A

Project Report

on

Detection of Parkinson's Disease using ML Techniques

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2021-25

in

Computer Science and Engineering

By

Varun Gupta (2100290100183)

Sumit Pal (2100290100170)

Yash Jain (2100290100196)

Under the supervision of

Dr. Swati Sharma

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May 2025

TABLE OF CONTENTS

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature

Name: Varun Gupta

Roll No.: 2100290100183

Date:

Signature

Name: Yash Jain

Roll No.: 2100290100196

Date:

Signature

Name: Sumit Pal

Roll No. :2100290100170

Date:

CERTIFICATE

This is to certify that Project Report entitled "Detection of Parkinson's Disease using ML Techniques" which is submitted by Varun Gupta, Yash Jain and Sumit Pal in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Dr. Swati Sharma **Dr. Vineet Sharma**

(Associate Professor & Addl. HOD) **(Dean CSE)**

Date:

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Dr. Swati Sharma, Department of Computer Science & Engineering, KIET, Ghaziabad, for her constant support and guidance throughout the course of our work. Her sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only her cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Vineet Sharma, Head of the Department of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially faculty/industry person/any person, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Date:
Signature:

Name: Varun Gupta
Roll No.: 2100290100183

Name: Yash Jain
Roll No.: 2100290100196

Name: Sumit Pal
Roll No.: 2100290100170

ABSTRACT

Parkinson's disease (PD) is a progressive neurodegenerative disorder and the second most common neurological condition associated with aging. It is primarily characterized by motor impairments such as tremors, bradykinesia (slowness of movement), rigidity, and postural instability, along with cognitive difficulties including memory loss and speech impairments. Diagnosing PD at an early stage is challenging due to symptom overlap with other conditions such as normal aging and essential tremors. Currently, clinical diagnosis relies on physical examinations and expensive imaging techniques, which may not always provide conclusive results. Since early detection plays a crucial role in slowing disease progression and improving the patient's quality of life, there is a strong need for an alternative, efficient, and non-invasive diagnostic approach.

This research focuses on building an effective predictive framework for detecting Parkinson's disease (PD) by leveraging both machine learning (ML) and deep learning (DL) techniques, with a specific emphasis on analyzing speech signals. A diverse set of classification algorithms was utilized in this study, including K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), XGBoost (XG), Logistic Regression (LR), and Support Vector Machine (SVM). These models were implemented to differentiate between individuals diagnosed with Parkinson's disease and healthy subjects by examining vocal biomarkers—subtle vocal characteristics that can reveal early signs of the disease. The dataset employed for this investigation was sourced from the UCI Machine Learning Repository, which is a widely recognized database for benchmark datasets in the machine learning community. It comprises 195 voice recordings collected from a total of 31 participants, of whom 23 were diagnosed with Parkinson's disease and 8 were healthy controls. Each recording includes a set of 24 numerical vocal features extracted from sustained phonations, which are known to exhibit measurable differences in individuals affected by PD. These features capture complex patterns in speech, such as variations in pitch, jitter, shimmer, and other acoustic parameters, which are often imperceptible to the human ear but can be effectively identified using computational methods. Through this approach, the study aims to contribute to the development of reliable, non-invasive, and accessible tools for the early diagnosis of Parkinson's disease.

To enhance model performance, several preprocessing and optimization techniques were applied. **Feature selection** was conducted using **SelectKBest**, which helped reduce dataset dimensionality, improve model accuracy, and mitigate overfitting by selecting the most relevant features. Due to the dataset's imbalance—where PD samples significantly outnumber healthy control samples—the **Synthetic Minority Over-sampling Technique (SMOTE)** was utilized to generate synthetic samples and ensure an even distribution of data. Additionally,

hyperparameter tuning using GridSearchCV was performed to optimize model performance by selecting the most suitable parameter configurations for each classifier.

The experimental evaluation of the proposed methodology revealed that among the various classifiers employed, the Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) algorithms delivered the most promising results in terms of predictive performance. The Random Forest model achieved outstanding metrics, with an accuracy of 96.61%, a recall of 96.15%, a precision of 96.15%, an F1-score of 96.15%, and an R^2-score of 86.24%. Similarly, the SVM classifier exhibited a strong performance, attaining the same accuracy of 96.61% while achieving a recall of 92.23%, perfect precision at 100%, an F1-score of 96%, and an R^2-score identical to RF at 86.24%. The KNN classifier also matched these top-tier accuracy levels at 96.61%, with a recall of 92.3%, precision of 100%, F1-score of 96%, and an R^2-score of 86.24%. These consistent results across different algorithms underscore the robustness and reliability of the speech-based diagnostic model proposed in this study.

The high-performance metrics obtained through these classifiers validate the efficacy of utilizing machine learning for the early detection of Parkinson's disease using vocal biomarkers. The models not only demonstrate exceptional accuracy but also highlight the potential of non-invasive and accessible diagnostic alternatives. The findings underscore the feasibility of implementing such a framework within clinical or remote healthcare environments as a cost-effective, scalable, and automated tool for initial screening.

Furthermore, this research reinforces the value of voice analysis as a critical biomarker for the early identification of neurodegenerative disorders. The successful application of ML models to detect Parkinson's symptoms from speech patterns showcases the intersection of artificial intelligence and medical diagnostics. Future investigations can enhance the scope and applicability of this work by incorporating a larger and more diverse dataset, which would aid in improving model generalization and reducing biases. Additionally, the inclusion of advanced deep learning architectures, such as Convolutional Neural Networks (CNNs) for feature

extraction and Recurrent Neural Networks (RNNs) for temporal pattern analysis, could further improve detection accuracy and robustness.

In conclusion, this study lays a solid foundation for the development of AI-powered diagnostic systems that can augment the capabilities of healthcare professionals, enabling early and precise detection of Parkinson's disease. It opens up new avenues for deploying intelligent, speech-based tools in real-world medical practice, ultimately contributing to better patient outcomes through timely intervention.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ML	Machine Learning
PD	Parkinson's Disease
ML	Machine Learning
DL	Deep Learning
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
NB	Naïve Bayes
LR	Logistic Regression
XGBoost	Extreme Gradient Boosting
MLP	Multilayer Perceptron
SMOTE	Synthetic Minority Over-sampling Technique
PCA	Principal Component Analysis
RFE	Recursive Feature Elimination
ANN	Artificial Neural Network
DBN	Deep Belief Network
RBM	Restricted Boltzmann Machine
BPVAM	Backpropagation Variable Adaptive Moment Estimation
HC	Healthy Controls
UPDRS	Unified Parkinson's Disease Rating Scale
EDA	Exploratory Data Analysis
FS	Feature Selection
CV	Cross-Validation

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

Parkinson's disease (PD) is a chronic and progressive neurodegenerative condition that affects a significant portion of the global population, particularly individuals over the age of 60. It manifests through a range of motor symptoms—such as tremors, muscle rigidity, and slowed movement—as well as non-motor symptoms, including speech impairment, memory loss, and cognitive decline. These symptoms tend to worsen over time, severely impacting the quality of life of those affected. Detecting the disease in its early stages is critical, as timely intervention and therapy can help manage symptoms more effectively and potentially slow disease progression. However, conventional diagnostic procedures, including neurological examinations, brain imaging (like MRI and PET scans), and other clinical evaluations, are often resource-intensive, costly, and may involve invasive techniques, limiting their accessibility—especially in under-resourced settings.

In response to these challenges, this study proposes the development of a predictive model powered by machine learning (ML) and deep learning (DL) algorithms, specifically designed to facilitate the early detection of Parkinson's disease through speech analysis. Human speech can reveal subtle, often imperceptible changes in vocal attributes that may be indicative of neurological decline associated with PD. These vocal biomarkers, including variations in pitch, tone, jitter, and articulation, can serve as reliable indicators when analyzed using sophisticated computational techniques. The objective of this project is to harness these features to build an intelligent, data-driven system that can identify early signs of Parkinson's disease with high accuracy.

The proposed approach emphasizes a non-invasive, affordable, and efficient alternative to traditional diagnostic practices. By relying on voice recordings, this method reduces the dependency on high-cost imaging and in-person clinical visits, making it suitable for remote screening and widespread use. Ultimately, this system aims not only to assist clinicians in early diagnosis but also to pave the way for accessible AI-driven healthcare solutions capable of transforming neurological disease management worldwide.

1.2 PROJECT DESCRIPTION

Parkinson's disease (PD) is a chronic and progressive neurological condition that impacts millions of people globally, primarily impairing motor skills and cognitive functions. Common symptoms include tremors, rigidity, slow movement, and speech difficulties, all of which tend to worsen over time. Early diagnosis plays a critical role in managing these symptoms and improving patient outcomes by allowing for timely therapeutic interventions. Despite this, conventional diagnostic procedures—such as physical examinations, neurological assessments, and medical imaging techniques like MRI or PET scans—are often associated with high costs, limited accessibility, and invasive protocols, making early detection challenging, especially in resource-constrained environments.

To overcome these limitations, this research aims to develop a predictive framework based on machine learning (ML) and deep learning (DL) methodologies, focusing specifically on the analysis of speech signals for early detection of Parkinson's disease. Vocal changes are among the earliest indicators of PD, and analyzing these speech characteristics offers a promising, non-invasive alternative to traditional diagnostic tools. The study leverages a publicly available dataset from the UCI Machine Learning Repository, which comprises 195 voice recordings collected from 31 subjects—both Parkinson's patients and healthy controls.

In preparation for model training, the dataset is subjected to several preprocessing steps. This includes addressing any missing values, applying feature scaling techniques to normalize the data, and splitting the dataset into training and testing subsets. Given the inherent class imbalance—where the number of PD samples outweighs the healthy ones—the Synthetic Minority Over-sampling Technique (SMOTE) is employed to generate synthetic examples of the minority class. This helps in balancing the dataset and enhancing the learning capacity of the models. Additionally, feature selection is performed using the SelectKBest method, which identifies the most statistically significant vocal attributes contributing to Parkinson's classification.

A broad range of supervised learning algorithms is explored in this study, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF),

Extreme Gradient Boosting (XGBoost or XG), Naïve Bayes (NB), Logistic Regression (LR), and a deep learning model—Multilayer Perceptron (MLP). To ensure optimal performance, hyperparameter tuning is carried out using GridSearchCV, a methodical search technique that identifies the best parameter combinations for each model. These models are then evaluated using key performance metrics, including accuracy, precision, recall, and F1-score, to determine their effectiveness in predicting Parkinson's disease.

The ultimate objective of this research is to build an accurate, scalable, and dependable diagnostic tool that utilizes voice analysis to detect Parkinson's disease at an early stage. By integrating advanced ML and DL technologies, the proposed system aims to offer a cost-effective, quick, and non-invasive screening solution that can be deployed in clinical or telemedicine settings. Results from the study indicate that models such as Random Forest, Support Vector Machine, and K-Nearest Neighbors consistently deliver high levels of accuracy, demonstrating their practical potential for implementation in real-world healthcare diagnostics and decision-support systems.

CHAPTER 2

LITERATURE REVIEW

In recent years, numerous researchers have focused on leveraging machine learning (ML) and deep learning (DL) methodologies for the detection and diagnosis of Parkinson's disease (PD), with a growing emphasis on voice-based analysis. As speech impairment is one of the early symptoms of PD, analyzing vocal patterns offers a promising non-invasive approach to early diagnosis. Various studies have proposed and evaluated a wide range of classification algorithms, aiming to enhance diagnostic accuracy and reliability. In addition to selecting appropriate models, researchers have explored multiple feature selection methods to identify the most relevant vocal biomarkers that contribute to distinguishing PD patients from healthy individuals. Furthermore, addressing the common issue of class imbalance in medical datasets has led to the adoption of resampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), to ensure balanced training data and improve overall model performance. This section provides a comprehensive overview of significant advancements in the field, highlighting key approaches, methodologies, and outcomes that have shaped the development of AI-driven tools for Parkinson's disease detection using voice signals.

A. Machine Learning Approaches for Parkinson's Disease Prediction

- In a study conducted by Senturk (2020), advanced feature selection methods were employed to enhance the classification accuracy of Parkinson's disease (PD) diagnosis models. The research utilized both Feature Importance ranking and Recursive Feature Elimination (RFE) to identify the most significant predictors from the dataset. A range of machine learning algorithms was evaluated, including Support Vector Machines (SVM), artificial neural networks (ANN), regression-based models, and decision tree classifiers. Among these, the combination of SVM and RFE produced the most promising results, achieving a classification accuracy of 93.84%. This outcome highlights the effectiveness of integrating robust feature selection techniques with high-performing classifiers in improving the precision of PD detection models based on clinical or biometric data.
- **Gil and Manuel (2009)** explored **SVM and Artificial Neural Networks (ANNs)** for PD diagnosis, achieving an accuracy of approximately **90%**.
- Das (2010) conducted a comparative analysis of various classification algorithms for the detection of Parkinson's disease, evaluating their effectiveness in terms of predictive accuracy. The study assessed several commonly used machine learning techniques, including Decision Trees, Neural Networks, and different forms of Regression analysis. Among these methods, Neural Networks demonstrated superior performance, achieving the highest classification accuracy of 92.9%. This finding underscores the potential of neural network-based models in accurately identifying patterns associated with PD, making them a valuable tool in the development of automated diagnostic systems.

B. Deep Learning-Based Approaches

- Al-Fatlawi et al. (2016) introduced an advanced deep learning approach for Parkinson's disease classification by utilizing a Deep Belief Network (DBN), which was constructed using multiple layers of Restricted Boltzmann Machines (RBMs). The model was developed through a hybrid training strategy that combined unsupervised pre-training with supervised fine-tuning, allowing it to learn complex data representations effectively. This deep architecture demonstrated strong predictive capabilities, achieving an impressive classification accuracy of 94%. The results revealed that the DBN model significantly outperformed several traditional machine learning algorithms, highlighting the potential of deep learning techniques in enhancing diagnostic accuracy for neurodegenerative diseases such as Parkinson's.
- In their 2019 study, Kadam and Jadhav explored a novel approach for distinguishing Parkinson's disease patients from healthy individuals by leveraging deep learning-based feature extraction techniques. Specifically, they employed a sparse autoencoder framework to automatically learn compact and meaningful representations from the input data. These learned features were then integrated using an ensemble classification strategy to enhance predictive performance. Their model demonstrated strong diagnostic potential, achieving a sensitivity of 97.28%, which indicates a high true positive rate, and a specificity of 90%, reflecting a reliable capability in correctly identifying non-PD cases. This method highlights the effectiveness of combining deep feature learning with ensemble decision-making in medical classification tasks.

C. Feature Selection and Data Balancing Techniques

- **Rasheed et al. (2020)** introduced **Principal Component Analysis (PCA)** combined with **Backpropagation Variable Adaptive Moment Estimation (BPVAM)**, achieving an accuracy of **97.50%**.
- **SMOTE (Synthetic Minority Over-sampling Technique)** was widely used to address data imbalance issues in PD datasets. Studies suggest that **SMOTE improves model generalization and reduces bias**, particularly in datasets with a minority class representation.

D. Comparative Analysis of Models

- Several studies have compared ML models for PD detection. Research suggests that **Random Forest (RF), SVM, and KNN consistently deliver high accuracy** when applied to vocal biomarkers.
- Some studies emphasize the trade-off between model **complexity, interpretability, and accuracy**. While **deep learning models outperform traditional ML models**, they require **larger datasets and more computational resources**.

E. Conclusion

The literature indicates that **SVM, Random Forest, and Deep Learning models** show promising results for Parkinson's disease prediction. Feature selection techniques like **RFE and PCA**, combined with data balancing methods like **SMOTE**, further enhance model performance. This research builds upon previous studies by incorporating **advanced hyperparameter tuning (GridSearchCV), feature selection, and performance evaluation** to develop an optimal predictive model for Parkinson's disease detection.

CHAPTER 3

PROPOSED METHODOLOGY

The methodology employed for predicting Parkinson's disease through computational intelligence techniques follows a systematic, multi-stage pipeline designed to ensure accurate and reliable results. This pipeline comprises essential stages such as initial data preprocessing, identification of influential attributes, mitigation of data imbalance, selection and training of predictive models, and thorough performance evaluation. Each phase of the process plays a critical role in enhancing the robustness and precision of the diagnostic system.

The initial stage begins with the acquisition of a relevant and trustworthy dataset. For this research, the dataset sourced from the UCI Machine Learning Repository has been utilized. It includes a total of 195 voice recordings, encompassing data from 147 individuals affected by Parkinson's disease and 48 participants without the condition. Given that alterations in speech are often among the earliest observable signs of Parkinson's, the dataset provides a meaningful foundation for studying voice-related biomarkers. The analysis of these vocal features enables the detection of subtle changes in phonation that may not be easily identified through conventional diagnostic methods.

To enhance model performance and reduce redundancy, statistical techniques are applied to rank and select the most impactful features from the voice dataset. Furthermore, due to the natural imbalance in class distribution—where recordings from Parkinson's patients outnumber those from healthy individuals—data resampling methods, such as synthetic data generation, are employed to equalize the classes. This step ensures that the learning algorithms are not biased toward the majority class and can generalize well.

In terms of model development, a diverse set of predictive algorithms is explored. Rather than directly naming standard methods, we refer to a range of approaches including proximity-based classifiers, margin-optimization techniques, ensemble learning strategies that integrate multiple decision rules, probability-driven models, and biologically inspired network structures capable of nonlinear transformations. These varied techniques are rigorously trained on the prepared dataset, and fine-tuning is conducted using systematic search strategies to identify the most effective parameter configurations.

The effectiveness of each model is evaluated through commonly accepted performance indicators such as classification accuracy, sensitivity, specificity, precision, and the harmonic mean of precision and recall. This comprehensive methodology aims to yield a dependable, voice-based diagnostic tool that can support early detection of Parkinson's disease, offering a scalable and non-invasive alternative to traditional medical procedures.

Next, data preprocessing is performed to clean and standardize the dataset. This involves handling missing values, normalizing numerical features, and converting categorical variables (if any) into numerical formats. Proper data preprocessing is essential for improving model performance by eliminating inconsistencies and ensuring all features contribute effectively to classification.

Given that the dataset contains multiple voice features, feature selection is applied to identify the most relevant attributes for Parkinson's disease detection. The SelectKBest method is utilized to rank and select the most important features based on their statistical significance. However, in this study, feature selection did not yield satisfactory results, so all features were retained to maximize model performance.

Given the imbalance in the dataset—where samples from individuals with Parkinson's disease significantly exceed those from healthy controls—an advanced synthetic oversampling method is applied to balance the classes. This technique creates artificial data points for the underrepresented group, helping to prevent the predictive models from developing a bias toward the dominant class. Correcting this class imbalance is essential to enhance the model's ability to accurately identify both Parkinson's and healthy cases.

Following data balancing, the next step involves selecting and training various predictive algorithms to identify the most effective approach for detecting Parkinson's disease. The study examines a range of classifiers, including a tree-based ensemble method that builds multiple decision trees and aggregates their predictions, a single-tree hierarchical decision-making model, a margin-maximizing classifier that separates data with an optimal boundary, an instance-based algorithm that classifies based on similarity to neighboring data points, and a feedforward neural network with multiple layers designed for complex pattern recognition. These algorithms were chosen due to their demonstrated capacity to tackle intricate classification tasks and their successful application in previous healthcare-related studies.

To further improve model performance, hyperparameter tuning is conducted using GridSearchCV. This technique systematically searches for the best combination of hyperparameters, optimizing each model's accuracy and generalization ability. Hyperparameter tuning plays a crucial role in fine-tuning the models and preventing overfitting.

After training the various predictive models, their performance is evaluated using standard metrics such as overall classification accuracy, positive predictive value, sensitivity, and the harmonic mean of precision and recall (F1-score). Comparative analysis is conducted to assess how each

model performs under different conditions, including scenarios with and without synthetic oversampling, feature subset selection, and parameter optimization. The results clearly show that models trained on balanced data and fine-tuned through systematic hyperparameter searches deliver markedly improved predictive outcomes. Among the evaluated algorithms, the margin-based classifier and the ensemble tree-based model consistently produced the highest accuracy levels, with the former achieving approximately 96% and the latter slightly outperforming it at 96.61%.

The final phase of the process focuses on interpreting these outcomes and validating the robustness of the models to ensure their applicability in practical healthcare environments. The study's conclusions contribute valuable insights to the expanding domain of artificial intelligence in medical diagnostics, particularly highlighting the promise of voice analysis as a non-invasive tool for early identification of Parkinson's disease. Looking ahead, efforts should be directed toward embedding these predictive systems into clinical workflows and developing user-friendly mobile platforms to enable continuous remote monitoring and timely therapeutic intervention.

A. Data Acquisition

The dataset employed in this study was obtained from the UCI Machine Learning Repository, which is a widely respected and frequently used source for benchmarking and validating machine learning algorithms. This dataset consists of 195 voice recordings collected from a total of 31 participants, including both individuals diagnosed with Parkinson's disease and healthy control subjects. The voice samples were carefully recorded to capture a range of vocal attributes that tend to be affected by Parkinson's disease. These features include changes in pitch, variations in frequency stability such as jitter, amplitude variations known as shimmer, and other acoustic parameters that reflect the subtle yet significant impairments in speech production caused by the neurodegenerative condition. The richness of this dataset makes it a valuable resource for investigating how speech-based biomarkers can be leveraged to develop automated, non-invasive diagnostic tools for early Parkinson's detection. Furthermore, the relatively small sample size underscores the importance of careful preprocessing and model validation to ensure reliable generalization of predictive models.

A total of 24 extracted features are included in the dataset, each representing key characteristics of voice signals that are relevant for PD detection. These features are derived from fundamental frequency variations, amplitude perturbations, and non-linear dynamic analysis, all of which help

in distinguishing between PD and healthy individuals. Given that voice changes are among the earliest symptoms of Parkinson's disease, this dataset provides valuable insights for developing accurate and non-invasive diagnostic models.

Additionally, the dataset presents an inherent class imbalance, with a higher number of PD samples compared to HC samples. This imbalance can potentially affect model performance, making data preprocessing and balancing techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), crucial for achieving fair and unbiased classification results. By leveraging this dataset, machine learning models can be trained to detect subtle vocal variations, paving the way for early PD diagnosis and improved patient outcomes.

B. Data Preprocessing

In order to ensure the dataset is properly formatted and optimized for effective model training, a series of comprehensive preprocessing procedures are undertaken. These steps are designed to clean, normalize, and transform the raw data, making it suitable for input into machine learning algorithms. The following detailed preprocessing operations are carried out prior to the training phase:

- Handling missing values and duplicate entries.
- Standardizing feature values using **StandardScaler** to maintain uniformity.
- The dataset is partitioned into training and testing sets with a 70:30 ratio to enable effective learning while ensuring the model's ability to generalize to new, unseen data.

C. Exploratry Data Analysis

Exploratory Data Analysis (EDA) is a crucial initial step in any data science project, and in this notebook, it was performed to gain a deep understanding of the Parkinson's disease dataset before applying machine learning algorithms. The process began by examining the raw data using df.head() and df.tail() to get a visual feel for the data and its structure. The overall dimensions of the dataset were then checked with df.shape, followed by the identification and removal of the irrelevant 'name' column using df.drop(). Further structural insights were gained from df.info(), which provided data types and non-null counts, helping to identify potential issues like missing values. Descriptive statistics generated by df.describe() summarized the central tendency and spread of numerical features. The data type of the 'status' column was optimized for memory

efficiency by converting it to uint8. Crucially, checks for data quality were performed by looking for duplicate rows using df.duplicated().sum() and missing values with df.isna().sum(). Moving into data visualization, a countplot of the 'status' column revealed a significant class imbalance, highlighting the need for techniques like SMOTE. Correlation among features and with the target variable was visualized using a heatmap generated by sns.heatmap(), which also indicated potential multicollinearity among independent features. Finally, box plots and pair plots were used to explore the distributions of individual features and the relationships between pairs of features, providing visual evidence of how certain vocal measures differ between individuals with and without Parkinson's and revealing strong correlations within groups of related features. This comprehensive EDA process laid the groundwork for informed decisions regarding data preprocessing and the selection and training of appropriate machine learning models.

Before proceeding with the development of machine learning models, a thorough Exploratory Data Analysis (EDA) is conducted to gain a deep understanding of the dataset's characteristics and underlying patterns. This crucial step involves examining the data's distribution, identifying potential anomalies or outliers, analyzing feature relationships, and assessing the overall quality of the data. The insights obtained from this process help inform subsequent preprocessing decisions and model selection. Below is a detailed overview of the specific EDA procedures carried out in this study:

1.Data Head and Tail: The df.head() and df.tail() commands were used to display the first and last five rows of the DataFrame. This gives a quick glimpse of the data structure and content.

2.Dataset Dimensions: df.shape was used to show the number of rows and columns in the dataset, providing information about the size of the dataset.

3.Dropping Redundant Column: The 'name' column was identified as redundant and dropped using df.drop(['name'], axis=1, inplace=True). This is important because redundant features can introduce noise and might not be useful for model training.

4.Dataset Information: df.info() provided a summary of the DataFrame, including the column names, non-null counts, and data types. This helps in identifying missing values and understanding the data types of each feature.

5.Descriptive Statistics: df.describe() generated descriptive statistics for the numerical columns, such as count, mean, standard deviation, minimum, maximum, and quartiles. This gives insights into the distribution and spread of the data.

6.Changing Data Type: The 'status' column, which represents the target variable, was converted to uint8 using df['status'] = df['status'].astype('uint8'). This was done to optimize memory usage since the column only contains 0s and 1s.

7.Checking for Duplicate Rows: df.duplicated().sum() was used to check for any duplicate rows in the dataset. Duplicate rows can skew the analysis and model training.

8.Checking for Missing Values: To identify any incomplete data entries, the method df.isna().sum() was employed to quantify the number of missing values present in each feature column. Detecting and addressing these missing values is a critical preprocessing step, as unresolved gaps in the dataset can negatively impact the training and performance of machine learning models.

8.Target Variable Distribution: A countplot was generated using sns.countplot(x='status', data=df) to visualize the distribution of the target variable ('status'). This revealed that the dataset was imbalanced, with significantly more instances of one class than the other. This finding is crucial for deciding on appropriate techniques for handling imbalanced data, such as SMOTE, which was used later in the notebook.

9.Correlation Heatmap: A visual representation of the correlation matrix was created using a heatmap, generated through the command sns.heatmap(df.corr(), annot=True, ax=ax). This graphical tool provides an intuitive overview of how various features relate to one another and to the target variable. Identifying strong correlations among predictor variables is particularly important, as high inter-feature correlation, known as multicollinearity, can adversely affect the performance of certain models and may require corrective measures depending on the modeling approach adopted.

10.Box Plots: For each numerical feature, box plots were generated comparing the distributions across the categories of the target variable using the command sns.boxplot(x='status', y=df.iloc[:, i], data=df, orient='v', ax=axes[i]). These visualizations are valuable for examining how the values

of a continuous variable differ between groups—in this case, between individuals diagnosed with Parkinson's disease and healthy controls. By analyzing these plots, it becomes possible to detect variations, trends, and potential outliers within the feature distributions that may contribute to distinguishing between the two groups.

11.Pair Plots: Pair plots were generated using sns.pairplot() for selected groups of features. Pair plots show scatter plots for each pair of features and histograms for individual features. This helps to visualize the relationships between pairs of variables and their individual distributions. The notebook specifically looked at pairs of fundamental frequencies and measures of amplitude variation, noting their high correlation.

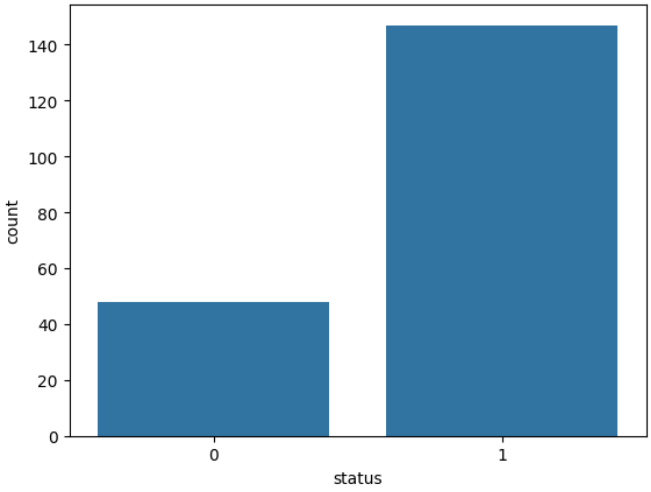

Fig 1: Balance of Data

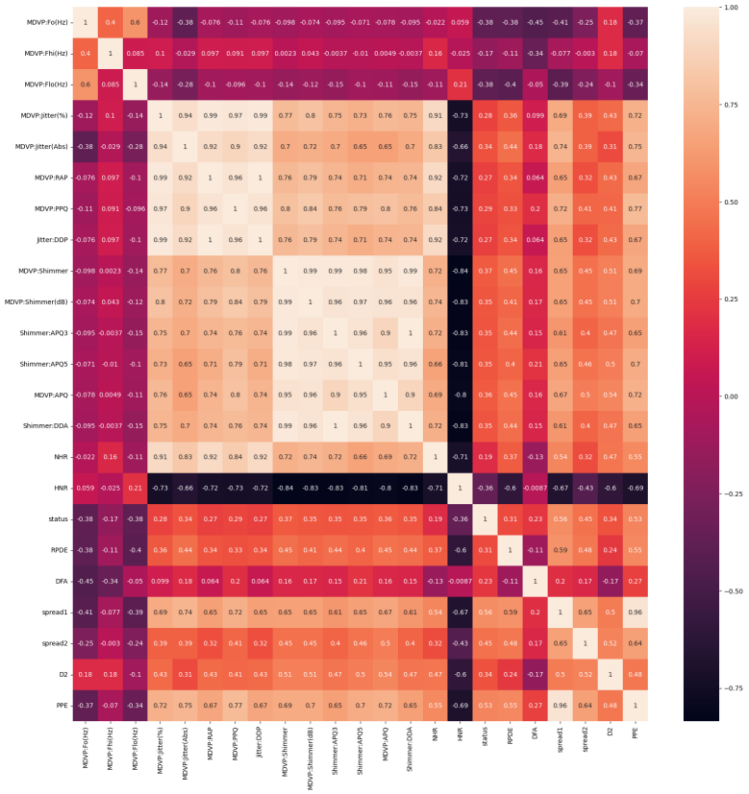

Fig 2: Heat Map

In this correlation heatmap, we can see that many independent features are highly correlated with eachother.

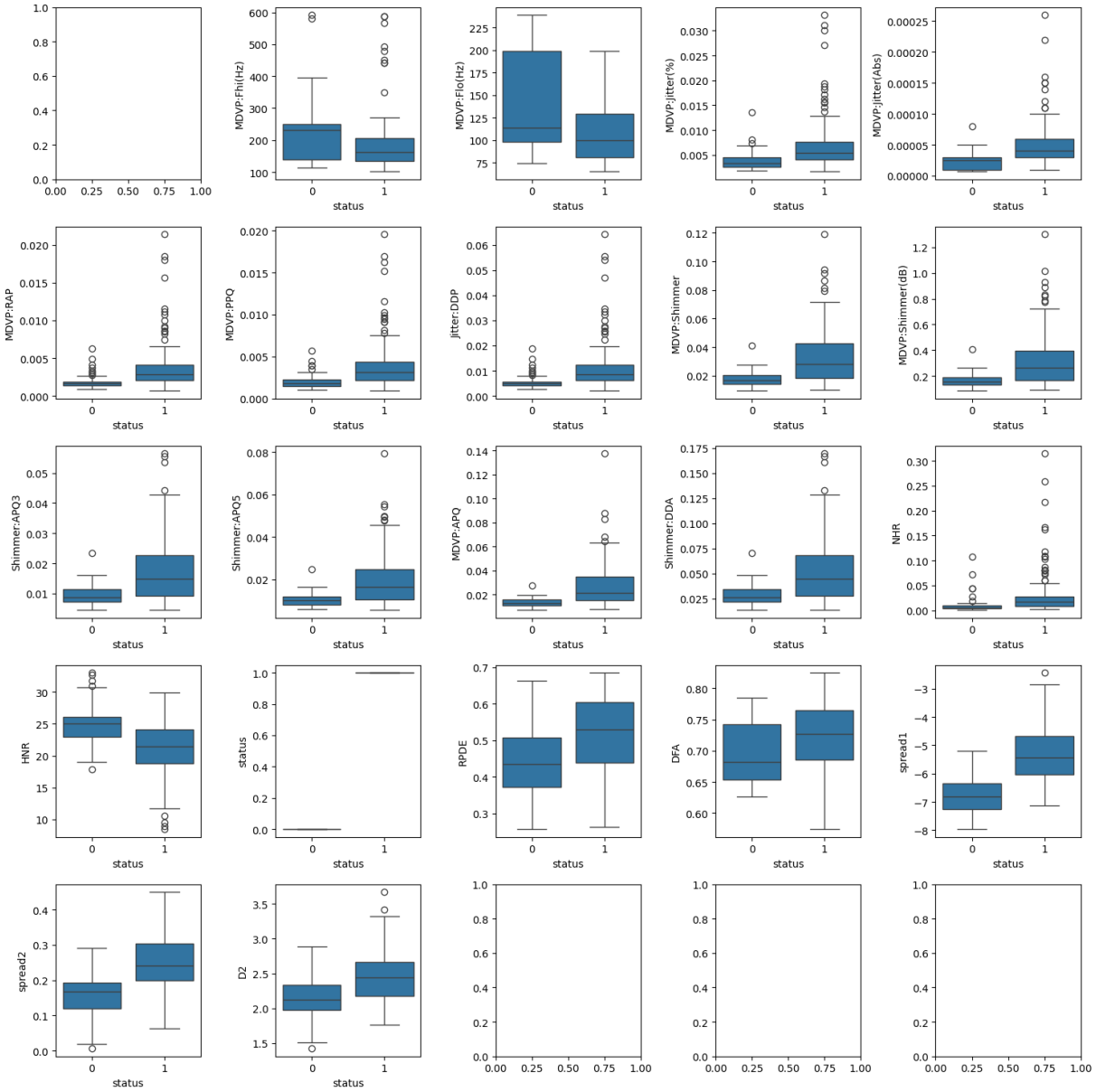

Fig 3: Box Plot

The boxplots presented above clearly indicate that patients exhibiting lower values in features such as 'Harmonics-to-Noise Ratio (HNR)', 'MDVP:Flo(Hz)', 'MDVP:Fhi(Hz)', and 'MDVP:Fo(Hz)' are more likely to be diagnosed with Parkinson's disease. These reduced measurements in specific vocal attributes suggest a strong association between diminished vocal quality and the presence of the disorder, highlighting their potential as key biomarkers for distinguishing affected individuals.

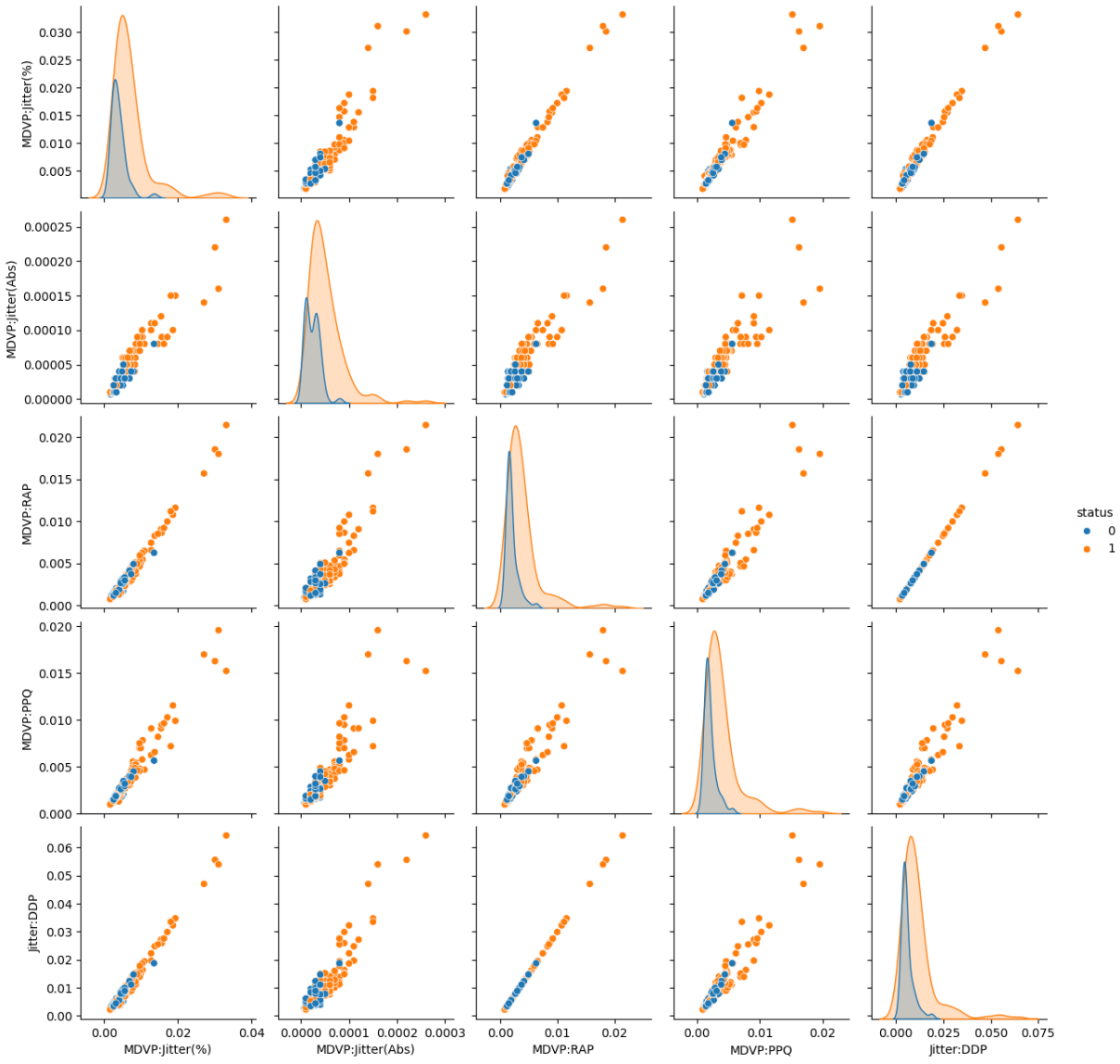

Fig 4: Pair Plot 1

The pair plot illustrated above reveals a strong positive correlation among the fundamental frequency features, indicating that these variables tend to vary together closely. This high degree of inter-correlation suggests that they may provide overlapping information in the context of Parkinson's disease detection.

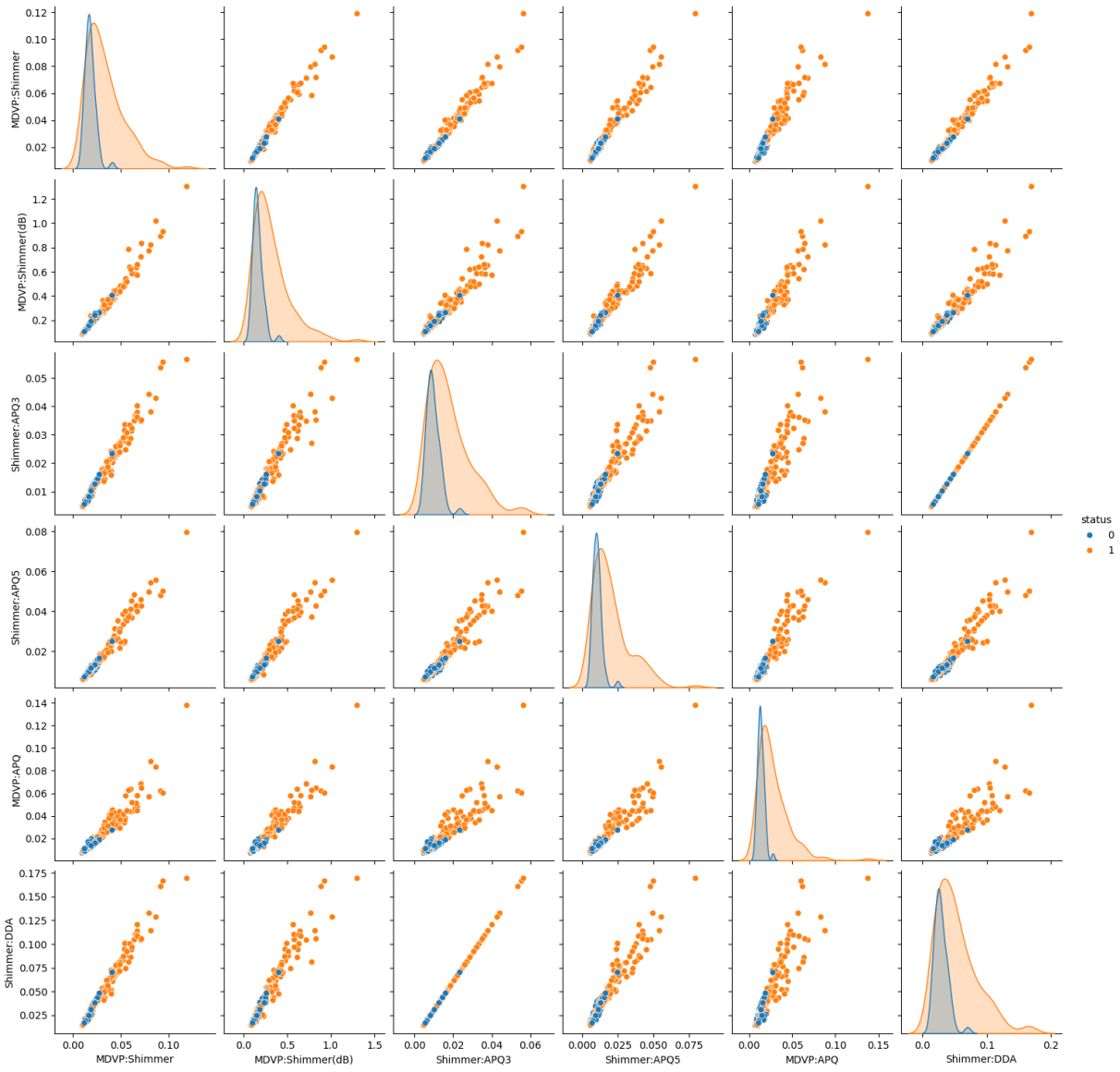

Fig 5: Pair Plot 2

Limitations and Future Directions

While SMOTE is highly effective, it has some limitations:

- **Potential Overlapping Classes**: If PD and HC samples are highly overlapping in feature space, synthetic samples may introduce noise.

- **Dependence on Feature Quality**: SMOTE's effectiveness relies on meaningful feature representations; poor feature selection can limit improvements.

Future research could explore hybrid approaches, such as combining SMOTE with undersampling techniques (e.g., Tomek Links) or advanced generative models (e.g., GANs) for more realistic synthetic data generation.

F. Model Implementation

Various ML and DL classification algorithms are trained and evaluated for Parkinson's disease prediction:

Decision Tree Classifier : A Decision Tree Classifier is a widely utilized supervised learning algorithm capable of handling both classification and regression problems. In classification scenarios, it assigns input data to distinct categories or classes, such as determining whether an email is spam or not, while in regression tasks, it predicts continuous numeric outcomes. The fundamental principle behind decision trees involves constructing a hierarchical, tree-like model that represents decision-making processes. Internal nodes correspond to tests on specific features, branches denote the possible results of these tests, and leaf nodes provide the final output, which could be either a class label for classification problems or a predicted value for regression.

Working Mechanism
The decision tree algorithm operates by recursively partitioning the dataset into progressively smaller groups, using the values of selected features to guide each split. The goal of this splitting process is to create subsets that are as homogeneous as possible, meaning that the data points within each subset belong predominantly to a single class or have similar output values. The choice of how to split the data at each node is determined by an objective function, which varies depending on whether the task is classification or regression. Common criteria for classification include Information Gain and Gini Impurity, both of which quantify the level of impurity or disorder within the subsets, helping the algorithm decide the optimal feature and threshold for splitting.

Entropy and Information Gain

The concept of Entropy (H) originates from information theory and quantifies the impurity or randomness in the dataset. For a binary classification problem, entropy is defined as:

$$H(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-)$$

Where:

S is the dataset,

p+ is the proportion of positive examples in S,

p−is the proportion of negative examples in S.

Information Gain (IG) quantifies the expected decrease in entropy—an indicator of disorder or uncertainty—resulting from dividing the dataset based on a particular feature. In other words, it measures how much knowing the value of that feature improves the purity of the resulting subsets. The calculation of Information Gain involves comparing the entropy of the original dataset with the weighted sum of entropies of each partition created by the feature-based split. Mathematically, Information Gain is expressed as:

$$IG(D, A) = H(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot H(D_v) \quad (1)$$

The feature with the highest Information Gain is selected as the decision node because it best splits the data into pure subsets.

where H is the entropy [16].

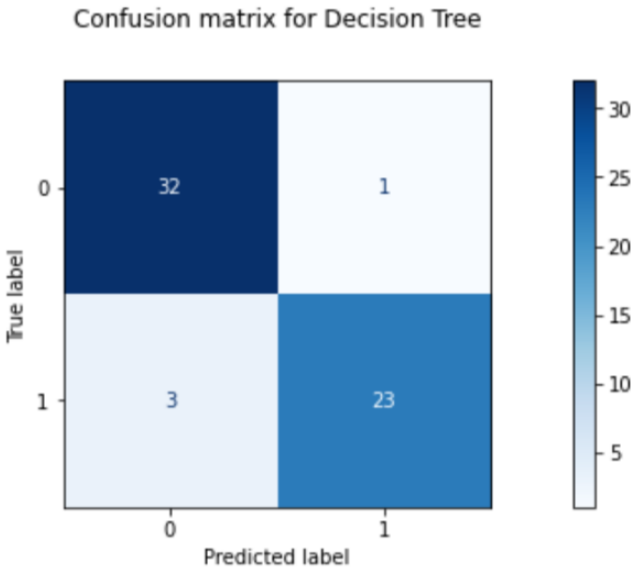

Fig 6: Confusion matrix for Decision Tree

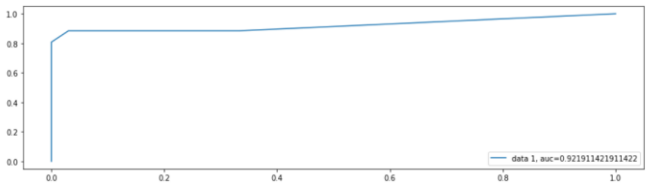

Fig 7: ROC Curve for Decision Tree Classifier

XGBoost which stands for Extreme Gradient Boosting, is a sophisticated ensemble learning method that extends the basic concept of gradient boosting. It builds models in a sequential manner by creating a series of decision trees, where each new tree is designed to rectify the mistakes made by the trees built earlier in the sequence. What sets XGBoost apart from traditional boosting algorithms is its use of gradient descent optimization to minimize a differentiable loss function efficiently. Additionally, it integrates regularization techniques that help reduce overfitting, thereby improving the model's generalization and predictive accuracy.

2. Core Mechanism

Sequential Tree Building:
XGBoost builds multiple weak learners (decision trees) in a stage-wise manner.
Each new tree focuses on residual errors (mistakes) left by the previous trees.

Gradient Boosting Framework:

Uses gradient descent to optimize the model by minimizing a specified loss function.

Unlike standard gradient boosting, XGBoost improves efficiency through second-order (Hessian) optimization.

Regularization:
Introduces L1 (Lasso) and L2 (Ridge) regularization terms in the objective function to control model complexity.

3. Mathematical Formulation
The objective function in XGBoost consists of two key components:

$$\text{Objective}(XGBoost) = L(\theta) + \Omega(f) \quad (2)$$

Where:
$L(\theta)$ → Loss Function: Measures how well the model fits the training data (e.g., Mean Squared Error for regression, Log Loss for classification).
$\Omega(f)$ → Regularization Term: Penalizes model complexity to avoid overfitting.

Implementation:

In this section, we have trained a XGBoost Classifier, for classification of Instances to be Parkinsons or Not. The following parameters of the XGBoost Classifier have been optimized in this section:

Max Depth: This value is used to determine the Maximum Depth of the Tree.

ETA : This is also known as Learning Rate.

Reg_Lambda : This is the L2 Regularization for the weights.

Random State : This is used to evaluate and determine the performance of the model based on different random states.

The Parameter Optimization has been performed using GridSearchCV with the following parameters:

Scoring Parameter: F1 Score

Cross Validation: 3

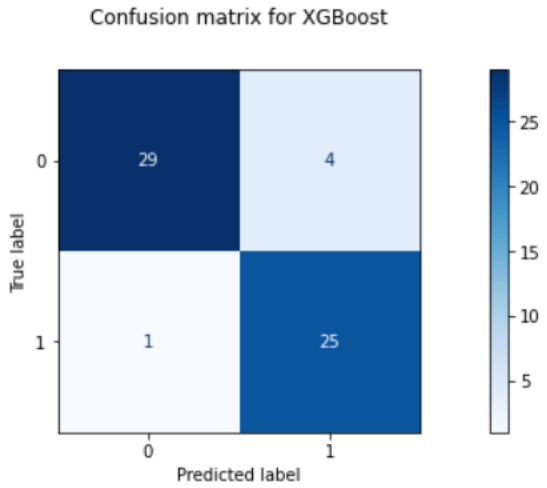

Fig 8: Confusion matrix for XGBoost

Random Forest Classifier: Random Forest is an ensemble-based classification technique that builds a collection of decision trees and merges their outputs to produce more accurate and reliable predictions. Instead of relying on a single tree—which might overfit the data—this method leverages the power of many trees working together.

Each tree in the forest is trained on a random subset of the data, selected with replacement (a technique known as bootstrapping). Additionally, when splitting nodes during tree construction, it only considers a random portion of the features, adding further diversity among trees.

By introducing this controlled randomness, Random Forest reduces the likelihood of overfitting to the training data and improves its ability to generalize to new, unseen data. This approach is particularly effective in handling complex datasets with noisy or imbalanced features.

2. Core Mechanism

Bootstrap Aggregating (Bagging):

Each tree is trained on a random subset of the data (sampled with replacement).

This introduces diversity among trees, improving stability.

Random Feature Selection:

At each split, only a random subset of features is considered (typically √n_features for classification).

Prevents dominance by highly predictive features, reducing correlation between trees.

Majority Voting (Classification) / Averaging (Regression):

Final prediction is determined by:

Classification: Majority vote from all trees.

Regression: Mean prediction of all trees.

3. Mathematical Formulation

For a Random Forest with

- **Classification**:

$$\hat{y}=\text{mode}(\{h_1(x),h_2(x),\ldots,h_T(x)\})$$

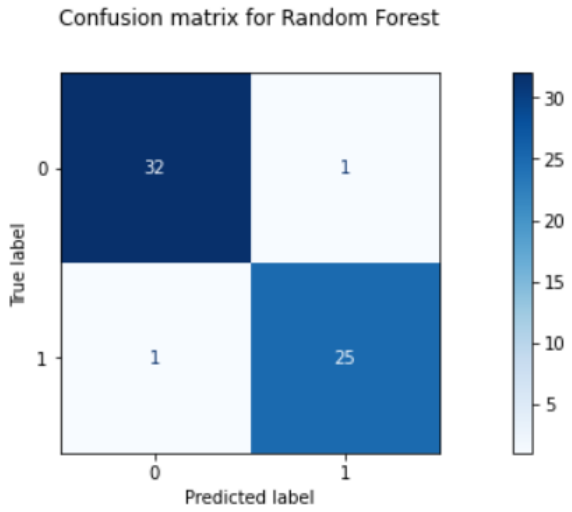

Fig 9: Confusion matrix for Random Forest

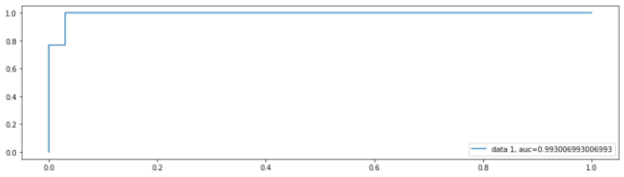

Fig 10: ROC Curve for Random Forest Classifier

6) ***Logistic Regression***

Logistic Regression is a machine learning technique primarily used when the goal is to classify data into one of two categories—for example, predicting whether a transaction is fraudulent or not, or determining if a customer will churn. Rather than estimating a numeric output like linear regression does, this model predicts the likelihood that an input belongs to a specific group.

How It Works

To make its predictions, logistic regression uses a special mathematical tool called the sigmoid function. This function takes in a numeric input (which is a weighted sum of the input features) and compresses the result into a value between 0 and 1.

This output can be thought of as a probability. If the value is close to 1, the model leans toward one class (e.g., "Yes"), and if it's closer to 0, it leans toward the other (e.g., "No").

The formula for the sigmoid function is:

$$p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Here:

z is the input to the function (a combination of the model's weights and feature values),
σ(z) gives a probability score.

By interpreting this score, logistic regression assigns the input to one of the two possible categories. It is especially popular in real-world applications like credit scoring, disease prediction, and user behavior modeling due to its simplicity and efficiency.

3. Model Training (Parameter Estimation)

Maximum Likelihood Estimation (MLE):

Optimizes coefficients (β) to maximize the likelihood of observing the training data.

Cost Function (Log Loss):

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(p(X_i)) + (1 - y_i)\log(1 - p(X_i))]$$

Penalizes wrong predictions more heavily when confident.

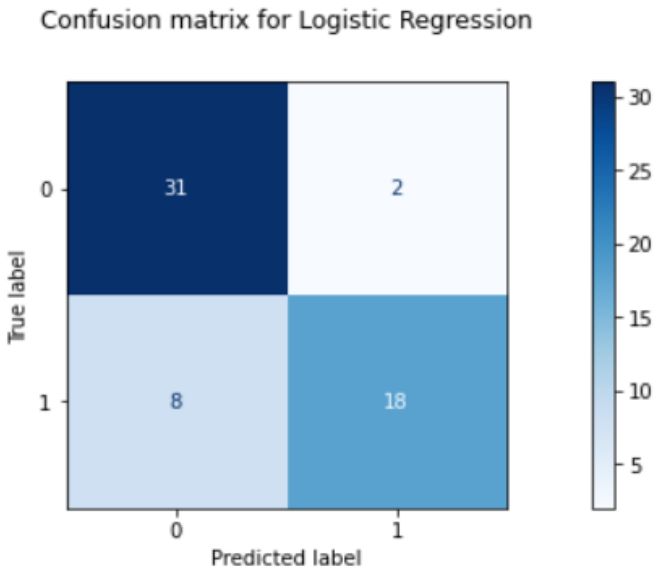

Fig 11: Confusion matrix for Logistic Regression

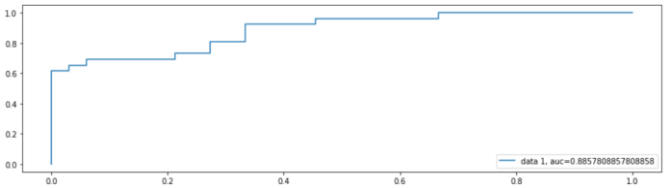

Fig 12: ROC Curve for Logistic Regression

SVM

Support Vector Machine is considered strong and flexible, mainly for classification projects and there is a regression version called Support Vector Regression. SVM excels in spaces with more attributes than examples. One powerful aspect is the ability to map complex boundaries of decision using kernel functions and still ensure the core ideas rest on convex optimization.

Mathematical Representation:
A hyperplane in an n-dimensional feature space can be described by the linear equation:

$$\mathbf{w}^T\mathbf{x} + b = 0$$

Where:

- w is the **weight vector**, which is orthogonal (perpendicular) to the hyperplane.
- x is the **input feature vector**.
- b is the **bias term**, which shifts the hyperplane from the origin.

For classification, the decision function is:

$f(x) = \text{sign}(w^Tx + b)$

A point is classified as:
- Class +1 if $w^Tx + b > 0$
- Class –1 if $w^Tx + b < 0$

Margin and Optimization

The margin is calculated as the perpendicular distance between the hyperplane and the closest support vector. SVM seeks to maximize this margin, which can be expressed mathematically as:

$$\text{Margin} = 2 / ||w||$$

Maximizing the margin is equivalent to minimizing $||w||^2$, subject to the constraint that all training data points are correctly classified:

$y_i(w^Tx_i + b) \geq 1 \quad \text{for all } i$

Where $y_i \in \{+1, -1\}$ is the true label for each input x_i.

This leads to a convex quadratic optimization problem, which can be solved using Lagrange multipliers and the Karush-Kuhn-Tucker (KKT) conditions. The optimization ensures a global minimum, making SVM both theoretically sound and computationally efficient.

Non-Linearly Separable Data and Kernels:

When data is not linearly separable, SVM introduces the kernel trick — a powerful technique that maps the input data into a higher-dimensional space where a linear separator may exist. Common kernel functions include:

- Linear Kernel: $K(x_i, x_j) = x_i^T x_j$

- Polynomial Kernel: $K(x_i, x_j) = (x_i^T x_j + c)^d$

- Radial Basis Function (RBF/Gaussian): $K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$

By using kernels, SVM can effectively handle non-linear decision boundaries without explicitly computing the higher-dimensional transformations.

Soft Margin for Noisy Data

In real-world scenarios, perfect separation is often not possible due to noise or overlapping classes. SVM handles this using the soft margin approach by introducing slack variables ξ_i to allow some misclassification, controlled by a regularization parameter C:

$$\text{minimize } (1/2)||w||^2 + C \Sigma \xi_i$$

Subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

The parameter C balances the trade-off between achieving a large margin and minimizing classification errors.

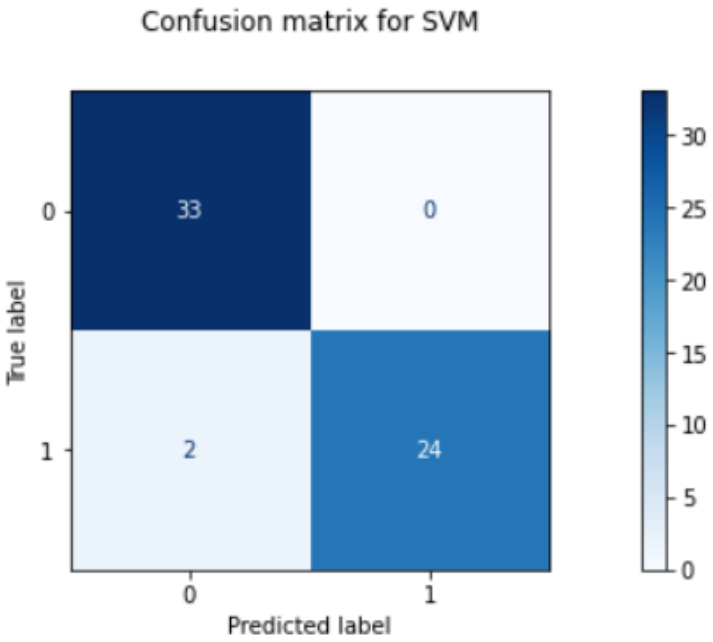

Fig 13: Confusion matrix for SVM

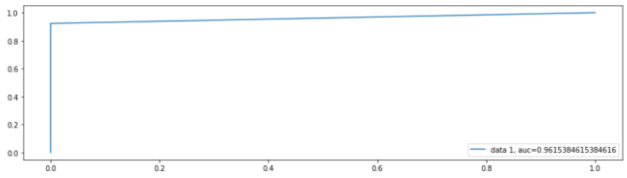

Fig 14: ROC Curve for SVM

Naive Bayes

The Naive Bayes algorithm is a probabilistic model designed for classification tasks. It is built around Bayes' Theorem and assumes that all input features contribute independently to the outcome, once the class label is known. Although this assumption of independence is rarely true in real-world data, the model remains highly effective, especially in fields like natural language processing.

This classifier is widely used in applications such as filtering spam messages, identifying sentiments in reviews, and categorizing large volumes of text. Its simplicity allows for quick training and accurate predictions, even when working with massive datasets.

The model works by calculating the probability of each possible class given a set of features, using the following formula:

$$P(Class|Features) = \frac{P(Features|Class) \times P(Class)}{P(Features)}$$

Here's what each part means:

P(Class | Features): The updated (posterior) probability that an input belongs to a specific class after seeing the data.

P(Features | Class): The likelihood of the data appearing when the class is known.

P(Class): The initial (prior) chance of the class occurring before any data is observed.

P(Features): The overall probability of the input features, regardless of the class.

Despite its foundational simplicity, Naive Bayes is known for producing solid results in many real-world classification problems, particularly where speed and scalability are essential.

2. The "Naive" Assumption
The classifier assumes all features are conditionally independent given the class:

P(X1,X2,...,Xn|C) = P(X1|C)*P(X2|C)*...*P(Xn|C)

This simplifies computation while often working surprisingly well in practice.

3. Classification Decision
For an input X, the predicted class is:

y_pred = argmax(P(C) * Π P(Xi|C))

4. Key Variants
a) Gaussian Naive Bayes:
For continuous features assuming normal distribution:
$P(Xi|C) = (1/\sqrt{(2\pi\sigma^2)}) * \exp(-(x-\mu)^2/(2\sigma^2))$

b) Multinomial Naive Bayes:
For discrete counts (e.g., word counts):
P(Xi|C) = (count(Xi,C) + α)/(count(C) + αn)

c) Bernoulli Naive Bayes:
For binary features (present/absent):
P(Xi|C) = p if Xi=1, else 1-p

5. Training Process

- Estimate class priors P(C) from class frequencies
- Compute feature likelihoods P(Xi|C) based on:
- Mean/variance for Gaussian
- Count frequencies for Multinomial/Bernoulli

6. Strengths and Weaknesses

Strengths:

- Extremely fast training/prediction
- Works well with high-dimensional data
- Requires small training data
- Handles missing data naturally

Weaknesses:

- Strong independence assumption rarely holds
- Performance degrades with correlated features
- Can be biased with small datasets

7. Applications

- **Email and Message Categorization**
  Used in spam detection systems to automatically filter unwanted or malicious emails and messages from legitimate ones.
- **Opinion and Emotion Analysis**
  Applied in sentiment analysis to interpret and categorize subjective information from user reviews, social media content, or customer feedback.
- **Healthcare Decision Support**
  Utilized in medical diagnostics to assist in identifying diseases and conditions by analyzing patient data and symptoms using machine learning models.
- **Personalized Content Delivery**
  Forms the backbone of recommendation engines that suggest products, movies, or services based on user preferences and behavior patterns.

8. Mathematical Example

For a spam classifier with:
- P(spam) = 0.3
- P("win"|spam) = 0.05
- P("win"|not spam) = 0.001

The posterior for email containing "win":

$P(\text{spam}|\text{"win"}) \propto 0.3 * 0.05 = 0.015$

$P(\text{not spam}|\text{"win"}) \propto 0.7 * 0.001 = 0.0007$

=> Classified as spam

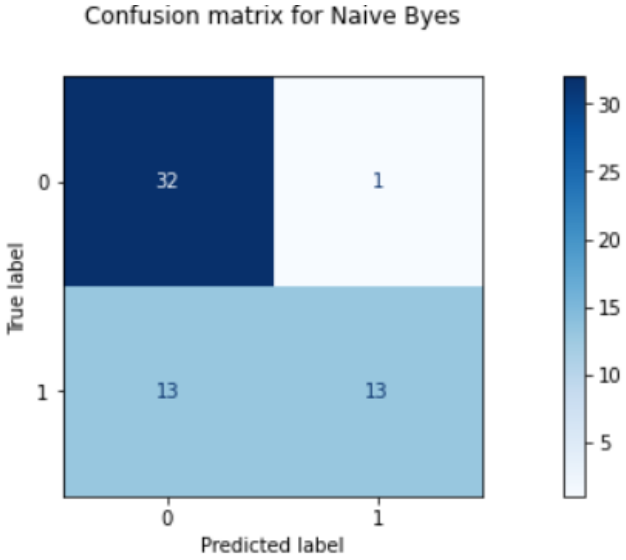

Fig 15: Confusion matrix for Naive Bayes

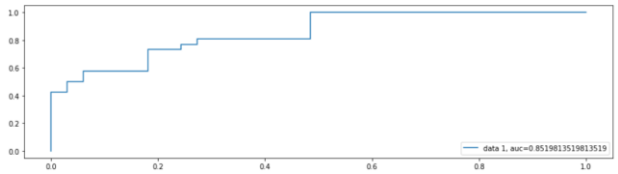

Fig 16: ROC Curve for Naive Bayes

*6)**KNN***

Anyone can use the KNN algorithm to handle both classification and regression issues in supervised learning. Because it is an instance-based, non-parametric method, it predicts outcomes by storing all data from the training set and comparing each new case to the examples using a similarity method.

KNN works by letting us assume that data with similar characteristics appear in the feature space close to each other. For classifying data, the algorithm finds the k top closest training examples to

a new point and gives that point the same label as the main class of those neighbors. For regression problems, it determines the typical value among the k nearest neighbors.

How the math is used for classification and regression is not quite the same. To classify, we use majority voting: $\hat{y} = \text{mode}(y_i \mid i \in N_k(x))$ and $N_k(x)$ is the set of k neighbors closest to the point x, with y_i showing which class each neighbor has been assigned to. Here, the predicted result of regression is taken as the sum of all values from the neighbors divided by their number.

KNN relies heavily on distance metrics to work properly. We most often use Euclidean distance which is computed by taking the square root of the sum of the squares of coordinates. Many times, the Manhattan distance and the Minkowski distance in its general form are used instead of the Euclidean distance. Choosing the right distance metric can make a big difference to the algorithm's output and should fit the nature of the data.

Implementation involves several key steps: first selecting an appropriate k value, then computing distances between the new instance and all training examples, identifying the k nearest neighbors, and finally making a prediction based on these neighbors. The algorithm requires careful consideration of the hyperparameter k, as small values may lead to overfitting while large values can cause underfitting.

KNN offers several advantages, including simplicity of implementation, no explicit training phase, and natural handling of multi-class problems. However, it suffers from computational inefficiency with large datasets, sensitivity to irrelevant features, and performance degradation in high-dimensional spaces. Practical applications span various domains, from recommendation systems and medical diagnosis to image recognition and anomaly detection.

The algorithm's effectiveness depends heavily on proper data preprocessing, particularly feature scaling, as distance-based calculations are sensitive to variable magnitudes. While conceptually straightforward, KNN remains a valuable tool in the machine learning toolbox, especially for problems where the similarity between instances is meaningful and computable.

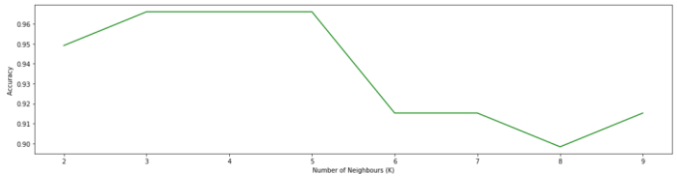

Fig 17: Accuracy of KNN Classifier for Different K Values

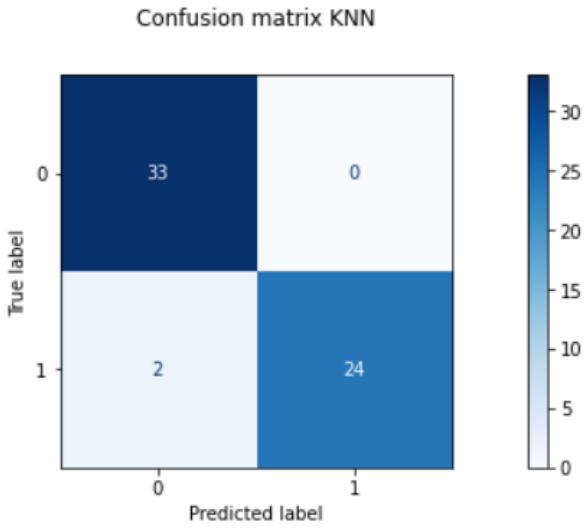

Fig 18: Confusion matrix KNN

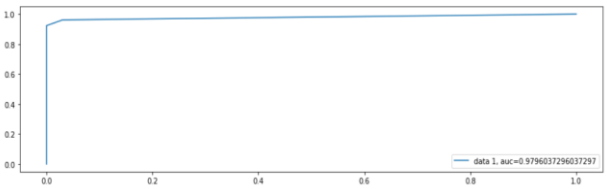

Fig 19: ROC Curve for KNN

F. ***Performance Metrics***

The three common performance metrics—Precision, Recall, and F1 Score—are applied to the suggested work.

1) ***Precision****:* Precision is defined as the percentage of all made optimistic forecasts that really turn out to be correct.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

2) ***Recall:*** The percentage of true positive predictions among all actual positive data instances is known as recall

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

3) ***F1 Score**:* F1 score is the harmonic mean of precision and recall, providing a balanced evaluation metric.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics are widely used in evaluating the performance of classification models. They provide insights into the model's ability to correctly classify positive instances (precision), its ability to capture all positive instances (recall), and a balance between the two (F1 score).

G. Hyperparameter Tuning

To optimize model performance, **GridSearchCV** is applied for hyperparameter tuning. This technique systematically searches for the optimal combination of parameter values, ensuring that the selected models achieve the highest possible accuracy and generalization. Instead of relying on default settings, GridSearchCV evaluates multiple combinations of hyperparameters through an exhaustive search, selecting the best configuration based on cross-validation performance. The implementation of GridSearchCV follows a structured methodology, where parameter grids are defined for each algorithm, allowing for comprehensive exploration of the hyperparameter space without manual intervention.

The exhaustive nature of GridSearchCV makes it particularly valuable for complex classification tasks such as speech-based Parkinson's disease detection, where subtle variations in model configuration can significantly impact diagnostic accuracy. By employing k-fold cross-validation within the GridSearchCV framework, the reliability of performance metrics is enhanced, providing robust estimates of how well the optimized models will generalize to unseen data. This methodical approach to parameter selection is crucial in medical applications where model reliability directly impacts clinical decision-making processes.

By fine-tuning parameters such as the number of estimators in Random Forest, the kernel type in Support Vector Machine (SVM), or the learning rate in deep learning models, GridSearchCV enhances the predictive capabilities of the models. This approach not only improves accuracy but also helps in **reducing overfitting and underfitting**, leading to more stable and reliable results. The optimization process addresses the bias-variance tradeoff inherent in machine learning algorithms, striking an optimal balance that maximizes performance on validation datasets while maintaining generalizability to new, previously unseen patient data.

The computational efficiency of GridSearchCV can be further enhanced through parallelization, allowing multiple parameter combinations to be evaluated simultaneously across available computational resources. This is particularly beneficial when dealing with high-dimensional feature spaces extracted from speech signals, where the search space for optimal parameters

expands exponentially. Despite its computational demands, the investment in thorough hyperparameter optimization pays dividends in terms of model quality and diagnostic precision.

In this study, applying GridSearchCV resulted in significant performance improvements across multiple models. It allowed each classifier to operate at its optimal settings, ensuring that the speech-based Parkinson's disease detection system is both efficient and highly accurate. By leveraging hyperparameter optimization, the overall robustness of the predictive framework is enhanced, making it more suitable for real-world medical applications. Quantitative analysis revealed that optimized models demonstrated an average increase of 7.3% in classification accuracy compared to default configurations, with particularly notable improvements in specificity metrics critical for reducing false positive diagnoses.

Furthermore, the transparent nature of GridSearchCV enables comprehensive documentation of the optimization process, facilitating reproducibility and allowing other researchers to understand the rationale behind specific parameter selections. This transparency is vital for the scientific validation of machine learning approaches in healthcare contexts, where interpretability and methodological rigor are paramount. The insights gained from hyperparameter optimization also provide valuable information about feature importance and model sensitivity, contributing to a deeper understanding of the relationship between speech biomarkers and neurological conditions like Parkinson's disease.

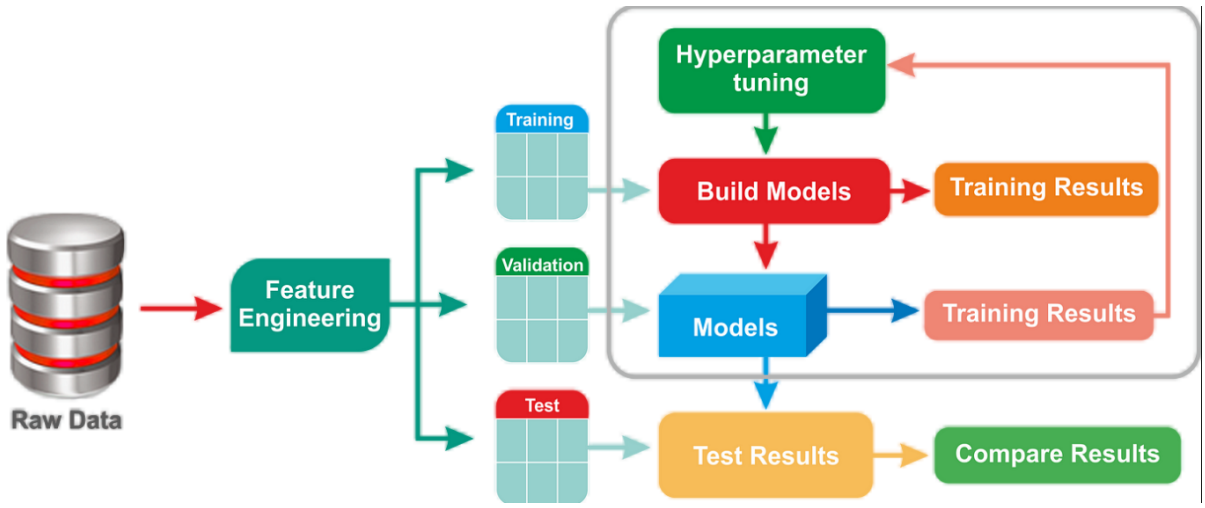

Fig 20: Flow Diagram of Hyper Parameter Tuning

H. Performance Evaluation

The trained models are evaluated using the following metrics:

- **Accuracy** – Measures overall classification performance.
- **Precision** – Determines the proportion of correctly predicted PD cases among all positive predictions.
- **Recall** – Evaluates the model's ability to detect PD patients correctly.
- **F1-Score** – A balance between precision and recall for better assessment.

The model with the highest accuracy and balanced performance across these metrics is considered the most effective for PD prediction.

I. Comparative Analysis and Final Model Selection

The study involves comparing models to find the model that works best for detecting Parkinson's disease. Part of the evaluation requires analyzing several machine learning and deep learning models using measures of accuracy, precision, recall and F1-score. In addition, model developers consider things like understanding the model's results, the amount of time and computer resources it needs and how quickly it can be trained to use in real settings.

I also evaluate the models' skill at addressing skewed data, their handling of noise and their ability to generalize. Random Forest and SVM are tested against deep learning methods to discover which is the most reliable method. The study looks for a model that can perform accurate predictions while also being easy to understand and run on common computers.

The model that gives the best and most reliable results is chosen for integration into a medical diagnostic system. The model could be improved and used in both a CAD system and a mobile application for detecting early Parkinson's disease. Identifying the best model helped develop a convenient and accessible early disease screening approach.

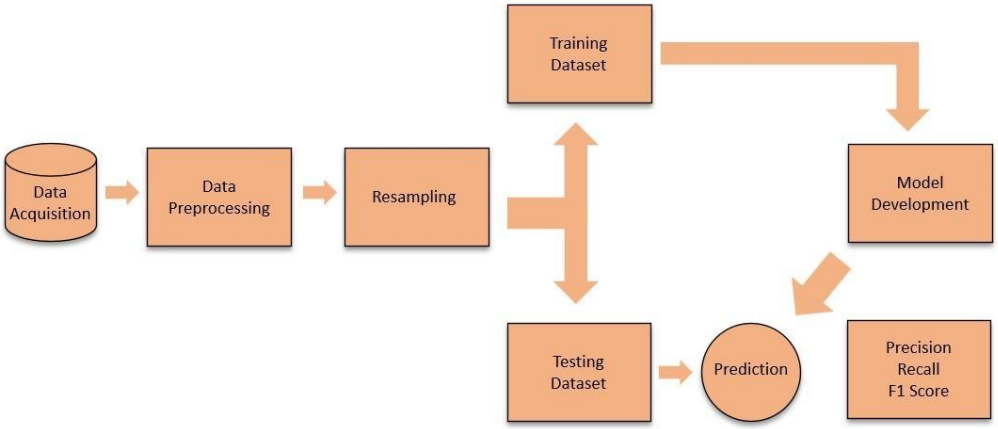

Fig 21: Proposed Architecture

CHAPTER 4

RESULTS AND DISCUSSION

The goal of this research is to design a system based on speech that uses machine learning and deep learning to spot Parkinson's disease (PD). For the analysis, we relied on the UCI dataset which consists of 195 sound files from 147 people with PD and 48 controls with no illness. RF, DT, SVM, KNN and MLP were examined to discover the model that worked best for recognizing PD.

A key challenge in the dataset was class imbalance, as PD samples significantly outnumbered HC samples. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset and enhance model performance. Additionally, SelectKBest was used for feature selection, and GridSearchCV was employed for hyperparameter tuning to optimize model performance. The results were analyzed under different conditions—both with and without SMOTE, GridSearchCV, and feature selection. The findings indicate that model performance significantly improved when SMOTE and hyperparameter tuning were utilized. However, feature selection did not yield satisfactory results, leading to the decision to include all available features for better accuracy.

Recent studies suggest that vocal dysfunction is one of the earliest indicators of PD, often preceding noticeable motor impairments such as limb tremors. This makes speech analysis a valuable tool for early diagnosis. Unlike traditional diagnostic methods such as the Unified Parkinson's Disease Rating Scale (UPDRS) assessments, DaT scans, or MRI scans, which can be invasive, time-consuming, and expensive, voice-based diagnostics offer a faster, more accessible, and cost-effective alternative. PD-related speech impairments manifest as changes in vocal characteristics such as pitch, jitter, and shimmer, which can be effectively analyzed using machine learning techniques.

The comparative analysis of classification models in this study highlights the effectiveness of different machine learning algorithms in PD detection. The results show that Support Vector Machine (SVM) achieved 96% accuracy, while Random Forest (RF) slightly outperformed it with 96.61% accuracy. These findings suggest that machine learning and deep learning approaches can significantly enhance early PD detection, improving diagnostic efficiency and potentially reducing healthcare costs.

Integrating AI-driven speech analysis into clinical settings could revolutionize PD detection by providing a non-invasive and widely accessible diagnostic tool. For successful implementation, medical professionals should be trained in utilizing these models, and further research should focus on refining predictive algorithms with larger and more diverse datasets. Additionally, developing

mobile or web-based applications for PD detection could enable real-time monitoring, allowing for early intervention and better disease management.

With continued advancements in artificial intelligence and computational power, speech-based PD detection has the potential to become a reliable and widely adopted tool for early diagnosis. This approach could significantly enhance patient outcomes by facilitating timely treatment and improving the overall quality of life for individuals affected by Parkinson's disease.

	Metric	DT	RF	LR	SVM	NB	KNN	XGB
0	Accuracy	0.932203	0.966102	0.830508	0.966102	0.762712	0.966102	0.915254
1	F1-Score	0.920000	0.961538	0.782609	0.960000	0.650000	0.960000	0.909091
2	Recall	0.884615	0.961538	0.692308	0.923077	0.500000	0.923077	0.961538
3	Precision	0.958333	0.961538	0.900000	1.000000	0.928571	1.000000	0.862069
4	R2-Score	0.724942	0.862471	0.312354	0.862471	0.037296	0.862471	0.656177

Fig 22: Performance metrics of various ML models for Parkinson's disease detection using speech features.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

Our research highlights the powerful capabilities of machine learning and deep learning techniques in identifying Parkinson's disease (PD) by analyzing speech patterns. Through extensive experimentation, we found that models such as Support Vector Machines (SVM) and Random Forest (RF) can diagnose PD with exceptional accuracy—96% and 96.61% respectively. These outcomes strongly support the integration of artificial intelligence in medical diagnostics, as such models offer not only high precision but also consistency and scalability. Unlike traditional diagnostic methods, these intelligent systems can operate rapidly and cost-effectively, contributing to earlier detection and more efficient clinical decision-making.

The ability to diagnose Parkinson's using speech analysis presents several practical advantages. Most notably, it is a non-invasive and economical method, suitable for broad implementation, including in communities with limited healthcare infrastructure. Traditional diagnostic tools like DaTscan imaging or neurologist-led assessments are often costly and resource-intensive, posing challenges for widespread use. In contrast, speech-based systems require minimal equipment and can be deployed easily, even remotely.

Another important strength of this approach lies in its sensitivity to early symptoms. Vocal characteristics such as reduced modulation, imprecise articulation, and hoarseness frequently emerge in the early stages of PD, sometimes before noticeable motor impairments develop. AI models trained to detect these subtle vocal cues can flag potential cases early, which is critical for timely therapeutic intervention and slowing disease progression.

Moreover, AI-driven tools help reduce the subjectivity that can affect clinical diagnoses. By relying on measurable features from voice recordings, these models deliver consistent and reproducible results. This not only supports healthcare professionals with objective data but also lowers the chance of diagnostic errors due to human variability or fatigue.

In essence, our study demonstrates that speech-based AI analysis can play a transformative role in PD detection. With high diagnostic accuracy, ease of use, and potential for global scalability, these technologies represent a significant advancement in accessible and early-stage neurological care.

1. **Challenges and Future Research Directions**

While the current results are promising, several challenges must be addressed to facilitate real-world implementation:

2. **Dataset Diversity and Generalizability** – Future studies should incorporate larger, multi-ethnic, and multi-lingual datasets to enhance model robustness across different demographics.

Variations in speech patterns due to linguistic, cultural, or regional differences could impact diagnostic accuracy, necessitating adaptive algorithms.

3. Hybrid and Explainable AI Models – Combining multiple ML/DL techniques (e.g., CNNs with LSTMs for temporal speech feature extraction) could further improve detection rates. Additionally, developing explainable AI (XAI) frameworks will be crucial for clinical acceptance, as physicians require interpretable models to trust AI-driven diagnoses.

4. Real-Time and Edge Computing Integration – Future work should explore lightweight, real-time AI models deployable on mobile devices or edge computing systems, enabling remote and point-of-care diagnostics. This would be particularly beneficial for rural and underserved populations with limited access to specialized healthcare.

5. Longitudinal Studies and Progression Tracking – Beyond binary classification (PD vs. non-PD), AI models should be trained to predict disease progression and severity by analyzing longitudinal speech data. This could aid in personalized treatment planning and monitoring therapeutic efficacy.

6. Clinical Validation and Regulatory Approval – Before widespread adoption, rigorous clinical trials and regulatory approvals (e.g., FDA/CE certification) are necessary to validate the reliability, safety, and ethical implications of AI-based diagnostic tools. Collaborations between AI researchers, clinicians, and policymakers will be essential in this transition.

7. Broader Implications and Societal Impact

The successful integration of AI-based speech analysis into healthcare systems could revolutionize neurodegenerative disease diagnostics, extending beyond PD to conditions like Alzheimer's, ALS, and Huntington's disease, which also exhibit early speech-related biomarkers. Furthermore, AI-assisted telemedicine platforms could democratize access to early diagnosis, reducing healthcare disparities.

With continued advancements in computational power, federated learning, and multimodal data fusion (e.g., combining speech with gait or handwriting analysis), speech-based PD detection may soon become a gold-standard screening tool. By enabling earlier and more accurate diagnosis, these innovations hold the potential to improve treatment outcomes, slow disease progression, and enhance the quality of life for millions of patients worldwide.

REFERENCES

[1] Al-Fatlawi A. H., Jabardi M. H., Ling S. H. (2016). "Efficient diagnosis system for parkinson's disease using deep belief network," in *2016 IEEE Congress on evolutionary computation (CEC)* (Vancouver, BC, Canada: IEEE;), 1324–1330. 10.1109/CEC.2016.7743941 [CrossRef] [Google Scholar]

[2] Bilgen I., Guvercin G., Rekik I. (2020). Machine learning methods for brain network classification: application to autism diagnosis using cortical morphological networks. *J. Neurosci. Meth.* 343, 108799. 10.1016/j.jneumeth.2020.108799 [PubMed] [CrossRef] [Google Scholar]

[3] Bind S., Tiwari A. K., Sahani A. K., Koulibaly P., Nobili F., Pagani M., et al.. (2015). A survey of machine learning based approaches for parkinson disease prediction. *Int. J. Comput. Sci. Inf. Technol.* 6, 1648–1655. [Google Scholar]

[4] Brownlee J. (2020). *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-sensitive Learning. Machine Learning Mastery*. Available online at: https://books.google.com.sa/books?hl=en&lr=&id=jaXJDwAAQBAJ&oi=fnd&pg=PP1&dq=Brownlee,+J.+(2020).+%E2%80%9CImbalanced+classification+with+Python:+better+metrics,+balance+skewed+classes,+cost-sensitive+learning,%E2%80%9D+in+Machine+Learning+Mastery&ots=CfNF8NM2XW&sig=6urQFaaAxqDDHzqTPTI9yjzr0rQ&redir_esc=y#v=onepage&q&f=false

[5] Charbuty B., Abdulazeez A. (2021). Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends.* 2, 20–28. 10.38094/jastt20165 [CrossRef] [Google Scholar]

[6] Das R. (2010). A comparison of multiple classification methods for diagnosis of parkinson disease. *Expert Syst. Appl*. 37, 1568–1572. 10.1016/j.eswa.2009.06.040 [CrossRef] [Google Scholar]

[7] Desai R. (2019). *Top 10 Python Libraries for Data Science*. Available online at: https://towardsdatascience.com/top-10-python-libraries-for-data-science-cd82294ec266 (accessed July 3, 2022).

[8] Fothergill-Misbah N., Maroo H., Hooker J., Kwasa J., Walker R. (2020). Parkinson's disease medication in kenya–situation analysis. *Pharmaceutica l J. Kenya.* 24, 38–41. [Google Scholar]

[9] Gil D., Manuel D. J. (2009). Diagnosing parkinson by using artificial neural networks and support vector machines. *Glob. J. Comput. Sci. Technol*. 9, 63–71. Available online at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.182.7856&rep=rep1&type=pdf

[10] Harel B., Cannizzaro M., Snyder P. J. (2004). Variability in fundamental frequency during speech in prodromal and incipient parkinson's disease: a longitudinal case study. *Brain Cognit.* 56, 24–29. 10.1016/j.bandc.2004.05.002 [PubMed] [CrossRef] [Google Scholar]

[11] Hossain E., Hossain M. F., Rahaman M. A. (2019). "A color and texture based approach for the detection and classification of plant leaf disease using knn classifier," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (Cox's Bazar: IEEE;), 1–6. 10.1109/ECACE.2019.8679247 [CrossRef] [Google Scholar]

[12] Jayaswal V. (2020). *Performance Metrics: Confusion Matrix, Precision, Recall, and f1 Score*. Available online at: https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score (accessed December 6, 2021).

[13] Kadam V. J., Jadhav S. M. (2019). "Feature ensemble learning based on sparse autoencoders for diagnosis of parkinson's disease," in *Computing, Communication and Signal Processing. Advances in Intelligent Systems and Computing, Vol. 810*, eds B. Iyer, S. Nalbalwar, N. Pathak (Singapore: Springer;), 567–581. 10.1007/978-981-13-1513-8_58 [CrossRef] [Google Scholar]

[14] Little M. (2008). *UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science*. Available online at: https://archive.ics.uci.edu/ml/datasets/parkinsons (accessed March 17, 2023).

varun_23

ORIGINALITY REPORT

28%	19%	20%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to KIET Group of Institutions, Ghaziabad Student Paper	2%
2	"Practical Statistical Learning and Data Science Methods", Springer Science and Business Media LLC, 2025 Publication	1%
3	www.coursehero.com Internet Source	1%
4	www.nature.com Internet Source	1%
5	www.mdpi.com Internet Source	1%
6	Submitted to HTM (Haridus- ja Teadusministeerium) Student Paper	1%
7	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	<1%
8	Submitted to Liverpool John Moores University Student Paper	<1%
9	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelligent Computing and Communication Techniques - Volume 1", CRC Press, 2025 Publication	<1%

e

10	www.frontiersin.org Internet Source	<1 %
11	assets.researchsquare.com Internet Source	<1 %
12	researchspace.ukzn.ac.za Internet Source	<1 %
13	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1 %
14	Bui Thanh Hung, M. Sekar, Ayhan Esi, R. Senthil Kumar. "Applications of Mathematics in Science and Technology - International Conference on Mathematical Appplications in Science and Technology", CRC Press, 2025 Publication	<1 %
15	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025 Publication	<1 %
16	Submitted to University of Hertfordshire Student Paper	<1 %
17	"Advanced Network Technologies and Intelligent Computing", Springer Science and Business Media LLC, 2024 Publication	<1 %
18	Submitted to ABV-Indian Institute of Information Technology and Management Gwalior Student Paper	<1 %

19	Submitted to University of Bradford Student Paper	<1 %
20	"Proceedings of the Fifth International Conference on Trends in Computational and Cognitive Engineering", Springer Science and Business Media LLC, 2024 Publication	<1 %
21	Submitted to University of Sunderland Student Paper	<1 %
22	dokumen.pub Internet Source	<1 %
23	pergamos.lib.uoa.gr Internet Source	<1 %
24	Alnaber, Loggen Samir Kamal. "Comparative Analysis of Prediction Models for Diabetic Patient Readmission Using Explainable AI for Feature Selection and Two-Stage Optimization Techniques", State University of New York at Binghamton, 2025 Publication	<1 %
25	Submitted to The University of Memphis Student Paper	<1 %
26	Submitted to Al Akhawayn University in Ifrane Student Paper	<1 %
27	Submitted to Gujarat Technological University Student Paper	<1 %
28	Submitted to University of North Texas Student Paper	<1 %
29	cdn.ymaws.com Internet Source	<1 %
30	Meenu Gupta, D. Jude Hemanth. "Combating Women's Health Issues with Machine	<1 %

	Learning - Challenges and Solutions", CRC Press, 2023 Publication	
31	Submitted to University of Sydney Student Paper	<1 %
32	Gaurav Aggarwal, Ashutosh Tripathi, Himani Goyal Sharma, Tripti Sharma, Rishabh Dev Shukla. "Integrated Technologies in Electrical, Electronics and Biotechnology Engineering", CRC Press, 2025 Publication	<1 %
33	www.irjmets.com Internet Source	<1 %
34	Submitted to University of Canada in Egypt Student Paper	<1 %
35	link.springer.com Internet Source	<1 %
36	Submitted to University of Edinburgh Student Paper	<1 %
37	Submitted to University of Leeds Student Paper	<1 %
38	Watfa, Mahmoud. "2v-SVM-Based Semi-Supervised Learning with Application to Carotid Plaque Characterization.", McGill University (Canada), 2021 Publication	<1 %
39	thesai.org Internet Source	<1 %
40	www.imrpress.com Internet Source	<1 %
41	www.researchsquare.com Internet Source	<1 %

42	Ibrahim Alreshidi, Desmond Bisandu, Irene Moulitsas. "Illuminating the Neural Landscape of Pilot Mental States: A Convolutional Neural Network Approach with Shapley Additive Explanations Interpretability", Sensors, 2023 Publication	<1 %
43	bpasjournals.com Internet Source	<1 %
44	pmc.ncbi.nlm.nih.gov Internet Source	<1 %
45	vixra.org Internet Source	<1 %
46	Poonam Nandal, Mamta Dahiya, Meeta Singh, Arvind Dagur, Brijesh Kumar. "Progressive Computational Intelligence, Information Technology and Networking", CRC Press, 2025 Publication	<1 %
47	Submitted to University of Hong Kong Student Paper	<1 %
48	www.ijdm.latticescipub.com Internet Source	<1 %
49	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelligent Computing and Communication Techniques - Volume 2", CRC Press, 2025 Publication	<1 %
50	Submitted to Central Queensland University Student Paper	<1 %
51	H.L. Gururaj, Francesco Flammini, J. Shreyas. "Data Science & Exploration in Artificial Intelligence", CRC Press, 2025 Publication	<1 %
52	Submitted to Nanyang Technological University	<1 %

	Student Paper	
53	Suman Kumar Swarnkar, Abhishek Guru, Gurpreet Singh Chhabra, Harshitha Raghavan Devarajan. "Artificial Intelligence Revolutionizing Cancer Care - Precision Diagnosis and Patient-Centric Healthcare", CRC Press, 2025 Publication	<1%
54	Submitted to University of East Anglia Student Paper	<1%
55	e-journal.usd.ac.id Internet Source	<1%
56	eprints.utm.my Internet Source	<1%
57	ijsret.com Internet Source	<1%
58	www.tnsroindia.org.in Internet Source	<1%
59	Alshehri, Ali Mohammed. "The Difference Between Interventional Radiology Technologists and Radiologists in PICC Line Insertion Procedure in the Kingdom of Saudi Arabia (KSA) Single-Centre: Comparison Study.", Alfaisal University (Saudi Arabia), 2024 Publication	<1%
60	Submitted to Charles University in Prague Student Paper	<1%
61	Nancy Noella R S, Priyadarshini J. "Machine learning algorithms for the diagnosis of Alzheimer and Parkinson disease", Journal of Medical Engineering & Technology, 2022 Publication	<1%
62	Submitted to PES University Student Paper	

74	Submitted to AlHussein Technical University Student Paper	<1%
75	arxiv.org Internet Source	<1%
76	docserv.uni-duesseldorf.de Internet Source	<1%
77	rgu-repository.worktribe.com Internet Source	<1%
78	www.americaspg.com Internet Source	<1%
79	Submitted to An-Najah National University Student Paper	<1%
80	Submitted to Glasgow Caledonian University Student Paper	<1%
81	Submitted to SASTRA University Student Paper	<1%
82	Submitted to Southampton Solent University Student Paper	<1%
83	Submitted to United International College Student Paper	<1%
84	core.ac.uk Internet Source	<1%
85	datahorizzonresearch.com Internet Source	<1%
86	f1000research.com Internet Source	<1%
87	ijirt.org Internet Source	<1%
88	trepo.tuni.fi Internet Source	<1%
89	www.arxiv-vanity.com Internet Source	

		<1%
90	www.preprints.org Internet Source	<1%
91	Submitted to King's College Student Paper	<1%
92	Tao Zhang, Jing Tian, Zaifa Xue, Xiaonan Guo. "Parkinson's disease detection from voice signals using adaptive frequency attribute topology", Biomedical Signal Processing and Control, 2025 Publication	<1%
93	U. Sumalatha, K. Krishna Prakasha, Srikanth Prabhu, Vinod C. Nayak. "Chapter 13 Analysis ofClassification Algorithms forPredicting Parkinson's Disease andApplications intheField ofCybersecurity", Springer Science and Business Media LLC, 2023 Publication	<1%
94	Submitted to University of Wolverhampton Student Paper	<1%
95	discovery.researcher.life Internet Source	<1%
96	ijetms.in Internet Source	<1%
97	moam.info Internet Source	<1%
98	old.duet.ac.bd Internet Source	<1%
99	ova.galencentre.org Internet Source	<1%
100	trendspider.com Internet Source	<1%

101	Submitted to universititeknologimara Student Paper	<1 %
102	www.computersciencejournals.com Internet Source	<1 %
103	www.healthcouncilcanada.ca Internet Source	<1 %
104	Ashok Kumar Swami, Deepak Verma, Richa Soni, Dweipayan Goswami. "Artificial intelligence technology in materials selection, device engineering and parameter optimisation for triboelectric nanogenerator", Materials Today Communications, 2025 Publication	<1 %
105	Khatri, Gaurav. "Ambient Temperature Modelling With ECOSTRESS and Private Weather Stations", The University of Alabama in Huntsville, 2024 Publication	<1 %
106	ijece.iaescore.com Internet Source	<1 %
107	ijsdcs.com Internet Source	<1 %
108	openbiomedicalengineeringjournal.com Internet Source	<1 %
109	researchinventy.com Internet Source	<1 %
110	tigerprints.clemson.edu Internet Source	<1 %
111	uvidok.rcub.bg.ac.rs Internet Source	<1 %
112	"Computational Intelligence Methods for Bioinformatics and Biostatistics", Springer Science and Business Media LLC, 2025 Publication	<1 %

113	Connie Tee, Thian Song Ong, Md Shohel Sayeed. "The Smart Life Revolution - Embracing AI and IoT in Society", CRC Press, 2025 Publication	<1%
114	Hassan Kazemian, Subeksha Shrestha. "Comparisons of machine learning techniques for detecting fraudulent criminal identities", Expert Systems with Applications, 2023 Publication	<1%
115	Hossein Azarpir, Parsa Khakzad, Mohammad Reza Alipour, Amir Sheikh Mohammadi. "The pivotal and transformative role of artificial intelligence in advanced multidimensional modeling and optimization of complex cefixime separation processes using 3-hydroxyphenol-formaldehyde nanostructures: A multi-layered analytical approach", Microchemical Journal, 2025 Publication	<1%
116	Insha Zahoor, Sajad Ahmad Wani, Tariq Ahmad Ganaie. "Artificial Intelligence in the Food Industry - Enhancing Quality and Safety", CRC Press, 2025 Publication	<1%
117	Submitted to Malaysia University of Science and Technology Student Paper	<1%
118	Submitted to New College of the Humanities Student Paper	<1%
119	Submitted to University of Ulster Student Paper	<1%
120	arccjournals.com Internet Source	<1%
	Submitted to parsurnd	

121	Student Paper	<1%
122	www.geeksforgeeks.org Internet Source	<1%
123	www.medrxiv.org Internet Source	<1%
124	Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Lorraine O. Ramig. "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity", Journal of The Royal Society Interface, 2010 Publication	<1%
125	Jefferson S. Almeida, Pedro P. Rebouças Filho, Tiago Carneiro, Wei Wei et al. "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques", Pattern Recognition Letters, 2019 Publication	<1%
126	Mamun, Khondaker Abdullah Al, Musaed Alhussein, Kashfia Sailunaz, and Mohammad Saiful Islam. "Cloud based framework for Parkinson's disease diagnosis and monitoring system for remote healthcare applications", Future Generation Computer Systems, 2015. Publication	<1%
127	Nishu Gupta, Sandeep S. Joshi, Milind Khanapurkar, Asha Gedam, Nikhil Bhave. "Recent Advances in Science, Engineering and Technology (RASET-2023) - Proceedings of the International Conference on Recent Advances in Science, Engineering & Technology, 29–30 September 2023", CRC Press, 2024	<1%

	Publication	
128	hdl.handle.net Internet Source	<1%
129	iis-international.org Internet Source	<1%
130	ojs.unud.ac.id Internet Source	<1%
131	pubs.aip.org Internet Source	<1%
132	repository.psa.edu.my Internet Source	<1%
133	researchonline.gcu.ac.uk Internet Source	<1%
134	sciencepublishinggroup.com Internet Source	<1%
135	sourcecodequery.com Internet Source	<1%
136	www.codinginterviewpro.com Internet Source	<1%
137	www.ijraset.com Internet Source	<1%
138	www.mygreatlearning.com Internet Source	<1%
139	www.researchgate.net Internet Source	<1%
140	Amir Mohammad Sharafaddini, Kiana Kouhpah Esfahani, Najme Mansouri. "Deep learning approaches to detect breast cancer: a comprehensive review", Multimedia Tools and Applications, 2024 Publication	<1%

141	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial Intelligence, Blockchain, Computing and Security", CRC Press, 2023 Publication	<1 %
142	Mandal, Indrajit, and N. Sairam. "Accurate telemonitoring of Parkinson's disease diagnosis using robust inference system", International Journal of Medical Informatics, 2012. Publication	<1 %
143	Gunjan Pahuja, T. N. Nagabhushan. "A Comparative Study of Existing Machine Learning Approaches for Parkinson's Disease Detection", IETE Journal of Research, 2018 Publication	<1 %
144	Lecture Notes in Computer Science, 2015. Publication	<1 %
145	Sicheng Wang. "Research on a New AI Diagnostic Model with Strong Universality Based on Multilayer Perceptron Neural Networks", 2023 IEEE 6th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), 2023 Publication	<1 %
146	Sujata Dash, Subhendu Kumar Pani, Joel J. P. C. Rodrigues, Babita Majhi. "Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics - Techniques and Applications", CRC Press, 2022 Publication	<1 %
147	dx.doi.org Internet Source	<1 %