**Project Synopsis**

**on**

# "Disease Prediction"

**Submitted in partial fulfillment of the requirement for the degree of Bachelors of Engineering by**

**Aditi Jadhav**
**Ananya Rawool**
**Anjali Chauhan**
**Roshni Kane**

Under the guidance of
**Ms. Dhanashri Dhawale**

**LOKMANYA TILAK COLLEGE OF ENGINEERING**
Affiliated to
**UNIVERSITY OF MUMBAI**

**Department of Computer Science Engineering (Data Science)**
**Academic Year – 2024-2025**

# CERTIFICATE

This is to certify that the mini project II entitled **"DISEASE PREDICTION"** is a bonafide work of **Aditi Jadhav(101), Ananya Rawool(103), Anjali Chauhan(104), Roshni Kane(108)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award degree of **"Bachelor of Engineering"** in **"Computer Science and Engineering (Data Science)"**.

**Ms.Dhanashri Dhawale** **Dr. Nandini C Nag** **Dr.Subhash K Shinde**

(Project Guide) (Head of Department) (Principal)

**External Examiner**

Place: Lokmanya Tilak College of Engineering

Date:

# MINI PROJECT APPROVAL

This Mini Project entitled "DISEASE PREDICTION" by Aditi Jadhav (101), Ananya Rawool (103), Anjali Chauhan (104) and Roshni Kane (108) is approved for the degree of Bachelor of Engineering in Computer Science and Engineering (DataScience).

**Examiner**

1. . . . . .

(Internal Examiner Name Sign)

2. . . . . .

(External Examiner Name Sign)

Date:

Place:

# ACKNOWLEDGEMENT

We would like to acknowledge and extend our heartfelt gratitude to all those people who have been associated with this project and have helped us with it thus making worthwhile experience.

Firstly we extend our thanks to various people which include our project guide **Ms.Dhanashri Dhawale** who has shared her opinions and experiences through which we received the required information crucial for our project synopsis. We are also thankful to head of the department **Dr. Nandini C Nag** and all the staff members of Computer Science and Engineering(Data Science) for their highly co-operative and encouraging attitudes, which have always boosted us.

We also take this opportunity with great pleasure to thank our esteemed Principal **Dr. Subhash K. Shinde** whose timely support and encouragement has helped us succeed in our venture.

| Name of Candidate | Signature |
|---|---|
| **1.Aditi Jadhav** | |
| **2.Ananya Rawool** | |
| **3.Anjali Chauhan** | |
| **4.Roshni Kane** | |

# ABSTRACT

This is a report on disease prediction which aim to develop a robust predictive model for neurodegenerative diseases, specifically Parkinson's and Alzheimer's disease, utilizing advanced machine learning techniques. Given the rising prevalence of these conditions and their profound impact on individuals and society, early detection is crucial for effective intervention and management. Through the application of various algorithms, including support vector machines, random forests, and deep learning models, we evaluate their predictive accuracy and interpretability. Our model aims to identify key risk factors and early symptoms associated with both diseases. By integrating these insights. Ultimately, this project aspires to contribute to the broader understanding of neurodegenerative diseases and enhance the quality of life for affected individuals. Furthermore, our findings will contribute to the ongoing research efforts in understanding the underlying mechanisms of neurodegeneration and may pave the way for future therapeutic interventions. By advancing predictive capabilities, we aspire to improve the quality of life for individuals at risk and foster a proactive approach to neurodegenerative disease management.

# Contents

# List of Figures

# Chapter 1

# INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting around 60 percent of individuals over the age of 50. It leads to significant challenges in mobility and speech, making it difficult for patients with Parkinson's (PWP) to access regular treatment and monitoring. Early detection of Parkinson's is critical, as it allows for timely intervention that can slow the disease's progression and enable patients to maintain a better quality of life. Similarly, Alzheimer's disease, another neurodegenerative disorder, gradually diminishes cognitive abilities, particularly memory. Alzheimer's severely impairs the brain's functioning, leading to memory loss, confusion, and eventually an inability to perform daily tasks. Unlike Parkinson's, early detection of Alzheimer's remains difficult in the current medical landscape, with symptoms often being mistaken for normal aging. While there is no cure for either disease, detecting them early can help manage their progression, allowing for more targeted treatments that improve patient outcomes.

Disease prediction involves using data-driven models to forecast the likelihood of someone developing a particular condition, like Parkinson's or Alzheimer's. This approach helps healthcare professionals identify high-risk individuals before the full onset of symptoms, enabling personalized interventions and treatments. By analyzing various factors such as genetics, lifestyle, and medical history, predictive models can provide early warning, giving patients a better chance of managing these conditions. Additionally, this method helps optimize healthcare resources

by focusing on prevention and early treatment, reducing the burden of advanced disease management. Ultimately, disease prediction enhances the overall quality of care by allowing for proactive, patient-centered treatment plans that can delay the onset of debilitating symptoms and improve long-term outcomes.

## 1.1 Motivation

The motivation behind a project focused on disease prediction, particularly for conditions like Parkinson's and Alzheimer's, neurodegenerative diseases, particularly Parkinson's and Alzheimer's, pose significant challenges to individuals, families, and healthcare systems worldwide. The growing prevalence of these conditions necessitates innovative approaches for early detection and intervention. Enhancing Quality of Life By predicting the likelihood of developing these conditions, we empower patients and their families to make informed decisions about their health and care strategies. This proactive approach can lead to improved emotional and psychological well-being, fostering a sense of control over their future. Through this project, we aim to make a meaningful impact on the lives of individuals at risk for Parkinson's and Alzheimer's. By harnessing the power of predictive analytics, we strive to enhance early detection, improve care strategies, and ultimately contribute to a healthier future for those affected by these debilitating diseases.

## 1.2 Problem Statement

The rising incidence of neurodegenerative diseases like Parkinson's and Alzheimer's poses major challenges for healthcare. These diseases are often diagnosed late, limiting treatment options and reducing patients' quality of life. The lack of reliable early detection methods delays interventions and adds emotional and financial strain on caregivers. Additionally, current predictive tools are underutilized or lack integration with diverse datasets, limiting the ability to identify at-risk individuals and fully understand these diseases.

## 1.3 Objectives

Disease prediction focuses on improving healthcare outcomes and resource management.Key objectives include early detection for timely intervention and better patient outcomes.Risk assessment helps tailor personalized medicine and preventive strategies.It also aids in managing high-risk populations to improve public health.Predicting disease trends enables efficient resource allocation and reduces hospitalizations.Overall, it optimizes healthcare delivery and enhances the care quality.

## 1.4 Organisation of the Report

The report Disease Prediction presented here is organized into Six chapters. After list of abbreviations and figures, Chapter 1 is the introductory part that describes the topic of the report, states the purpose of the report which is followed by motivation for choosing this particular topic. Further this chapter contains problem definition and objectives. Chapter 2 is literature survey that summarizes all past research papers based on this model of work which gives brief information of past work and the gaps which were identified in the work. It is followed by our contribution towards this modal of work in the project. In Chapter 3 we will see the proposed system that contains key elements of the report body includes architecture of the project, flow of algorithm how project actually works followed by its process of designing. Also contains the details of hardware and software that has been used. Chapter 4 is followed by experiment part which was executed for making this project successful and last but not the least it also contains result. Chapter 5 is all about the Result and Discussion that helps us to narrow the whole topic down to a simple point. Chapter 6 is all about key highlight of our project and it is further followed by future scope of this project topic.

# Chapter 2

# LITERATURE SURVEY

## 2.1 Survey of Existing System

Existing systems for Parkinson's disease (PD) detection have explored various methods, primarily focusing on MRI scans, gait, and genetic data. Studies such as Bilal et al. have used genetic data with an SVM model, achieving an accuracy of 0.889, while others like Raundale, Thosar, and Rane employed keystroke data to predict PD severity using Random Forest classifiers. Audio data has also been used, as seen in Cordella et al.'s research, though their models heavily relied on MATLAB, unlike the open-source, Python-based models in recent studies. Majority of the research emphasizes the use of deep learning, like Ali et al.'s ensemble deep learning models applied to phonation data, though feature selection was lacking in their work. Wang et al. implemented 12 machine learning models on voice biomarkers, achieving a high accuracy of 96.45 percent with a deep learning model, but it required substantial memory. Other methods, such as decision trees, random forest, and KNN, have been used on brain MRI scans for mild cognitive impairment detection in PD patients. However, many existing systems either lack feature selection or require expensive, memory-intensive models, highlighting the need for more efficient and accurate methods like KNN or logistic regression, which are explored in newer research.

## 2.2 Limitation in Existing System

The existing systems for Parkinson's disease (PD) detection have several limitations. While studies like Bilal et al. achieved respectable accuracy using genetic data with an SVM model, these approaches rely on data that may not be readily available for all patients, limiting their widespread applicability. Similarly, Raundale, Thosar, and Rane's use of keystroke data is restricted to specific datasets, and Cordella et al.'s reliance on MATLAB for audio data classification introduces dependency on proprietary software, reducing accessibility. Deep learning models, such as those developed by Ali et al. and Wang et al., have achieved high accuracy, but they often lack feature selection, which could improve performance, and tend to be resource-intensive, requiring significant memory and computational power. Additionally, some models, like those using brain MRI scans, suffer from small datasets, necessitating artificial data augmentation, which can introduce bias or overfitting. Overall, these limitations underscore the need for more efficient, accessible, and less resource-heavy methods, as well as a greater emphasis on feature selection to enhance model performance.

# Chapter 3

# PROPOSED METHODOLOGY

## 3.1 Introduction

In the realm of healthcare, the application of machine learning algorithms has shown significant promise in predicting diseases, particularly through a structured approach that encompasses data collection, preprocessing, feature engineering, model selection, training, evaluation, and ongoing monitoring. Initially, data such as MDVP, Shimmer, HNR, DFA, and various other clinical measurements are gathered to create a robust dataset. This data undergoes thorough preprocessing, ensuring cleanliness and addressing any missing values while encoding categorical variables to facilitate analysis.

Feature engineering plays a crucial role in identifying the most relevant predictors of disease, drawing from demographic details, lifestyle factors, medical history, clinical measurements, cognitive assessments, and observable symptoms. Following this, appropriate machine learning models, such as support vector machines and logistic regression, are selected based on their suitability for the task. The training phase involves using labeled data, where the dataset is divided into training and testing sets to evaluate the model's predictive capabilities.

Model evaluation is conducted using key performance metrics like accuracy, precision, recall, and F1-score, supplemented by ROC curve analysis to provide a

comprehensive view of the model's effectiveness. Once a reliable model is developed, it can be deployed in clinical practice or public health settings to aid in disease prediction. Continuous monitoring and maintenance of the model ensure it remains accurate and relevant, adapting to new data and shifts in the disease landscape. Through this systematic approach, we can achieve high accuracy in predicting health outcomes, thereby enhancing decision-making in healthcare environments.

## 3.2 Methodology

### A. Architecture

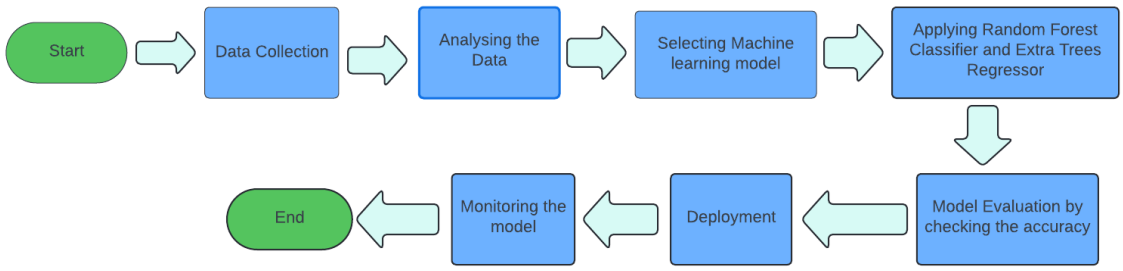


Figure 3.1: Block Diagram

### B. Framework

Disease prediction plays a crucial role in identifying individuals at risk of conditions such as Parkinson's disease or Alzheimer's disease. By leveraging machine learning algorithms and analyzing various data points such as clinical measurements, demographic information, lifestyle factors, and cognitive assessments, healthcare professionals can detect patterns that indicate the early stages of these neurodegenerative diseases.

**Data Collection** - The data collection process for disease prediction involves gathering diverse vocal and clinical features, including MDVP (Multi-Dimensional Voice Program), Shimmer (jitter and shimmer measures), HNR (Harmonics-to-

Noise Ratio), and DFA (Detrended Fluctuation Analysis). This data can be sourced from patient records, clinical assessments, and voice analysis tools. By systematically collecting these parameters, researchers can create a comprehensive dataset that serves as the foundation for predictive modeling.

**Data Preprocessing** - Data preprocessing involved cleansing the dataset by handling missing values, outliers, and inconsistencies. This process involves cleaning the dataset by removing duplicates and correcting inconsistencies in the collected vocal and clinical features such as MDVP, Shimmer, HNR, and DFA.

**Exploratory Data Analysis (EDA)** - EDA was conducted to understand the distribution of variables, correlations, and patterns within the data.

**Feature Engineering** - Feature engineering techniques selects the most relevant features (variables) from the dataset that are likely to be predictive of the disease like Demographic Details, Lifestyle Factors, Medical History, Clinical Measurements, Cognitive and Functional Assessments and Symptoms.

**Model Selection** - Various machine learning algorithms including train-test-split,random forests and SVM were evaluated for disease prediction.

**Model Training** - The dataset was split into training and validation sets. Models were trained on the training data which was 20 percent.

**Model Evaluation** - Assess the model's performance using metrics like accuracy score and confusion matrix.

**Deployment** - The chosen model was deployed into production systems for real-time disease prediction.

**Monitoring and Maintenance** - Ongoing monitoring of model performance in production was conducted. The model was retrained and updated periodically with new data. Continuous evaluation of the model's effectiveness was performed, with adjustments made as needed.

# Chapter 4

# IMPLEMENTATION

## 4.1 Dataset

The dataset utilized in this project comprises various types of data essential for predicting Parkinson's and Alzheimer's diseases.The ranges for each parameter in voice analysis can vary based on the dataset and population. However, general ranges for these parameters, particularly in studies of Parkinson's disease and similar conditions, can be described as follows:

MDVP (Hz): Typically ranges from 70 Hz to 300 Hz. The exact range can depend on the speaker's age and sex.

MDVP (Hz): Usually ranges from around 150 Hz to 450 Hz, again depending on the individual's voice characteristics.

MDVP (Hz): Often falls between 60 Hz and 250 Hz.

MDVP (Abs): Generally ranges from 0.001 to 0.01 seconds, representing the absolute variation in pitch.

MDVP (Relative Average Perturbation): Commonly ranges from 0.1

MDVP (Pitch Perturbation Quotient): Generally ranges from 0.1

Jitter (Difference of Differences of Pitch): Often ranges from 0.0005 to 0.01 sec-

onds.

MDVP : Typically ranges from 0.1 to 1.5, where higher values indicate more amplitude variation.

MDVP (dB): Generally falls between 0.1 dB and 3 dB.

Shimmer : Usually ranges from 0.1 to 1.5.

Shimmer : Commonly ranges from 0.1 to 1.5.

MDVP : Often ranges from 0.1 to 2.0.

Shimmer : Generally falls between 0.1 and 1.0.

NHR (Noise-to-Harmonics Ratio): Typically ranges from 0.1 to 0.7, with higher values indicating more noise.

HNR (Harmonics-to-Noise Ratio): Usually ranges from 5 dB to 30 dB.

RPDE (Recurrence Period Density Entropy): Ranges from 0 to 1, with higher values indicating more complexity.

D2: Typically ranges from 1 to 10, reflecting the complexity of the voice signal.

DFA (Detrended Fluctuation Analysis): Ranges from 0 to 2, with values near 0.5 indicating random noise and higher values indicating long-range correlations.

spread1, spread2: Usually range from 0 to 1, with higher values indicating greater variability in fundamental frequency.

## 4.2 Software and Hardware Used

### 4.2.1 Hardware

- Computer – Dell inspiron 13
- Processor – 11th Gen Intel(R) Core(TM) i5-11300H @ 3.10 GHz 2.61 GHz
- RAM - 8 GB
- Memory Space – 256 GB

### 4.2.2 Software

- Jupyter Notebook
- PyCharm for GUI
- Sklearn

# Chapter 5

# RESULTS AND DISCUSSION

## 5.1 Results and Discussion

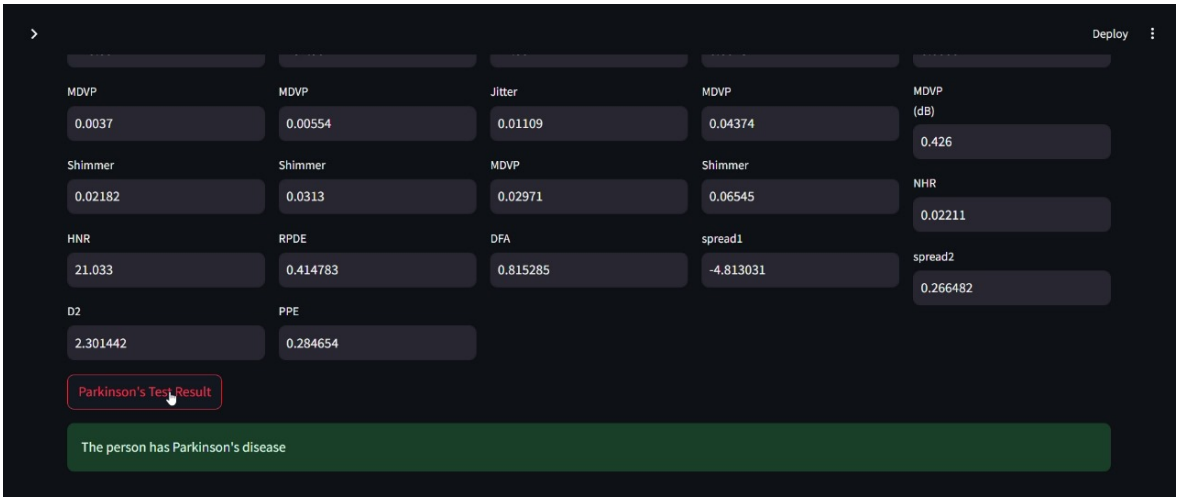


Figure 5.1: Result 1

As shown in Figure 5.1, in this model we are collecting the data such as MDVP, DFA, PPE, RPDE and more.These data is use for identifying critical risk factors and early symptoms, the findings contribute to the ongoing efforts to improve early diagnosis and intervention strategies, ultimately aiming to enhance patient care and quality of life.

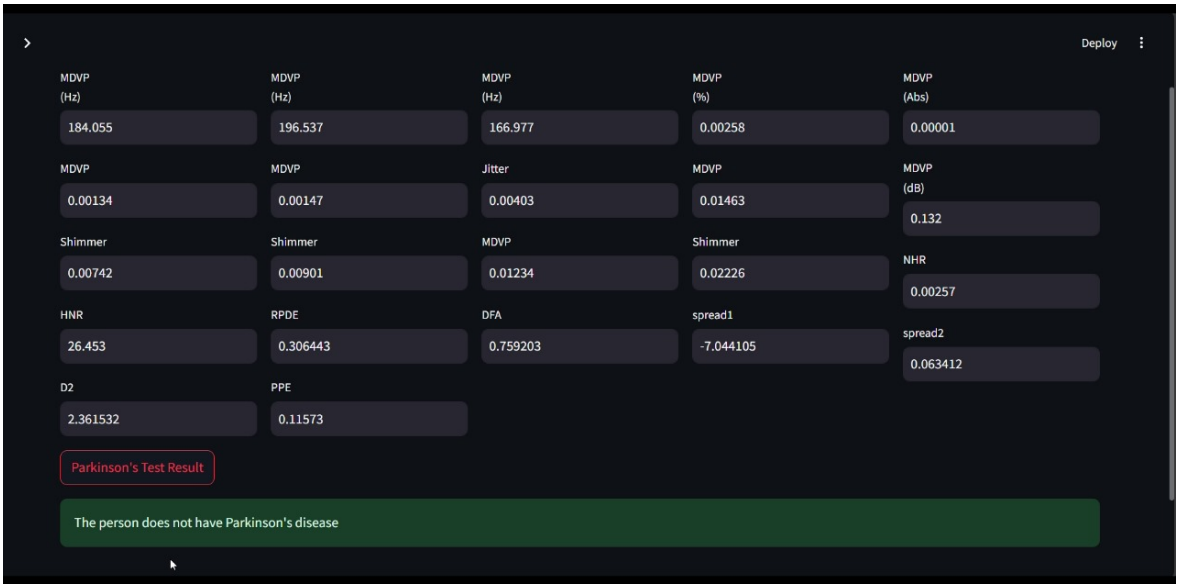


Figure 5.2: Result 2

So by applying above two machine learning models and by calculating confusion matrix we got the accuracy of 87 percent and by using these algorithm we get the result as user can detect parkinson disease or not.

# Chapter 6

# CONCLUSIONS

In this project, we've created a model that predicts whether a person has a specific disease based on their input data. First, we have our diabetes prediction model, after entering the relevant health information, such as glucose levels, blood pressure and more our model processes this data and quickly determines if the person has disease or not.

As shown in Figure 5.2 after inputting the data, the model predicted that the person does not have Parkinson disease. Next we move on to heart disease prediction. This model uses different inputs, like cholesterol levels age and lifestyle factors by entering these details on model and analyzes the data and indicates whether heart disease is present. Its efficient and can provide results in just a few moments. Next we have Parkinson's disease prediction model. This requires inputs such as different brain, MRI measurements and after entering these inputs, the model processes the information to predict if the person has Parkinson's disease. Lastly, we have the breast cancer prediction model this asks for inputs like tumor size age and other relevant factors. Once these details are entered our model analyzes the data and gives the result. Each model is tailored to specific conditions, providing critical insights for early detection and better health management. This technology has the potential to save lives by identifying diseases early and allowing for timely intervention.

## 6.1 Project Highlights

Advantages:

- Multimodal Data Integration
- Early Detection

Limitations:

• Dataset Diversity: The dataset may not fully represent all demographic groups, affecting generalizability.

• Potential Biases: Reliance on existing datasets may introduce biases related to missing variables or underrepresented populations.

Features:

1) Advanced Machine Learning Techniques: Utilizes a variety of state-of-the-art algorithms, including deep learning, to enhance predictive accuracy and model robustness.

2) Focus on Early Detection: Emphasizes identifying early signs and risk factors for Parkinson's and Alzheimer's diseases, aiming to facilitate timely interventions.

## 6.2 Future Scope

Partner with healthcare providers to implement the predictive model in clinical settings, facilitating real-world application and feedback for continuous improvement.

Utilize predictive insights to develop personalized treatment and intervention strategies tailored to individual risk profiles and disease trajectories.

# REFERENCES

[1] Aditi Govindu and Sushila Palwe, “Early detection of Parkinson’s disease using machine learning”, 2023, Procedia Computer Science, vol. 218, pp. 249-261

[2] M. Sudharsan, G. Thailambal,”Alzheimer’s disease prediction using machine learning techniques and principal component analysis (PCA)”,2023, vol. 81, pp.182-190.

[3] James M. Ellison, MD, MPH,”Alzheimer’s vs. Parkinson’s: A Comparison”,2023.

[4] Kaggle Bank Dataset -
https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset
https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set