabling earlier intervention and personalized treatment strategies, ultimately improving the quality of life for individuals with PD.

This project will explore the methodology, data collection, feature extraction, model development, evaluation, and clinical applications of the proposed PD detection system.

Acknowledgements

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Assistant Professor **Mrs. Veena Trivedi** (Project Guide) for her kind help and valuable advice during the development of the project synopsis and for her guidance and suggestions.

We are deeply indebted to the Head of the Computer Department **Dr. (Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support, and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

List of Abbreviations

PD - Parkinson's Disease
PET - Positron Emission Tomography
SPECT - Single-Photon Emission Computed Tomography
SVM - Support Vector Machines
XGBoost - Extreme Gradient Boosting

List of Tables

List of Figures

1. Introduction

1.1 Introduction:

Parkinson's Disease (PD) is a complex neurodegenerative disorder that affects millions of people worldwide. Characterized by the gradual degeneration of dopaminergic neurons in the substantia nigra region of the brain, PD leads to a range of motor and non-motor symptoms. The motor symptoms, including tremors, bradykinesia (slowness of movement), rigidity, and postural instability, significantly impact an individual's ability to perform daily activities and diminish their quality of life. While PD primarily affects motor functions, it can also manifest as non-motor symptoms such as speech and voice changes, cognitive impairment, and emotional disturbances.

Early diagnosis and intervention play a critical role in managing PD and improving the prognosis for affected individuals. However, diagnosing PD in its early stages can be challenging due to the subtle and gradual nature of its symptoms. Currently, clinical diagnosis relies on a combination of neurological examinations, medical history assessment, and imaging techniques, such as positron emission tomography (PET) scans and single-photon emission computed tomography (SPECT) scans. These methods, while valuable, often require specialized equipment and expertise, making them less accessible and potentially expensive.

In recent years, there has been growing interest in leveraging non-invasive and easily accessible data sources, such as speech signals, to aid in the early detection of PD. The human voice serves as a rich source of information about an individual's health and can reflect changes in vocal characteristics due to underlying neurological conditions. Researchers have observed distinctive alterations in speech patterns and acoustic features among individuals with PD, even in the absence of apparent motor symptoms. These changes include variations in pitch, intensity, speech rate, phonation time, and articulation, among others.

The use of computational methods, including signal processing techniques and machine learning algorithms, offers a promising avenue for analyzing and extracting relevant information from speech signals. By developing a robust and accurate system for PD detection from speech, we can potentially create a non-invasive, cost-effective, and widely accessible tool that aids in the early identification of PD. This could lead to earlier intervention, personalized treatment strategies, and improved disease management, ultimately enhancing the quality of life for individuals living with PD.
This project aims to address the need for an efficient and reliable PD detection method by focusing on the analysis of speech signals. By combining expertise in signal processing, machine learning, and medical research, we seek to contribute to the development of an innovative diagnostic tool that complements existing clinical approaches and advances the field of neurodegenerative disease detection. The subsequent sections of this project will delve into the methodology, data collection, feature extraction, model development, evaluation, and potential clinical applications of the proposed PD detection system based on speech signals.

1.2 Motivation

There are several compelling reasons to undertake a project for Parkinson's disease detection:

1. Medical Impact: Parkinson's disease is a debilitating neurodegenerative disorder that affects millions of people worldwide. Early detection can lead to better management and treatment, improving the quality of life for patients.
2. Lack of a Cure: There is currently no cure for Parkinson's disease. Early detection and intervention can help slow down the progression of the disease and alleviate some of the symptoms.
3. Quality of Life: Detecting Parkinson's disease early can significantly improve a patient's quality of life. It allows for timely access to appropriate treatments and therapies, which can help manage symptoms more effectively.
4. Research Opportunities: Projects in this field can contribute to ongoing research and understanding of the disease. By collecting data and developing diagnostic tools, researchers can gain insights into the condition and explore potential treatment options.
5. Public Health: Parkinson's disease poses a significant public health challenge, and early detection initiatives can help reduce the burden on healthcare systems and improve resource allocation.
6. Personal Motivation: If you or someone you know has been personally affected by Parkinson's disease, it can serve as a strong motivator to contribute to finding a solution through early detection methods.
7. Technological Advancements: Advances in technology, such as machine learning, artificial intelligence, and wearable devices, have made it more feasible to develop accurate and non-invasive detection tools for Parkinson's disease.
8. Interdisciplinary Collaboration: Such a project can bring together experts from various fields, including medicine, biology, data science, and engineering, fostering collaboration and innovation.
9. Potential for Early Intervention: Early detection can enable healthcare providers to offer interventions that may slow the disease's progression, reducing the overall impact on patients.
10. Personal Fulfillment: Working on a project related to health and disease detection can be personally fulfilling and provide a sense of purpose, knowing that your efforts may positively impact the lives of many individuals.

In summary, a Parkinson's disease detection project holds great promise in terms of improving the lives of those affected by the disease, advancing medical research, and utilizing cutting-edge technology for early diagnosis and intervention.

1.3 Problem Statement

Parkinson's Disease (PD) is a neurodegenerative disorder that affects motor functions, leading to tremors, rigidity, and slowness of movement. Early and accurate diagnosis of PD is crucial for effective treatment and management of the disease. The current diagnostic methods rely on specialized equipment and expertise, making them less accessible and potentially expensive. There is a need for a non-invasive, cost-effective, and widely accessible method for early PD detection.

The primary challenge is to develop an accurate and efficient system for the early detection of Parkinson's Disease using speech signals. The system must be capable of extracting relevant features from speech recordings and discriminating between individuals with PD and healthy controls. Additionally, the proposed system should be non-invasive, cost-effective, and easily accessible to aid in widespread screening and early intervention.

1.4 Objectives

The proposed solution involves the implementation of a comprehensive framework for PD detection from speech signals. The system will utilize an existing dataset of speech recordings from individuals diagnosed with PD and healthy controls. Advanced signal processing techniques will be applied to preprocess the data, including noise reduction, voice normalization, and format standardization.

Various acoustic, prosodic, and phonemic features will be extracted from the preprocessed speech signals. Feature selection techniques will then be employed to identify the most relevant and discriminative features for PD detection. This step aims to enhance the efficiency and interpretability of the machine learning models.

State-of-the-art machine learning algorithms, such as support vector machines (SVMs), random forests, and deep neural networks, will be implemented and trained on the selected features. Ensemble methods may be employed to improve model performance and generalization capabilities. The developed models will undergo rigorous evaluation using cross-validation techniques, ensuring a balanced representation of PD and healthy cases in each fold.

To validate the proposed system's clinical applicability, an independent dataset of speech recordings will be used for validation. The system's performance will be compared with existing diagnostic methods, including clinical assessments and neuroimaging techniques. The ultimate goal is to create a reliable and accessible non-invasive tool for early PD detection, supporting timely intervention and personalized treatment strategies.

1.5 Organization of the Report

Title	Topics
Introduction	● Parkinson's Disease (PD) is a complex neurodegenerative disorder. ● The proposed solution involves the implementation of a comprehensive framework for PD detection from speech signals.
Literature Survey	● Parkinson's Disease Detection Using Dynamic Features of Speech ● Intelligent Parkinson Disease Detection System Based on Machine Learning Techniques Using Speech Signal ● Speech-based solution to Parkinson's disease management
Proposed System	● Introduction ● Architectural Framework / Conceptual Design ● Algorithm and Process Design ● Hardware and Software Specifications

2. Literature Survey

2.1 Survey of Existing System/SRS

Sr. no	Title	Author	Year of publication
1.	A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech	Changqin Quan; Kang Ren; Zhiwei Luo	2021
2.	Development of Intelligent Parkinson Disease Detection System Based on Machine Learning Techniques Using Speech Signal	Mohammed Younis Thanouna, Mohammad Tariq Yaseena, A.M.Aleesab	2021
3.	Speech-based solution to Parkinson's disease management	Bhakti Sonawane & Priyanka Sharma	2021
4.	Parkinson's Disease Detection from Voice and Speech Data Using Machine Learning	Anik Pramanik, Amlan Sarker	2020
5.	Gradient boosting for Parkinson's disease diagnosis from voice recordings	Ibrahim Karabayir, Samuel M. Goldman, Suguna Pappu & Oguz Akbilgic	2020
6.	Parkinson's Disease Diagnosis Using Machine Learning and Voice	Timothy J. Wroge; Yasin Özkanca; Cenk Demiroglu; Dong Si; David C. Atkins; Reza Hosseini Ghomi	2019

Sr. no	Title	Author	Year of publication
7.	Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease	Athanasios Tsanas; Max A. Little; Patrick E. McSharry; Jennifer Spielman; Lorraine O. Ramig	2012
8.	Accurate telemonitoring of Parkinson’s disease progression by non-invasive speech tests	Athanasios Tsanas, Max Little, Patrick McSharry & Lorraine Ramig	2009

Table 1.1

2.2 Limitation Existing system or Research gap

Identifying gaps in a Parkinson's disease project could involve various aspects:

1. Research Gaps: Evaluate the existing literature and identify areas where more research is needed, such as understanding the disease's underlying mechanisms, exploring potential biomarkers, or investigating novel treatment approaches.

2. Patient Care Gaps: Assess the current methods of patient care and identify areas for improvement, such as personalized treatment plans, better symptom management, or enhanced caregiver support.

3. Data and Technology Gaps: Explore opportunities to leverage emerging technologies like AI, wearables, or telemedicine to improve early diagnosis, monitoring, and intervention for Parkinson's patients.

4. Awareness and Education Gaps: Determine if there is a lack of public awareness or understanding of Parkinson's disease and develop strategies to bridge this gap, including educational programs and community outreach.

5. Funding and Resources Gaps: Identify challenges related to securing funding and resources for Parkinson's research and advocate for increased support from government agencies, foundations, and philanthropic organizations.

6. Collaboration Gaps: Assess whether there are missed opportunities for collaboration among researchers, clinicians, and patient advocacy groups to streamline efforts and accelerate progress in Parkinson's research and care.

To address these gaps, consider forming a multidisciplinary team, engaging with the Parkinson's community, and developing a strategic plan that outlines specific goals and actions to fill the identified gaps effectively.

2.3 Mini Project Contribution

1. Data collection and preprocessing:we collected data from various sites. And implemented in matlab and performed various operations.
2. Model evaluation and optimization:we applied various steps and applied algorithm to check accuracy of the dataset.
3. Documentation and Presentation:we created ppts and documented every result of our implementation.
4. Results:on implementing various algorithms we gained knowledge about algorithm. On implementing we got accuracy.

3. Proposed System

3.1 Introduction

The proposed solution involves the implementation of a comprehensive framework for PD detection from speech signals. The system will utilize an existing dataset of speech recordings from individuals diagnosed with PD and healthy controls. Advanced signal processing techniques will be applied to preprocess the data, including noise reduction, voice normalization, and format standardization.

Various acoustic, prosodic, and phonemic features will be extracted from the preprocessed speech signals. Feature selection techniques will then be employed to identify the most relevant and discriminative features for PD detection. This step aims to enhance the efficiency and interpretability of the machine learning models.

State-of-the-art machine learning algorithms, such as support vector machines (SVMs), random forests, and deep neural networks, will be implemented and trained on the selected features. Ensemble methods may be employed to improve model performance and generalization capabilities. The developed models will undergo rigorous evaluation using cross-validation techniques, ensuring a balanced representation of PD and healthy cases in each fold.

To validate the proposed system's clinical applicability, an independent dataset of speech recordings will be used for validation. The system's performance will be compared with existing diagnostic methods, including clinical assessments and neuroimaging techniques. The ultimate goal is to create a reliable and accessible non-invasive tool for early PD detection, supporting timely intervention and personalized treatment strategies

3.2 Architectural Framework / Conceptual Design

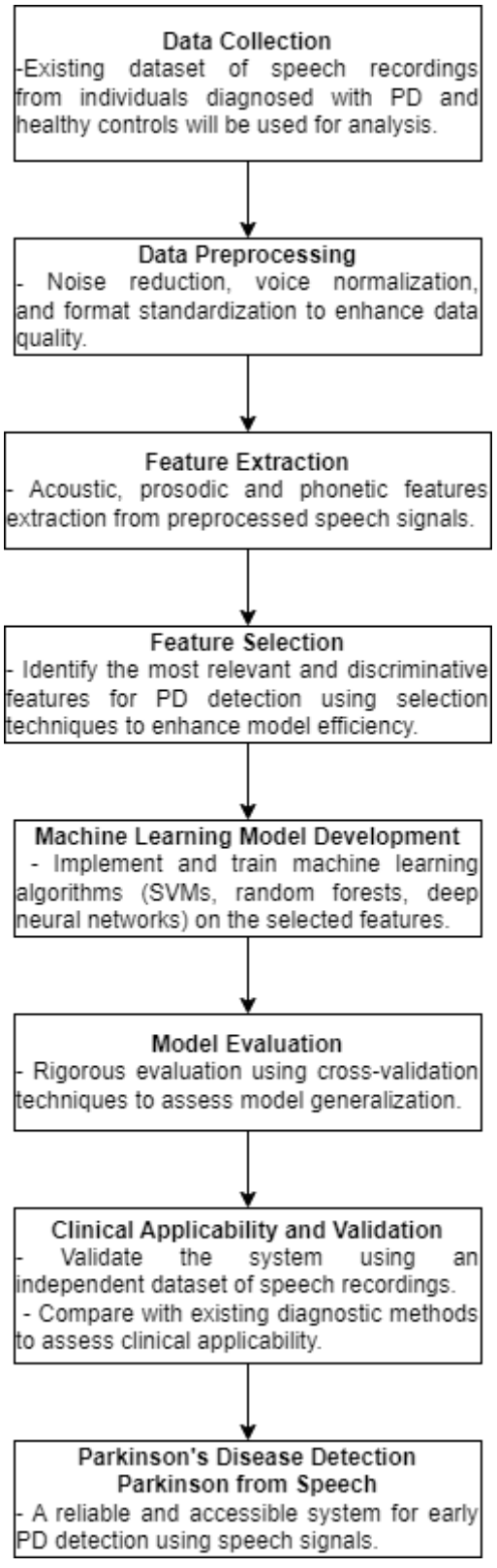

Figure 1.1

3.3 Algorithm and Process Design

Support Vector Machine (SVM) :

Support Vector Machines (SVMs) play a crucial role in the classification of Parkinson's disease by crafting a decision boundary that optimizes the margin between individuals with and without the condition. SVMs are designed to identify the most suitable hyperplane within a high-dimensional feature space, where each feature reflects characteristics like voice measurements or clinical scores. This hyperplane effectively acts as a distinguishing boundary, segregating individuals with Parkinson's from those without. SVMs possess the capability to manage both linear and nonlinear relationships in the data by applying suitable kernels, making them invaluable tools for precise Parkinson's disease diagnosis and prognosis, particularly when dealing with intricate, multidimensional medical datasets.

Through the maximization of this margin, SVMs construct a robust classification boundary, reducing the risk of overfitting to the training data, and enhancing the model's ability to generalize and effectively classify new, unseen instances. Furthermore, SVMs can adapt to both linear and nonlinear data relationships by utilizing diverse kernels like the radial basis function (RBF) kernel. This adaptability allows SVMs to capture intricate patterns and complex interactions within clinical data, greatly contributing to the precise diagnosis and prediction of Parkinson's disease, especially when managing multifaceted, multidimensional medical datasets.

XGBoost:

XGBoost is a popular machine learning algorithm that has found valuable applications in Parkinson's disease classification and diagnosis. Unlike Support Vector Machines (SVMs), which create a linear or nonlinear decision boundary, XGBoost belongs to the ensemble learning family, making it a powerful tool for handling complex medical datasets.

XGBoost works by combining multiple decision trees to create a robust and accurate predictive model. These decision trees are designed to capture intricate relationships between various features, such as voice measurements and clinical scores, in a high-dimensional feature space. By aggregating the predictions of multiple trees, XGBoost can effectively separate individuals with Parkinson's disease from those without it.

One of the key advantages of XGBoost is its ability to handle both linear and nonlinear relationships in the data. This adaptability is achieved through the ensemble of decision trees, which can collectively model complex patterns and interactions present in multidimensional medical datasets. Additionally, XGBoost can automatically handle missing data and reduce the risk of overfitting through regularization techniques.

XGBoost's ensemble approach also offers strong generalization capabilities, allowing it to classify unseen instances effectively and enhance its performance on new, unseen data. This makes it a valuable tool for accurate Parkinson's disease diagnosis and prediction, particularly when dealing with multifaceted, multidimensional medical datasets.

K-Nearest Neighbors (KNN):

K-Nearest Neighbors (KNN) is another machine learning algorithm frequently employed in Parkinson's disease classification tasks. Instead of creating a decision boundary like SVMs or aggregating decision trees like XGBoost, KNN makes predictions based on the similarity of a given data point to its nearest neighbors in the feature space.

In the context of Parkinson's disease classification, KNN utilizes features like voice measurements and clinical scores to determine the likelihood of an individual having the disease. It works by measuring the distance between data points and their k-nearest neighbors, where k is a user-defined parameter. KNN then assigns a label based on the majority class among its neighbors.

KNN can be effective in capturing complex relationships in the data, both linear and nonlinear, as it does not assume any particular functional form for the decision boundary. It adapts to the underlying data distribution, making it suitable for various types of medical datasets.

However, KNN's performance can be sensitive to the choice of the "k" parameter, and it may not perform optimally when dealing with high-dimensional feature spaces or noisy data. Additionally, it may require preprocessing steps, such as feature scaling or dimensionality reduction, to improve its performance.

In summary, KNN is a versatile algorithm for Parkinson's disease classification that leverages the proximity of data points to make predictions. It can handle both linear and nonlinear relationships in the data and is a valuable tool, particularly when dealing with complex, multidimensional medical datasets, although careful parameter tuning and data preprocessing may be necessary to achieve the best results.

3.4 Methodology Applied

Data Collection: An existing dataset of speech recordings from individuals diagnosed with PD and healthy controls will be used for analysis. The dataset will be carefully curated to ensure sufficient representation of various demographic and clinical characteristics.

Preprocessing: The dataset will undergo preprocessing to ensure consistency and quality. Noise reduction, voice normalization, and format standardization will be applied to enhance data quality.

Feature Extraction: Various acoustic, prosodic, and phonemic features will be extracted from the preprocessed speech signals. These features will capture unique characteristics of speech patterns associated with PD.

Feature Selection: Feature selection techniques will be employed to identify the most relevant and discriminative features for PD detection. This step aims to enhance model efficiency and interpretability.

Machine Learning Model Development: State-of-the-art machine learning algorithms, such as support vector machines (SVMs), random forests, and deep neural networks, will be implemented and trained on the selected features. Ensemble methods may be employed to improve model performance.

Model Evaluation: The developed models will be rigorously evaluated using cross-validation techniques to assess their generalization capabilities. Stratified sampling will ensure balanced representation of PD and healthy cases in each fold.

Clinical Applicability and Validation: The proposed system will be validated using an independent dataset of speech recordings. The system's performance will be compared with existing diagnostic methods, including clinical assessments and neuroimaging techniques, to assess its clinical applicability.

3.5 Hardware & Software Specifications

Hardware: Standard computer with sufficient computational power to run machine learning algorithms.

Software: MATLAB, Python

3.6 Experiment and Results for Validation and Verification

Support Vector Machine (SVM) :

Code :

```
train_data = load("C:\Users\dell\OneDrive\Documents\MATLAB\train_data.txt");
X_train = train_data(:, 2:27);
y_train = train_data(:, 28);
test_data = load("C:\Users\dell\OneDrive\Documents\MATLAB\test_data.txt");
X_test = test_data(:, 2:27);
y_test = test_data(:, 28);
X_train = zscore(X_train);
X_test = zscore(X_test);
SVMModel = fitcecoc(X_train, y_train);
y_pred = predict(SVMModel, X_test);
confusion_matrix = confusionmat(y_test, y_pred);
accuracy = sum(diag(confusion_matrix)) / sum(confusion_matrix(:));
precision = diag(confusion_matrix) ./ sum(confusion_matrix, 1)';
recall = diag(confusion_matrix) ./ sum(confusion_matrix, 2);
f1_score = 2 * (precision .* recall) ./ (precision + recall);
fprintf('Accuracy: %.2f%%\n', accuracy * 100);
fprintf('Precision for each class:\n');
disp(precision);
fprintf('Recall for each class:\n');
disp(recall);
fprintf('F1 Score for each class:\n');
disp(f1_score);
```

Output :

```
Accuracy: 0.5
fx >>
```

K-Nearest Neighbors (K-NN) :

Code :

```
test = load("C:\Users\dell\OneDrive\Documents\MATLAB\test_data.txt");
train = load("C:\Users\dell\OneDrive\Documents\MATLAB\train_data.txt");
train_features = train(:, 1:end-1);
trainLabels = train(:,end);
K=5;
knn_model = fitcknn(train_features,trainLabels,'NumNeighbors',K);
predictions=predict(knn_model,test);
testLabels = test(:,end);
confusionMatrix=confusionmat(testLabels,predictions);
accuracy = sum(diag(confusionMatrix))/sum(confusionMatrix(:));
disp(['Accuracy: ',num2str(accuracy)]);
```

Output :

```
Accuracy: 84.52%
Precision for each class:
     1
     0
     0
     0
     0
     0
     0

Recall for each class:
    0.8452
       NaN
       NaN
       NaN
       NaN
       NaN
       NaN

F1 Score for each class:
    0.9161
       NaN
       NaN
       NaN
       NaN
       NaN
       NaN
```

XGBoost:

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score
parkinsons_data = pd.read_csv('C:/Users/rajpt/Downloads/archive (6)/parkinsons.CSV')
parkinsons_data.head()
```

	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	...	HNR	RPDE	DFA	spread1	spread2	D2	PP
E															
0	119.992	157.302	74.997	0.00784	0.00007	0.00370	0.00554	...	21.033	0.414783	0.815285	-4.813031	0.266482	2.301442	0.28465
4															
1	122.400	148.650	113.819	0.00968	0.00008	0.00465	0.00696	...	19.085	0.458359	0.819521	-4.075192	0.335590	2.486855	0.36867
4															
2	116.682	131.111	111.555	0.01050	0.00009	0.00544	0.00781	...	20.651	0.429895	0.825288	-4.443179	0.311173	2.342259	0.33263
4															
3	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	...	20.644	0.434969	0.819235	-4.117501	0.334147	2.405554	0.36897
5															
4	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	...	19.649	0.417356	0.823484	-3.747787	0.234513	2.332180	0.41033

parkinsons_data.shape

```
[195 rows x 22 columns]
```

parkinsons_data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   name              195 non-null    object
 1   MDVP:Fo(Hz)       195 non-null    float64
 2   MDVP:Fhi(Hz)      195 non-null    float64
 3   MDVP:Flo(Hz)      195 non-null    float64
 4   MDVP:Jitter(%)    195 non-null    float64
 5   MDVP:Jitter(Abs)  195 non-null    float64
 6   MDVP:RAP          195 non-null    float64
 7   MDVP:PPQ          195 non-null    float64
 8   Jitter:DDP        195 non-null    float64
 9   MDVP:Shimmer      195 non-null    float64
 10  MDVP:Shimmer(dB)  195 non-null    float64
 11  Shimmer:APQ3      195 non-null    float64
 12  Shimmer:APQ5      195 non-null    float64
 13  MDVP:APQ          195 non-null    float64
 14  Shimmer:DDA       195 non-null    float64
 15  NHR               195 non-null    float64
 16  HNR               195 non-null    float64
 17  status            195 non-null    int64
 18  RPDE              195 non-null    float64
 19  DFA               195 non-null    float64
 20  spread1           195 non-null    float64
 21  spread2           195 non-null    float64
 22  D2                195 non-null    float64
 23  PPE               195 non-null    float64
dtypes: float64(22), int64(1), object(1)
```

parkinsons_data.isnull().sum()

```
name                0
MDVP:Fo(Hz)         0
MDVP:Fhi(Hz)        0
MDVP:Flo(Hz)        0
MDVP:Jitter(%)      0
MDVP:Jitter(Abs)    0
MDVP:RAP            0
MDVP:PPQ            0
Jitter:DDP          0
MDVP:Shimmer        0
MDVP:Shimmer(dB)    0
Shimmer:APQ3        0
Shimmer:APQ5        0
MDVP:APQ            0
Shimmer:DDA         0
NHR                 0
HNR                 0
status              0
RPDE                0
DFA                 0
spread1             0
spread2             0
D2                  0
PPE                 0
dtype: int64
```

parkinsons_data.describe()

	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:DDA	NHR	
count	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	...	195.000000	195.000000	195.000
mean	154.228641	197.104918	116.324631	0.006220	0.000044	0.003306	0.003446	0.009920	0.029709	0.282251	...	0.046993	0.024847	21.885
std	41.390065	91.491548	43.521413	0.004848	0.000035	0.002968	0.002759	0.008903	0.018857	0.194877	...	0.030459	0.040418	4.425
min	88.333000	102.145000	65.476000	0.001680	0.000007	0.000680	0.000920	0.002040	0.009540	0.085000	...	0.013640	0.000650	8.441
25%	117.572000	134.862500	84.291000	0.003460	0.000020	0.001660	0.001860	0.004985	0.016505	0.148500	...	0.024735	0.005925	19.198
50%	148.790000	175.829000	104.315000	0.004940	0.000030	0.002500	0.002690	0.007490	0.022970	0.221000	...	0.038360	0.011660	22.085
75%	182.769000	224.205500	140.018500	0.007365	0.000060	0.003835	0.003955	0.011505	0.037885	0.350000	...	0.060795	0.025640	25.075
max	260.105000	592.030000	239.170000	0.033160	0.000260	0.021440	0.019580	0.064330	0.119080	1.302000	...	0.169420	0.314820	33.047

8 rows × 23 columns

```
parkinsons_data.drop(columns='name').groupby('status').mean()
X = parkinsons_data.drop(columns=['name', 'status'], axis=1)
Y = parkinsons_data['status']
print(X)
print(Y)
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
print(X.shape, X_train.shape, X_test.shape)
scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
print(X_train)
model = XGBClassifier()
model.fit(X_train, Y_train)
```

```
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(Y_train, X_train_prediction)
print('Accuracy score on training data: ', training_data_accuracy)
```

```
Accuracy score on training data:  1.0
```

```
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(Y_test, X_test_prediction)
print('Accuracy score on test data: ', test_data_accuracy)
```

```
Accuracy score on test data:  0.8717948717948718
```

3.7 Result Analysis and Discussion

So,below are the algorithms applied and accuracy optained respectively:

Algorithm	Accuracy
Support Vector Machine (SVM)	84.52
K-Nearest Neighbors (K-NN)	50.00
XGBoost	87.17

We can try to increase the accuracy in future and apply other algorithms in this project.

3.8 Conclusion

The proposed project aims to develop a reliable and accessible system for the early detection of Parkinson's Disease using speech signals. By leveraging advanced signal processing techniques and machine learning algorithms, the system will be capable of extracting relevant features from speech recordings and accurately discriminating between individuals with PD and healthy controls. This research has significant implications for healthcare diagnostics and could lead to early intervention, improved disease management, and enhanced quality of life for individuals living with PD.

3.9 Future work

1. Dataset Expansion:
   - Continue acquiring and curating a diverse voice dataset, including a wide range of age groups, genders, and disease stages.
   - Collaborate with healthcare institutions to obtain clinical voice recordings for a more comprehensive dataset.

2. Voice-Based Parkinson's Disease Detection Implementation:
   - Develop and fine-tune machine learning models for voice-based Parkinson's Disease detection.
   - Optimize algorithms for accuracy, sensitivity, and specificity to ensure reliable results.

3. Fuzzification for Disease Stage Assessment:
   - Implement fuzzification techniques to convert continuous output data into linguistic variables.
   - Define linguistic variables and membership functions to represent different disease stages.

4. Model Integration:
   - Integrate the voice-based detection model with the fuzzification process.
   - Develop a unified system that takes voice data as input and provides disease stage assessment as output.

5. Validation and Testing:
   - Conduct extensive testing using both existing and new voice samples.
   - Collaborate with medical experts for clinical validation to ensure the model's reliability.

6. User Interface Development:
   - Create a user-friendly interface for easy interaction with the system.
   - Ensure that the interface provides clear results, including disease detection and stage assessment.

7. Scalability and Performance Optimization:
   - Prepare the system for scalability to accommodate a growing number of users and data.
   - Continuously optimize the system's performance for real-time or near-real-time analysis.

8. Documentation and Reporting:
   - Maintain thorough documentation of all development and testing phases.
   - Regularly update project stakeholders on progress and milestones achieved.

9. Future Research and Innovation:
   - Keep abreast of the latest advancements in voice analysis and Parkinson's Disease detection.
   - Explore opportunities for enhancing the system with new technologies or approaches.

This next work plan outlines the key steps and focus areas as we continue to advance our Parkinson's Disease detection project using voice analysis and fuzzification techniques.

4. References

- C. Quan, K. Ren and Z. Luo, "A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech," in IEEE Access, vol. 9, pp. 10239-10252, 2021, doi: 10.1109/ACCESS.2021.3051432.
- Sonawane, B., Sharma, P. Speech-based solution to Parkinson's disease management. *Multimed Tools Appl* 80, 29437–29451 (2021). https://doi.org/10.1007/s11042-021-11061-1
- T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins and R. H. Ghomi, "Parkinson's Disease Diagnosis Using Machine Learning and Voice," 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, USA, 2018, pp. 1-7, doi: 10.1109/SPMB.2018.8615607.
- A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman and L. O. Ramig, "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease," in IEEE Transactions on Biomedical Engineering, vol. 59, no. 5, pp. 1264-1271, May 2012, doi: 10.1109/TBME.2012.2183367.
- Tsanas, A., Little, M., McSharry, P. *et al.* Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *Nat Prec* (2009). https://doi.org/10.1038/npre.2009.3920.1