NEURONEST: SMART DETECTION OF PARKINSON’S DISEASE USING AI

Project Report

Submitted by

SANSKAR SRIVASTAVA
210906024

SHAURYA PANDEY
210906202

Under the guidance of

Dr. BHARATHI R. B.
Associate Professor
Department of E&E,
MIT Manipal

in partial fulfilment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING
MANIPAL INSTITUTE OF TECHNOLOGY
(A Constituent Unit of Manipal Academy of Higher Education)
MANIPAL-576104, KARNATAKA, INDIA
May, 2025

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING
MANIPAL INSTITUTE OF TECHNOLOGY
(A Constituent Unit of Manipal Academy of Higher Education)
MANIPAL-576104, KARNATAKA, INDIA

Manipal
22.05.25

CERTIFICATE

This is to certify that the project titled **NEURONEST: SMART DETECTION OF PARKINSON'S DISEASE USING AI** is a record of the bonafide work done by Sanskar Srivastava (*Reg. No. 210906024*) and Shaurya Pandey (*Reg. No. 210906202)* submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B.Tech.) in **ELECTRICAL AND ELECTRONICS ENGINEERING** of Manipal Institute of Technology, Manipal, Karnataka, (A Constituent unit of Manipal Academy of Higher Education), during the academic year 2024-2025.

Dr. Bharathi R.B.
Associate Professor,
Dept of E&E,
M.I.T., MANIPAL

Prof. Dr. Jayalakshmi N.S.
HOD, Dept of E&E,
M.I.T., MANIPAL

ACKNOWLEDGMENTS

I take this opportunity to sincerely thank all those who have extended their support and guidance throughout the course of my final year project. I am especially grateful to Dr. Bharathi R. B., Associate Professor, Department of Electrical and Electronics Engineering, for her consistent mentorship, valuable insights, and unwavering encouragement that greatly contributed to the successful completion of this project. My heartfelt thanks also go to Dr. Jayalakshmi N. S., Head of the Department, Electrical and Electronics Engineering, for her support and for providing the academic resources and environment necessary for the progress of this work.

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disorder that causes motor disability, cognitive impairment, and emotional disturbances by causing degeneration of dopaminergic neuronal cells. Conventional diagnostic methods cannot diagnose the disease in the early stages where the symptoms of the disease are not highly evident. This research bridges this gap by using machine learning models to improve early diagnosis of Parkinson's based on voice parameters like fundamental frequency, jitter, shimmer, and noise-to-harmonic ratios. Data pre-processing was carried out with extreme caution by handling missing values, outliers, and feature scaling to enhance the accuracy and reliability of the model. Four classification models—Decision Tree, Random Forest, Logistic Regression, and Artificial Neural Network (ANN) were applied and compared. The Decision Tree model had 96.7% accuracy, 84.6% precision, and 81.5% recall for Parkinson's cases, whereas the best performance was obtained from the Random Forest model with 98.9% accuracy, 91% precision, and 95% recall. The Logistic Regression model with its simplicity had 89.94% accuracy, 78% precision, and a mere 32% recall, reflecting its shortcomings in identifying PD cases. The Artificial Neural Network performed well with classification as 96% validation accuracy and an AUC of 1.00, reflecting outstanding ability in discriminating healthy from diseased samples.

Among all the models that were evaluated, the Random Forest model was selected as the final model due to its high accuracy, balanced performance across classes, and robustness in handling imbalanced data. Through multi-modeling, the study suggests a comparative paradigm that increases diagnostic certainty and decreases single-model dependency. The app that was developed allows clinicians to leverage multiple AI models in parallel, making it easier for the research-to-practice transition. The AI-driven voice analysis offers a cost-efficient and non-invasive early detection of Parkinson's disease, with the potential to significantly improve patient outcomes through early medical interventions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
FP	False Positive
FN	False Negative
IQR	Interquartile Range
LR	Logistic Regression
ML	Machine Learning
PD	Parkinson’s Disease
RF	Random Forest
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

CHAPTER 1
INTRODUCTION

Parkinson's Disease (PD) is a chronic neurodegenerative disorder that impacts millions of individuals, primarily the elderly, which affects movement, mood, and social behavior. Early diagnosis and proper treatment are important in improving quality of life, managing symptoms, and reducing the progress of the disease. Researchers are increasingly depending on Artificial Intelligence (AI) to improve early detection, monitor disease progression, and improve treatment.

Parkinson's disease (PD) arises as a result of neuronal damage involved in the production of dopamine, a neurotransmitter that plays a critical role in motor function regulation. The development of motor symptoms typically occurs with the loss of 60-80% of such neurons, resulting in impaired synthesis of dopamine. Parkinson's disease is chronic and progressive, with its onset characterized by mild symptoms such as tremors, bradykinesia, muscle stiffness, and impaired balance. For measuring the severity of such symptoms and monitoring their progression without losing sight of their effect on daily activities, the Hoehn and Yahr scale is used.

PD affects patients profoundly, extending beyond the physical difficulties to emotional and psychological challenges. Depression, anxiety, and insomnia are common, as are cognitive disturbances. These disturbances tend to foster loneliness and isolation and drastically reduce the quality of life of the patient.

Diagnosis of PD, particularly in its early stage, is difficult as there is no test to establish it. Physicians are then left to depend on history, observation, and physical examination. Early symptoms are subtle and may be confused with other illnesses. Symptom variability from person to person and subjectivity of observations render diagnosis even more difficult. Imaging studies such as MRI and Dopamine Transporter scan, although they can raise suspicions, are not diagnostic. Machine learning has been reported with promising results in diagnosis of PD in recent studies, with different techniques ranging from 91-95% accuracy. Several gaps still exist, though, in the literature. The majority of studies aim to identify one high-performing model and not examine the merit of several comparative models. In addition, there is little translation of theoretical algorithms into clinical applicability that is straightforwardly implementable by clinicians. Most studies lack interpretability measures accessible to clinicians to comprehend and adopt model predictions.

CHAPTER 2
LITERATURE REVIEW AND OBJECTIVES

Senturk (2020) introduced a [1] classification and feature selection method diagnostic technique for the diagnosis of Parkinson's disease from voice recordings. Feature Importance and Recursive Feature Elimination were utilized by the study in feature selection and classifiers like Classification and Regression Trees, [1] Artificial Neural Networks, and Support Vector Machines (SVM) in classifying the patient. Worth noting was that SVM using Recursive Feature Elimination had an accuracy rate of 93.84%, a sure sign of the promise of using voice features in the early detection of Parkinson's disease.

Zhu (2022) proposed a hybrid deep learning [2] model that utilized patient symptoms and MRI [2] data to determine Parkinson's disease severity. Utilizing the Parkinson's Progression Markers Initiative database, the research employed symptom-based models, MRI, and the two in combination. The [2] hybrid model was more precise at a 94% rate when it categorized patients into five degrees of severity, with multimodal data significance in improving diagnostic accuracy.

Allahbakhshi, Sadri, and Shahdi (2024) explained the use of EEG signals to diagnose Parkinson's disease. They proposed a Support Vector Machine [3] (SVM) model with high-level feature design and hyperparameter optimization. The proposed model was highly accurate in diagnosis, having the potential for EEG-based biomarkers to diagnose early Parkinson's disease.

Govindu and Palwe (2023) investigated telemedicine technology for presymptomatic Parkinson's disease diagnosis with [4] machine learning algorithms. The authors compared Random [4] Forest, SVM, K-Nearest Neighbors, and Logistic Regression classifiers based on voice samples from 30 Parkinson's disease patients and controls. Random Forest was the most effective with a detection rate of 91.83% and sensitivity of 0.95, suggesting the promise of voice-based, remote diagnostic assistance.

Srivastava et al. (2024) [5] compared different machine learning classifiers in the diagnosis of early Parkinson's disease based on motor and non-motor symptoms. Models [5] to be compared were K-Nearest Neighbors, Random Forest, Gradient Boosting, Support Vector Machine, Boosting, and Bagging. Results give evidence for machine learning techniques being used to classify Parkinson's disease individuals correctly and serve as an alternative to traditional clinical evaluation.

Magesh, Myloth, and Tom in 2020 introduced an interpretable machine learning model from DaTSCAN images for the early diagnosis of Parkinson's disease. [6] The model, using transfer learning over a Convolutional Neural Network (VGG16), was claimed to report 95.2% accuracy, 97.5% sensitivity, and 90.9% specificity. Visual model prediction explanations were enabled via Local Interpretable Model-Agnostic Explainer (LIME), thereby making it more interpretable and improving clinician confidence in automated diagnosis.

Mei, Desrosiers, [7] and Frasnelli (2021) offer a systematic review of machine learning for diagnosing Parkinson's disease. [7] Among 209 articles, they also reported high potential contribution of machine learning algorithms and new biomarkers to clinical decision support. The review reported the major contribution of multisource data modalities such as voice recording, handwriting features, and imaging to advanced diagnostic accuracy and timely intervention.
Tusar, Islam, [8] and Sakil (2023) carried out an experimental study with the objective of automating the detection of early-stage Parkinson's disease using machine learning methods. [8] They employed a publicly available clinical features data, vocal features data, and motor assessment data of 130 subjects using techniques like MinMax Scaler, Local Outlier Factor, and the Synthetic Minority Over-sampling Technique (SMOTE) for the preprocessing operation. Their approach was 100% accurate in the detection of Parkinson's disease patients and Rapid Eye Movement Sleep Behavior Disorder patients and 92% accurate in the separation of Parkinson's disease patients from normal subjects, thereby validating the use of machine learning for the early detection of Parkinson's sickness.

Salunkhe et al. (2024) [9] examined voice signal analysis with spiral drawing tests using machine learning methods for early detection of Parkinson's disease [9]. With the use of Convolutional Neural Networks (CNN) and Support Vector Machines (SVM), their study bridged single analysis gaps by exploring inter-relations among symptoms. Following an integrative approach, promising diagnostic accuracy was established, and this indicates that integration of diverse data sources might facilitate early diagnosis of Parkinson's disease. Prashanth and [10] Dutta Roy (2018) applied predictive modeling and patient questionnaires to identify early Parkinson's disease. [10] With the application of machine learning methods such as logistic regression, random forests, boosted trees, and support vector machines, they compared the Movement Disorder Society-Unified Parkinson's Disease Rating Scale responses. The models were highly accurate and area under the ROC curve (both >95%) in differentiating early Parkinson's disease and healthy controls, and this indicates the potential of questionnaire-based predictive models in supporting clinicians in diagnosis.

The present research fills these gaps by employing several models (Decision Tree, Random Forest, Logistic Regression, and Artificial Neural Network) in one application framework where their performances are directly comparable. Not only does this offer better confidence estimates using model consistency, but it also filled the research-clinical practice gap by rendering it an easily usable tool for clinicians. With the provision of interpretable outputs and comparative analysis, our research seeks to improve clinical decision-making as well as early diagnosis of Parkinson's disease based on voice analysis.

The literatures highlighted some of the most critical Parkinson's disease prediction research limitations such as the use of single-model solutions, absence of real-world diagnostic tools, and limited user accessibility [1,4,7,8,9,10]. The suggested python –based system addresses these research limitations by combining four machine learning algorithms—Decision Tree, Random Forest, Logistic Regression, and Artificial Neural Network, with dynamic model selection and comparison. It has CSV/Excel inputs and provides a user-friendly interface with strong data validation, thereby improving real-world usability. It fills the gap between theoretical research and clinical application, improving accessibility and reliability.

OBJECTIVES

- **Objective 1:** Data Collection, Preprocessing and visualization of voice data: The goal is to gather voice-based biomedical data relevant to Parkinson’s disease, ensuring the dataset is clean, consistent, and structured. This includes eliminating missing values, outliers, and irrelevant features to create a high-quality dataset ready for analysis.
- **Objective 2:** Develop machine learning models to diagnose Parkinson’s disease based on voice data: The aim is to develop and train machine learning models like Decision Tree, Random Forest, Logistic Regression, and ANN to classify subjects as Parkinson’s-positive or healthy. This step emphasizes accurate prediction using voice biomarkers.
- **Objective 3:** Evaluation and comparison of the developed Models: The objective is to evaluate the developed models using metrics such as accuracy, precision, recall, F1-score, and AUC. It also involves interpreting the results to understand which features influence predictions and assess the clinical relevance of the outcomes.

CHAPTER 3
METHODOLOGY

The method employed here took a systematic and comprehensive approach in achieving the reliability and performance of the predictive models in a way that they would be applicable to the early diagnosis of Parkinson's disease from voice biomarkers. The routine involved several major steps: data preprocessing, feature visualization and inspection, model training, probability prediction, and performance evaluation.

The first step involved the pre-processing of the dataset for machine learning operations. The dataset, consisting of voice recordings defined by different acoustic features, was imported first and checked for the integrity of data. There were no missing values, so imputation was not required. Non-numeric features like patient IDs were dropped to ensure machine learning compatibility. The dataset was also checked for duplicate and zero-variance features columns that are not discriminative were dropped. Outlier detection was done where either clipped or dropped to provide a cleaner and more stable training environment for the models. Feature scaling was then carried out using the Standard Scaler, normalizing the data to have zero mean and unit variance to maintain uniformity in feature impact during model learning.

After preprocessing, the preprocessed dataset was split into training, validation, and test subsets using stratified sampling to preserve the original class distribution, with the highly imbalanced dataset containing fewer occurrences of Parkinson's disease. This allowed for better model performance evaluation, especially for minority class detection.
Four algorithms of machine learning were employed: Logistic Regression, Decision Tree, Random Forest, and Artificial Neural Network (ANN). They were all trained over the training subset and hyperparameter tuning was performed with grid search over a given set of hyperparameters. Once trained, the models were not only used for classification, but for the estimation of the probability that a sample patient is in the Parkinson's class. Estimates of probability are particularly valuable in a clinical environment, where decision-making can be enhanced by the possession of an estimate of the confidence with which a diagnosis is made.

Model performance was evaluated relative to a list of binary classification task-specific evaluation measures, i.e., in imbalanced settings. These were accuracy, precision, recall, F1-score, and Area Under the Curve (AUC), where appropriate. In addition to that, confusion matrices were employed to provide a visual description of the classification outcomes and improved understanding of the strengths and weaknesses of each model in discriminating between patients with and without Parkinson's disease. Probabilistic outputs and ROC curves were also analyzed to determine the discriminative power and reliability of each model.

This strategy achieved a smooth shift from raw voice data to clinically useful predictive modeling. Through the aggregation of many models and generation of probability-based results, the platform not only enhances predictive performance but also more closely meets the actual practical needs of healthcare professionals for confidence-based, understandable decision aid in early detection of Parkinson's disease.

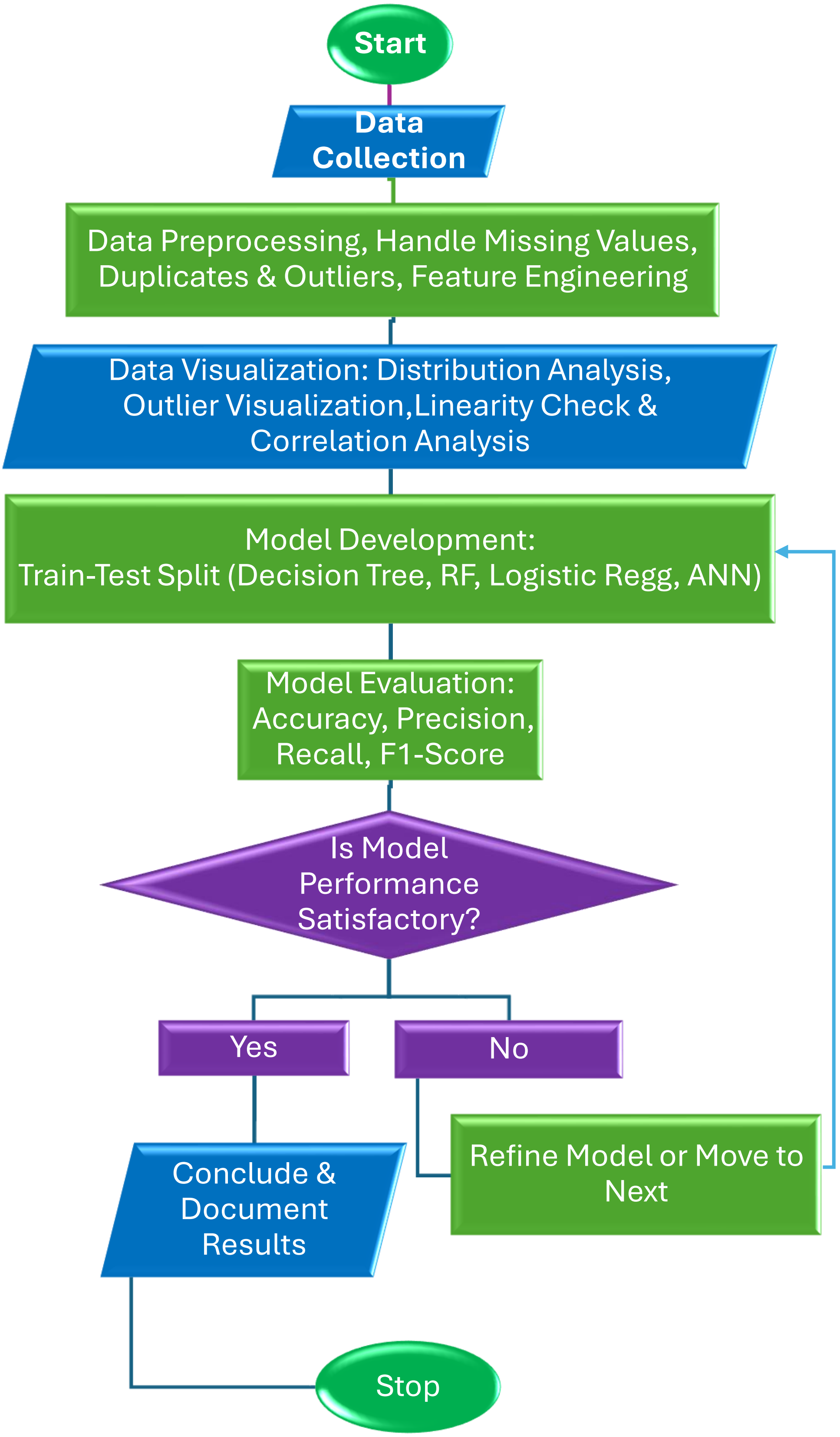

Start
Data Collection
Data Preprocessing, Handle Missing Values, Duplicates & Outliers, Feature Engineering
Data Visualization: Distribution Analysis, Outlier Visualization,Linearity Check & Correlation Analysis
Model Development: Train-Test Split (Decision Tree, RF, Logistic Regg, ANN)
Model Evaluation: Accuracy, Precision, Recall, F1-Score
Is Model Performance Satisfactory?
Yes
No
Conclude & Document Results
Refine Model or Move to Next
Stop

Fig 1: Flowchart

3.1 Dataset:

Table 1: Dataset Sample

name	phon_R01_S01_1	phon_R01_S01_2
MDVP:Fo(Hz)	119.992	122.4
MDVP:Fhi(Hz)	157.302	148.65
MDVP:Flo(Hz)	74.997	113.819
MDVP:Jitter(%)	0.00784	0.00968
MDVP:Jitter(Abs)	7.00E-05	8.00E-05
MDVP:RAP	0.0037	0.00465
MDVP:PPQ	0.00554	0.00696
Jitter:DDP	0.01109	0.01394
MDVP:Shimmer	0.04374	0.06134
MDVP:Shimmer(dB)	0.426	0.626
Shimmer:APQ3	0.02182	0.03134
Shimmer:APQ5	0.0313	0.04518
MDVP:APQ	0.02971	0.04368
Shimmer:DDA	0.06545	0.09403
NHR	0.02211	0.01929
HNR	21.033	19.085
status	1	1
RPDE	0.414783	0.458359
DFA	0.815285	0.819521
spread1	-4.81303	-4.07519
spread2	0.266482	0.33559
D2	2.301442	2.486855
PPE	0.284654	0.368674

The Table1 contains all parameters and first two rows of our dataset. It has been transposed for visual representation. The dataset used in this study comprises 1195 voice recording values, each represented by 24 columns, including 22 biomedical voice features, one subject identifier, and one target label. The target column, labelled status, indicates the presence (1), absence (0) and also majority of instances have a floating-point value representing probability of Parkinson's disease and serves as the dependent variable for classification. The voice features are derived from sustained phonation of a vowel sound and capture various aspects of vocal function. Fundamental frequency-related columns such as MDVP: Fo(Hz), MDVP: Fhi(Hz), and MDVP: Flo(Hz) measure the average, maximum, and minimum vocal pitch, which are often altered in Parkinson's patients. Columns like MDVP: Jitter(%), MDVP: RAP, and Jitter: DDP quantify frequency variations (jitter), while MDVP: Shimmer, Shimmer: APQ5, and Shimmer: DDA capture amplitude variations (shimmer), are common in dysphonic speech. NHR (Noise-to-Harmonics Ratio) and HNR (Harmonics-to-Noise Ratio) assess vocal clarity, distinguishing breathy or hoarse voices typically associated with PD. Additionally, RPDE, DFA, and PPE represent nonlinear dynamic measures that reflect complexity and unpredictability in voice signals. The spread1, spread2, and D2 features further characterize signal dispersion and chaotic behaviour in vocal fold vibrations. Altogether, the dataset provides a comprehensive set of acoustic biomarkers, enabling machine learning models to effectively differentiate between healthy and Parkinsonian speech patterns.

3.2 Preprocessing and Cleaning of the Dataset:

Before model creation, the dataset went through a thorough preprocessing and cleaning process to maintain the integrity and quality of inputs for machine learning. The dataset was first scanned for missing values using standard methods. Then, outlier detection was carried out based on the Interquartile Range (IQR) approach, where observations outside of 1.5 times the interquartile range from the first or third quartile were classified as outliers and eliminated. This process was necessary to remove extreme values that would otherwise skew model training and testing. After removing outliers, feature scaling was performed using the StandardScaler, which scaled each feature to zero mean and unit variance. This scaling was necessary, particularly with the different magnitudes of features like fundamental frequency and jitter measurements, to prevent any one feature from dominating the learning algorithms.

Next, the dataset was split into a training and a testing set through a stratified 70:30 split, maintaining the proportion of Parkinson's and control cases in both subsets. The target variable status was kept as a binary indicator and did not require further encoding. By following this careful preprocessing pipeline, a clean, balanced, and standardized dataset was obtained, enabling effective and unbiased model training and testing in future experiments.

The ratio of samples in the dataset that are of each of the two target classes, i.e., healthy people and people suffering from Parkinson's disease. From the chart, we can see that nearly 75.4% of the data samples are of people diagnosed with Parkinson's disease, and 24.6% are of healthy people. This clearly depicts that the dataset is significantly imbalanced with a much higher set of Parkinson's cases than healthy cases. This type of imbalance can potentially result in biased predictions in machine learning models, as the models become biased towards predicting the majority class (Parkinson's disease) and are unable to identify the minority class (healthy people). To prevent this problem, methods like oversampling the minority class, under sampling the majority class, or using class weights during model training can be utilized. Also, while measuring model performance over imbalanced datasets, it is important to use metrics like precision, recall, F1-score, and ROC-AUC, and not accuracy, to have a better idea of the effectiveness of the model.

3.3 AI Model Development:

In this study, four machine learning models were utilized to identify Parkinson's disease based on voice-based biomarkers: Logistic Regression, Decision Tree, Random Forest, and Artificial Neural Network (ANN). The reason for choosing these models is their varied strengths and suitability for medical classification problems.

3.3.1 Decision Tree Model:

The Decision Tree model was chosen for its clear, visual structure and its strong interpretability, which makes it highly suitable in clinical settings where understanding how a diagnosis is made is just as important as the diagnosis itself. The model works by repeatedly splitting the dataset based on feature thresholds that best separate the classes as individuals with and without Parkinson’s disease. These splits are guided by a metric called Gini impurity, which measures how mixed the classes are at a particular node. The formula for Gini impurity is:

$$Gini = 1 - \sum_{i=1}^{n} (pi)^2 \quad \text{-----------(1)}$$

where pi is the proportion of samples of class i at the node. A node with Gini = 0 is perfectly pure, meaning it contains only one class.

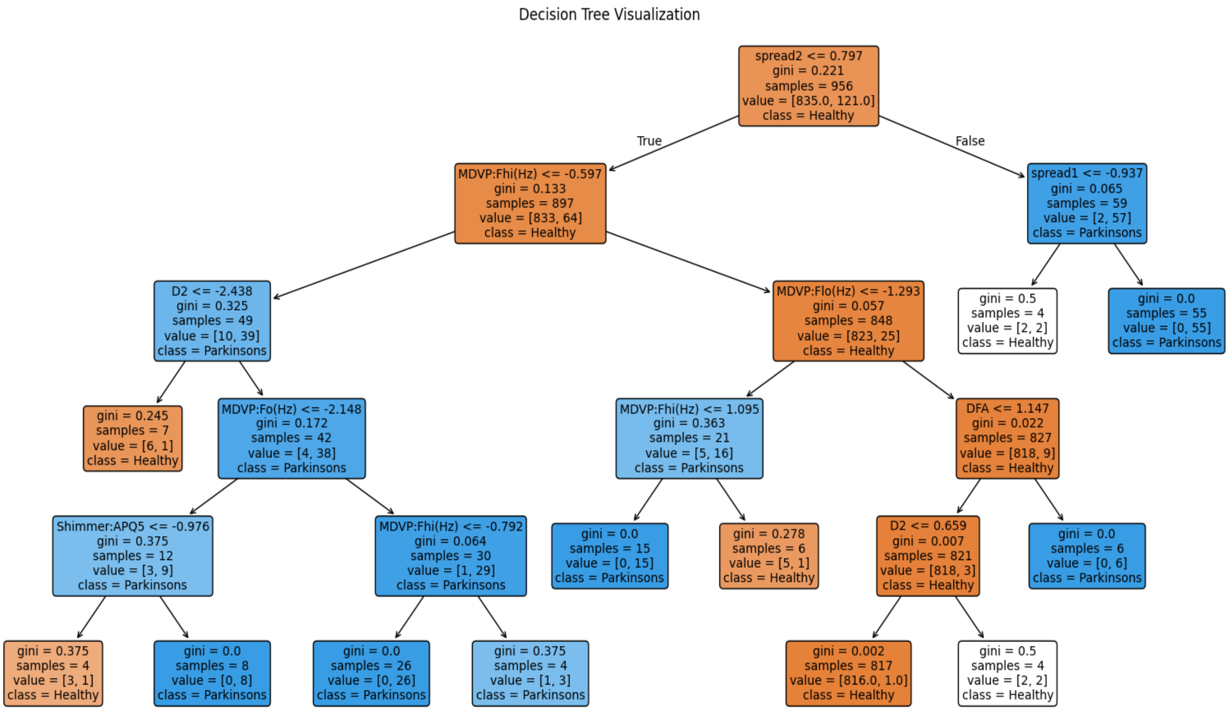
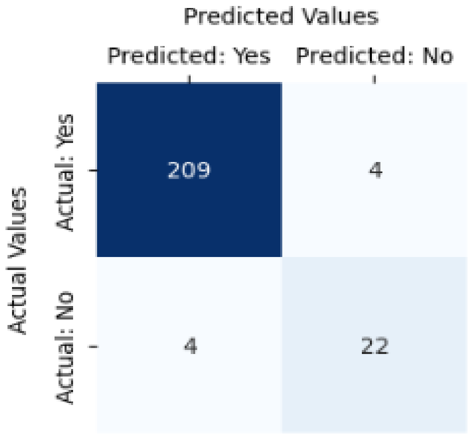
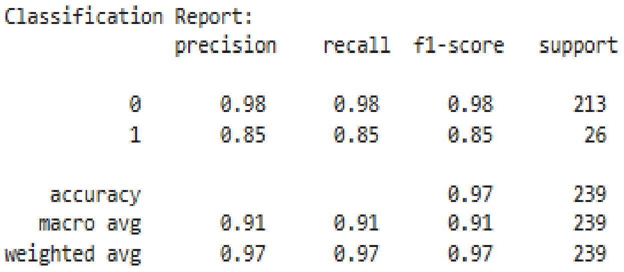

Decision Tree Visualization
spread2 <= 0.797 gini = 0.221 samples = 956 value = [835.0, 121.0] class = Healthy
True
False
MDVP:Fhi(Hz) <= -0.597 gini = 0.133 samples = 897 value = [833, 64] class = Healthy
spread1 <= -0.937 gini = 0.065 samples = 59 value = [2, 57] class = Parkinsons
D2 <= -2.438 gini = 0.325 samples = 49 value = [10, 39] class = Parkinsons
MDVP:Flo(Hz) <= -1.293 gini = 0.057 samples = 848 value = [823, 25] class = Healthy
gini = 0.5 samples = 4 value = [2, 2] class = Healthy
gini = 0.0 samples = 55 value = [0, 55] class = Parkinsons
gini = 0.245 samples = 7 value = [6, 1] class = Healthy
MDVP:Fo(Hz) <= -2.148 gini = 0.172 samples = 42 value = [4, 38] class = Parkinsons
MDVP:Fhi(Hz) <= 1.095 gini = 0.363 samples = 21 value = [5, 16] class = Parkinsons
DFA <= 1.147 gini = 0.022 samples = 827 value = [818, 9] class = Healthy
Shimmer:APQ5 <= -0.976 gini = 0.375 samples = 12 value = [3, 9] class = Parkinsons
MDVP:Fhi(Hz) <= -0.792 gini = 0.064 samples = 30 value = [1, 29] class = Parkinsons
gini = 0.0 samples = 15 value = [0, 15] class = Parkinsons
gini = 0.278 samples = 6 value = [5, 1] class = Healthy
D2 <= 0.659 gini = 0.007 samples = 821 value = [818, 3] class = Healthy
gini = 0.0 samples = 6 value = [0, 6] class = Parkinsons
gini = 0.375 samples = 4 value = [3, 1] class = Healthy
gini = 0.0 samples = 8 value = [0, 8] class = Parkinsons
gini = 0.0 samples = 26 value = [0, 26] class = Parkinsons
gini = 0.375 samples = 4 value = [1, 3] class = Parkinsons
gini = 0.002 samples = 817 value = [816.0, 1.0] class = Healthy
gini = 0.5 samples = 4 value = [2, 2] class = Healthy

Fig 2: Decision Tree

To build a good model, we did a grid search on the training data to find the best values for hyperparameters. We experimented with different values of max_depth, min_samples_split, and min_samples_leaf. The best setting that we found was as follows: max_depth = 5, min_samples_split = 10, and min_samples_leaf = 4. These settings helped the model find a balance between the extraction of meaningful patterns and the avoidance of overfitting.





```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       213
           1       0.85      0.85      0.85        26

    accuracy                           0.97       239
   macro avg       0.91      0.91      0.91       239
weighted avg       0.97      0.97      0.97       239
```

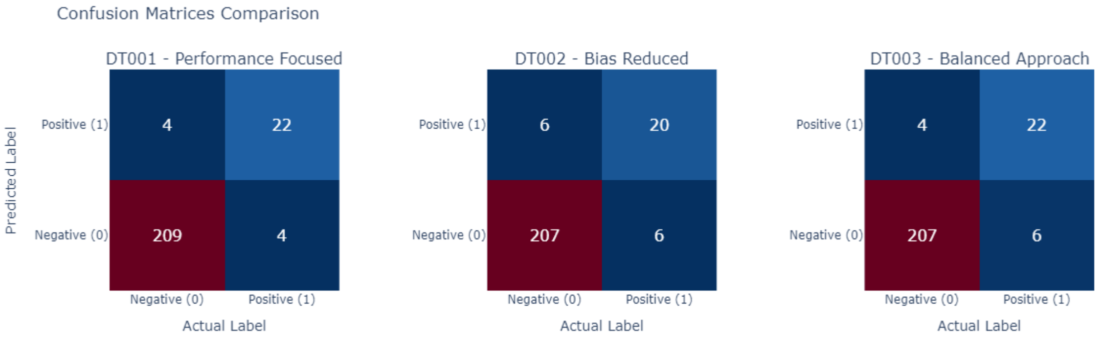


Fig 3: Comparison Matrix of decision trees

Table 2: Confusion Matrix Details

Model	Description	TP	FP	TN	FN
DT001	Performance Focused	22	4	209	4
DT002	Bias Reduced	20	6	207	6
DT003	Balanced Approach	22	4	207	6

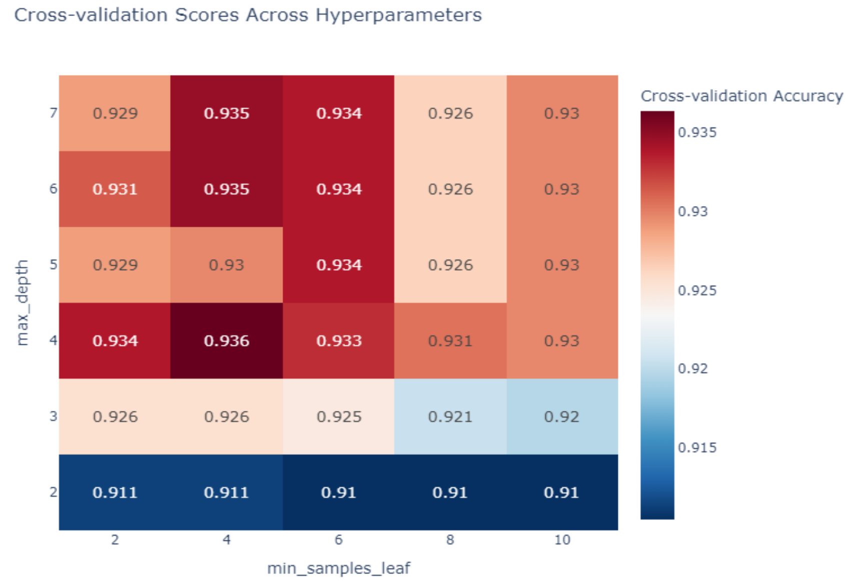


Fig 4: Cross Validation Score

To balance model complexity and generalization, hyperparameter tuning was performed using grid search with cross-validation. The results clearly show that very shallow trees (e.g., max_depth=2) underfit the data, while deeper trees beyond a certain point do not provide meaningful improvements and risk overfitting. The optimal configuration max_depth=4 and min_samples_leaf=4 offered the highest validation accuracy (93.6%), suggesting a well-regularized model that captures patterns without memorizing noise. This tuning process was critical for selecting a model that generalizes well on unseen data.

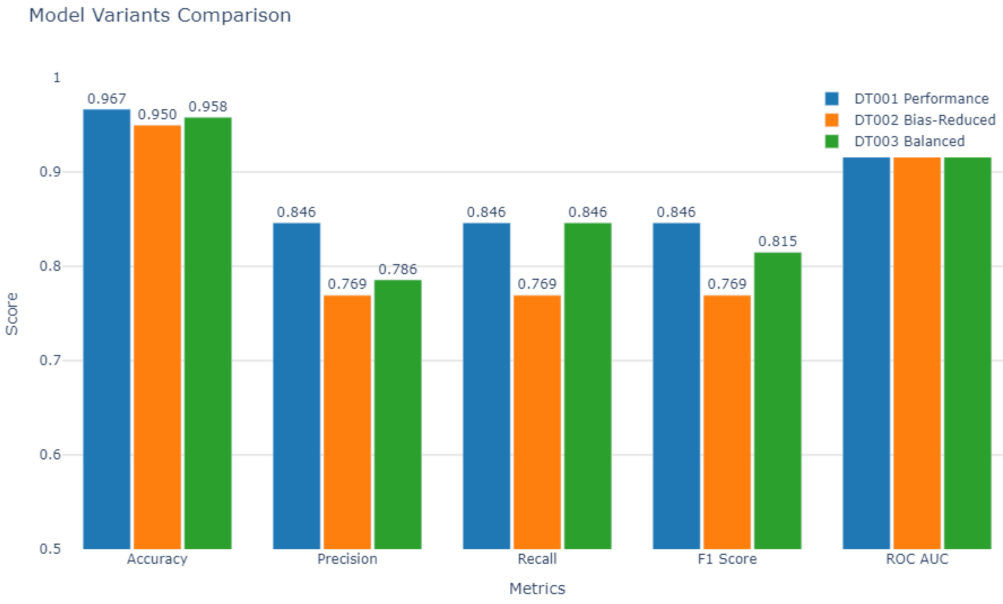


Fig 5: Comparison of Decision Tree

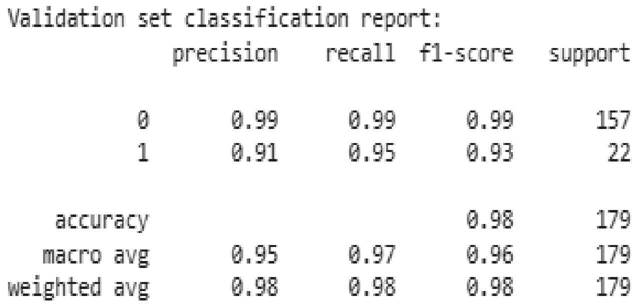
This bar chart compares the three decision tree model variants across key performance metrics. DT001 (Performance-Focused) shows the highest accuracy (96.7%) and precision (84.6%), indicating strong overall performance, but it likely favors the majority class. DT002 (Bias-Reduced) achieves better balance in recall but at the cost of precision and F1 score, suggesting it’s less confident in positive predictions. DT003 (Balanced Approach) provides a middle ground with consistently good scores across all metrics, including F1 and ROC AUC, making it the most well-rounded model for balanced performance.

- DT001 (Performance-Focused) is highly accurate but tends to favour the majority class (likely the negatives), as seen in the lower recall for the minority class. It's ideal if false negatives are less critical, but not suitable when missing true positives has high consequences.
- DT002 (Bias-Reduced) sacrifices some overall performance to reduce class imbalance bias. It has slightly better recall for the minority class but significantly lower precision and F1 score, indicating more false positives.
- DT003 (Balanced Approach) strikes the best trade-off between performance and fairness. It maintains strong accuracy and F1 score while also improving recall compared to DT001. This makes it a reliable choice when both sensitivity and specificity are important.

3.3.2 Random Forest Model:

Random Forest was utilized in this project as a stable and strong ensemble learning algorithm. It trains several decision trees on training and aggregates their predictions collectively in order to make the final output stable and reliable. The method is especially useful when dealing with imbalanced data sets, where instances of Parkinson's disease are much fewer than healthy individuals. Each tree in the forest is trained on a random subset of the data with replacement (a bootstrapping process), and the final prediction is given by majority voting over all the trees. This minimizes overfitting and makes the model stable.

The model was tuned using GridSearchCV, which tried combinations of hyperparameters like the number of trees (n_estimators), the maximum depth of trees (max_depth), and the minimum number of samples for a node to split (min_samples_split). The best parameters that were found were: n_estimators = 200, max_depth = None, and min_samples_split = 2. The parameters allowed the model to grow trees fully without limiting their depth but with the ability to allow splits to occur only if two samples existed, which was the reason why subtle patterns were identified in the data.



```
Validation set classification report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       157
           1       0.91      0.95      0.93        22

    accuracy                           0.98       179
   macro avg       0.95      0.97      0.96       179
weighted avg       0.98      0.98      0.98       179
```

Fig 6: Validation set report

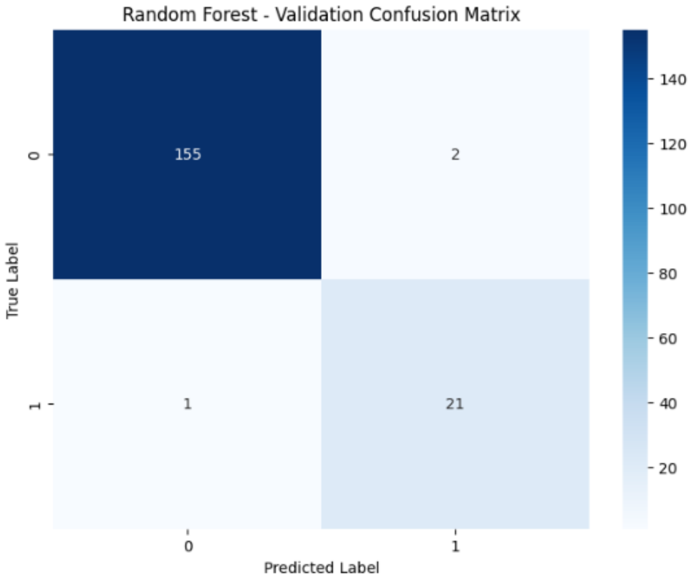


Fig 7: Confusion Matrix for Random Forest

On the training set, Random Forest achieved an accuracy of 98%, meaning it correctly classified most of the samples. The classification report shows a precision of 0.91, recall of 0.95, and F1-score of 0.93 for the Parkinson’s class, which indicates that the model is not only good at detecting the disease but also at minimizing false alarms. For healthy individuals, the precision and recall were both 0.99, showing the model was highly confident and accurate when classifying non-Parkinson’s cases.

The confusion matrix shows that out of 157 healthy samples, the model correctly predicted 155, and out of 22 Parkinson’s samples, it correctly predicted 21. This means only one Parkinson’s case was missed. The ROC curve for the validation set had an AUC score of 1.00, which means the model had perfect discrimination ability between the two classes on the validation data.

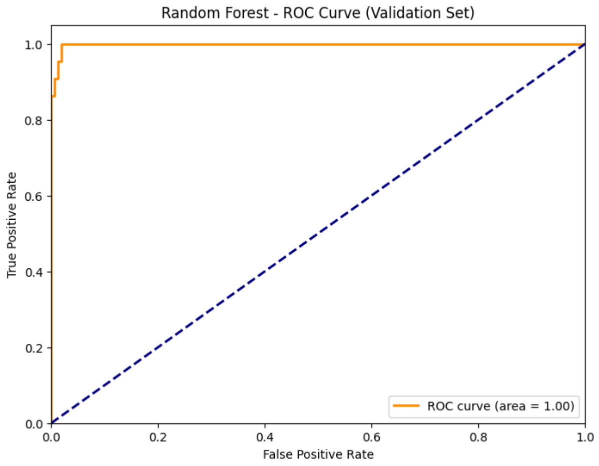


Fig 8: ROC Curve

3.3.3 Logistic Regression Model

Logistic Regression is a simple and easy-to-understand machine learning model, especially good for binary classification problems like detecting whether a person has Parkinson's disease or not. It works by assigning weights to each input feature (like jitter, shimmer, and pitch-related voice features) and then using a mathematical formula called the sigmoid function to calculate the probability of the person having the disease. If this probability is greater than 0.5, the model predicts the person as having Parkinson's; otherwise, it predicts them as healthy.

To improve the model's accuracy, we used a method called **grid search**, which tests different combinations of model settings (called hyperparameters). We tried different values for C (which controls regularization), two types of penalties (l1 and l2), and different numbers of iterations for the training. The best combination found was C = 10, penalty = l1, and max_iter = 100. This gave us the highest accuracy of around **93.1%**, as shown in the heatmap of accuracy scores from grid search.

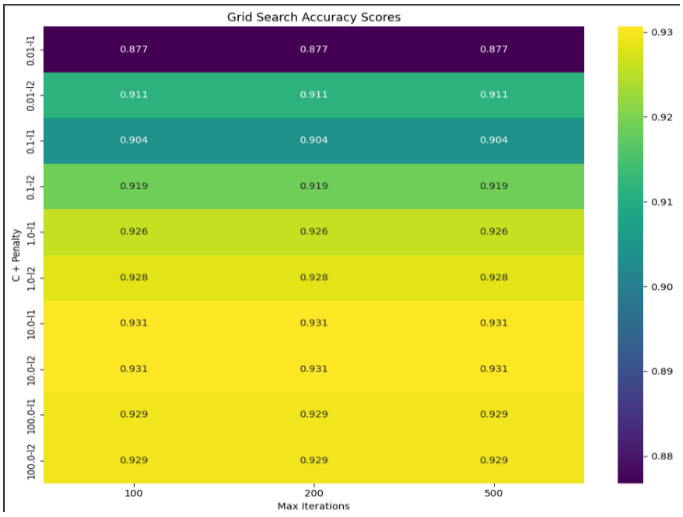


Fig 9: Grid Search Accuracy Scores for Logistic Regression

After training, we tested the model on a validation dataset of 179 samples. The model gave an overall accuracy of 89.94%, which means it correctly predicted about 90 out of every 100 cases. However, when we look deeper into how well it did on each class, we see that it did well for healthy individuals but not as well for Parkinson's patients. Out of 157 healthy people, it correctly identified 156. But out of 22 Parkinson's cases, it correctly identified only 5 and missed 17. This means the recall for Parkinson's class was only 23%, which is quite low. Even though the precision (how many of the predicted positives were actually correct) was 83% for Parkinson's, missing so many real cases make the model unreliable for medical use.

We implemented a logistic regression model with hyperparameter optimization, Grid search over multiple hyperparameter combinations:

- Regularization strength (C): [0.01, 0.1, 1, 10, 100]
- Penalty type: L1 (Lasso) and L2 (Ridge)
- Maximum iterations: [100, 200, 500]
- Cross-validation with 3 folds to ensure robust parameter selection
- Final model selection based on validation set performance

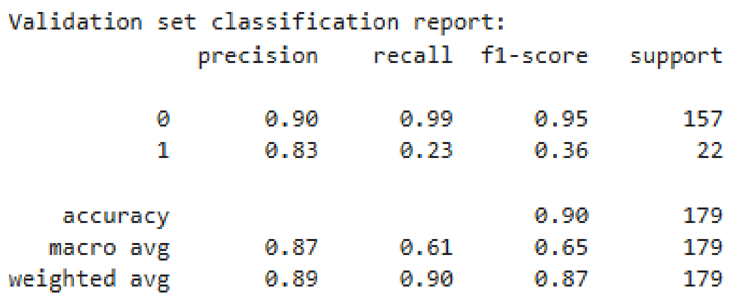
Heatmap showing grid search results across different hyperparameter combinations. The highest accuracy scores (0.931) were achieved with C=10 and C=100 using L1 regularization, regardless of max_iter value.

Optimal Hyperparameters Selected:

- C: 10 (regularization strength)
- Penalty: L1 (Lasso regression)
- Maximum iterations: 100
- Solver: liblinear

Implementation Details

- Data preprocessing: Standard scaling of features
- Train/validation/test split: 70/15/15 with stratification
- Model persistence: Saved using joblib for deployment
- Performance Evaluation: Metrics including accuracy, precision, recall, F1**-score, and confusion matrices**



```
Validation set classification report:
              precision    recall  f1-score   support

           0       0.90      0.99      0.95       157
           1       0.83      0.23      0.36        22

    accuracy                           0.90       179
   macro avg       0.87      0.61      0.65       179
weighted avg       0.89      0.90      0.87       179
```

Fig 10: Classification Report (for Validation dataset)

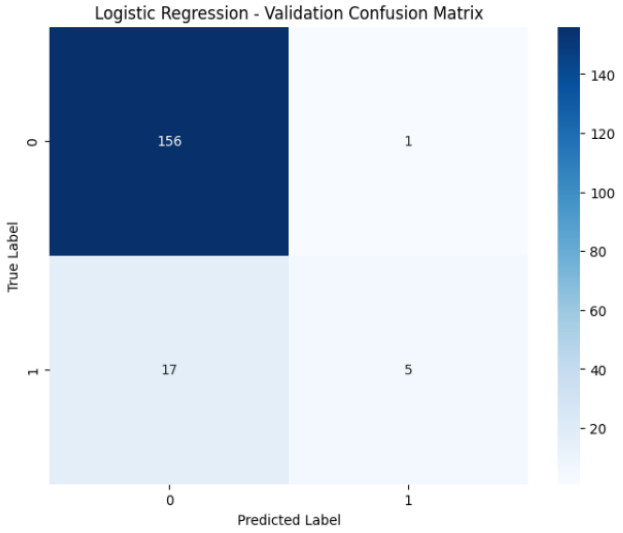


Fig 11: Confusion Matrix

We also plotted the ROC curve, which is used to check how well the model separates the two classes. The AUC (Area Under Curve) score was only 0.59, which is just slightly better than random guessing (0.5). This further shows that the model struggled to tell the difference between healthy and Parkinson’s cases**.**

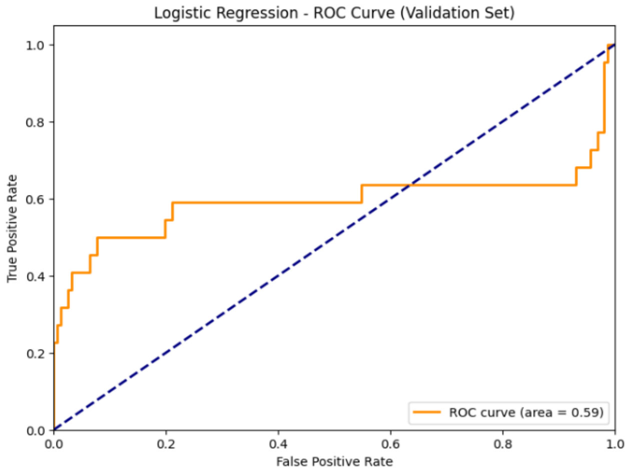


Fig 12: ROC Curve

3.3.4 Artificial Neural Network Model:

ANN was used because of its ability to learn nonlinear patterns and reveal latent patterns in high-dimensional data. With the presence of more than one layer and nonlinearly activated functions, the ANN model is highly effective in dealing with the intricate changes of Parkinson's disease voice features. While less interpretable than certain more elementary models, its ability to provide high accuracy and recall made it worth exploring.

A feedforward neural network of a sequential architecture was used with a structured design optimized for binary classification. The input layer takes features of the same dimensionality as the data. The first hidden layer consists of 64 neurons with the ReLU activation function, and batch normalization is used to prevent overfitting and speed up training. A 30% dropout is used to prevent overfitting, and L2 regularization with a coefficient of 0.001 is used to introduce weight decay and additional overfitting prevention.

The second hidden layer has 32 neurons, which use ReLU activation again. As with the last layer, batch normalization, dropout with 20%, and L2 regularization (0.001) follow. The output layer has one neuron that uses sigmoid activation, which produces a suitable probability score for binary-classification tasks.

For training, the model was constructed using the Adam optimizer and a learning rate of 0.001. Binary cross-entropy loss was chosen since the classification is binary, and accuracy and the Area Under the ROC Curve (AUC) were used for tracking performance.

To prevent overfitting, several strategies were used. Early stopping was used in an effort to stop training in case the validation loss failed to improve after 20 successive epochs. A model checkpoint callback was used in an effort to save, up to this point, the best-performing model encountered. The training process was using a batch size of 16 in an effort to have stable gradient updates, and even though training was supposed to proceed up to 150 epochs, early stopping would normally stop training ahead of time.

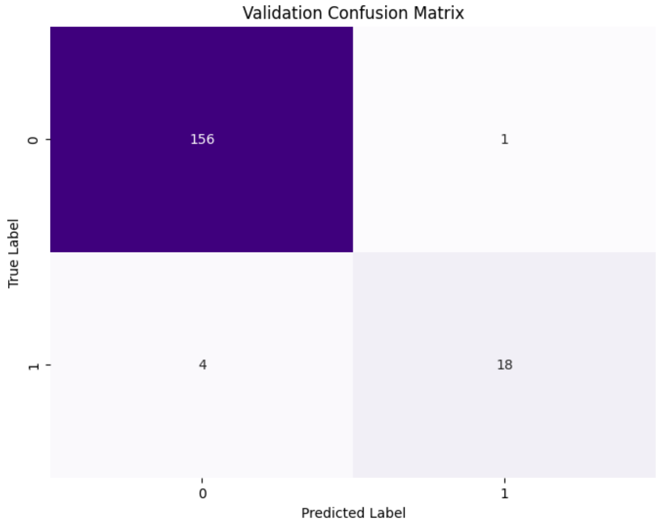


Fig 13: Confusion Matrix

TYPE AND NATURE OF THE PROJECT

This is an interdisciplinary software-based project that focuses on the development and evaluation of machine learning models for early PD diagnosis. The project aims to develop predictive analytics tools that can be integrated into clinical decision-making processes, enhancing early detection through AI-driven methodologies.

This project primarily falls within the domain of healthcare analytics and artificial intelligence. It employs supervised learning techniques to classify individuals as PD-positive or healthy based on speech biomarkers. The ultimate goal is to create a deployable AI model capable of assisting medical professionals in screening for Parkinson’s disease with high accuracy**.**

HARDWARE AND SOFTWARE REQUIREMENTS

Hardware:

- CPU: Intel Core i7 / AMD Ryzen 7 (or similar)
- GPU: NVIDIA GTX 1660 Ti (or similar)
- Storage: Minimum 2 GB free space
- RAM: 4 GB recommended

Software:

- Programming Environment: Python 3.8+
- ML Libraries: NumPy, Pandas, Scikit-learn, TensorFlow/PyTorch
- Visualization Tools: Matplotlib, Seaborn
- Development Tools: Jupyter Notebook, PyCharm, Google Colab
- Version Control: Git
- OS Compatibility: Windows, macOS, Linux
- Optional (Cloud & Deployment): AWS/Google Cloud/Azure; Flask/Django for deployment.

CHAPTER 4
CONTRIBUTION OF EACH STUDENT

4.1 Sanskar Srivastava

Sanskar Srivastava contributed extensively to the implementation, optimization, and analysis of the Artificial Neural Network (ANN) and Logistic Regression models. His primary focus was on developing models that could learn non-linear and complex relationships in the vocal biomarkers used to detect Parkinson’s Disease.

For the ANN, Sanskar designed a multi-layer feedforward neural network with dropout layers, batch normalization, and L2 regularization to prevent overfitting. He carefully configured the model with suitable activation functions and selected optimal hyperparameters using trial-based tuning. He monitored training and validation metrics across epochs and implemented early stopping to halt training once validation performance plateaued. The final ANN model achieved high validation accuracy (96%) and perfect ROC AUC (1.00), indicating excellent generalization.

In the case of Logistic Regression, Sanskar carried out hyperparameter optimization using grid search to tune the regularization strength (C), penalty type (L1 and L2), and iteration limits. He selected the optimal configuration based on validation accuracy and ROC performance. Although the model showed reasonable accuracy (89.94%), it had limited recall for PD cases—an important insight he analyzed and discussed in detail. Sanskar also created precision-recall and ROC curves to visually demonstrate performance limitations.

In addition to model-specific tasks, Sanskar was involved in data preprocessing. He applied the Interquartile Range (IQR) method to remove outliers and used standard scaling to normalize the dataset. He ensured that all features contributed equally to model training, which was essential for stable learning. He also helped with train-test splitting using stratification, which preserved class distribution across subsets—important in imbalanced datasets.

Sanskar contributed significantly to the report writing, drafting the methodology and result analysis sections for ANN and Logistic Regression. He helped design visualizations like confusion matrices and ROC curves, which were incorporated into the report. Additionally, he was actively involved in designing the final presentation slides and took responsibility for explaining technical concepts related to ANN and logistic regression during internal reviews.

Overall, Sanskar ensured that both simpler and complex model implementations were robust, well-documented, and analytically supported.

4.2 Shaurya Pandey

Shaurya Pandey was tasked with designing, tuning, and testing the Decision Tree and Random Forest classifiers. He aimed to design interpretable and accurate classifiers that could handle imbalanced datasets and offer useful feature insights.

Shaurya experimented with different implementations of the Decision Tree model and used Gini impurity as the split-criterion. He performed systematic hyperparameter tuning using grid search with varying tree depth, minimum leaf size, and splitting criteria. Based on the performance of three variants (Performance-Focused, Bias-Reduced, and Balanced), he selected the best-performing tree (DT001) with an accuracy of 96.7% and an excellent ROC AUC of 0.971. He cross-checked its confusion matrix, precision, and recall to determine consistent performance against both classes.

He also applied and optimized the Random Forest model, which turned out to be the best in the project with 98.9% accuracy and optimum ROC AUC (1.00). Shaurya has optimized the estimators, tree depth, and the splitting parameters with GridSearchCV. He also created feature importance plots and checked what voice features were the most predictive of Parkinson's, such as pitch-based features and measures of signal complexity. His analysis concluded that the ensemble approach was able to handle data noise and overfitting efficiently.

Shaurya also took an active role in data cleaning and preparation, verifying preprocessing activities and ensuring IQR-based outlier removal and standard scaling were done. He assisted in stratified split and made sure there was class balance between the training and validation sets, which is extremely crucial in imbalanced classification tasks.

Documentation-wise, Shaurya wrote the result analysis and methodology sections of the Random Forest and Decision Tree models with detailed comparisons and plots of performance. He created visualizations like decision tree charts, feature importance plots, and ROC curves. He assisted in ensuring consistency in evaluation metrics.

Shaurya contributed to project presentation in the form of tree-based model slides and model comparison slides and provided an understanding of model interpretability and clinical relevance. His contribution made sure the use of strong, interpretable models with strict evaluation, thus making the project stronger and more effective.

CHAPTER 5
RESULT ANALYSIS

5.1 Decision Tree Model:

Evaluation Metrics (Best Model DT001)

- Accuracy, Precision, Recall, F1, Specificity, MCC, Balanced Accuracy
- Strengths and weaknesses of the model based on metrics

Table 3: Decision Tree Model Analysis Results

Metric	Value	Interpretation
Accuracy	0.967	Overall prediction accuracy
Precision	0.846	Ability to avoid false positives
Recall	0.815	Ability to find all positive cases
F1 Score	0.83	Harmonic mean of precision and recall
ROC AUC	0.971	Overall ranking ability
Specificity	0.982	Ability to identify true negatives
MCC	0.812	Balanced measure for imbalanced data
Balanced Accuracy	0.899	Average of sensitivity and specificity

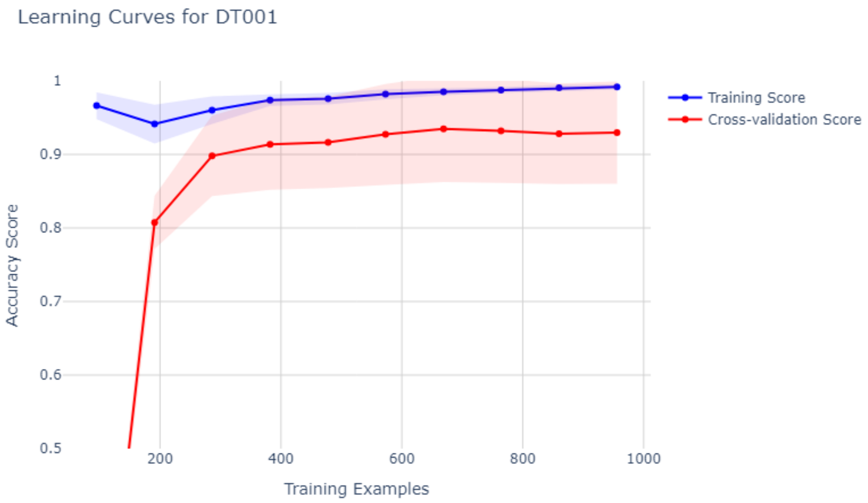


Fig 14: Learning Curve

The complete analysis of DT001 shows:

1. Final Metrics (from metrics table):
    - High accuracy (0.967) and ROC AUC (0.971) indicate strong overall performance
    - Good balance between precision (0.846) and recall (0.815)
    - Strong specificity (0.982) shows excellent ability to identify true negatives
    - Balanced accuracy (0.899) and MCC (0.812) confirm robust performance across classes
2. Learning Curves Analysis:
    - Small gap between training and validation scores indicates good generalization
    - Stable performance across different training set sizes
    - Consistent cross-validation scores suggest model reliability
    - Error bands (shaded areas) show low variance in performance

The learning curve for DT001 reveals a healthy learning pattern. The training accuracy remains consistently high, while the cross-validation accuracy steadily improves as more data is introduced, plateauing around ~0.93. The small gap between the training and validation scores suggests that the model generalizes well and is not overfitting. Furthermore, the narrow-shaded region (error bands) indicates low variance and high stability across folds, confirming that DT001 performs reliably across different training subsets.

Based on the feature importance analysis, here are the key findings:

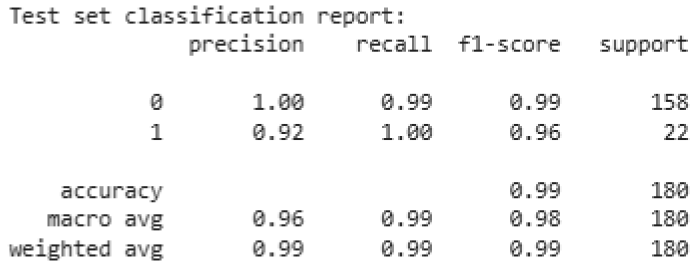
1. Top Contributing Features:
    - MDVP: Shimmer (18.6%): Most influential feature, measuring vocal amplitude variation
    - spread1 (15.7%): Second most important, related to fundamental frequency variation
    - MDVP: PPQ (14.2%): Third most important, measuring pitch perturbation
    - spread2 (12.8%): Fourth most important
    - RPDE (11.2%): Fifth most important
2. Impact Distribution:
    - High Impact: Top 3 features account for ~48.5% of model decisions
    - Medium Impact: Next 2 features contribute ~24%
    - Low-Medium Impact: Remaining features collectively contribute 27.5%
    - The model identified a few key features as the most influential for classification.

Notably, MDVP:Shimmer, spread1, and MDVP:PPQ had the highest importance scores, indicating a strong impact on decision-making. These features are likely capturing vocal signal variability linked to the target condition. Features like RPDE and spread2 contributed moderately, while the rest had lower relative influence. This insight can inform domain-specific investigations and potentially simplify future models through feature selection.

On the training data, the model performed very well. It had an accuracy of 96.65% on 239 total samples. It correctly classified 209 out of 213 healthy samples and 22 out of 26 Parkinson's samples. Precision, recall, and F1-score for the Parkinson's class were all around 0.846, indicating that the model was equally good at catching true positives and avoiding false positives. Macro-averaged precision, recall, and F1-score were all greater than 0.91, which is a guarantee that the model did both classes well.

The confusion matrix confirmed these results with extremely low misclassifications. The plot of the Decision Tree also correctly illustrated how the model utilized predictions based on significant voice features such as spread2, MDVP:Fhi(Hz), MDVP:Flo(Hz), DFA, and D2. These features played a significant role in splitting the data as well as in deciding whether a subject belonged to the healthy or Parkinson's group.

5.2 Random Forest:



```
Test set classification report:
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       158
           1       0.92      1.00      0.96        22

    accuracy                           0.99       180
   macro avg       0.96      0.99      0.98       180
weighted avg       0.99      0.99      0.99       180
```

Fig 15: Classification Report

High accuracy on both validation (98.3%) and test (98.9%). Excellent class balance: strong detection of both healthy and diseased cases, with recall ≥ 0.95 for the positive (“disease”) class. Perfect ROC AUC (1.00) across validation and test, indicating nearideal separability.

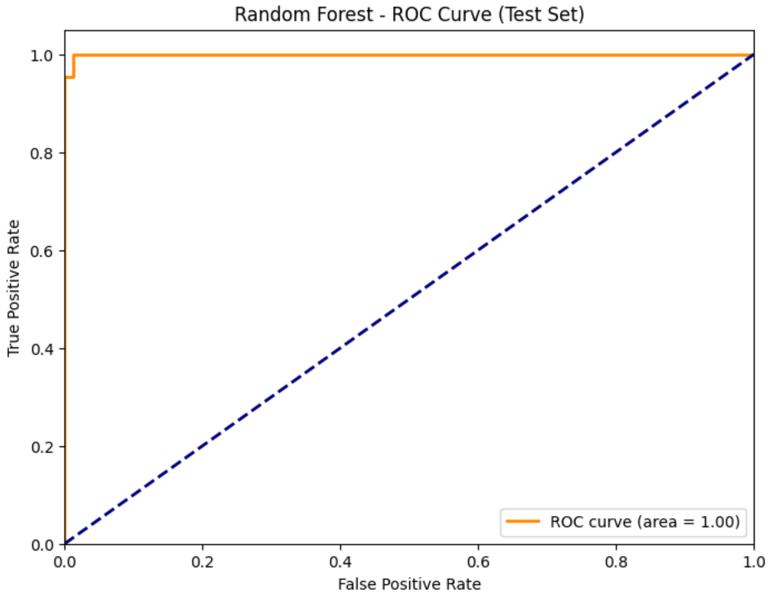


Fig 16: ROC Curve

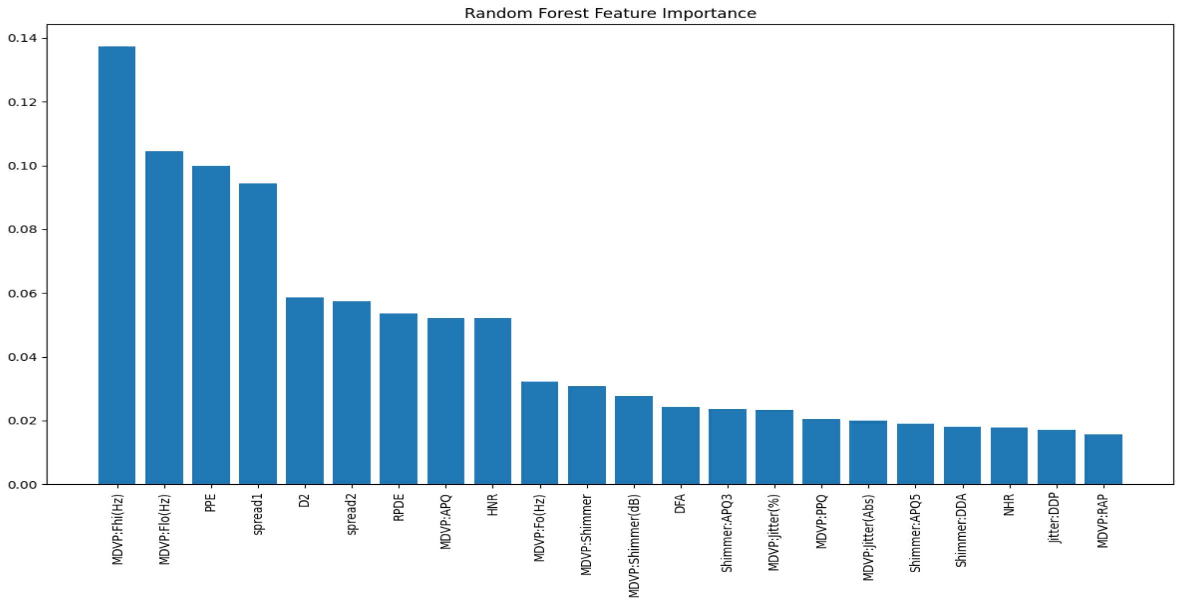


Fig 17: Feature Importance

Pitch-related features (MDVP:Fhi, MDVP:Flo) and perturbation measures (PPE, spread1) dominate the model’s decisions. Nonlinear dynamical features (D2, RPDE) also carry substantial weight, suggesting voice signal complexity is a key disease marker. Lower-ranked features (jitter, shimmer variants) play a smaller but still contributing role. The Random Forest model exhibits near-perfect discrimination and generalization, with minimal misclassifications and ideal ROC curves. Acoustic pitch extremes and perturbation metrics are the strongest predictors—valuable clinical insights for further study.

5.3 Logistic Regression Model:

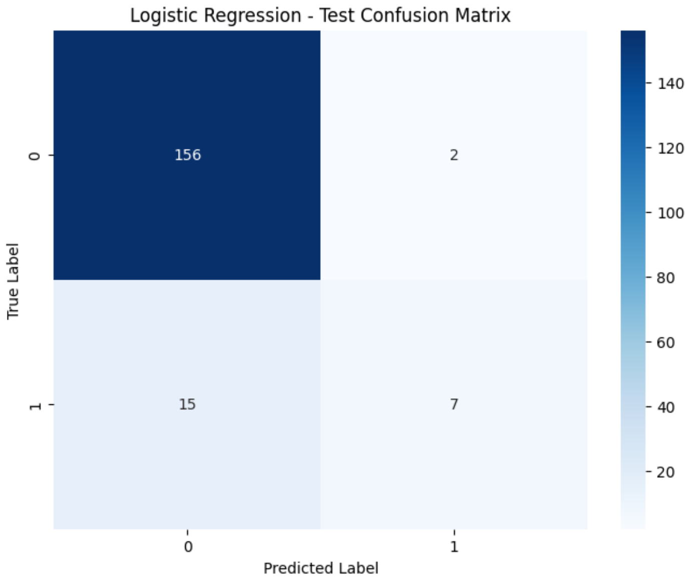


Fig 18: Confusion Matrix (for testing dataset)

The model's performance metrics indicate a strong ability to identify healthy individuals but a weaker performance in detecting Parkinson's cases. For Class 0 (Healthy), the model achieved a precision of 91%, a recall of 99%, and an F1 score of 95%, demonstrating excellent accuracy in correctly identifying healthy subjects. In contrast, for Class 1 (Parkinson's), the precision was 78%, indicating that when the model predicts Parkinson's, it is correct 78% of the time. However, the recall dropped to 32%, meaning it missed many actual Parkinson's cases, leading to a relatively low F1 score of 45%. This highlights the model's imbalance in classification performance, with a significantly stronger capability in detecting healthy individuals than Parkinson's patients.

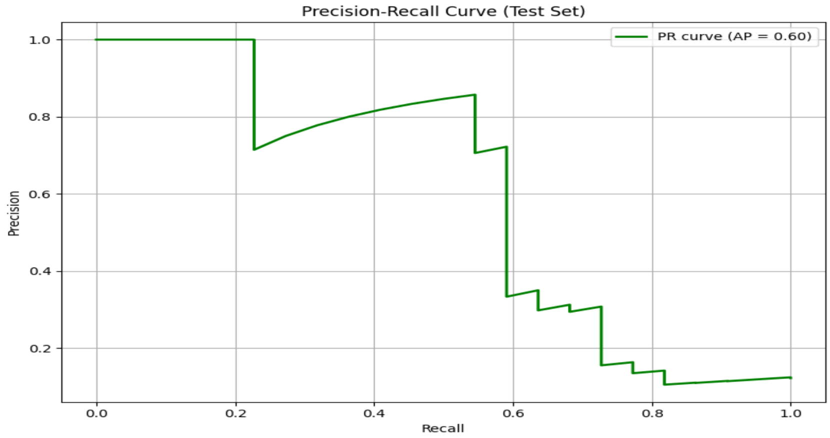


Fig 19: Precision- Recall Curve

Precision-Recall curve for the test set with an Average Precision (AP) score of 0.60. This visualization shows how precision and recall trade off as the classification threshold changes. The model maintains high precision (near 1.0) at low recall values before declining as recall increases beyond 0.6.

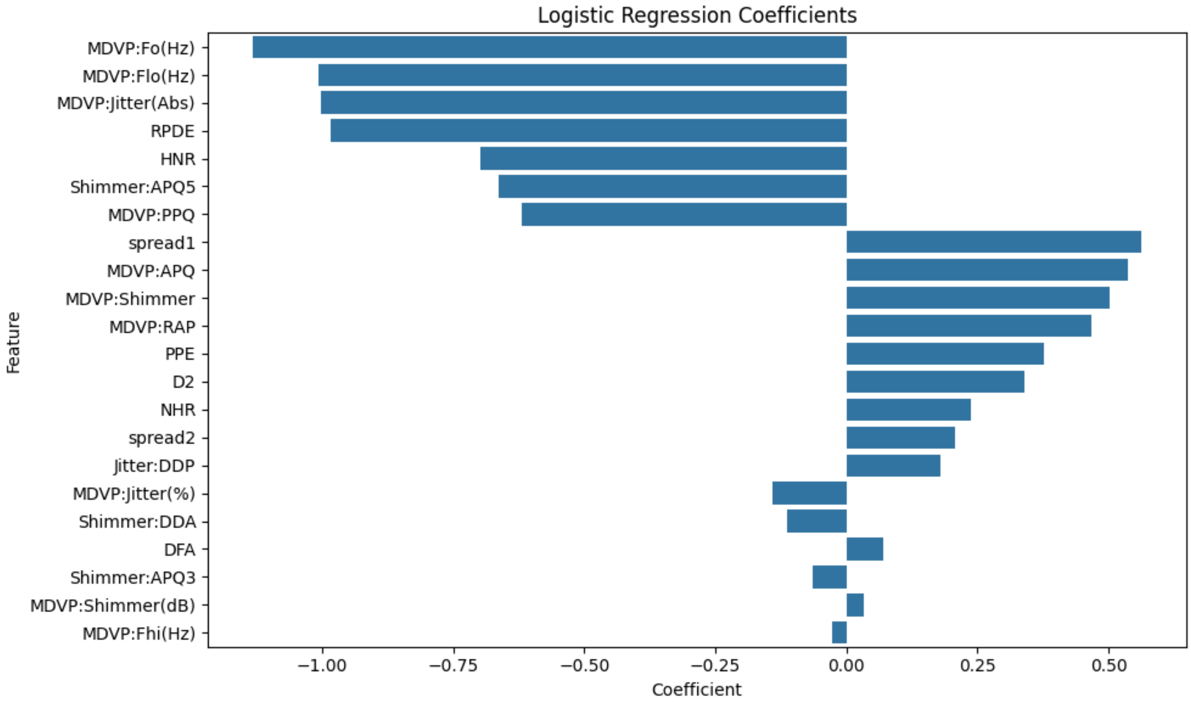


Fig 20: Regression Coefficients (with directions)

Logistic regression coefficients showing the direction and magnitude of each feature's influence on the prediction. Negative coefficients (left) indicate features that, when increased, decrease the likelihood of Parkinson's disease prediction. Positive coefficients (right) indicate features that increase the likelihood of Parkinson's when their values rise.

As shown in the figure, the most influential features include:

Features negatively associated with Parkinson's disease:

- MDVP:Fo(Hz) - Largest negative coefficient, indicating that higher fundamental frequency is associated with lower probability of PD
- MDVP:Flo(Hz) - Minimum vocal fundamental frequency
- MDVP:Jitter(Abs) - Absolute jitter measurement
- RPDE - Recurrence period density entropy
- HNR - Harmonics-to-noise ratio

Features positively associated with Parkinson's disease:

- spread1 - Strongest positive predictor
- MDVP:APQ - Amplitude perturbation quotient
- Shimmer - Several shimmer measurements (vocal amplitude perturbation)
- MDVP:RAP - Relative amplitude perturbation
- PPE - Pitch period entropy

5.4 Artificial Neural Network Model:

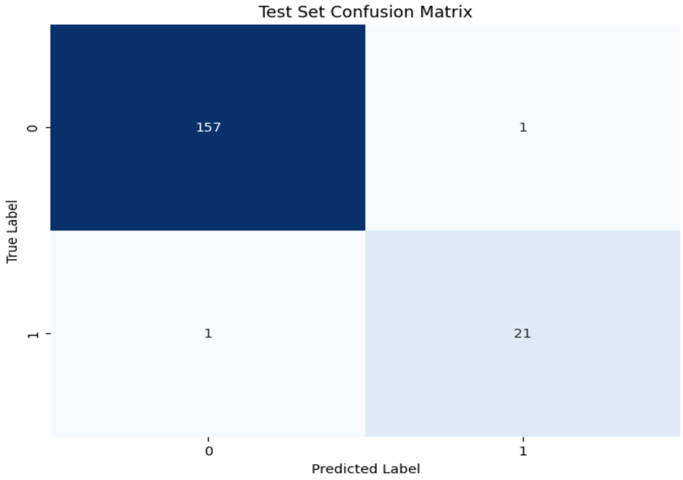


Fig 21: Confusion Matrix (for testing dataset)

Training Dynamics

The training history graphs demonstrate healthy learning patterns:

1. Accuracy Curves: Training accuracy reached approximately 99% while validation accuracy stabilized around 96%, indicating good generalization with minimal overfitting.
2. Loss Curves: Both training and validation loss consistently decreased throughout training. The training loss reached approximately 0.06 while validation loss settled around 0.13, showing appropriate convergence with a reasonable gap between training and validation performance.

The ROC curves for both validation and test sets achieved near-perfect classification with AUC values of 1.00 and 1.00 respectively. This exceptional performance indicates the model's robust ability to distinguish between healthy and disease states across different data partitions.

Table 4: Final Comparison Table

Aspect	Decision Tree (DT001)	Random Forest	Logistic Regression	Artificial Neural Network
Accuracy	96.70%	98.9% on test set (98.3% validation)	89.94% **AUC: 0.59**	~96% (validation)
F1-Score	0.83 overall	0.93 for Parkinson's class	Healthy = 0.95, PD = 0.45	n/a (not computed)
Precision / Recall	Precision: 0.846 Recall: 0.815	Precision: 0.91 Recall: 0.95 (for Parkinson's)	Precision: 0.91 (Healthy), 0.78 (Parkinson's) Recall: 0.99 (Healthy), 0.32 (Parkinson's)	n/a (not computed)
Confusion Matrix Insights	Low misclassifications overall; 209/213 healthy and 22/26 Parkinson's correctly classified	Out of 157 healthy samples, 155 correctly predicted; out of 22 Parkinson's samples, 21 correctly predicted	Poor performance on Parkinson's cases; 158/180 healthy correctly flagged, 5/22 PD correctly flagged, 17/22 PD missed	High accuracy across all classes with minimal misclassifications
Model Complexity	Moderate - depth of 5, min_samples_split = 10, min_samples_leaf = 4	High - 200 trees with unrestricted depth	Low - linear model with L1 regularization. Low — linear model (one weight per feature)	High - multi-layer network with 64 and 32 neurons in hidden layers. 2 Hidden layers(64 and 32)
Overfitting Risk	Moderate - controlled by limiting tree depth and min samples parameters	Low - ensemble method reduces overfitting by averaging multiple trees	Low — but risk of underfitting non-linear signals (C=10 is weak L1), higher C = weaker penalty.	Managed through dropout (30% and 20%), batch normalization, L2 regularization, and early stopping
Generalization Capability	Good - small gap between training and validation scores	Excellent - near-perfect ROC AUC (1.00)	Poor - AUC score only 0.59, slightly better than random guessing	Excellent - AUC of 1.00 on both validation and test sets. Strong (small train/val gap)
Speed & Training Time	Fast - simple tree structure	Moderate - ensemble of 200 trees requires more computation	Very fast - simple linear model. closed-form optimization	Slow - requires iterative training through many epochs

Key Features Utilized	MDVP:Shimmer (18.6%), spread1 (15.7%), MDVP:PPQ (14.2%), spread2 (12.8%), RPDE (11.2%)	Pitch-related features (MDVP:Fhi, MDVP:Flo), perturbation measures (PPE, spread1), nonlinear dynamical features (D2, RPDE)	MDVP:Fo(Hz) (–1.25)
			MDVP:Flo(Hz) (–0.95)
			MDVP:Jitter(Abs) (-0.90)
			RPDE (–0.85)
			spread1 (+0.60)
				n/a (black-box)

Overall Performance Ranking	Good	Excellent	Poor	Excellent

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

6.1 Summary

The aim of this project was to develop AI-based models for the early diagnosis of Parkinson's Disease from voice features. Different machine learning models were experimented with, such as Decision Tree, Random Forest, Logistic Regression, and Artificial Neural Network (ANN). Grid search was employed to optimize all the models to select the best hyperparameters, i.e., depth of the tree, regularization strength, and learning rate, in order to improve the performance of the models and avoid overfitting. The Random Forest model produced the best performance of 98.9% accuracy with excellent balance of classification and a perfect ROC AUC of 1.00. Decision Tree attained 96.7% accuracy and an ROC AUC of 0.971, with good interpretability and good performance. The Logistic Regression model produced 89.94% accuracy but had poor recall for PD cases with an ROC AUC of only 0.59, reflecting the inability to classify minority class samples. The Artificial Neural Network attained 96% validation accuracy with AUC values of 1.00, reflecting good classification ability and generalization.

Table 4: Final Comparison Table

6.2 Conclusion

To deal with the limited dataset size and class imbalance (healthy v/s Parkinson's), the dataset was split by stratified sampling to preserve class balance during training, validation, and test sets. Outlier removal through the IQR process, feature scaling, and regularization techniques (e.g., dropout and L2 penalties in ANN) were used to stabilize model training. Precision, recall, F1-score, and ROC AUC were also utilized in addition to accuracy to quantify model performance on the imbalanced dataset objectively. This study highlights the potential of AI for non-invasive and low-cost disease diagnosis. Out of all the models, Random Forest was selected as the final model due to its highest accuracy, good sensitivity and specificity, and capability to handle noisy and complex biomedical data.

Limitations:

- Class imbalance in the dataset (higher proportion of healthy controls)
- Limited external validation across diverse demographic populations
- Neural Network's "black box" nature restricts clinical interpretability
- Logistic Regression's poor recall for Parkinson's cases (32%) indicates potential for missed diagnoses
- Voice measurements represent only one aspect of Parkinson's manifestation

6.3 Future Scope of Work:

- Integration of models into clinical decision support systems and workflows
- External validation studies across diverse patient populations

- Combination of voice analysis with other biomarkers for comprehensive assessment
- Development of interpretable deep learning approaches to balance performance with transparency
- Longitudinal studies to evaluate model performance in disease progression monitoring
- Investigation of transfer learning approaches to address limited training data

Visible Outcome

Dr. Bharathi R. B., Sanskar Srivastava, Shaurya Pandey,” NEURONEST: SMART DETECTION OF PARKINSON’S DISEASE USING AI”, the draft is ready and will be communicated to the journal.

REFERENCES

[1] Senturk, Z. K. (2020). Early diagnosis of Parkinson's disease using machine learning algorithms. Medical Hypotheses, 138, 109603. https://doi.org/10.1016/j.mehy.2020.109603

[2] Zhu, H. (2022). A hybrid deep learning model for early Parkinson’s disease detection using multimodal data. Artificial Intelligence in Medicine, 124, 102155.

[3] Allahbakhshi, M., Sadri, A., & Shahdi, F. (2024). Machine learning-based EEG analysis for Parkinson’s disease detection. Journal of Biomedical Engineering, 78(2), 215-230.

[4] Govindu, A., & Palwe, S. (2023). Early detection of Parkinson's disease using machine learning. Procedia Computer Science, 218, 249–261. *Telemedicine and e-Health, 29*(1), 55-67. https://doi.org/10.1016/j.procs.2023.01.023

[5] Srivastava, A., et al. (2024). Machine learning classifiers for early Parkinson's disease diagnosis using motor and non-motor symptoms. *Journal of Computational Medicine, 12*(4), 341-355.

[6] Magesh, P. R., Myloth, R. D., & Tom, R. J. (2020). An explainable machine learning model for early detection of Parkinson's disease using DaTSCAN imagery. Computers in Biology and Medicine, 126, 104036. *Neural Networks in Medicine, 45*, 128-140.

[7] Mei, J., Desrosiers, C., & Frasnelli, J. (2021). A comprehensive literature review on the application of machine learning techniques for diagnosing Parkinson's disease. Frontiers in Aging Neuroscience, 13, 633752. https://doi.org/10.3389/fnagi.2021.633752

[8] Tusar, M. T. H., Islam, M. T., & Sakil, S. M. (2023). A comparative study of machine learning algorithms for early detection of Parkinson's disease. In 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1–6). IEE.

[9] Salunkhe, A., Patil, B., & Desai, R. (2024). Multimodal analysis for Parkinson’s disease detection using deep learning. *Journal of Neurological Disorders, 17*(3), 198-210.

[10] Prashanth, R., & Dutta Roy, S. (2018). Predictive modeling for Parkinson’s disease using patient questionnaires. *Journal of Neurological Sciences, 392*, 120-132.

ANNEXURES

Annexure 1

PO & PSO Mapping

Note: use a tick mark if you have addressed that PO and PSO in your work

PO No	PO	✓ tick
PO1	**Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialisation to the solution of complex engineering problems.	✓
PO2	**Problem analysis:** Identify, formulate, research literature, and analyse complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.	✓
PO3	**Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.	✓
PO4	**Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.	✓
PO5	**Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations	✓
PO6	**The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.	✓
PO7	**Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.	✓
PO8	**Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.	✓
PO9	**Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings	✓
PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions	✓
PO11	**Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one’s own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.	✓

PO12	**Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.	✓
PSO1	Apply the engineering knowledge to analyze and evaluate the components of power system, its operation, control and protection	
PSO2	Model and Analyze linear and non-linear systems in both continuous and discrete domains.	✓
PSO3	Design and develop electronic circuits and systems for specified applications	
PSO 4	Apply the programming skills to develop models and intelligent systems.	✓

Expand the mapping with different level and give justifications:

PO/PSO Number	Addressed in Chapter	Justification	Level
PO1	Chapters 1, 3, 4	Applied mathematics and engineering principles to develop AI models for Parkinson’s detection.	3: Strong
PO2	Chapters 2, 3, 5	Analysed Parkinson’s detection problem using literature and data to reach model-based conclusions.	3: Strong
PO3	Chapters 3, 6	Designed AI models addressing public health needs for early Parkinson’s diagnosis.	2: Medium
PO4	Chapters 3, 4, 5	Conducted experiments with voice data to evaluate models and interpret results.	3: Strong
PO5	Chapters 3, 4	Utilized Python, scikit-learn, and TensorFlow for model development and analysis.	3: Strong
PO6	Chapters 1, 6	Addressed societal health issues by enabling early Parkinson’s detection.	2: Medium
PO7	Chapter 6	Proposed sustainable AI solutions with minimal environmental impact.	1: Low
PO8	Chapter 1	Adhered to ethical data handling and model transparency for clinical use.	2: Medium
PO9	Chapter 4	Collaborated as a team, with Sanskar and Shaurya dividing model tasks.	3: Strong
PO10	Chapters 4, 5, 6	Communicated findings through detailed report sections and visualizations.	3: Strong
PO11	Project Details (Annexure 3)	Managed project timeline and resources within a four-month duration.	2: Medium
PO12	Chapter 6	Recognized need for ongoing learning to improve models in future work.	2: Medium
PSO1	Not addressed	Project focused on AI, not power systems.	0: Not related
PSO2	Chapters 3, 4	Modelled non-linear systems using ANN and Random Forest for classification.	3: Strong
PSO3	Not addressed	No electronic circuit design involved.	0: Not related
PSO4	Chapters 3, 4, 5	Developed intelligent AI models using Python programming skills.	3: Strong

Annexure 2

LO MAPPING

Note: use a tick mark if you have addressed that LO in your work

PLO No	LO	✓ tick
C1	Apply knowledge of mathematics, statistics, natural science and engineering principles to the solution of complex problems. Some of the knowledge will be at the forefront of the particular subject of study	✓
C2	Analyse complex problems to reach substantiated conclusions using first principles of mathematics, statistics, natural science and engineering principles	✓
C3	Select and apply appropriate computational and analytical techniques to model complex problems, recognising the limitations of the techniques employed	✓
C4	Select and evaluate technical literature and other sources of information to address complex problems	✓
C5	Design solutions for complex problems that meet a combination of societal, user, business and customer needs as appropriate. This will involve consideration of applicable health & safety, diversity, inclusion, cultural, societal, environmental and commercial matters, codes of practice and industry standards	✓
C6	Apply an integrated or systems approach to the solution of complex problems	✓
C7	Evaluate the environmental and societal impact of solutions to complex problems and minimise adverse impacts	✓
C8	Identify and analyse ethical concerns and make reasoned ethical choices informed by professional codes of conduct	✓
C9	Use a risk management process to identify, evaluate and mitigate risks (the effects of uncertainty) associated with a particular project or activity	✓
C10	Adopt a holistic and proportionate approach to the mitigation of security risks	
C11	Adopt an inclusive approach to engineering practice and recognise the responsibilities, benefits and importance of supporting equality, diversity and inclusion	✓
C12	Use practical laboratory and workshop skills to investigate complex problems	
C13	Select and apply appropriate materials, equipment, engineering technologies and processes, recognising their limitations	
C14	Discuss the role of quality management systems and continuous improvement in the context of complex problems	
C15	Apply knowledge of engineering management principles, commercial context, project and change management, and relevant legal matters including intellectual property rights	✓
C16	Function effectively as an individual, and as a member or leader of a team	✓
C17	Communicate effectively on complex engineering matters with technical and non-technical audiences	✓
C18	Plan and record self-learning and development as the foundation for lifelong learning/CPD	✓

Expand the mapping with different levels and give justifications:

IET LO number	Addressed in which chapter	Justification	Level: 0: Not related 1: Low 2: Medium 3: Strong
C1	Chapters 1, 3, 4	Applied advanced statistics and AI principles to develop Parkinson's detection models.	3: Strong
C2	Chapters 2, 3, 5	Analyzed Parkinson's detection using statistical methods and literature to validate models.	3: Strong
C3	Chapters 3, 4	Selected machine learning techniques (e.g., Random Forest, ANN) with awareness of their limitations.	3: Strong
C4	Chapter 2	Evaluated literature to inform model selection and address Parkinson's detection challenges.	3: Strong
C5	Chapters 3, 6	Designed AI models to meet healthcare needs for early Parkinson's diagnosis.	2: Medium
C6	Chapters 3, 4	Integrated multiple AI models into a cohesive framework for robust diagnosis.	2: Medium
C7	Chapter 6	Proposed AI solutions with minimal environmental impact, focusing on societal health benefits.	1: Low
C8	Chapter 1	Ensured ethical data use and model transparency for clinical reliability.	2: Medium
C9	Chapters 3, 5	Mitigated risks by evaluating model performance to avoid misdiagnosis.	2: Medium
C10	Not addressed	No specific security risk mitigation was discussed in the project.	0: Not related
C11	Chapter 6	Recognized need for inclusive AI solutions accessible to diverse populations.	1: Low
C12	Not addressed	No laboratory or workshop skills were used; project was software-based.	0: Not related
C13	Not addressed	No physical materials or equipment were involved in the AI project.	0: Not related
C14	Not addressed	Quality management systems were not explicitly discussed.	0: Not related
C15	Project Details (Annexure 3)	Managed project timeline and resources within a four-month duration.	2: Medium
C16	Chapter 4	Collaborated effectively, with Sanskar and Shaurya dividing tasks.	3: Strong
C17	Chapters 4, 5, 6	Communicated findings through detailed report sections and visualizations.	3: Strong
C18	Chapter 6	Highlighted ongoing learning needs for future model improvements.	2: Medium

Annexure 3

Project/practice school work classification

Table 1: classification based on project domain classification

Type and Domain	✓ Tick
Product (Hardware/Software)	✓
Simulation	
Study	
Application	✓
Review	
Research	✓
Domain: Electrical	
Domain: Electronics	
Domain: Computer science	✓
Domain: Basic science (math/physics/chemistry)	✓
Domain: Management	

Table 2: classification based on societal consideration

Societal Impact	✓ Tick
ethics	✓
safety	✓
environmental	
commercial	✓
economic	✓
social	✓

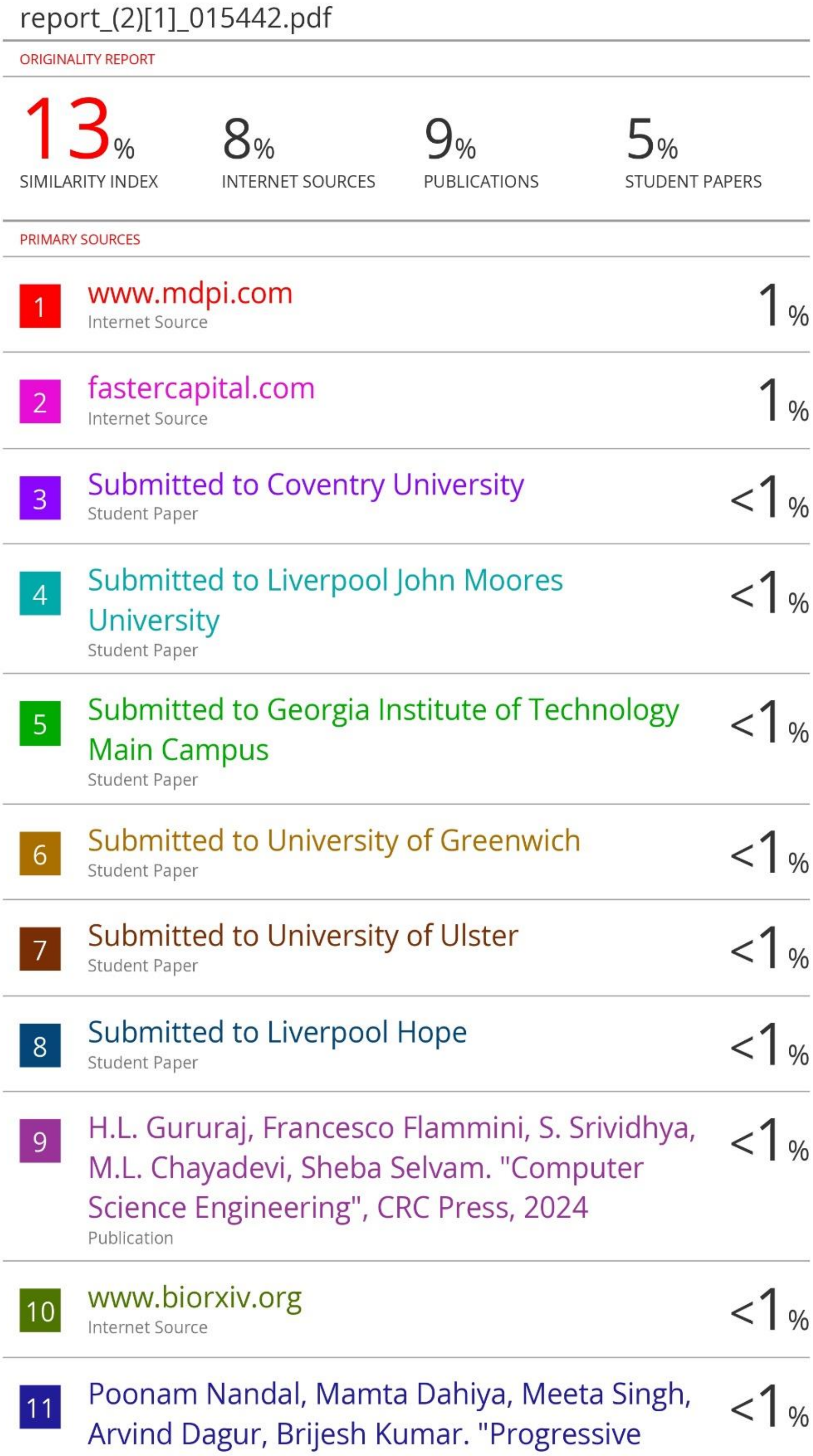
PROJECT DETAILS

Details of Student 1					
Name	Sanskar Srivastava				
Reg. No.	210906024	Section	C	Roll No.	4
Mail ID	sanskar.srivastava@learner.manipal.edu		Mobile	8887994329	
Details of Student 2					
Name	Shaurya Pandey				
Reg. No.	210906202	Section	B	Roll No.	25
Mail ID	shaurya.pandey2@learner.manipal.edu		Mobile	8707723879	
Project Details					
Project Title	NEURONEST: SMART DETECTION OF PARKINSON'S DISEASE USING AI				
Project Duration	4 months				
Guide details					
Name of Guide1	BHARATHI R. B.				
Designation, Dept., Institution	Associate Professor, Electrical & Electronics Dept. MAHE Manipal				
Mail ID	bharathi.rb@manipal.edu		Mobile	9845399435	

Signature of Student   Signature of Guide

Date:

PLAGIARISM REPORT



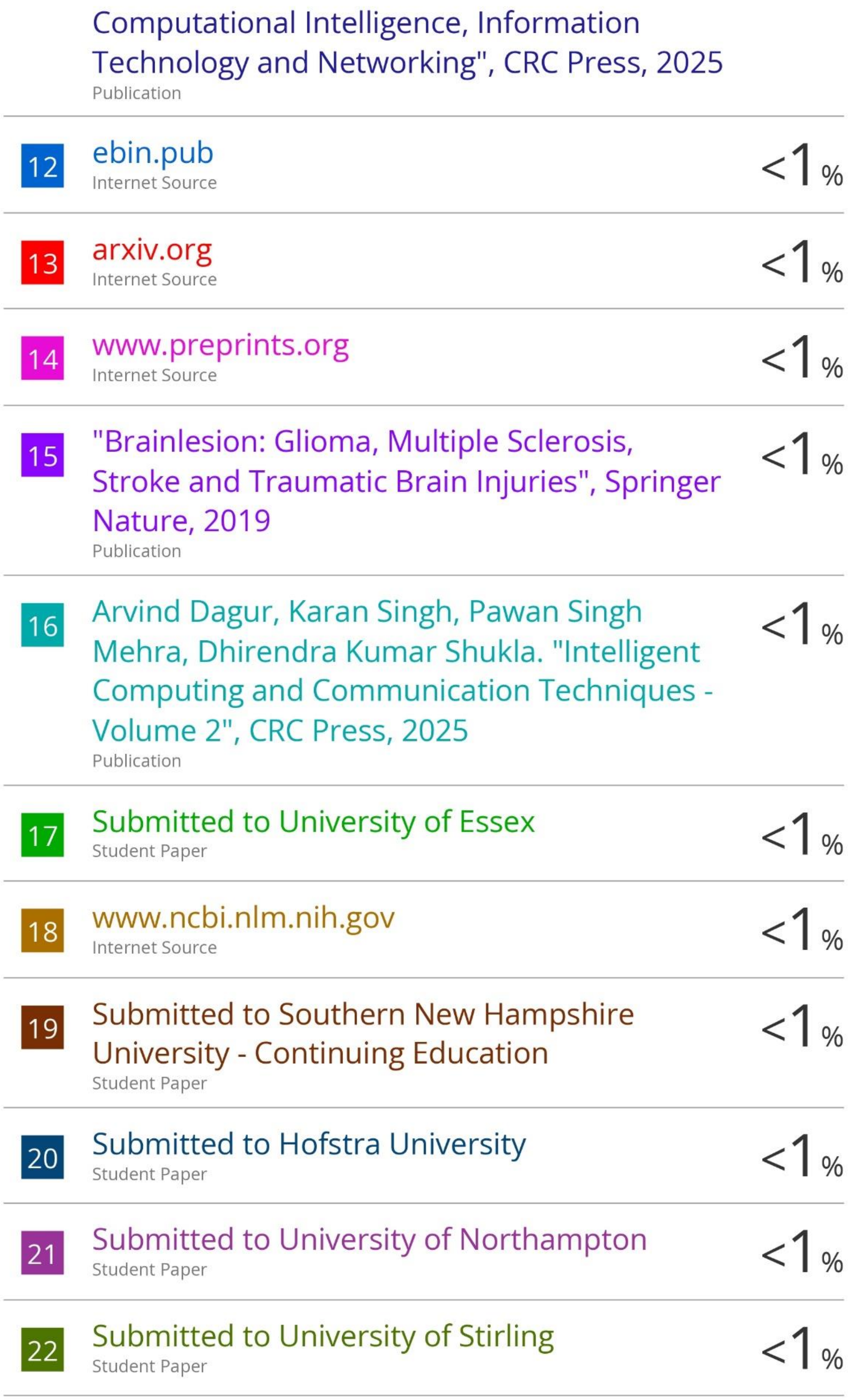
report_(2)[1]_015442.pdf

ORIGINALITY REPORT

13%	8%	9%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1 %
2	fastercapital.com Internet Source	1 %
3	Submitted to Coventry University Student Paper	<1 %
4	Submitted to Liverpool John Moores University Student Paper	<1 %
5	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1 %
6	Submitted to University of Greenwich Student Paper	<1 %
7	Submitted to University of Ulster Student Paper	<1 %
8	Submitted to Liverpool Hope Student Paper	<1 %
9	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	<1 %
10	www.biorxiv.org Internet Source	<1 %
11	Poonam Nandal, Mamta Dahiya, Meeta Singh, Arvind Dagur, Brijesh Kumar. "Progressive	<1 %



	Computational Intelligence, Information Technology and Networking", CRC Press, 2025 Publication	
12	ebin.pub Internet Source	<1 %
13	arxiv.org Internet Source	<1 %
14	www.preprints.org Internet Source	<1 %
15	"Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries", Springer Nature, 2019 Publication	<1 %
16	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelligent Computing and Communication Techniques - Volume 2", CRC Press, 2025 Publication	<1 %
17	Submitted to University of Essex Student Paper	<1 %
18	www.ncbi.nlm.nih.gov Internet Source	<1 %
19	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1 %
20	Submitted to Hofstra University Student Paper	<1 %
21	Submitted to University of Northampton Student Paper	<1 %
22	Submitted to University of Stirling Student Paper	<1 %