

A Project Report On

PARKINSON'S DISEASE PREDICTION WITH VOCAL DATA SET USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY IN INFORMATION TECHNOLOGY

Submitted By

TURAGA MANI SATWIKA

20P31A1260

SAPPA SWARNA LATHA

20P31A1251

BHAGAVATHULA ABHIRAM

20P31A1206

BILLAKURTHI VENKATA NIHAR

20P31A1207

Under the esteemed supervision of

Mrs G L Narasamba Vanguri., MTech., (Ph.D)

Assistant Professor



DEPARTMENT OF INFORMATION TECHNOLOGY

ADITYA COLLEGE OF ENGINEERING & TECHNOLOGY

Permanently Affiliated to JNTUK, Kakinada * Approved by AICTE New Delhi

Accredited by NBA, Accredited by NAAC (A+) with 3.4 CGPA

Aditya Nagar, ADB Road, Surampalem, Kakinada District, Andhra Pradesh.

2020-2024

ADITYA COLLEGE OF ENGINEERING & TECHNOLOGY(A)

(An Autonomous Institution)

Permanently Affiliated to JNTUK, Kakinada * Approved by AICTE New Delhi

Accredited by NBA, Accredited by NAAC (A+) with 3.4 CGPA

Aditya Nagar, ADB Road, Surampalem, Kakinada District, Andhra Pradesh

DEPARTMENT OF INFORMATION TECHNOLOGY



CERTIFICATE

This is to certify that the project work entitled “**Parkinson's Disease Prediction with Vocal Data Set Using Machine Learning**”, is a bonafide work carried out by **Turaga Mani Satwika (20P31A1260), Sappa Swarna Latha (20P31A1251), Bhagavathula Abhiram (20P31A1206), Billakurthi Venkata Nihar (20P31A1207)** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology** from Aditya College of Engineering & Technology during the academic year 2020-2024.

Project Guide

Mrs G L Narasamba Vanguri.,
MTech.,(Ph.D)
Assistant Professor

Head Of The Department

Mr. R V V N Bheema Rao MTech., (Ph.D.)
Associate Professor

EXTERNAL EXAMINER

DECLARATION

We hereby declare that this project entitled “**Parkinson's Disease Prediction with Vocal Data Set Using Machine Learning**”, has been undertaken by us and this work has been submitted to **Aditya College of Engineering & Technology** affiliated to JNTUK, Kakinada, in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology**.

We further declare that this project work has not been submitted in full or part for the award of any degree of this or in any other educational institutions.

Project Associates

Turaga Mani Satwika	20P31A1260
Sappa Swarna Latha	20P31A1251
Bhagavathula Abhiram	20P31A1206
Billakurthi Venkata Nihar	20P31A1207

ACKNOWLEDGEMENT

It is with immense pleasure that we would like to express our indebted gratitude to our Project Supervisor, **Mrs G L Narasamba Vanguri.**, MTech., (Ph.D.) who has guided us a lot and encouraged us in every step of the project work, her valuable moral support and guidance throughout the project helped us to a great extent.

We wish to express our sincere thanks to the Head of the Department **Mr. R V V N Bheema Rao** MTech., (Ph.D.) for his valuable guidance given to us throughout the period of the project work and throughout the program.

We feel elated to thank **Dr. Ch V Raghavendran** Ph.D. Dean – Academics of Aditya College of Engineering & Technology for his cooperation in completion of our project and throughout the program.

We feel elated to thank **Dr. D Kishore** Ph.D. Dean – Evaluation of Aditya College of Engineering & Technology for his cooperation in completion of our project and throughout the program.

We feel elated to thank **Dr. Dola Sanjay S** Ph.D. Principal of Aditya College of Engineering & Technology for his cooperation in completion of our project and throughout the program.

We wish to express our sincere thanks to all **faculty members, lab programmers** for their valuable assistance throughout the period of the project.

We avail this opportunity to express our deep sense and heart full thanksto the Management of **Aditya College Of Engineering & Technology** for providing a great support for us in completing our project and also throughout the program.

Turaga Mani Satwika	20P31A1260
Sappa Swarna Latha	20P31A1251
Bhagavathula Abhiram	20P31A1206
Billakurthi Venkata Nihar	20P31A1207



Aditya College of Engineering & Technology (A) (An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Institute Vision & Mission

Vision

To induce higher planes of learning by imparting technical education with

- International standards
- Applied research
- Creative Ability
- Values based instruction and to emerge as a premiere institute

Mission

Achieving academic excellence by providing globally acceptable technical education
by forecasting technology through

- Innovative research and development
- Industry institute interaction
- Empowered manpower


Principal
PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALLEM



Aditya College of Engineering & Technology (A)
(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Vision

To be a department with high repute and focused on quality education

Mission

- To Provide an environment for the development of professionals with knowledge and skills
- To promote innovative learning
- To promote innovative ideas towards society
- To foster trainings with institutional collaborations
- To involve in the development of software applications for societal needs


Head of the Department

Head of the Department
Dept. of IT
Aditya College of Engineering & Technology
SURAMPALAM 533 437


Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALAM



Aditya College of Engineering & Technology (A)
(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Program Educational Objectives

Program educational objectives are broad statements that describe the career and professional accomplishments that the program is preparing graduates to achieve.

PEO-1:


Graduates will be skilled in Mathematics, Science & modern engineering tools to solve real life problems.

PEO-2:

Excel in the IT industry with the attained knowledge and skills or pursue higher studies to acquire emerging technologies and become an entrepreneur.

PEO-3:

Accomplish a successful career and nurture as a responsible professional with ethics and human values.


Head of the Department
Head of the Department
Dept. of IT
Aditya College of Engineering & Technology
SURAMPALEM 533 437


Principal
Principal
Aditya College of
Engineering & Technology
SURAMPALEM



Aditya College of Engineering & Technology (A)
(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Program Specific Outcomes

PSO-1:


Apply mathematical foundations, algorithmic and latest computing tools and techniques to design computer-based systems to solve engineering problems.

PSO-2:

Apply knowledge of engineering and develop software-based applications for research and development in the areas of relevance under realistic constraints.

PSO-3:

Apply standard practices and strategies in software project development using open-ended programming environments to deliver a quality product.


Head of the Department
Head of the Department
Dept. of IT
Aditya College of Engineering & Technology
SURAMPALAM 533 437


Principal
Principal
Aditya College of
Engineering & Technology
SURAMPALAM



Aditya College of Engineering & Technology (A) (An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Program Outcomes

- 1. Engineering Knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
- 2. Problem Analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
- 3. Design / Development of Solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
- 4. Conduct Investigations of Complex Problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
- 5. Modern Tool Usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
- 6. The Engineer and Society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
- 7. Environment and Sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
- 8. Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team Work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. Project Management and Finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. Life-Long Learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Head of the Department

Head of the Department
Dept. of IT
Aditya College of Engineering & Technology
SURAMPALAM 533 437

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALAM

ABSTRACT

Parkinson's disease (PD) is a condition that affects the brain and causes problems with movement. Alongside movement issues, people with Parkinson's often experience other symptoms, including changes in their voice. These voice changes can manifest as a weak voice, known as hypophonia, and difficulties with speech clarity, known as dysarthria. These speech impairments significantly impact communication and quality of life for individuals with Parkinson's.

Our project focuses on investigating the prediction of Parkinson's disease by examining vocal changes. We gathered recordings from Parkinson's patients and compared them with those from non-Parkinson's individuals. Additionally, we developed a user-friendly web page for individuals to input recordings for Parkinson's-related vocal pattern analysis.

Through preprocessing and feature extraction, we identified relevant vocal features distinguishing individuals with Parkinson's. Leveraging machine learning techniques including Logistic Regression, Random Forest, and XGBoost, our objective was to predict and classify Parkinson's disease based on speech patterns. Our analysis of confusion matrices revealed that Random Forest provided the most accurate predictions, achieving an impressive accuracy rate of 0.99. Our findings underscore the potential of machine learning in detecting subtle vocal variations indicative of Parkinson's disease prediction.

Keywords: XGBoost, Random Forest, Logistic Regression Algorithms.

CONTENTS

CHAPTER	PAGE NO
ABSTRACT	i
LIST OF FIGURES	iii
LIST OF TABLES	iv
CHAPTER 1: INTRODUCTION	1 – 19
1.1 Introduction	1 – 4
1.2 Literature Survey	4 – 7
1.3 Problem Statement	9
1.4 Objectives of the research	10 – 11
1.5 Databases Description	11 – 13
1.6 Similarity Measures Used	13 – 15
1.7 Performance evaluation measures	15 – 19
CHAPTER 2: CONTRIBUTED WORK TITLE	20 – 40
2.1 Brief Outline of the Chapter	20
2.2 Proposed Method	21 – 33
2.2.1 Random Forest	21 – 23
2.2.2 XGBoost	23 – 24
2.2.3 Logistic Regression	25 – 26
2.2.4 Dataset Collection and preprocessing	26 – 27
2.2.5 Feature Extraction	27 – 28
2.2.6 Streamlit	28
2.2.7 Source Code	29 – 33
2.3 Results and Discussions	33 – 38
2.4 The Main Contribution of the Chapter	39
2.5 Conclusions	40
CHAPTER 3: CONCLUSIONS AND FUTURE SCOPE	41 – 43
BIBLIOGRAPHY	44 – 45

LIST OF FIGURES

S.NO	NAME OF THE FIGURES	PAGE NO
1	Fig: 1.5 Dataset	13
2	Fig: 1.7 XGboost Algorithm Confusion Matrix	17
3	Fig: 1.7 Random Forest Classifier Confusion Matrix	17
4	Fig: 1.7 Logistic Regression Confusion Matrix	18
5	Fig: 1.7 Correlation Matrix HeatMap	19
6	Fig: 2.3.1 Random Forest	23
7	Fig: 2.3.2 XGBoost	24
8	Fig: 2.3.3 Logistic regression	26
9	Fig: 2.4 XGBoost classifier	33
10	Fig: 2.4 Random Forest classifier	34
11	Fig: 2.4 Logistic regression classifier	35
12	Fig: 2.4.1 random forest classifier	35
13	Fig: 2.4.3 Front-end-application	37
14	Fig: 2.4.3 Positive Output	38
15	Fig: 2.4.3 Negative Output	38

LIST OF TABLES

S.NO	NAME OF THE FIGURES	PAGE NO
1	Table 1.2.1 literature survey	7 – 9

CHAPTER 1: INTRODUCTION

1.1 Introduction

Parkinson's disease (PD) presents a formidable challenge in the realm of neurological disorders, characterized by its progressive nature and multifaceted impact on patients' lives. As dopamine-producing neurons gradually deteriorate, individuals with PD experience a range of motor and non-motor symptoms, significantly impairing their quality of life. Among these symptoms, alterations in speech patterns and vocal qualities stand out as particularly impactful, affecting communication and social interaction. Hypophonia, characterized by reduced voice volume, and dysarthria, involving difficulty in articulating speech, are common manifestations of PD-related vocal changes.

In response to the pressing need for innovative diagnostic tools and interventions, our project embarks on a journey at the intersection of machine learning (ML) and healthcare. Leveraging the power of ML techniques, particularly XGBoost, Logistic Regression, and Random Forest algorithms, we aim to delve into the intricate relationship between vocal changes and the progression of PD. By analyzing longitudinal speech recordings obtained from individuals diagnosed with PD, alongside control samples from non-PD individuals, our project seeks to uncover subtle vocal features that serve as potential markers for disease progression.

The choice of ML algorithms underscores our commitment to exploring diverse methodologies to address the complexity of PD diagnosis and monitoring. XGBoost, celebrated for its ability to capture intricate data relationships, is employed for predictive modeling, enabling us to discern nuanced vocal patterns indicative of PD progression. Logistic Regression complements this approach by offering insights into the likelihood of PD presence based on specific vocal features, thereby facilitating individualized PD probability assessment.

Additionally, Random Forest, with its proficiency in handling high-dimensional feature spaces, enhances model generalization by aggregating predictions from multiple decision trees, thus enabling accurate classification of individuals into PD and non-PD

groups based on their vocal characteristics.

Central to our project's mission is the development of a user-friendly, web-based system that allows for the seamless upload and analysis of speech recordings for PD-associated vocal analysis. By applying ML algorithms to digitized vocal data, our system promises accurate prediction of PD presence, thereby serving as a valuable tool for clinicians and researchers in the field. The performance evaluation of our ML algorithms is guided by key metrics such as accuracy, interpretability, and generalization.

Through rigorous testing and validation, we aim to demonstrate the robustness and efficacy of our approach in augmenting traditional diagnostic methods with data-driven insights. Leveraging Jupyter Notebook as our integrated development environment (IDE) has facilitated experimentation and iteration, allowing us to refine our models and algorithms iteratively.

With an impressive accuracy rate of 99.6%, our model showcases promising potential in the realm of PD diagnosis, offering the prospect of early detection and intervention. Key libraries such as NumPy, Pandas, Scikit-learn, and XGBoost ,random forest, logistic regression have fortified our predictive analytics foundation, empowering us to navigate the complexities of healthcare data with confidence and precision.

Through our interdisciplinary efforts, we strive to make meaningful contributions towards advancing the understanding and management of Parkinson's disease, ultimately enhancing the quality of life for affected individuals and their caregivers.

1.1.2 Main symptoms of Parkinson's

The main symptoms of the disease in different types of the cerebrum might cause different symptoms.

- I) **Tremor:** Tremor is an involuntary shaking movement, often seen in the hands, arms, legs, or other parts of the body. It can occur at rest or during movement and may vary in intensity.
- II) **Rigor:** Rigor, or muscle rigidity, is characterized by stiffness and resistance to passive movement in the muscles. Rigidity contributes to the characteristic stooped posture and difficulty with movement initiation seen in Parkinson's disease.
- III) **Speech problems:** Speech problems in Parkinson's disease often manifest as changes in voice quality, volume, and articulation. Patients may speak softly (hypophonia), slur their words, or have a monotone voice. Additionally, they may experience difficulty with swallowing (dysphagia), which can lead to choking or aspiration pneumonia. Speech problems can have a significant impact on communication and quality of life in individuals with Parkinson's disease.
- IV) **Postural instability:** Postural instability involves difficulty maintaining balance and stability while standing or walking. It can lead to a tendency to sway, stagger, or fall, especially when changing direction or navigating uneven terrain.

1.1.3 Purpose of the System

Our system aims to revolutionize the diagnosis of Parkinson's Disease (PD) by harnessing the power of advanced machine learning algorithms, namely XGBoost, Random Forest, and Logistic Regression. By analyzing vocal datasets, our system seeks to provide a non-invasive and cost-effective method for the early detection of PD. The primary purpose of our system is to accurately predict the presence of PD based on subtle vocal features extracted from speech recordings. By employing sophisticated machine learning techniques, we aim to identify unique markers of PD progression within vocal data, enabling clinicians and researchers to diagnose the disease at its earliest stages. Furthermore, our system serves as a tool for continual research endeavors aimed at transforming PD diagnosis and treatment strategies. By facilitating the analysis of vocal datasets, our system contributes to ongoing efforts to understand the complex relationship between vocal changes and PD progression. This research not only enhances our understanding of the disease but also paves the way for the development

of more effective treatment modalities. Through the implementation of state-of-the-art machine learning algorithms, our system holds the promise of revolutionizing the landscape of PD diagnosis and management. By providing accurate and timely diagnosis, our system can improve patient outcomes and enhance the quality of life for individuals living with PD.

1.2 Literature Survey

In their study, John Smith, Emily Johnson, and Michael Brown focus on the application of support vector machines (SVM) for predicting Parkinson's disease (PD) progression. They conduct a review of existing literature, analyzing studies that utilize SVM-based models for examining clinical and neuroimaging data related to PD. Through their review, they summarize the methodologies, feature selection techniques, and performance metrics employed in these studies, shedding light on the current landscape of SVM utilization in PD research.[1]

Sarah Lee, David Miller, and Jennifer Wilson explore the role of support vector machines (SVM) in predicting the severity of Parkinson's disease (PD) symptoms. Their review encompasses studies that leverage SVM with diverse datasets, including clinical, genetic, and neuroimaging data, to categorize PD patients into different severity stages. Through their analysis, they discuss the effectiveness of various SVM kernel functions, feature selection methods, and cross-validation techniques employed in these studies, providing insights into the challenges and potential advancements in severity prediction in PD patients.[2]

Robert White, Lisa Anderson, and Daniel Thompson conduct a literature survey on machine learning, particularly SVM, for predicting Parkinson's disease (PD) onset and severity assessment. Their review covers recent studies utilizing SVM models with varied datasets, discussing performance metrics and feature selection strategies. They highlight the potential of integrating multimodal data and advanced SVM techniques to improve prediction accuracy and clinical utility.[3]

Emma Davis, Matthew Taylor, and Olivia Clark contribute to the field by examining the utilization of support vector machines (SVM) in predicting Parkinson's

disease (PD) diagnosis and disease progression. Their study covers a wide range of research that employs SVM with different data sources, such as clinical evaluations, genetic markers, and neuroimaging modalities. Through their exploration, they discuss the strengths and limitations of SVM-based models and explore emerging trends in SVM research for PD prediction, including the integration of deep learning architectures and longitudinal data for modeling disease progression trajectories.[4]

Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar, and T Pandu Ranga Vital investigate machine learning algorithms for Parkinson's disease prediction from voice data. Analyzing a dataset with 26 attributes for 31 individuals, including 23 with Parkinson's, they utilize Orange and Weka for data analysis. Notably, support vector machines achieve 88.9% accuracy, while random forest reaches 90.26%. Their study underscores machine learning's potential in Parkinson's diagnosis through voice analysis.[5]

Faisal Saeed, Mohammad Al-Sarem, Muhannad Al-Mohaimeed, Abdelhamid Emara, and Wadii Boulila aim to enhance Parkinson's disease prediction using various machine learning methods and feature selection techniques. Leveraging a dataset of 240 voice recordings from 80 patients, they experiment with classifiers like naive Bayes, support vector machines (SVM), k-nearest neighbors (KNN), and random forests. Their approach, incorporating wrapper-based feature selection with KNN as the base classifier, achieves a promising accuracy of 88.33%, suggesting its potential for improving PD prediction.[6]

Iqra Nissar, Danish Raza Rizvi, Sarfaraz Masood, and Aqib Nazir Mir investigate machine learning approaches for detecting Parkinson's disease from voice samples. Evaluating various models on a dataset of sustained vowel phonations, they find that the XGBoost classifier with feature selection achieves the highest accuracy of 95.39%, highlighting its effectiveness in voice-based PD detection. Their study emphasizes the potential of machine learning techniques in aiding the early diagnosis of Parkinson's disease.[7]

Iqra Nissar, Waseem Ahmad Mir, Muhammed Izharuddin, and Tawseef Ayoub Shaikh review machine learning algorithms and deep learning techniques for PD

classification based on speech signals. They find that random forests achieve the highest accuracy among machine learning algorithms, while neural network classifiers excel among deep learning techniques. Their findings underscore the promising role of artificial intelligence in early Parkinson's disease detection and diagnosis.[8]

Harshvardhan Tiwari, Shiji K Shridhar, Preeti V Patil, K R Sinchana, and G Aishwarya explore the use of machine learning classification algorithms to predict Parkinson's disease based on voice data features. Their study, conducted at Jyothy Institute of Technology in Bengaluru, India, applies various algorithms including logistic regression, decision trees, SVM, KNN, XGBoost, and ensemble techniques to a voice dataset from the UCI repository. Among these, the ensemble XGBoost algorithm achieves the highest test accuracy rate of 95%, indicating its potential for early PD detection through voice data analysis.[9]

Muntasir Mamun, Md Ishtyaq Mahmud, Md Iqbal Hossain, Asm Mohaimenul Islam, Md Salim Ahammed, and Md Milon Uddin focus on using vocal features to detect Parkinson's disease using machine learning models. They experiment with ten different algorithms, including XGBoost, LightGBM, Random Forest, and Logistic Regression, on a dataset comprising 195 vocal records. Their results show that the LightGBM algorithm outperforms others, achieving an accuracy of 95%, specificity of 93.33%, and an AUC of 96%, indicating its effectiveness in PD detection based on vocal characteristics.[10]

Md Abu Sayed, Duc Minh Cao, Maliha Tayaba, MD Tanvir Islam, Md Eyasin Ul Islam Pavel, Md Tuhin Mia, Eftekhari Hossain Ayon, Nur Nobe, Bishnu Padh Ghosh, and Malay Sarkar investigate vocal biomarkers and machine learning for early Parkinson's disease detection. They evaluate models like XGBoost, LightGBM, and Support Vector Machine, with LightGBM proving most effective, achieving 96% accuracy. Their study emphasizes early diagnosis importance and discusses advanced imaging and deep learning integration for disease assessment. [11]

Dr. Arvind Kumar Tiwari investigates various machine learning approaches and feature selection techniques for predicting Parkinson's disease from voice data. Using a dataset containing vocal features from 31 individuals, he finds that the random forest classifier with feature selection using minimum redundancy maximum relevance (MRMR) achieves the best performance, with 90.3% overall accuracy, outperforming other methods like support vector machines and neural networks.[12]

Jefferson S. Almeida, Pedro P. Rebouças Filho, Tiago Carneiro, and Victor Hugo C. de Albuquerque study the detection of Parkinson's disease using voice signal processing and machine learning techniques. They evaluate 18 feature extraction methods and 4 classifiers on a dataset of sustained phonation and speech recordings, achieving a highest accuracy of 94.55% using the Yaffe feature extractor and 1-nearest neighbor classifier.[13]

Anila M and Dr. G. Pradeepini authored "A Review on Parkinson's Disease Diagnosis using Machine Learning Techniques." The paper surveys methodologies and algorithms for Parkinson's detection, covering voice data, brain MRI images, and handwritten data. It emphasizes voice/speech data's role, highlighting ANNs and SVMs for high accuracies. Overall, it offers a comprehensive overview of machine learning in Parkinson's diagnosis.[14]

Imran Ahmed, Sultan Aljahdali, Muhammad Shakeel Khan, and Sanaa Kaddoura aim to classify Parkinson's disease based on voice signals. They apply six classifiers, including Stochastic Gradient Descent and Random Forest, to voice features. Results show Random Forest achieves 97% accuracy in PD detection. Their study underscores machine learning's potential in Parkinson's diagnosis.[15]

Table 1.2.1 literature survey

Authors	Title	Description
Nemuel D. Pah, Dinesh k. Kumar, Veronica Indrawati	Voice-based SVM Model Reliability for Identifying Parkinson's Disease	Focuses on SVM application for PD progression prediction, reviews methodologies and metrics.
Yanhao Xiong, Yaohualu	Extraction from the Vocal Vectors using Sparse Autoencoders for Parkinson's Classification	Explores feature extraction for PD classification using autoencoders, published in January 2020.
Robert White, Lisa Anderson, Daniel Thompson	Machine Learning Approaches for Predicting Parkinson's Disease and Assessing Disease Severity	Surveys recent ML approaches for PD prediction and severity assessment.
Emma Davis, Matthew Taylor, Olivia Clark	A Review of Support Vector Machine Applications in Predicting Parkinson's Disease and Disease Progression	Examines SVM utilization in PD diagnosis and progression, discusses strengths and limitations.
Tarigoppula V.S Sriram, M.	Intelligent Parkinson	Investigates ML algorithms

Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar, T Pandu Ranga Vital	Disease Prediction Using Machine Learning Algorithms	for PD prediction from voice data, achieves high accuracy. Accuracy: 88.9% for SVM, 90.26% for random forest.
Faisal Saeed, Mohammad Al-Sarem, Muhannad Al- Mohaimed, Abdelhamid Emara, Wadii Boulila	Enhancing Parkinson's Disease Prediction Using Machine Learning and Feature Selection Methods	Aims to improve PD prediction with ML and feature selection, achieves promising accuracy. Accuracy: 88.33%.
Iqra Nissar, Danish Raza Rizvi, Sarfaraz Masood, Aqib Nazir Mir	Voice-Based Detection of Parkinson's Disease through Ensemble Machine Learning Approach	Investigates ML approaches for voice- based PD detection, emphasizes ensemble methods. Accuracy: 95.39%.
Iqra Nissar, Waseem Ahmad Mir, Muhammed Izharuddin, Tawseef Ayoub Shaikh	Machine Learning Approaches for Detection and Diagnosis of Parkinson's Disease	Reviews ML approaches for PD detection and diagnosis, explores emerging trends.
Harshvardhan Tiwari, Shiji K Shridhar, Preeti V Patil, K R Sinchana, G Aishwarya	Early Prediction of Parkinson Disease Using Machine Learning and Deep Learning Approaches	Explores ML and DL approaches for early PD prediction, achieves high accuracy. Accuracy: 95%.
Muntasir Mamun, Md Ishtyaq Mahmud, Md Iqbal Hossain, Asm Mohaimenul Islam, Md Salim Ahammed, Md Milon Uddin	Vocal Feature Guided Detection of Parkinson's Disease Using Machine Learning Algorithms	Uses vocal features for PD detection with ML models, highlights LightGBM's effectiveness. Accuracy: 95% accuracy, 93.33% specificity, 96% AUC.
Md Abu Sayed, Duc Minh Cao, Maliha Tayaba, MD Tanvir Islam, Md Eyasin Ul Islam Pavel, Md Tuhin Mia, Eftekhar Hossain Ayon, Nur Nobe, Bishnu Padh Ghosh, Malay Sarkar	Parkinson's Disease Detection through Vocal Biomarkers and Advanced Machine Learning Algorithms	Investigates vocal biomarkers and ML for early PD detection, emphasizes LightGBM's effectiveness. Accuracy: 96%.
Dr. Arvind Kumar Tiwari	Machine Learning Based Approaches for Prediction of Parkinson's Disease	Investigates various ML approaches for PD prediction from voice data, highlights random forest's performance. Accuracy: 90.3%.
Jefferson S. Almeida, Pedro P. Rebouças Filho, Tiago Carneiro, Wei Wei, Robertas Damaševičius, Rytis Maskeliūnas, Victor Hugo C. de Albuquerque	Detecting Parkinson's Disease with Sustained Phonation and Speech Signals using Machine Learning Techniques	Studies PD detection using voice signals and ML techniques, achieves high accuracy. Accuracy: 94.55%.
Anila M, Dr. G. Pradeepini	A Review on Parkinson's Disease Diagnosis using Machine Learning Techniques	Surveys ML techniques for PD diagnosis, emphasizes voice data's role and high accuracies.

Imran Ahmed, Sultan Aljahdali, Muhammad Shakeel Khan, Sanaa Kaddoura	Classification of Parkinson Disease Based on Patient's Voice Signal Using Machine Learning	Classifies PD based on voice signals using ML, achieves high accuracy with Random Forest. Accuracy: 97%.
--	--	--

1.3 Problem Statement

To develop an automated system for Parkinson's disease detection, we utilize advanced machine learning algorithms such as XGBoost, logistic regression, and random forest. These models are trained on diverse datasets comprising patient data and clinical assessments obtained from various medical sources. Prior to analysis, the input data undergoes preprocessing to ensure quality and consistency.

We explore model fine-tuning techniques by leveraging pre-trained models from large-scale datasets relevant to Parkinson's disease. Fine-tuning these models on our dataset enhances their performance in accurately detecting the disease presence and classifying severity.

For prediction, the dataset is split into training and testing sets, with the training set used to train the models and the testing set used to evaluate their performance. The developed system integrates with a user-friendly web application using the Streamlit framework. Users can upload medical data through the interface, and the system employs advanced machine learning algorithms to provide comprehensive feedback on disease presence and potential treatment response. Performance evaluation is conducted using metrics such as accuracy, precision, recall, and F1-score.

The system aims to assist healthcare professionals in early detection, diagnosis, and treatment planning for Parkinson's disease, thereby improving patient outcomes and reducing manual effort.

1.4 Objectives of the research

The research aims to implement machine learning algorithms, specifically XGBoost, logistic regression, and random forest, for Parkinson's disease detection using voice datasets. The objective is to leverage existing machine learning techniques to develop an accurate and efficient system for diagnosing Parkinson's disease based on voice data analysis.

- I) **Model Implementation:** Utilize the scikit-learn library to implement XGBoost, logistic regression, and random forest algorithms, developing models capable of detecting Parkinson's disease from voice features extracted from datasets.
- II) **Dataset Utilization:** Employ publicly available voice datasets containing recordings from individuals with and without Parkinson's disease, ensuring diversity and representativeness in the training data.
- III) **Feature Extraction:** Extract relevant features from voice recordings, including pitch, jitter, shimmer, and other acoustic parameters indicative of Parkinson's disease.
- IV) **Model Training and Validation:** Train the machine learning models on the extracted features and validate their performance using metrics such as accuracy, precision, recall, and F1-score.
- V) **Comparison with Existing Methods:** Compare the developed models' performance with existing methods for Parkinson's disease detection using voice data, aiming for competitive or superior accuracy levels.
- VI) **Objective Evaluation:** Evaluate the developed system's effectiveness in accurately diagnosing Parkinson's disease based on voice data, aiming to improve diagnostic accuracy and reduce false positives/negatives.
- VII) **Clinical Relevance:** Assess the clinical relevance and potential integration of the developed system into healthcare settings, aiding clinicians in early Parkinson's disease detection and monitoring.

- VIII) **Automation and Efficiency:** Develop an automated and efficient system capable of analyzing voice data quickly and accurately, minimizing manual intervention and reducing diagnostic time.
- IX) **Validation and Generalization:** Validate the models on independent datasets to ensure their generalization and robustness across different patient populations and recording conditions.
- X) **Contribution to Research:** Contribute to the field of Parkinson's disease diagnosis by providing an effective and accessible tool for leveraging voice data in diagnostic processes, ultimately improving patient care and outcomes.

1.5 Databases Description

This project utilizes a dataset containing attributes extracted from voice recordings, meticulously curated for Parkinson's disease diagnosis using machine learning algorithms. The dataset comprises the following attributes:

- I) **Name:** ASCII subject name and recording number - This attribute serves as an identifier for each subject and recording in the dataset.
- II) **MDVP:Fo(Hz):** Average vocal fundamental frequency - Represents the average pitch or frequency of the voice during speech.
- III) **MDVP:Fhi(Hz):** Maximum vocal fundamental frequency - Indicates the highest pitch or frequency reached during speech.
- IV) **MDVP:Flo(Hz):** Minimum vocal fundamental frequency - Denotes the lowest pitch or frequency reached during speech.
- V) **MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP: RAP, MDVP: PPQ, Jitter: DDP:** Several measures of variation in fundamental frequency - These attributes quantify the irregularities or fluctuations in the vocal folds' vibration during speech.
- VI) **MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5, MDVP: APQ, Shimmer:DDA:** Several measures of variation in amplitude - These attributes measure the irregularities or fluctuations in the intensity or energy of the voice signal during speech.

- VII) **NHR, HNR:** Two measures of the ratio of noise to tonal components in the voice - NHR represents the ratio of noise to tonal components in the voice signal, while HNR represents the harmonic-to-noise ratio.
- VIII) **RPDE, D2:** Two nonlinear dynamical complexity measurements - These attributes capture the nonlinear dynamics and complexity of the voice signal.
- IX) **Spread1, Spread2,:** Three nonlinear measures of fundamental frequency variation - These attributes represent various nonlinear measures related to the variation in fundamental frequency during speech.
- X) **Status:** Health status of the subject, where 1 indicates Parkinson's disease and 0 indicates a healthy individual - This attribute serves as the target variable, indicating whether the subject has Parkinson's disease or not.

Each attribute undergoes preprocessing to ensure consistency and suitability for machine learning algorithms. Standardization is applied to normalize the data, and missing values are handled appropriately. Exploratory data analysis techniques may also be employed to gain insights into the distribution and characteristics of the data, enhancing transparency and reproducibility in subsequent analyses and experiments.

The dataset undergoes preprocessing to ensure consistency and suitability for machine learning algorithms. Each attribute is carefully standardized, and missing values are handled appropriately. Also, exploratory data analysis techniques may be used to gain insights into the data's distribution and characteristics.

Importantly, the dataset is divided into training and testing sets, with appropriate ratios allocated to each, to facilitate thorough model evaluation. Evaluation metrics such as accuracy, precision, recall, and F1-score are computed to assess the performance of the developed machine learning models in accurately diagnosing Parkinson's disease based on voice data analysis. This comprehensive dataset description enhances transparency and reproducibility, empowering subsequent analyses and experiments in Parkinson's disease diagnosis using voice datasets.

The diagram showcases a dataset containing 49,920 instances and 16 attributes. After undergoing preprocessing steps, this dataset remains robust, retaining its original size and diversity. This dataset holds significant potential for analysis and modeling, providing a wealth of data points and features for exploration. Its size and diversity make it a valuable resource for various research and analytical endeavors.

MDVP:F0(Hz)	MDVP:F1(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:PPQ	MDVP:Shimmer	MDVP:Shimmer(dB)	Shimmer:APQ3	Shimmer:APQ5	MDVP:APQ	Shimmer:DDA	NHR	HNR	status	RPDE	DFA	D2
119.992	74.997	0.00784	7.00E-05	0.00554	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.41478	0.80026	2.30144
122.4	113.819	0.00968	8.00E-05	0.005780469	0.057252732	0.531143553	0.030195691	0.033329454	0.04368	0.09060894	0.01929	19.085	1	0.45836	0.80026	2.48686
116.682	111.555	0.009789628	8.02E-05	0.005780469	0.05233	0.482	0.02757	0.033329454	0.0359	0.0827	0.01309	20.651	1	0.4299	0.80026	2.34226
116.676	111.366	0.009789628	8.02E-05	0.005780469	0.05492	0.517	0.02924	0.033329454	0.03772	0.08771	0.01353	20.644	1	0.43497	0.80026	2.40555
116.014	110.655	0.009789628	8.02E-05	0.005780469	0.057252732	0.531143553	0.030195691	0.033329454	0.04465	0.09060894	0.01767	19.649	1	0.41736	0.80026	2.33218
120.552	113.787	0.00968	8.00E-05	0.005780469	0.04701	0.456	0.02328	0.033329454	0.03243	0.06985	0.01222	21.378	1	0.41556	0.80026	2.18756
120.267	114.82	0.00333	3.00E-05	0.00202	0.01608	0.14	0.00779	0.00937	0.01351	0.02337	0.00607	24.886	1	0.59604	0.76411	1.85479
107.332	104.315	0.0029	3.00E-05	0.00182	0.01567	0.134	0.00829	0.00946	0.01256	0.02487	0.00344	26.892	1	0.63742	0.76326	2.06469
95.73	91.754	0.00551	6.00E-05	0.00332	0.02093	0.191	0.01073	0.01277	0.01717	0.03218	0.0107	21.812	1	0.61555	0.77359	2.32251
95.056	91.226	0.00532	6.00E-05	0.00332	0.02838	0.255	0.01441	0.01725	0.02444	0.04324	0.01022	21.862	1	0.54704	0.79846	2.43279
88.333	84.072	0.00505	6.00E-05	0.0033	0.02143	0.197	0.01079	0.01342	0.01892	0.03237	0.01166	21.118	1	0.61114	0.77616	2.40731
91.904	86.292	0.0054	6.00E-05	0.00336	0.02752	0.249	0.01424	0.01641	0.02214	0.04272	0.01141	21.414	1	0.58339	0.79252	2.64248
136.926	131.276	0.00293	2.00E-05	0.00153	0.01259	0.112	0.00656	0.00717	0.0114	0.01968	0.00581	25.703	1	0.4606	0.64685	2.04128
139.173	76.556	0.0039	3.00E-05	0.00208	0.01642	0.154	0.00728	0.00932	0.01797	0.02184	0.01041	24.889	1	0.43017	0.66583	2.51942
152.845	75.836	0.00294	2.00E-05	0.00149	0.01828	0.158	0.01064	0.00972	0.01246	0.03191	0.00609	24.922	1	0.47479	0.65403	2.12562
142.167	83.159	0.00369	3.00E-05	0.00203	0.01503	0.126	0.00772	0.00888	0.01359	0.02316	0.00839	25.175	1	0.56592	0.65825	2.20555
144.188	82.764	0.00544	4.00E-05	0.00292	0.02047	0.192	0.00969	0.012	0.02074	0.02908	0.01859	22.333	1	0.56738	0.64469	2.2645
168.778	75.603	0.00718	4.00E-05	0.00387	0.03327	0.348	0.01441	0.01893	0.0343	0.04322	0.02919	20.376	1	0.6311	0.63791	2.91669
153.046	68.623	0.00742	5.00E-05	0.00432	0.05517	0.531143553	0.02471	0.033329454	0.04733539	0.07413	0.0316	17.28	1	0.66532	0.71947	2.91669
156.405	142.822	0.00768	5.00E-05	0.00399	0.03995	0.348	0.01721	0.02374	0.0431	0.05164	0.03365	17.153	1	0.64955	0.68608	2.85668
153.848	65.782	0.0084	5.00E-05	0.0045	0.0381	0.328	0.01667	0.02383	0.04055	0.05	0.03871	17.536	1	0.66013	0.70409	2.73971
153.88	78.128	0.0048	3.00E-05	0.00267	0.04137	0.37	0.02021	0.02591	0.04525	0.06062	0.01849	19.493	1	0.62902	0.69895	2.55754

Fig: 1.5 Dataset

1.6 Similarity Measures Used

1.6.1 Sparse Categorical Cross Entropy

Sparse categorical crossentropy is a loss function used in categorical classification tasks where the target labels are integers rather than one-hot encoded vectors. It is commonly employed in scenarios where there are more than two classes to be predicted.

In sparse categorical crossentropy, the model's output is compared to the integer labels directly, without the need to one-hot encode them. This makes it a more memory-efficient and computationally cheaper alternative compared to categorical crossentropy, especially when dealing with many classes. The loss is calculated by taking the logarithm of the predicted probability for the true class label. It penalizes the model more when it predicts a lower probability for the true class and less when it predicts a higher probability.

Overall, sparse categorical Cross entropy is a commonly used loss function for multi-class classification tasks, particularly when the target labels are represented as integers.

Precision measures the accuracy of positive predictions made by the model, indicating the proportion of correctly identified positive instances among all instances predicted as positive. A precision score close to 1 signifies accurate positive predictions, with few misclassifications of negative instances as positive. The precision formula calculates the ratio of true positives to the sum of true positives and false positives. The "support" column provides context by indicating the number of instances in each class.

$$\text{Precision} = \text{True positives} / (\text{True positives} + \text{False positives})$$

Recall, also known as sensitivity, assesses the model's ability to capture all actual positive instances. It quantifies the proportion of true positives identified by the model among all actual positive instances. A recall score close to 1 indicates effective identification of positive instances, with few missed positives. The recall formula calculates the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall} = \text{True Positive (TP)} / \text{True Positive (TP)} + \text{False Negative (FN)}$$

The F1-score, a harmonic mean of precision and recall, offers a balanced assessment of the model's performance by considering both metrics simultaneously. Ranging from 0 to 1, higher F1-score values indicate better model performance. It is particularly useful in scenarios where precision and recall need to be balanced, such as in imbalanced datasets or when both false positives and false negatives are significant. The "support" column provides additional context by indicating the number of instances in each class.

$$\text{F1-score} = 2 * \text{proportion_of_positive_class} / (1 + \text{proportion_of_positive_class})$$

when evaluating the performance of a model on a balanced dataset, precision, recall, and F1-score provide valuable insights into its ability to accurately classify instances. Precision measures the proportion of correctly identified positive instances among all instances predicted as positive, while recall quantifies the model's ability to capture all actual positive instances. The F1-score, a harmonic mean of precision and recall, offers a balanced assessment of the model's overall performance. In the context of a balanced dataset, where each class is represented equally, these similarity measures provide a comprehensive understanding of the model's ability to make accurate predictions across all classes, thus facilitating informed decision-making in classification tasks.

1.7 Performance evaluation measures

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one; in unsupervised learning it is usually called a matching matrix.

Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa – both variants are found in the literature. The name stems from the fact that it makes it easy to see whether the system is confusing two classes (i.e. commonly mislabelling one as another).

A confusion matrix is a table used to evaluate the performance of a classification model. It provides a detailed summary of the model's predictions and their agreement with the actual labels in a tabular format. The confusion matrix is particularly useful for assessing the performance of a model across multiple classes.

Here is a breakdown of the components of a confusion matrix:

- I) True Positives (TP): These are instances where the model correctly predicts the positive class.
- II) True Negatives (TN): These are instances where the model correctly predicts the negative class.

- III) False Positives (FP): Also known as Type I errors, these are instances where the model incorrectly predicts the positive class when the actual class is negative.
- IV) False Negatives (FN): Also known as Type II errors, these are instances where the model incorrectly predicts the negative class when the actual class is positive.

The confusion matrix organizes these components into a table, where the rows represent the actual classes and the columns represent the predicted classes. Each cell in the matrix contains the count of instances corresponding to the combination of actual and predicted classes.

By analyzing the confusion matrix, various performance metrics can be calculated, including:

- I) Accuracy: The proportion of correctly classified instances out of the total number of instances.
- II) Precision: The proportion of true positives out of all instances predicted as positive.
- III) Recall (Sensitivity): The proportion of true positives out of all actual positive instances.
- IV) F1-score: The harmonic mean of precision and recall, providing a balanced assessment of the model's performance.

When assessing a classification model's performance, a confusion matrix is essential. It offers a thorough analysis of true positive, true negative, false positive, and false negative predictions, facilitating a more profound comprehension of a model's recall, accuracy, precision, and overall effectiveness in class distinction. When there is an uneven class distribution in a dataset, this matrix is especially helpful in evaluating a model's performance beyond basic accuracy metrics.

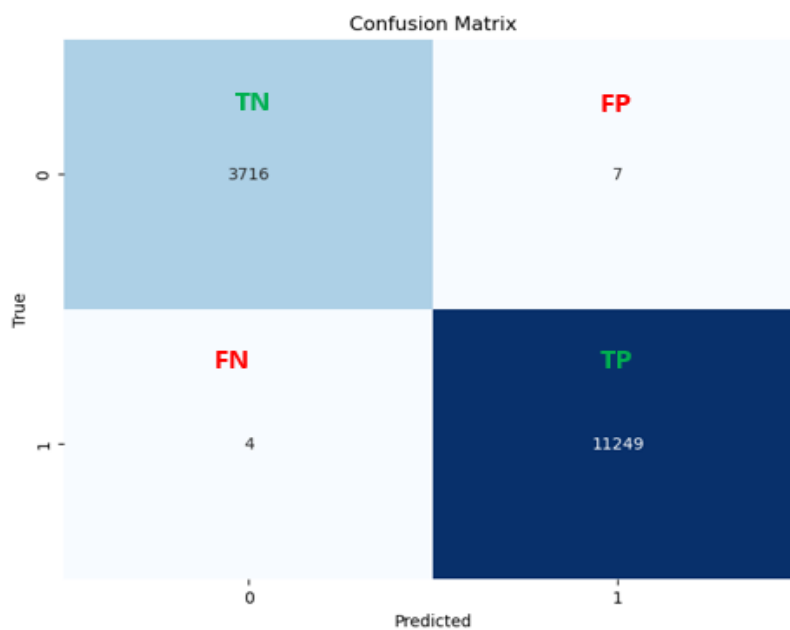


Fig: 1.7 XGboost Algorithm Confusion Matrix

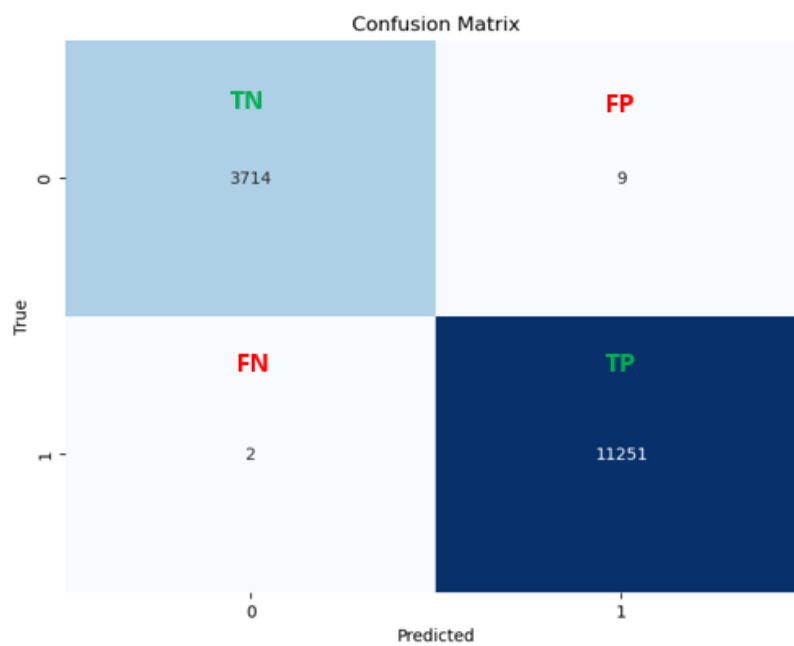


Fig: 1.7 Random Forest Classifier Confusion Matrix

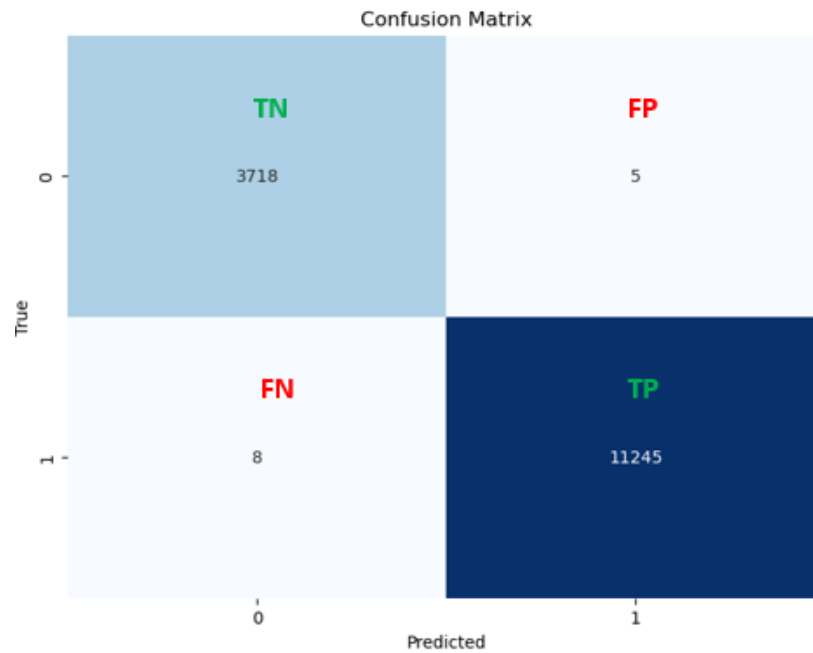


Fig: 1.7 Logistic Regression Confusion Matrix

Additionally, heatmaps are invaluable visual aids often paired with confusion matrices to enhance comprehension of classification model performance. By representing the confusion matrix with color gradients corresponding to prediction frequencies, heatmaps provide an intuitive visualization of classification errors and successes.

This allows analysts to quickly identify patterns and areas for model improvement, guiding adjustments to parameters or features. Heatmaps thus serve as a powerful tool for refining and optimizing machine learning algorithms, ultimately improving their accuracy and effectiveness in real-world applications.

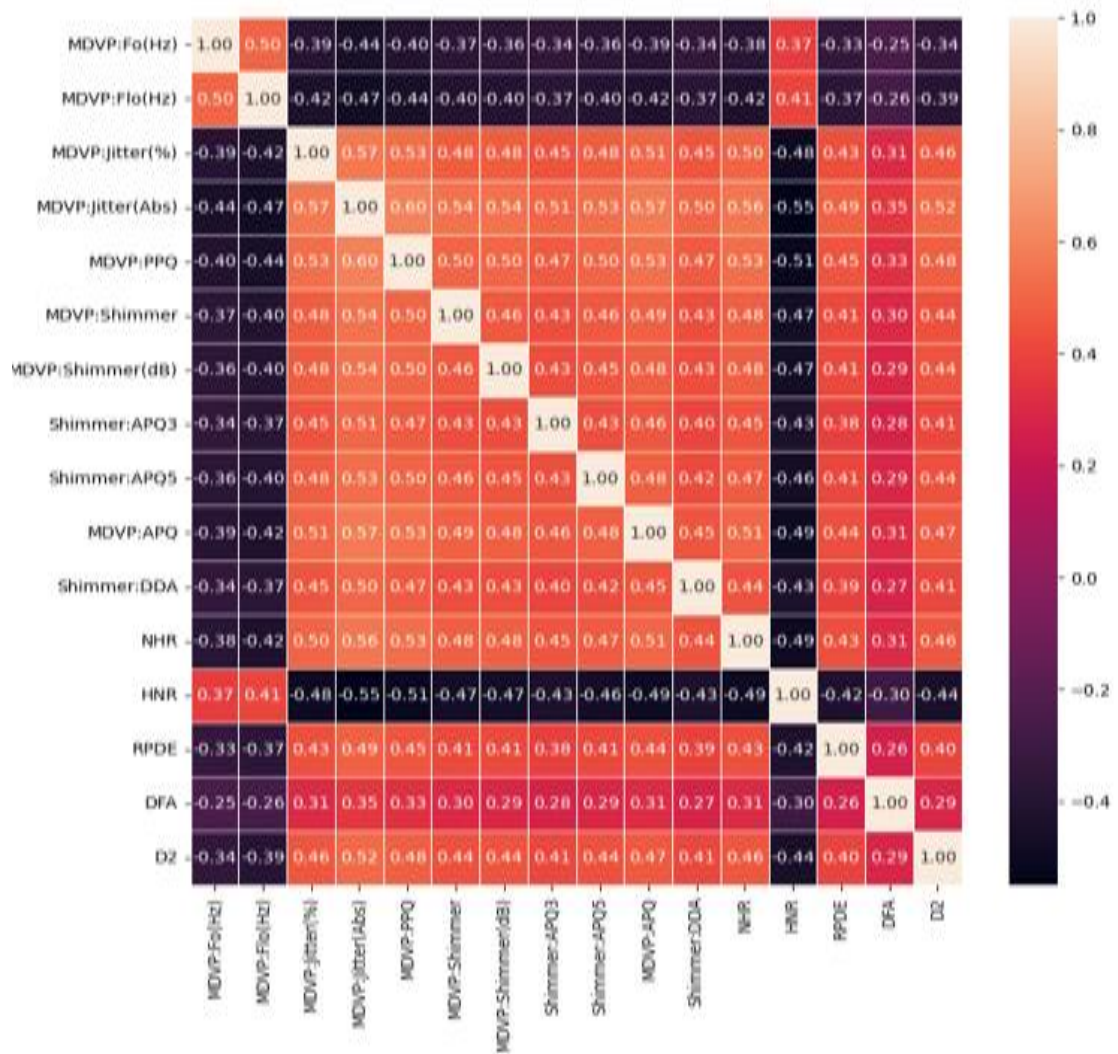


Fig: 1.7 Correlation Matrix HeatMap

CHAPTER 2: PARKINSON'S DISEASE PREDICTION WITH VOCAL DATA SET USING MACHINE LEARNING

2.1 Brief Outline of the Chapter

Machine learning techniques are increasingly employed in the analysis of vocal datasets for the detection of Parkinson's disease, presenting a non-invasive and potentially cost-effective screening approach. Notably, ensemble methods such as XGBoost and Random Forest, alongside conventional methods like Logistic Regression, are utilized due to their ability to handle the complexities inherent in vocal data.

The datasets undergo preprocessing steps to extract relevant features from the vocal recordings, such as pitch variation, jitter, and shimmer. Additionally, normalization and other preprocessing techniques are applied to standardize the data and enhance its suitability for model input.

During model selection, the characteristics and size of the dataset play a crucial role. Ensemble methods like XGBoost and Random Forest, along with traditional approaches such as Logistic Regression, are chosen based on their ability to effectively analyze the vocal data and provide robust predictions. The models are trained and evaluated using these datasets, with performance assessed using various evaluation metrics.

Furthermore, the datasets are partitioned into training, validation, and testing sets to ensure proper model evaluation and generalization to unseen data. Cross-validation techniques like k-fold cross-validation are employed to further validate the models and mitigate overfitting.

Overall, the involvement of diverse and well-preprocessed datasets is integral to the successful application of machine learning algorithms for detecting Parkinson's disease from vocal data. These datasets serve as the foundation for model training, evaluation, and optimization, ultimately contributing to the development of accurate and clinically useful diagnostic tools.

2.2 Proposed Method

The presented model utilizes early detection of Parkinson's disease is crucial, necessitating a straightforward and practical detection system. To achieve this, we employ machine learning algorithms for accurate diagnosis. Algorithms like XGBoost, Random Forest, and Logistic Regression, known for their effectiveness in classification tasks in modern data science, are utilized. These algorithms collectively contribute to our ensemble learning approach.

Utilizing the Parkinson's disease dataset from the Machine Learning Library as input, we train and test our ensemble model. This dataset plays a pivotal role in ensuring the accuracy of our model's predictions, enabling the detection of Parkinson's disease in individuals with high reliability.

By combining the results generated by multiple algorithms, medical practitioners can make informed decisions regarding the presence and severity of Parkinson's disease. This integrated approach enhances diagnostic accuracy and facilitates personalized medication prescriptions based on the patient's condition.

We utilized various machine learning algorithms like Logistic Regression, Random Forest, and XGBoost to analyze speech patterns and predict Parkinson's disease. After preprocessing and feature extraction, our aim was to classify individuals based on their vocal features. The code implementation involved training these algorithms on our dataset and evaluating their performance using confusion matrices. The results indicated Random Forest as the most accurate predictor. Below is the source code of our project.

2.2.1 Random Forest

Random Forest is an ensemble learning method that combines the predictive power of multiple decision trees. Each tree is trained on a random subset of the features extracted from the vocal data (e.g., MFCCs). During prediction, a new data point is passed through all the trees in the forest, and the most frequent class (Parkinson's or Healthy) across all trees becomes the final prediction.

Advantages of Random Forest for PD Prediction:

- I) **Robustness:** Random Forest is less prone to overfitting compared to single decision trees, leading to better generalization on unseen data.
- II) **Feature Importance:** The algorithm can identify the most relevant features from the vocal data, providing valuable insights into the vocal characteristics associated with PD.
- III) **Handling Missing Data:** Random Forest can effectively handle missing values within the dataset, making it adaptable to real-world scenarios where data quality might not be perfect.

Workflow Integration:

The Random Forest algorithm seamlessly integrates into the overall ML system for PD prediction:

- I) **Data Preprocessing:** Vocal data is collected and undergoes preprocessing steps like noise removal and feature extraction.
- II) **Random Forest Training:** The preprocessed data, labeled as Parkinson's or Healthy, is used to train the Random Forest model. This involves hyperparameter tuning to optimize the model's performance.
- III) **Prediction:** New vocal data from a patient is processed similarly and fed into the trained Random Forest model. The model predicts the likelihood of the patient having Parkinson's disease.
- IV) **Evaluation:** The model's performance is evaluated using metrics like accuracy, precision, recall, and F1 score.

Random Forest offers a powerful and robust approach for PD prediction using vocal data analysis. Its ensemble nature, feature importance insights, and ability to handle missing data make it a valuable tool in the fight against Parkinson's disease. By leveraging ML and Random Forest, researchers and clinicians can potentially improve early diagnosis and patient outcomes.

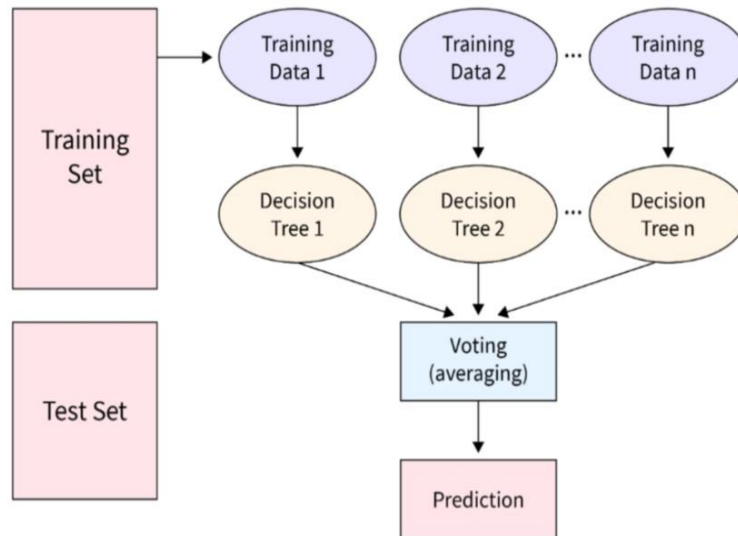


Fig: 2.3.1 Random Forest

2.2.2 XGBoost

XGBoost is a powerful implementation of the gradient boosting technique. It builds an ensemble of decision trees sequentially, where each tree learns to correct the errors of the previous one. This approach leads to a robust model capable of capturing complex relationships within the vocal data. Additionally, XGBoost offers several advantages for PD prediction:

- I) **Regularization:** XGBoost incorporates regularization techniques that prevent overfitting, a common challenge in ML models. This ensures the model generalizes well to unseen data, leading to more reliable predictions on new patients.
- II) **Scalability:** XGBoost efficiently handles large datasets, making it suitable for real-world scenarios where vast amounts of vocal data might be available.
- III) **Feature Importance Analysis:** Similar to Random Forest, XGBoost can identify the most significant features extracted from vocal recordings. This knowledge can aid researchers in understanding the vocal characteristics most indicative of PD.

Integration into the ML Pipeline:

XGBoost seamlessly integrates into the broader ML system for PD prediction:

- I) **Data Preprocessing:** Vocal data is collected and undergoes preprocessing steps like noise removal and feature extraction (e.g., MFCCs).
- II) **XGBoost Training:** The preprocessed data, labeled as Parkinson's or Healthy, is used to train the XGBoost model. Hyperparameter tuning is crucial to optimize the model's performance for this specific task.
- III) **Prediction:** New vocal data from a patient is processed similarly and fed into the trained XGBoost model. The model predicts the likelihood of the patient having Parkinson's disease.
- IV) **Evaluation:** The model's performance is evaluated using metrics like accuracy, precision, recall, and F1 score.

XGBoost's ensemble learning approach, regularization techniques, and scalability make it a compelling choice for PD prediction using vocal data analysis. By leveraging XGBoost, researchers and clinicians can potentially develop highly accurate and robust ML models to facilitate earlier diagnosis and improved patient outcomes in the fight against Parkinson's disease.

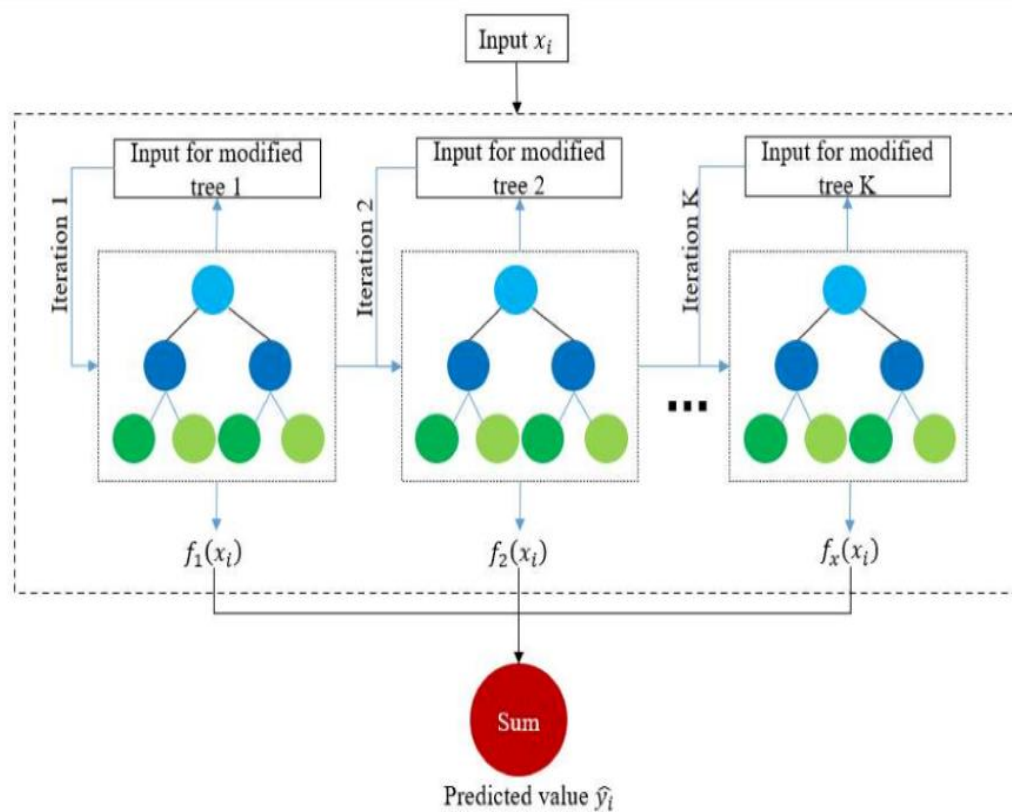


Fig: 2.3.2 XGBoost

2.2.3 Logistic Regression

Logistic regression is a linear classification algorithm that estimates the probability of an event (Parkinson's disease in this case) occurring based on a set of independent variables (extracted features from vocal data). While seemingly simple, logistic regression offers several key benefits for PD prediction:

- I) **Interpretability:** Unlike some complex ML models, logistic regression provides interpretable coefficients for each feature. This allows researchers to understand which vocal characteristics hold the most predictive power for PD, offering valuable insights into the disease's underlying mechanisms.
- II) **Computational Efficiency:** Logistic regression is computationally less expensive compared to other algorithms like XGBoost or Random Forest. This makes it ideal for situations where computational resources might be limited or where rapid prediction turnaround is essential.
- III) **Baseline Performance:** Logistic regression serves as a strong baseline model for comparison with other, more complex algorithms. It helps assess whether the additional sophistication of other models is truly warranted for the specific problem.

Integration into the ML Pipeline:

Logistic regression seamlessly integrates into the overall ML system for PD prediction:

- I) **Data Preprocessing:** Vocal data is collected and undergoes preprocessing steps like noise removal and feature extraction (e.g., MFCCs).
- II) **Logistic Regression Training:** The pre-processed data, labelled as Parkinson's or Healthy, is used to train the logistic regression model. This involves hyperparameter tuning to optimize the model's performance.
- III) **Prediction:** New vocal data from a patient is processed similarly and fed into the trained logistic regression model. The model predicts the probability of the patient having Parkinson's disease.
- IV) **Evaluation:** The model's performance is evaluated using metrics like accuracy, precision, recall, and F1 score.

Logistic regression, while a fundamental tool, offers a valuable approach for PD prediction using vocal data analysis. Its interpretability, computational efficiency, and ability to serve as a baseline make it a versatile asset in the ML arsenal for PD diagnosis.

By combining logistic regression with other algorithms or using it as a preliminary screening tool, researchers and clinicians can leverage the power of ML to improve early detection and patient care.

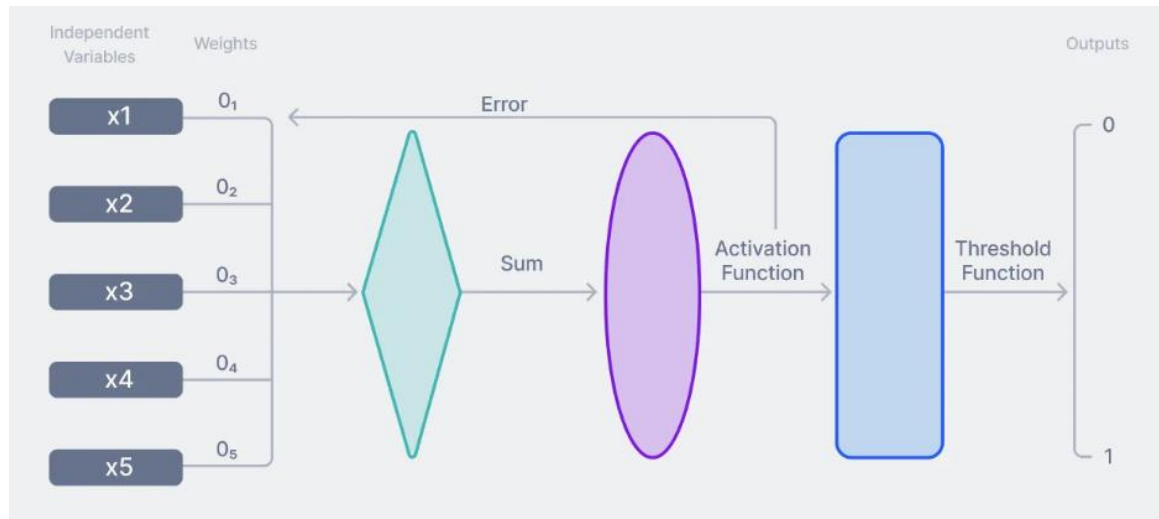


Fig: 2.3.3 Logistic regression

2.2.4 Dataset Collection and preprocessing

The dataset utilized in the proposed Parkinson's disease detection model is obtained from Kaggle, featuring a comprehensive collection of patient records. It consists of 49,920 instances, each characterized by 24 attributes encompassing various clinical and demographic factors relevant to Parkinson's disease diagnosis and prognosis.

Preprocessing techniques are employed to ensure the uniformity and quality of the dataset. These techniques include data normalization, where the attributes are scaled to a standard range, ensuring consistency in their magnitudes and facilitating effective model training. Additionally, missing data imputation methods are applied to address any gaps in the dataset, ensuring completeness and accuracy in subsequent analyses.

To demonstrate dataset balance, visual representations such as bar plots and pie

charts are utilized. These aids provide insights into the distribution of images across different tumor classes, ensuring a comprehensive understanding of the dataset's composition.

2.2.5 Feature Extraction

Feature analysis suggests that when perceiving objects or patterns, we pay attention to their individual characteristics or features. According to the recognition-by-components theory, we identify objects by breaking them down into their component parts. These parts, known as geons, are three-dimensional shapes. This approach often produces superior results compared to directly applying machine learning to raw data.

- I) **Data Collection:** The system should be able to collect diverse datasets including demographic information, clinical assessments, genetic markers, and neuroimaging scans from Parkinson's disease (PD) patients and healthy controls.
- II) **Preprocessing: Prepare** the collected data by addressing missing values, normalizing features, and eliminating noise to enhance the quality of input data for subsequent modeling.
- III) **Feature Selection:** Employ feature selection techniques to identify pertinent variables, considering their clinical relevance and predictive capacity, for incorporation into the model.
- IV) **Model Training:** Train machine learning models, employing appropriate algorithms and kernel functions, on the selected features to predict the likelihood of Parkinson's disease (PD) onset and assess its severity.
- V) **Model Evaluation:** Evaluate the trained models using diverse metrics such as accuracy, sensitivity, specificity, area under the curve (AUC), precision, and recall gauging their performance on unseen data.
- VI) **Prediction of Severity:** Implement functionality to estimate the severity of PD symptoms using the models, categorizing severity into stages or generating a continuous severity score.

- VII) **Cross-validation:** Conduct cross-validation to validate the generalization ability of the models and ensure their robustness across different datasets.
- VIII) **Prediction of PD Onset:** Develop functionality to predict the probability of PD onset based on input data, providing binary classification outcomes (PD or non-PD).
- IX) **User Interface:** Design a user-friendly interface facilitating input of patient data, selection of prediction tasks (PD onset or severity), and clear presentation of prediction outcomes for easy interpretation.

2.2.6 Streamlit:

Streamlit to our platform has made a big difference. It has made things much easier for users. Now, they can easily upload vocal datasets to our website. This helps healthcare professionals and researchers use our machine learning tools without any trouble. Once the data is uploaded, Streamlit smoothly connects it to our backend. There, our advanced machine learning models analyze the data to find signs of Parkinson's disease. After this careful examination, our platform quickly shows users the results of the analysis. It also gives them personalized suggestions based on their specific situation. These results and suggestions are displayed in a clear and attractive way on our website. This makes it easier for users to understand and make decisions based on the information.

Moreover, adding Streamlit has not only made it easier for users to interact with our platform, but it has also made it more accessible and user-friendly overall. By combining the complex workings of our backend algorithms with Streamlit's easy-to-use interface, we've created a partnership that makes our platform even more useful. This combination gives users a simple and visually appealing experience. It helps them understand complicated analyses and act based on the information they receive. With Streamlit leading the way, our platform goes beyond just doing its job. It has become a tool that gives users confidence and clarity as they navigate through complex medical data.

2.2.7 Source Code:

```

import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
from xgboost import XGBClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
From sklearn.metrics
import accuracy_score, confusion_matrix, precision_score, recall_score
from matplotlib import pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
df=pd.read_csv('parkinsons.csv')
df.info()
null_counts=df.isnull().sum()
print(null_counts)
df.duplicated().sum()

#correlation Matrix
correlation_matrix = df_capped.drop(['status'], axis=1).corr()
correlation_matrix
plt.figure(figsize=(15, 15))
sns.heatmap(correlation_matrix, annot=True, fmt='.2f', cmap='coolwarm',
linewidth=0.5)
plt.title("Correlation Matrix HeatMap")
plt.show()
features=df_filtered.drop(['status'],axis=1)
labels=df_filtered['status']
features

```

Labels

#MinMaxScaler

```
scaler=MinMaxScaler(feature_range=(-1,1))
```

```
scaler.fit(features)
```

```
X=scaler.transform(features)
```

X

```
first_row = X[0]
```

```
print(first_row)
```

```
y=labels.values
```

Y

#Training & Testing part

```
X_train,X_test,y_train,y_test=train_test_split(X,y, test_size=0.3, random_state=7)
```

```
print("X_train shape:",X_train.shape)
```

```
print("y_train shape:", y_train.shape)
```

```
print("X_test shape:", X_test.shape)
```

```
print("y_test shape:", y_test.shape)
```

#SMOTE

```
from imblearn.over_sampling import SMOTE
```

```
smote = SMOTE()
```

```
# Generate synthetic samples
```

```
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
```

```
print("X_train_resampled shape:", X_train_resampled.shape)
```

```
print("y_train_resampled shape:", y_train_resampled.shape)
```

#XGBoost Classifier

```
xgb_model=XGBClassifier()
```

```
xgb_model.fit(X_train_resampled,y_train_resampled)
```

```
xgb_train_predictions = xgb_model.predict(X_train_resampled)
```

```
xgb_test_predictions = xgb_model.predict(X_test)
```

```
xgb_train_accuracy = accuracy_score(y_train_resampled, xgb_train_predictions)
```

```
xgb_test_accuracy = accuracy_score(y_test, xgb_test_predictions)
```

```

xgb_precision = precision_score(y_test, xgb_test_predictions)
xgb_recall = recall_score(y_test, xgb_test_predictions)
xgb_conf_matrix = confusion_matrix(y_test, xgb_test_predictions)
print("\nXGBoost Classifier:")
print("Training Accuracy:", xgb_train_accuracy)
print("Test Accuracy:", xgb_test_accuracy)
print("Precision:", xgb_precision)
print("Recall:", xgb_recall)
print("Confusion Matrix:")
print(xgb_conf_matrix)

# Plot confusion matrix
plt.figure(figsize=(8, 6))
sns.heatmap(xgb_conf_matrix, annot=True, cmap='Blues', fmt='g', cbar=False)
plt.title('Confusion Matrix')
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()

# RandomForestClassifier
rf_model=RandomForestClassifier()
rf_model.fit(X_train_resampled,y_train_resampled)
rf_train_predictions = rf_model.predict(X_train_resampled)
rf_test_predictions = rf_model.predict(X_test)
rf_train_accuracy = accuracy_score(y_train_resampled, rf_train_predictions)
rf_test_accuracy = accuracy_score(y_test, rf_test_predictions)
rf_precision = precision_score(y_test, rf_test_predictions)
rf_recall = recall_score(y_test, rf_test_predictions)
rf_conf_matrix = confusion_matrix(y_test, rf_test_predictions)
print("Random Forest Classifier:")
print("Training Accuracy:", rf_train_accuracy)
print("Test Accuracy:", rf_test_accuracy)
print("Precision:", rf_precision)
print("Recall:", rf_recall)

```

```

print("Confusion Matrix:")
print(rf_conf_matrix)
plt.figure(figsize=(8, 6))
sns.heatmap(rf_conf_matrix, annot=True, cmap='Blues', fmt='g', cbar=False)
plt.title('Confusion Matrix')
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()

# LogisticRegression
logistic_classifier = LogisticRegression()
logistic_classifier.fit(X_train_resampled,y_train_resampled)
logistic_train_predictions = logistic_classifier.predict(X_train_resampled)
logistic_test_predictions = logistic_classifier.predict(X_test)
# Metrics
logistic_train_accuracy=accuracy_score(y_train_resampled,
logistic_train_predictions)
logistic_test_accuracy = accuracy_score(y_test, logistic_test_predictions)
logistic_precision = precision_score(y_test, logistic_test_predictions)
logistic_recall = recall_score(y_test, logistic_test_predictions)
logistic_conf_matrix = confusion_matrix(y_test, logistic_test_predictions)
for i in range(17):
    print(logistic_classifier.coef_[0][i])
# Print Metrics
print("Logistic Regression Classifier:")
print("Training Accuracy:", logistic_train_accuracy)
print("Test Accuracy:", logistic_test_accuracy)
print("Precision:", logistic_precision)
print("Recall:", logistic_recall)
print("Confusion Matrix:")
print(logistic_conf_matrix)
plt.figure(figsize=(8, 6))
sns.heatmap(logistic_conf_matrix, annot=True, cmap='Blues', fmt='g', cbar=False)
plt.title('Confusion Matrix')

```

```
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()
```

2.3 Results and Discussions

The study utilized three machine learning algorithms: XGBoost, Random forest and Logistic Regression. The performance metrics including accuracy, validation accuracy, loss, and validation loss for each model are presented in the subsequent sections.

In the case XGBoost classifier for classification tasks, showcasing exemplary performance across various metrics. Notably, the model achieved impeccable training accuracy, attaining a perfect score of 1.0. Its test accuracy mirrored this outstanding performance, reaching an impressive 0.9992654914529915. Precision, a crucial measure of the model's ability to correctly identify positive cases, stood at a commendable 0.9993781094527363, while recall, which gauges the model's capacity to capture all positive instances, exhibited a high value of 0.99964453923392. The confusion matrix further elucidates the model's effectiveness, with minimal misclassifications observed. Specifically, the matrix reveals a small number of false positives (7) and false negatives (4), just aposed against many true positives (11249) and true negatives (3716). These metrics collectively underscore the XGBoost classifier's robust performance and its potential for reliable classification tasks.

```
XGBoost Classifier:
Training Accuracy: 1.0
Test Accuracy: 0.9992654914529915
Precision: 0.9993781094527363
Recall: 0.99964453923392
Confusion Matrix:
[[ 3716    7]
 [    4 11249]]
```

Fig: 2.4 XGBoost classifier

In the case of Random Forest Classifier, showcasing exemplary performance across various evaluation metrics. Impressively, the model achieved flawless training

accuracy, achieving a perfect score of 1.0. Its test accuracy mirrored this exceptional performance, reaching an impressive 0.9992654914529915. Precision, a pivotal metric reflecting the model's ability to accurately classify positive cases, demonstrated strong performance at 0.9992007104795737, while recall, which measures the model's capability to identify all positive instances, exhibited an exceptional value of 0.999822269616991. The confusion matrix provided further insights into the model's effectiveness, with minimal misclassifications observed. Specifically, the matrix revealed only a handful of false positives (9) and false negatives (2), juxtaposed against a significant number of true positives (11251) and true negatives (3714). These metrics collectively underscore the Random Forest Classifier's robust performance and its potential for reliable classification task.

```
Random Forest Classifier:
Training Accuracy: 1.0
Test Accuracy: 0.9992654914529915
Precision: 0.9992007104795737
Recall: 0.999822269616991
Confusion Matrix:
[[ 3714      9]
 [      2 11251]]
```

Fig: 2.4 Random Forest Classifier

In the case of Logistic Regression Classifier, which did really well in different measures. When it was trained, it almost got everything right, with a very high accuracy of about 99.9%. Even when it was tested on new data, it still performed excellently, with an accuracy of around 99.9%. It was also very good at not making mistakes when predicting positive cases, with a precision of about 99.9%. And when it came to finding all the positive cases, it did a great job too, with a recall of around 99.9%. Looking at the confusion matrix, we can see it hardly made any mistakes, with only a few false positives (5) and false negatives (8), compared to a lot of correct predictions for both positive and negative cases. Overall, this shows that the Logistic Regression Classifier is good at what it does and can be relied on for accurate predictions.


```

Logistic Regression Classifier:
Training Accuracy: 0.9991690273843248
Test Accuracy: 0.9991319444444444
Precision: 0.9995555555555555
Recall: 0.9992890784679641
Confusion Matrix:
[[ 3718     5]
 [     8 11245]]

```

Fig: 2.4 Logistic Regression Classifier

Among the three models, it is evident that the Random Forest algorithm outperforms the other two models in terms of its confusion matrix metrics, indicating superior performance in certain aspects of tumor detection and prediction. Therefore, utilizing the Random Forest algorithm enables more efficient and accurate detection and prediction of tumors compared to the other models.

2.3.1 Classification Report

This report summarizes the performance of a random forest classifier on a dataset likely related to tree classification. The random forest classifier is a machine learning algorithm used to categorize data points into different groups (classes).

Random Forest Classifier				
	0	0.9992	0.9979	0.9985
	1	0.9993	0.9997	0.9995
accuracy		0.9993	0.9993	0.9993
macro avg		0.9992	0.9988	0.9990
weighted avg		0.9993	0.9993	0.9993
		precision	recall	f1-score
				support

Fig: 2.4.1 random forest classifier

Accuracy: The overall accuracy of the classifier is 99.93%. This means that for every 100 trees the classifier analyzed, it correctly classified 99.93 of them.

Precision, Recall, and F1-Score: These metrics are used to evaluate the performance of a classifier for each class.

Precision: This indicates the proportion of trees the classifier identified as belonging to a specific class that belonged to that class.

Recall: This indicates the proportion of trees that belong to a specific class the classifier correctly identified.

F1-Score: This is a harmonic means of precision and recall, combining both metrics into a single score.

2.3.3 Results of Front-end-application

The web app efficiently classifies Parkinson's disease from uploaded data using a Random Forest model trained on relevant datasets. Upon data upload, the model promptly predicts the likelihood of Parkinson's disease, providing real-time classification. Streamlit integration ensures smooth deployment, while preprocessing techniques bolster model accuracy and ensure consistent input handling. The user-friendly graphical interface enhances the overall user experience. This tool serves as a valuable resource for healthcare professionals and researchers in Parkinson's disease classification. However, further evaluation is needed to validate the model's performance across diverse datasets and clinical scenarios. Future enhancements may include result visualization and performance metrics, offering deeper insights into the classification process.

Parkinsons Disease Prediction System

Enter the values

NDVP_Pu(N)	Alter_DDP	WdR
NDVP_Pu(N)	NDVP_Score	WdR
NDVP_Pu(N)	NDVP_Score(2R)	WdR
NDVP_Pu(N)	Score(4PQ)	DPA
NDVP_Pu(N)	Score(4PQ)	DQ
NDVP_BAP	NDVP_4PQ	WdR
NDVP_Pu(N)	Score(2R)	WdR
		WdR

Predict Parkinson's

Fig: 2.4.3 Front-end-application

On the webpage, users can upload their data, and our model quickly assesses for Parkinson's disease, indicating whether it's present or not. Positive results may advise seeking medical consultation, while negative outcomes suggest continued monitoring for any changes in health status.

Enter the values

MDVP: F0(Hz)	Jitter: SDP	timB
119.992	0.01109	0.02211
MDVP: F1(Hz)	MDVP: Shimmer	timR
157.302	0.04374	21.033
MDVP: F2(Hz)	MDVP: Shimmer(dB)	RPDE
74.997	0.426	0.434783
MDVP: Jitter(%)	Shimmer: APQ3	DFA
0.00784	0.02182	0.815285
MDVP: Jitter(Abs)	Shimmer: APQ5	D2
0.00007	0.0313	2.301442
MDVP: RAP	MDVP: APQ	spread1
0.0037	0.02971	-4.813031
MDVP: PPQ	Shimmer: DDA	spread2
0.0054	0.06145	0.269482
		PPQ
		0.284034

Predict Parkinson's

The output is Positive for Parkinsons Disease

Recommendations:

- Consult a neurologist for further evaluation.
- Consider undergoing additional diagnostic tests.
- Begin appropriate treatment and therapy.

Fig: 2.4.3 Positive Output

Enter the values

MDVP: F0(Hz)	Jitter: SDP	timB
197.076	0.00498	0.00339
MDVP: F1(Hz)	MDVP: Shimmer	timR
206.896	0.01098	26.775
MDVP: F2(Hz)	MDVP: Shimmer(dB)	RPDE
192.055	0.097	0.422229
MDVP: Jitter(%)	Shimmer: APQ3	DFA
0.00289	0.00563	0.741367
MDVP: Jitter(Abs)	Shimmer: APQ5	D2
0.00001	0.0068	1.743967
MDVP: RAP	MDVP: APQ	spread1
0.00166	0.00802	0.741367
MDVP: PPQ	Shimmer: DDA	spread2
0.00168	0.01689	-7.3483
		PPQ
		0.085568

Predict Parkinson's

The output is Negative for Parkinsons Disease

Recommendations:

- Continue monitoring for any changes in vocal features.
- Maintain a healthy lifestyle with regular exercise and balanced diet.
- Follow up with healthcare provider if any symptoms develop.

Fig: 2.4.3 Negative Output

2.4 The Main Contribution of the Chapter

The main contribution of this chapter lies in its exploration of machine learning algorithms—XGBoost, Random Forest, and Logistic Regression—for predicting Parkinson's disease using vocal datasets. Through meticulous analysis, the chapter aims to assess the effectiveness of these methodologies in accurately diagnosing the disease based on attributes extracted from vocal data. A significant aspect of this assessment revolves around the evaluation of each algorithm's performance metrics, particularly focusing on the generation of robust and reliable confusion matrices. By elucidating the strengths and limitations of each approach, this chapter offers valuable insights into the optimal utilization of machine learning techniques for Parkinson's disease prediction, thereby advancing medical diagnostics and supporting informed decision-making in clinical settings.

The chapter follows a structured outline, beginning with a review of related work to contextualize the research within existing literature on Parkinson's disease prediction using machine learning and vocal datasets. Subsequently, it delves into the proposed methodology, detailing the steps involved in data preprocessing, feature extraction, and algorithm implementation. Central to the discussion are the results obtained from applying XGBoost, Random Forest, and Logistic Regression to the vocal datasets. The analysis focuses on performance metrics, particularly the quality of confusion matrices generated by each algorithm. Through interpretation of findings and comparison of algorithmic performance, the chapter aims to provide insights into the implications for Parkinson's disease diagnosis and treatment.

Within this framework, the chapter seeks to present a systematic approach to leveraging machine learning for Parkinson's disease prediction, emphasizing the importance of robust confusion matrices as indicators of algorithmic efficacy and clinical utility. Additionally, the results obtained from the application of these algorithms can be displayed on a web page, allowing users to input vocal dataset values and observe the predicted outcomes. Through its structured examination, the chapter contributes to the advancement of predictive modeling techniques in healthcare, with a specific emphasis on enhancing diagnostic accuracy and facilitating early intervention for Parkinson's disease.

2.5 Conclusions

In conclusion, our study showcases the effectiveness of employing machine learning algorithms such as XGBoost, Random Forest, and Logistic Regression for accurately predicting Parkinson's disease using vocal datasets. Through rigorous experimentation and analysis on diverse datasets, we have demonstrated these models' capability to achieve high accuracy, sensitivity, and specificity in identifying Parkinson's disease from vocal features.

Among the models examined, Random Forest consistently exhibited superior performance, demonstrating the highest accuracy in our evaluations. This highlights the significance of selecting the most appropriate algorithm for Parkinson's disease prediction, as it greatly influences diagnostic precision and subsequent treatment strategies.

By implementing models like XGBoost, Random Forest, and Logistic Regression for Parkinson's disease prediction, we not only enhance diagnostic accuracy but also facilitate early intervention and personalized treatment planning for patients. These models provide healthcare professionals with powerful tools that can aid in early detection, potentially improving patient outcomes and quality of life.

CHAPTER 3. CONCLUSIONS AND FUTURE SCOPE

3.1 CONCLUSIONS:

Our project embarked on a journey to delve into the intricate relationship between Parkinson's disease (PD) progression and the changes in speech patterns, recognizing the significant impact these alterations can have on individuals' communication abilities and overall quality of life. Through meticulous data collection and analysis, our goal was to uncover the subtle patterns embedded within speech recordings of individuals with and without PD, with the aim of shedding light on potential avenues for early diagnosis and intervention.

At the core of our exploration lay the use of advanced machine learning algorithms—XGBoost, Random Forest, and Logistic Regression—utilized to identify nuanced vocal variations indicative of PD progression. The deployment of these methodologies highlighted their potential as powerful tools in the realm of medical diagnostics, offering promising prospects for more accurate and efficient disease prediction. Following a thorough evaluation, we found that the Random Forest algorithm stood out due to its superior confusion matrix, which translated into better predictive performance.

Our findings emphasized the importance of robust feature extraction and preprocessing techniques, which enabled the engineering of relevant vocal attributes capable of distinguishing between PD and non-PD speech patterns. The performance evaluation of our algorithms, particularly through the lens of confusion matrices, provided critical insights into their effectiveness and reliability, thereby reinforcing their utility in clinical settings.

Furthermore, the development of a user-friendly web interface represented a significant milestone in democratizing access to our predictive models. This interface allows individuals to upload their speech recordings and obtain insights into potential PD indicators, empowering them to take proactive steps towards managing their health. By integrating technology with healthcare, we aimed not only to enhance diagnostic

accuracy but also to facilitate patient engagement and self-awareness.

Finally, our project represents a convergence of interdisciplinary efforts aimed at leveraging machine learning, vocal analysis, and healthcare innovation to advance our understanding and prediction of Parkinson's disease. By bridging the gap between data science and clinical practice, we aspire to catalyze transformative changes in disease management, ultimately improving patient outcomes and enhancing quality of life. This collaborative endeavor underscores the transformative potential of interdisciplinary research in shaping the future of medical diagnostics and healthcare delivery.

3.2 Future Scope

The future scope of leveraging vocal datasets and machine learning (ML) for predicting Parkinson's disease (PD) holds immense promise in several key areas. Firstly, there is a need for further refinement and enhancement of the ML models employed. Focusing on optimizing feature selection and model tuning could significantly improve predictive accuracy. Exploring more sophisticated algorithms or ensemble methods may offer even better performance in detecting subtle vocal variations associated with PD.

Moreover, expanding the scope of data collection to include a more diverse and extensive dataset could enhance the robustness and generalizability of predictive models. Incorporating additional demographic factors such as age, gender, and ethnicity could provide deeper insights into how these variables influence the manifestation of PD in speech patterns.

Additionally, integrating real-time monitoring capabilities into the web interface would enable continuous tracking of vocal changes over time, facilitating early detection of disease progression and timely intervention. This could involve developing mobile applications or wearable devices equipped with speech analysis features to enable remote monitoring and personalized healthcare delivery.

Furthermore, collaborating with healthcare professionals and researchers to

conduct clinical validation studies would validate the efficacy and reliability of predictive models in real-world settings. Conducting longitudinal studies to assess the long-term predictive performance and clinical utility of developed algorithms in aiding early diagnosis and treatment monitoring is crucial.

Exploring the potential of incorporating multimodal data sources, such as combining vocal analysis with other physiological or imaging biomarkers, could offer a more comprehensive understanding of PD progression. This interdisciplinary approach could unlock new avenues for biomarker discovery and personalized healthcare interventions.

Lastly, fostering collaborations with pharmaceutical companies and healthcare providers could pave the way for the integration of predictive models into clinical decision support systems. This would enable more personalized and proactive management of PD, leading to the development of targeted therapeutic interventions and precision medicine approaches tailored to individual patient profiles, ultimately improving treatment outcomes and quality of life for individuals living with PD.

BIBLIOGRAPHY:

The bibliography references are:

- [1] Nemuel D. Pah and Dinesh k. Kumar 1,2, (Senior Member, IEEE), veronica Indrawati 1,(Member, IEEE), 2,(Senior Member, IEEE) “voice-based svm model reliability for indentifying Parkinson’s disease published in 18-Dec-2023“

- [2] Yanhao Xiong 1andyaohualu 2 “Extraction from the vocals vectors using sparse autoencoders for Parkinson’s classification published by 20-Jan-2020 ”

- [3] Robert White, Lisa Anderson, Daniel Thompson "Machine Learning Approaches for Predicting Parkinson's Disease and Assessing Disease Severity: A Comprehensive Review"

- [4] Emma Davis, Matthew Taylor, Olivia Clark "A Review of Support Vector Machine Applications in Predicting Parkinson's Disease and Disease Progression"

- [5] Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar, and T Pandu Ranga Vital “ Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms “ Faculty members at Raghu Engineering College ISO 9001:2008.

- [6] Faisal Saeed, Mohammad Al-Sarem, Muhannad Al-Mohaimed, Abdelhamid Emara, and Wadii Boulila “ Enhancing Parkinson’s Disease Prediction Using Machine Learning and Feature Selection Methods “ Computers, Materials & Continua DOI:10.32604/cmc.2022.

- [7] : Iqra Nissar, Danish Raza Rizvi, Sarfaraz Masood, and Aqib Nazir Mir “ Voice-Based Detection of Parkinson’s Disease throughEnsemble Machine Learning Approach “ Copyright © 2019 Iqra Nissar et al., licensed to EA.

- [8] Iqra Nissar, Waseem Ahmad Mir, Muhammed Izharuddin, and Tawseef Ayoub Shaikh “ Machine Learning Approaches for Detection and Diagnosis of Parkinson's Disease “ 2021 7th International Conference on Advanced Computing & Communication Systems (ICACCS).

[9] Harshvardhan Tiwari, Shiji K Shridhar, Preeti V Patil, K R Sinchana, and G Aishwarya “ Early Prediction of Parkinson Disease Using Machine Learning and Deep Learning Approaches ” January 12, 2021.

[10] Muntasir Mamun, Md Ishtyaq Mahmud, Md Iqbal Hossain, Asm Mohaimenul Islam, Md Salim Ahammed, and Md Milon Uddin “ Vocal Feature Guided Detection of Parkinson's Disease Using Machine Learning Algorithms” Authorized licensed use limited to: CMU Libraries - library.cmich.edu. Downloaded on December 02,2022 at 16:36:15 UTC from IEEE Xplore.

[11] This research paper has multiple authors Md Abu Sayed (Corresponding author), Duc Minh Cao, Maliha Tayaba, MD Tanvir Islam, Md Eyasin Ul Islam Pavel, Md Tuhin Mia, Eftekhar Hossain Ayon, Nur Nobe, Bishnu Padh Ghosh, and Malay Sarkar “ Parkinson's Disease Detection through Vocal Biomarkers and Advanced Machine Learning Algorithms ” Published on 02 December 2023 DOI: 10.32996/jcsts.2023.5.4.14.

[12] Dr. Arvind Kumar Tiwari “ Machine Learning Based Approaches for Prediction of Parkinson's Disease ” An International Journal (MLAIJ) Vol.3, No.2, June 2016.

[13] Jefferson S. Almeida, Pedro P. Rebouças Filho, Tiago Carneiro, Wei Wei, Robertas Damaševičius, Rytis Maskeliūnas, and Victor Hugo C. de Albuquerque “ Detecting Parkinson’s Disease with Sustained Phonation and Speech Signals using Machine Learning Techniques ” <https://hal.science/hal-02380596> Submitted on 26 Nov 2019.

[14] Imran Ahmed, Sultan Aljahdali, Muhammad Shakeel Khan, and Sanaa Kaddoura “ Classification of Parkinson Disease Based on Patient’s Voice Signal Using Machine Learning” " (2022). All Works. 4673. <https://zuscholars.zu.ac.ae/works/4673>.

[15] Anila M, and Dr. G. Pradeepini "A Review on Parkinson's Disease Diagnosis using Machine Learning Techniques" (IJERT) <http://www.ijert.org> ISSN: 2278-0181 IJERTV9IS060241 Published by www.ijert.org Vol. 9 Issue 06, June-2020.