# Machine Learning CSE343

## End Semester Presentation

Vansh Yadav 2022559
Utkarsh Dhilliwal 2022551
Vaibhav Singh 2022555
Shamik Sinha 2022468

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
**DELHI**

# Motivation

- Importance of Early Diagnosis in Parkinson's Disease (PD):
  - By the time motor symptoms appear, over **60% of dopaminergic neurons** are damaged, limiting treatment effectiveness.
  - Early detection is crucial to slow disease progression and improve patient outcomes
- Personal Relevance:
  - Three team members are currently studying **Cognition of Motor Movements (PSY308)**, gaining insights into the challenges of PD diagnosis, deepening the motivation for improving diagnostic methodologies

# Literature Review

- Study 1: Alshammri et al. (2023)
  - **Objective:** Used telemedicine and machine learning to detect early-stage PD using the MDVP voice dataset.
  - **Approach:** Trained four models—Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression.
  - **Results:** Random Forest achieved the highest accuracy (91.83%) and sensitivity (0.95), indicating its potential for PD detection. However, the study faced limitations due to the **small sample size (31 participants)** and the **exclusive use of voice data**, which may not capture all PD symptoms

# Literature Review

- Study 2: Singh et al. (2016)
  - **Objective:** Investigated early PD detection using over 100 features extracted from brain MRI images.
  - **Approach:** Used SVM with Principal Component Analysis (PCA) for dimensionality reduction and Fisher Discriminant Ratio (FDR) for feature selection.
  - **Results:** Yielded very high classification accuracies but had high computational costs.

# Literature Review

- Study 3: Prashanth et al. (2016)
  - **Objective:** Investigated early PD detection using multimodal features (REM sleep behavior, olfactory loss, cerebrospinal fluid markers) from the PPMI database.
  - **Approach:** Used Naïve Bayes, Support Vector Machine (SVM), Boosted Trees and Random Forests classifiers.
  - **Results:** Achieved a 96.40% accuracy for early PD diagnosis with 401 PD cases and 183 healthy controls. This study showed the effectiveness of combining non-motor features and imaging data for accurate diagnosis.

# Dataset description

- **Source:** The dataset was sourced from the **Parkinson's Progression Markers Initiative (PPMI)**, an international effort to identify PD biomarkers.

- **Composition: 13,000+ records** from **3,096 participants**.

  1736 PD patients,

  1,018 prodromal cases,

  279 healthy controls,

  63 SWEDD cases

- **Key Attributes:**

  Combination of cognitive, non-motor features like sleep behavior disorder, olfactory loss, CSF proteins and cognitive-behavioral test results.

Parkinson's Progression Markers Initiative

# Key Features in Dataset

| Category | Feature | Description |
|---|---|---|
| Clinical | **UPSIT** | University of Pennsylvania Smell Identification Test, Used to measure olfactory dysfunction |
| Clinical | **REM** | Measure of REM sleep disorder in patients, is a strong early non-motor symptom linked to PD |
| Biologics | **Tau** | Protein in cerebrospinal fluid, related to cognitive decline & Alzheimer's |

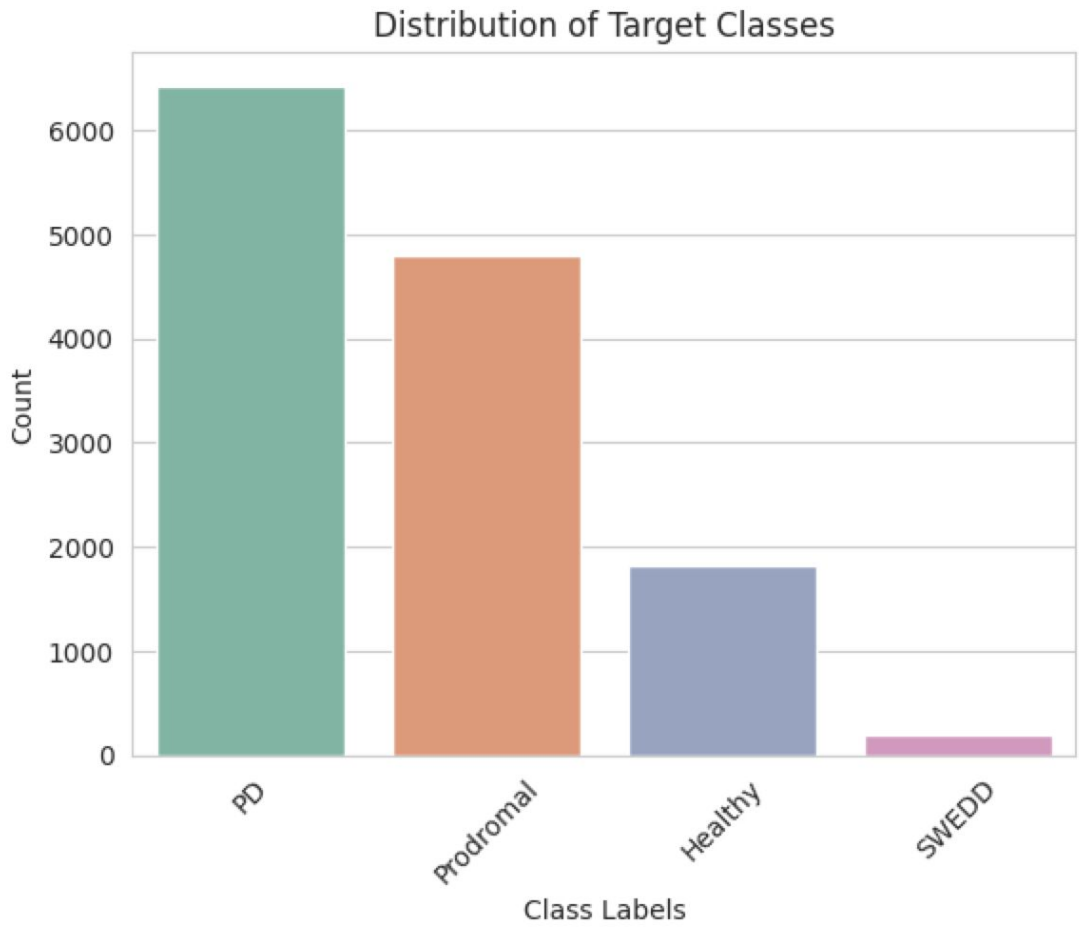
# Key Features in Dataset

| Category | Feature | Description |
| --- | --- | --- |
| Biologics | **Ptau** | Protein associated with cognitive decline and helps distinguish PD from other neurodegenerative diseases |
| Clinical | **MSEADLG** | Modified Schwab & England ADL Score, evaluates cognitive impairment impact on daily living activities |
| Demographics | **Age** | A major risk factor for PD, with its prevalence increasing in older populations |

# Key Features in Dataset

| Category | Feature | Description |
| --- | --- | --- |
| Clinical | **Epworth Sleepiness test** | A self-assessment test to help see how daytime sleepiness affects people |
| Clinical | **Montreal Cognitive Assessment** | A rapid screening instrument for mild cognitive dysfunction |
| Demographics | **fampd** | Family history of PD, higher risk of developing the condition due to hereditary factors |

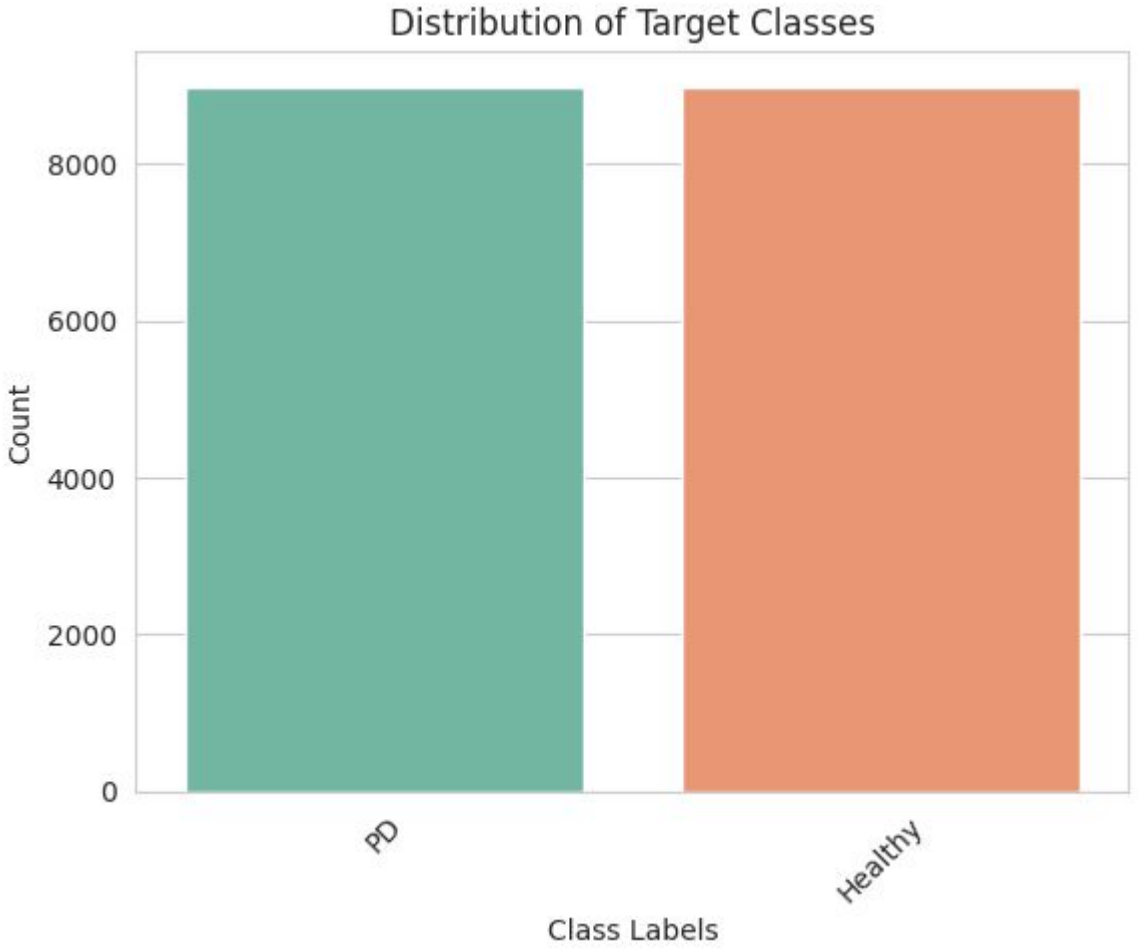
# Data Distribution




**Before Preprocessing**:

The dataset exhibited **class imbalance**, where certain classes (e.g., Parkinson's Disease and Prodromal cases) had significantly more samples compared to the Healthy Control group, making it harder for the model to learn effectively from the minority class.

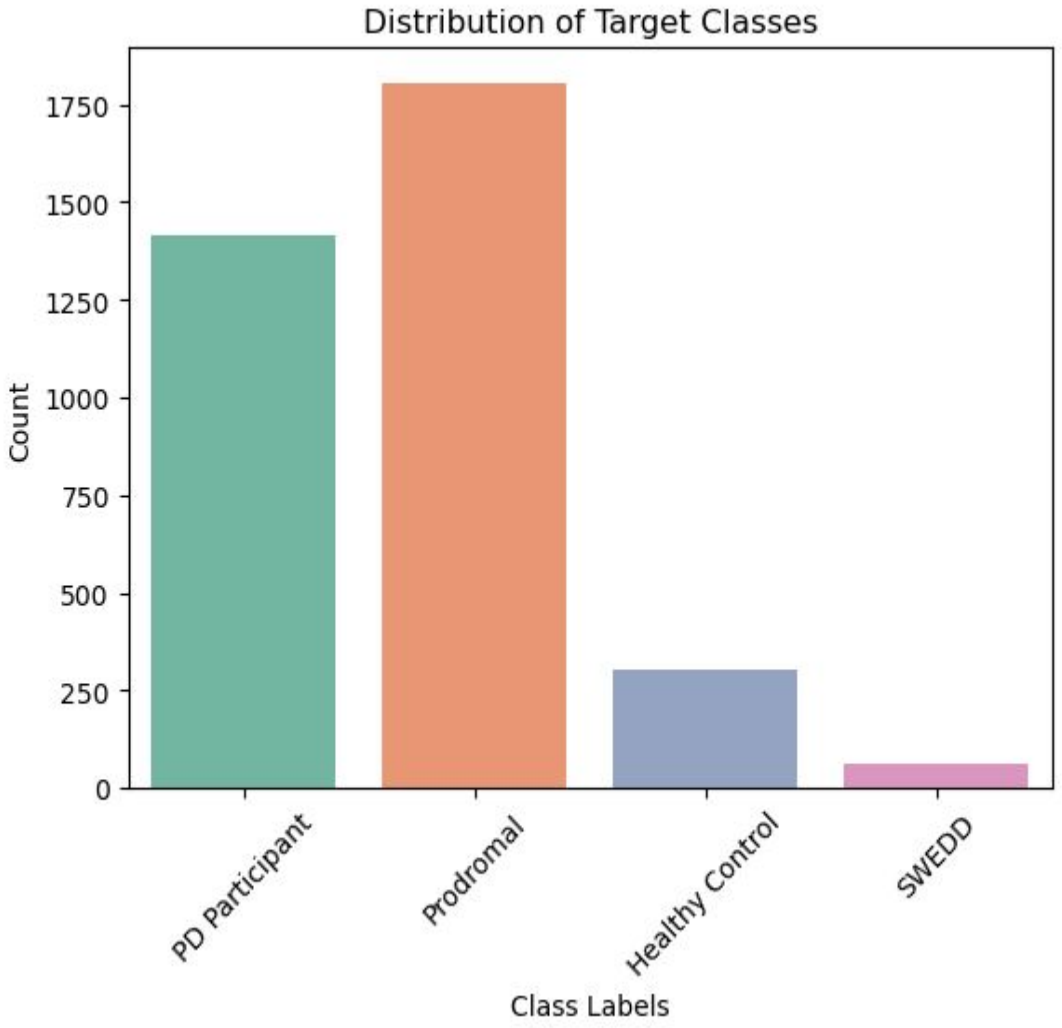
# Data Distribution

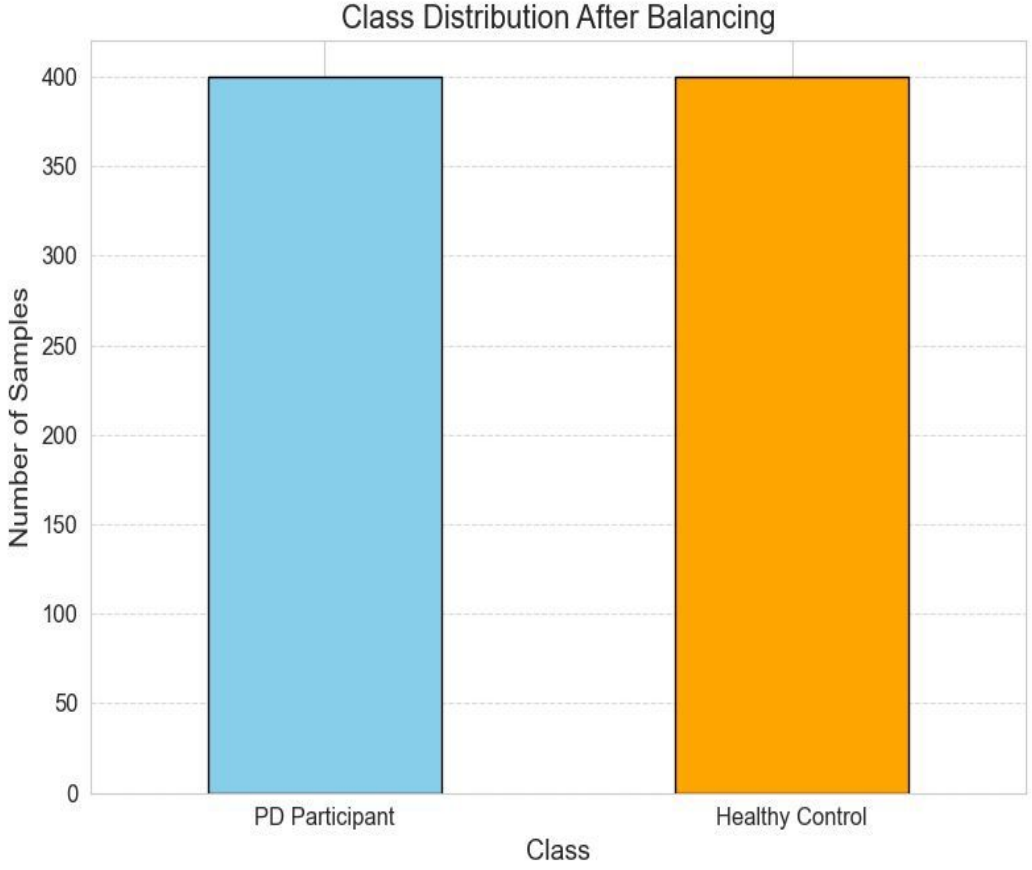



**After Preprocessing, Before Midsem**:

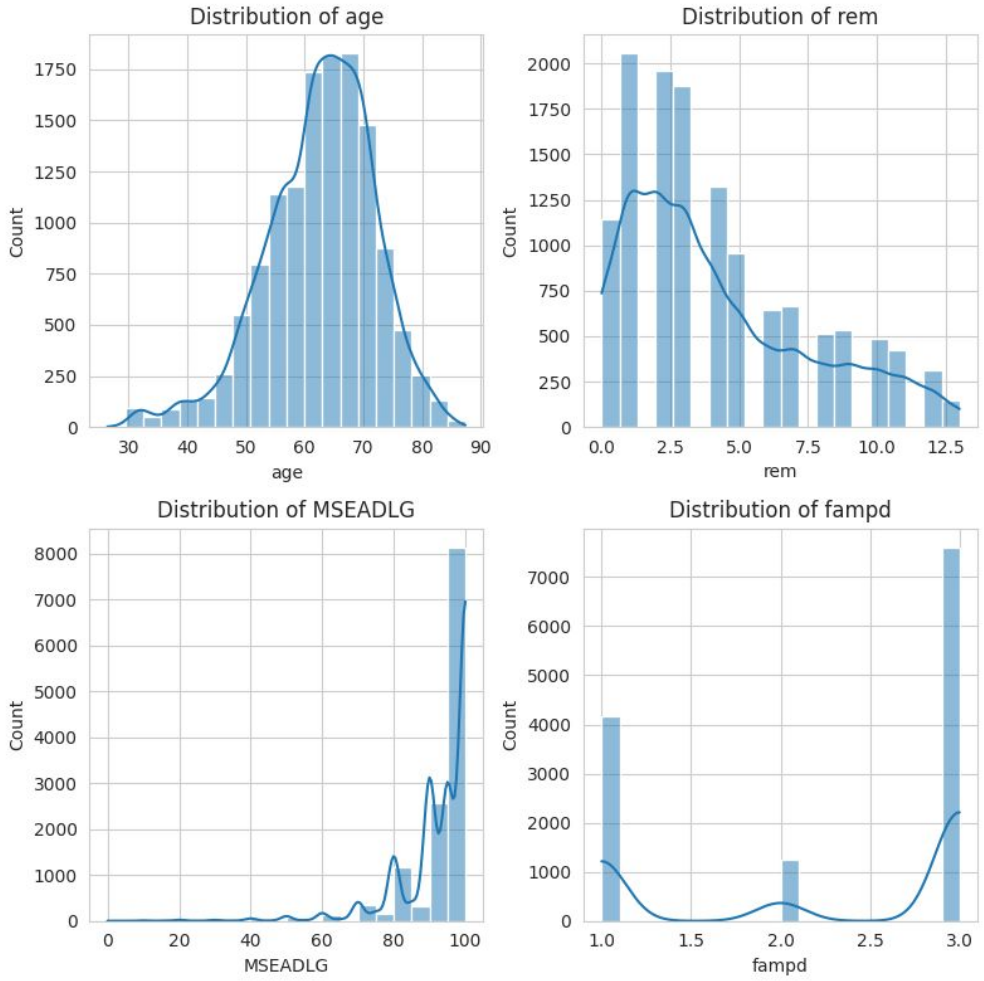
Class imbalance was reduced using **SMOTE (Synthetic Minority Over-sampling Technique)**, resulting in a more even distribution of samples across classes, improving the model's ability to learn from underrepresented classes.

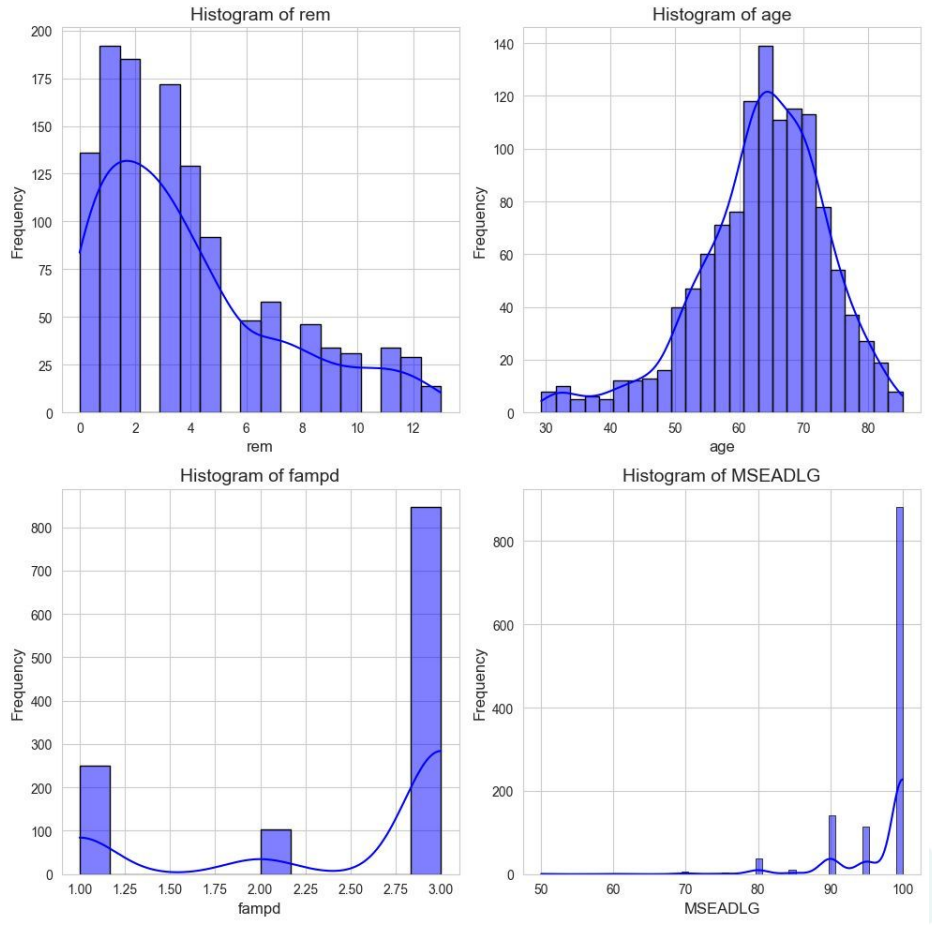
# Data Distribution



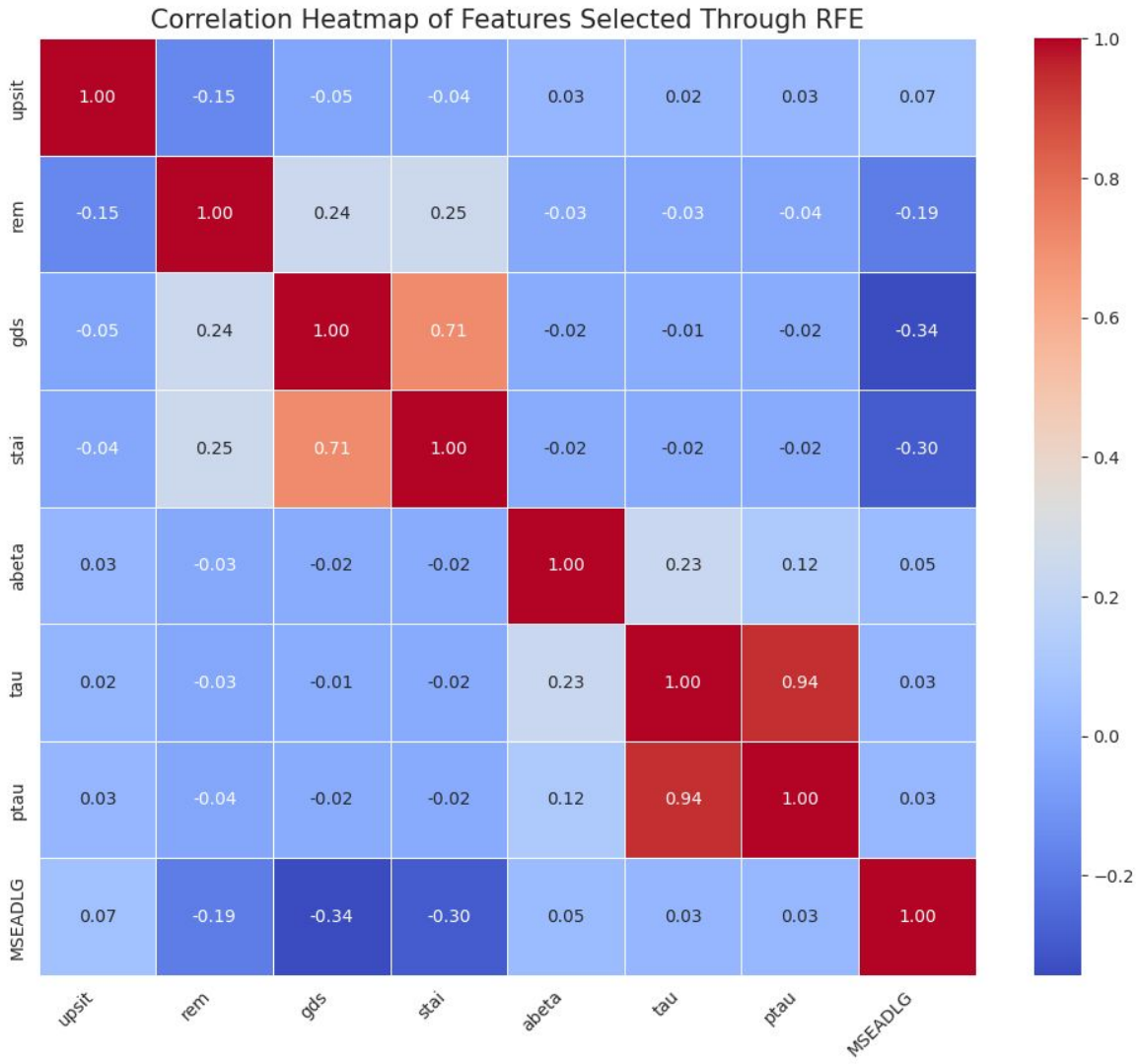



# Data Distribution



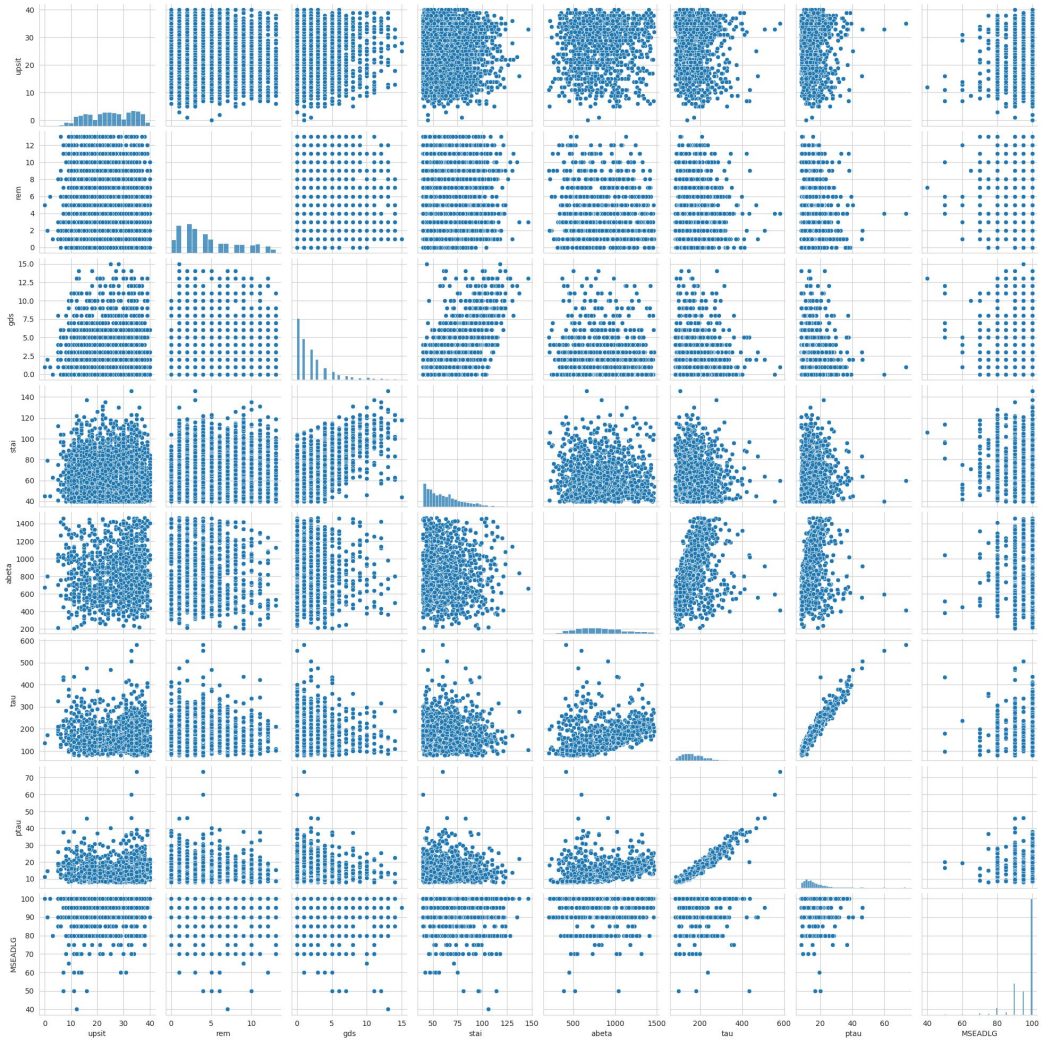
# Data Distribution




**Positive Correlation**: Strong positive correlation observed between MSEADLG and tau/ptau, indicating that as cognitive decline worsens (higher MSEADLG score), tau protein levels also increase.

**Negative Correlation**: Negative correlation found between MSEADLG and upsit/rem, suggesting that higher cognitive decline is associated with lower olfactory function (UPSIT) and REM sleep quality.

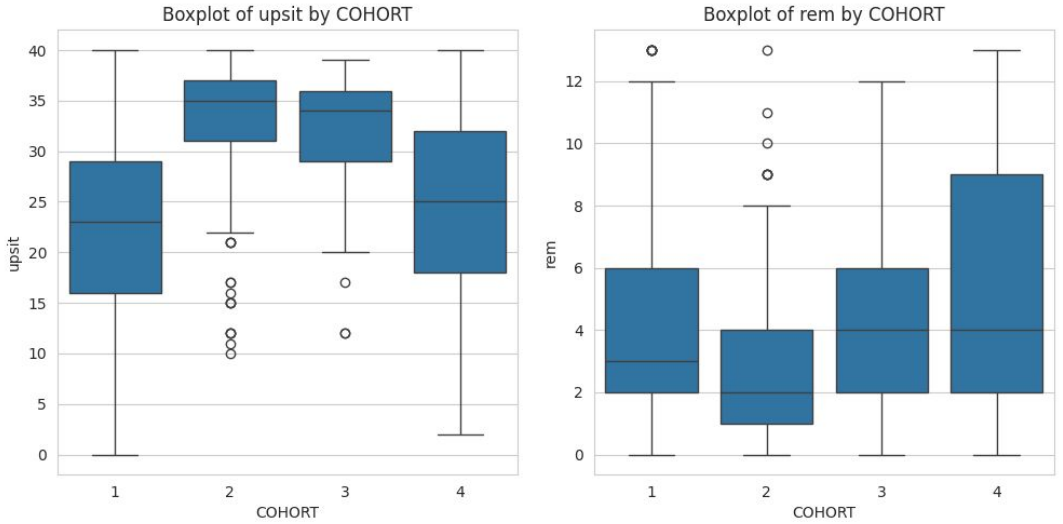
# Data Distribution




**Positive Correlation**: Strong positive correlation observed between MSEADLG and tau/ptau, indicating that as cognitive decline worsens (higher MSEADLG score), tau protein levels also increase.

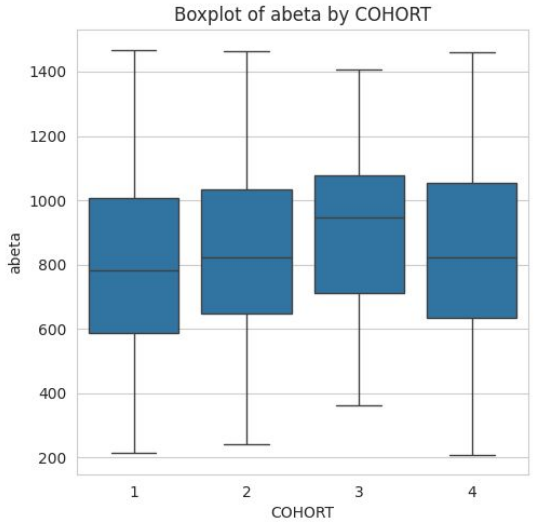
**Negative Correlation**: Negative correlation found between MSEADLG and upsit/rem, suggesting that higher cognitive decline is associated with lower olfactory function (UPSIT) and REM sleep quality.

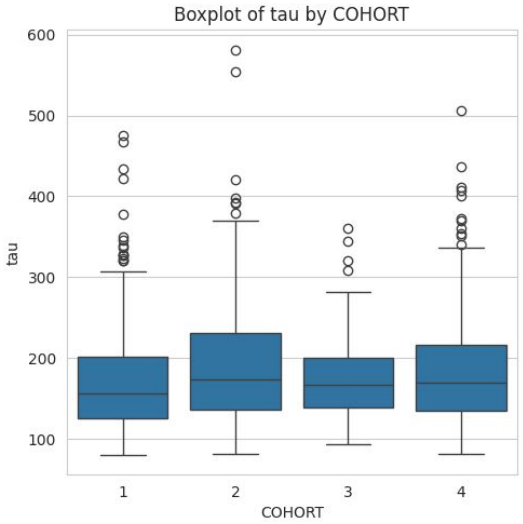
# Data Distribution

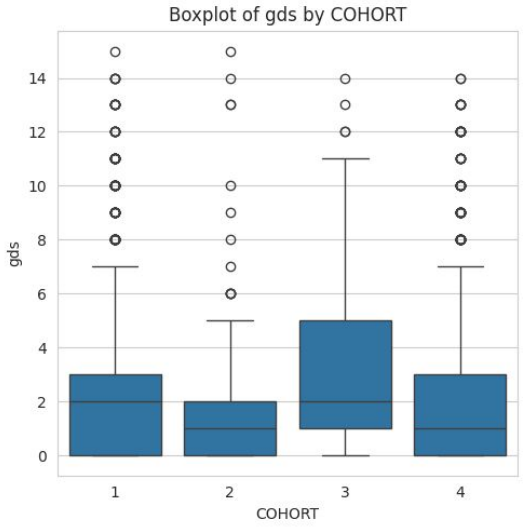


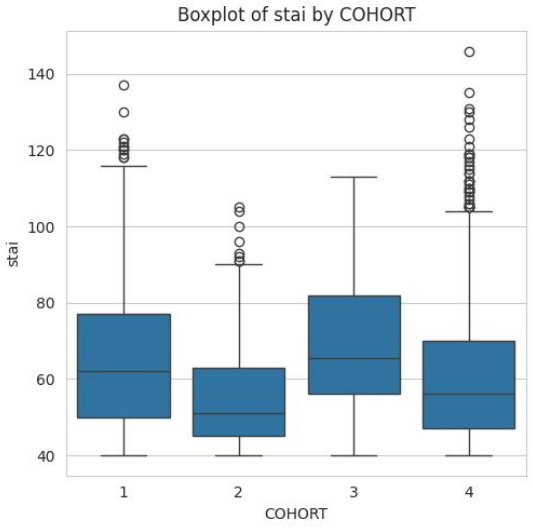


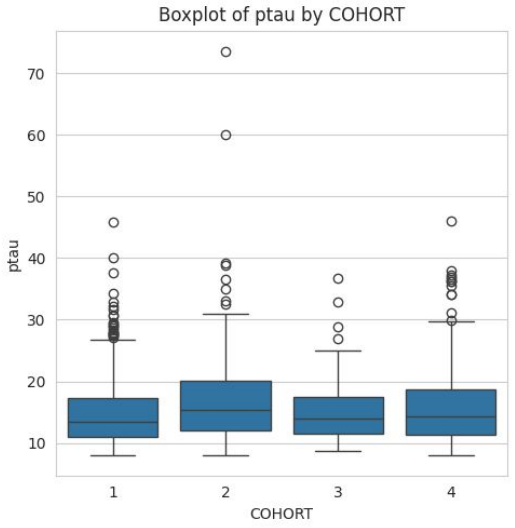


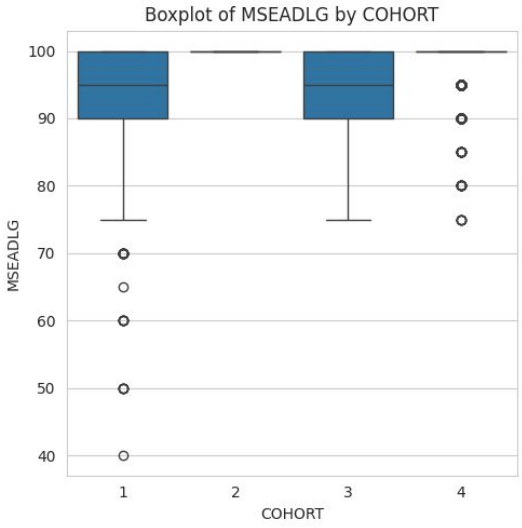


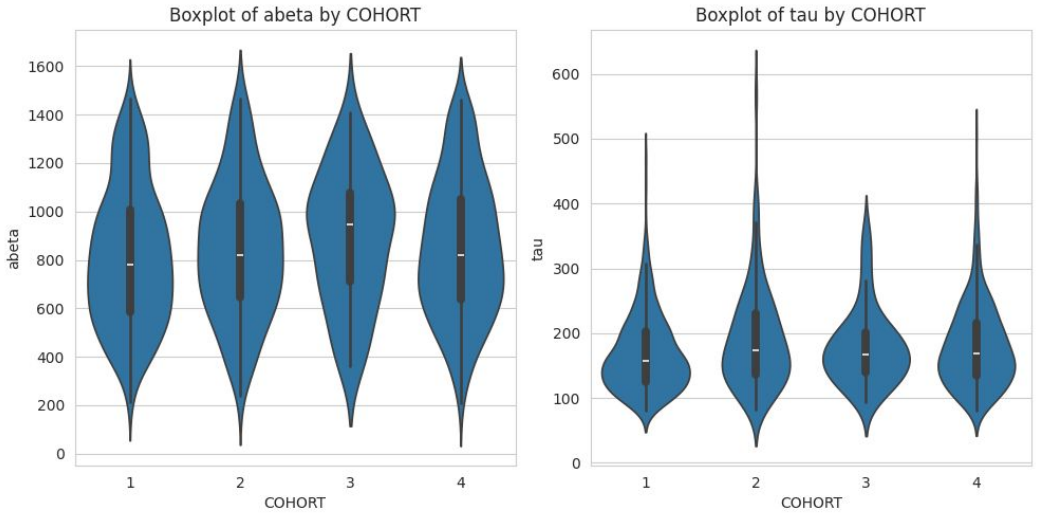



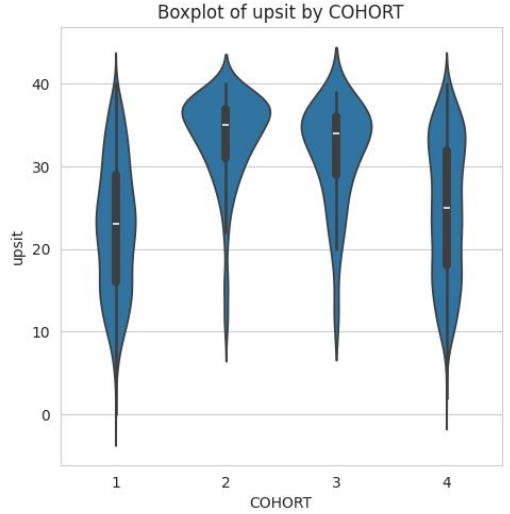



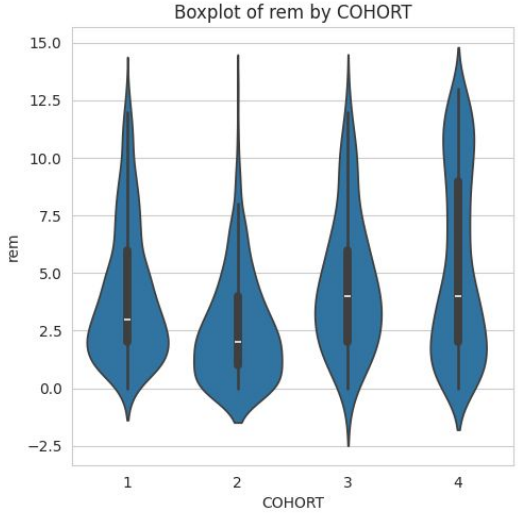
# Data Distribution

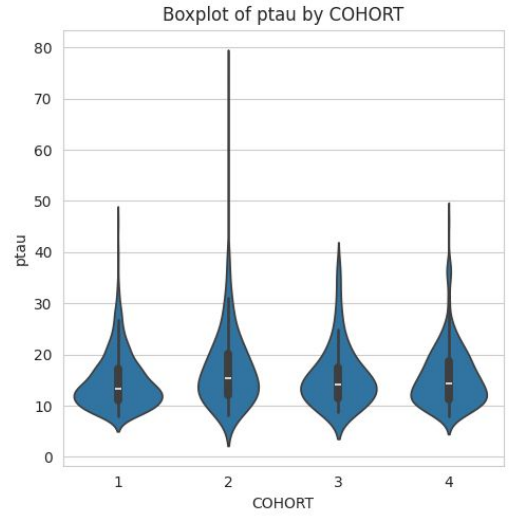


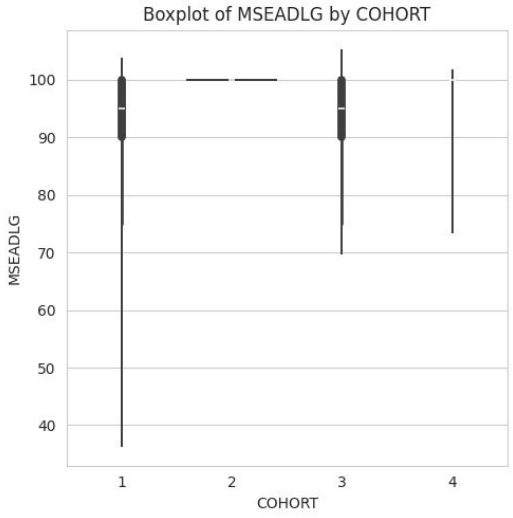


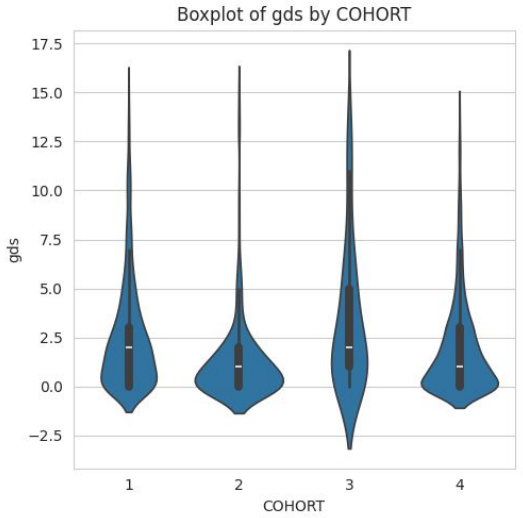


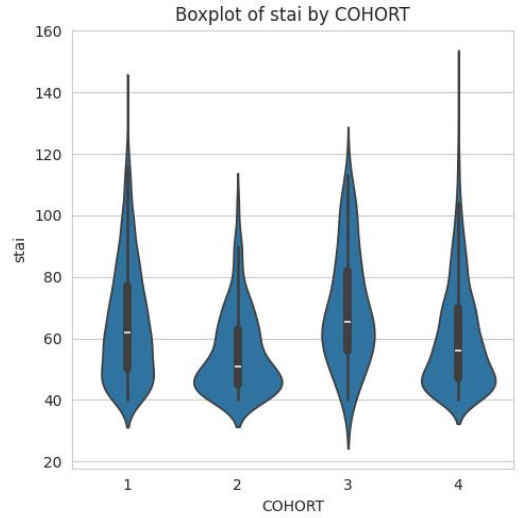


# Data Preprocessing

01

- Removed features with over **50% missing values.**
- **Imputed missing values** for numerical features using the mean and for categorical features using the mode.

**Handling Missing Data**

02

- Applied Recursive Feature Elimination (RFE) to **reduce features from 158 to 44**.
- Selected features include combination of motor, non-motor and genetic factors (e.g., REM, UPSIT, Tau, Abeta, MSEADLG).

**Feature Selection**

03

- **Initially a four-class problem** (PD, Prodromal, Healthy Controls, SWEDD).
- Simplified into a **binary classification t**ask by merging PD and Prodromal cases, and excluding SWEDD.

**Target Variable**

# Data Preprocessing

04

- Observed imbalance between PD/Prodromal cases and Healthy Controls.
- Addressed using SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes.

**Class Imbalance**

05

- Applied **standard scaling** to normalize the feature values for models like logistic regression, which are sensitive to data scale

**Feature Scaling**

# Methodology & Models

**Objective**: Predict Parkinson's disease prediction using clinical and biomarker data, **focusing on non-motor symptoms.**

**Data Collection**: Conducted literature survey and looked through multiple datasets online using tools like Google Scholar.

**Data Preprocessing**: Handling missing data, feature selection, Standard scaling for features

- **Feature Selection**: Used Recursive Feature Elimination (RFE) to identify critical features like UPSIT, REM sleep disorder, Abeta, Tau, Ptau, MSEADLG, age, and family history (FAMPD).
- **Class Imbalance Handling**: Applied random oversampling and undersampling to balance the dataset.
- **Data Split**: 80:20 split for training and testing.
- **Data Scaling**: Standard scaling applied for consistency in model performance.

**Models Used**: Evaluated Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine, Ensemble of Decision Trees.

# Results & Analysis

Table 1. Performance metrics for different models

| Metric | LR | NB | RF | SVM |
|---|---|---|---|---|
| **Train Accuracy (%)** | 85.80 | 91.06 | 100.00 | 95.58 |
| **Test Accuracy (%)** | 82.10 | 85.97 | 96.59 | 90.23 |
| **Train Sensitivity (%)** | 81.76 | 84.59 | 100.00 | 92.26 |
| **Test Sensitivity (%)** | 81.17 | 84.42 | 98.71 | 90.78 |
| **Train Specificity (%)** | 89.85 | 97.53 | 100.00 | 98.90 |
| **Test Specificity (%)** | 87.88 | 95.59 | 83.47 | 86.78 |
| **Train AUC (%)** | 92.13 | 96.03 | 100.00 | 98.90 |
| **Test AUC (%)** | 91.71 | 94.38 | 99.09 | 95.96 |

# Results Analysis Slide 1: Binary Classification Performance

- **Objective**: Improve generalization and combat overfitting.
- **Support Vector Machine (SVM):**
  - **Best Parameters**: C=10,γ=0.001,kernel=linearC = 10, \gamma = 0.001, \text{kernel} = \text{linear}C=10,γ=0.001,kernel=linear
  - **Performance**:
    - **Training Set**: Accuracy = 93.44%, Sensitivity = 91.56%, Specificity = 95.31%, ROC AUC = 97.29%
    - **Testing Set**: Accuracy = 92.50%, Sensitivity = 91.25%, Specificity = 93.75%, ROC AUC = 97.89%
  - **Observation**: Balanced metrics across training and testing sets; overfitting mitigated.
- **Random Forest (RF):**
  - **Best Parameters**: n_estimators=500,max_depth=None,class_weight=balanced_estimators = 500, \max_depth = \text{None}, \text{class_weight} = \text{balanced}n_estimators=500,max_depth=None,class_weight=balanced
  - **Performance**:
    - Training Accuracy: 95.46%
    - Testing Accuracy: 92.50%
  - **Observation**: Generalization improved; no significant overfitting.

| Metric | Training Set | Testing Set |
|---|---|---|
| Sensitivity | 91.56% | 91.25% |
| Specificity | 95.31% | 93.75% |
| ROC AUC | 97.29% | 97.89% |
| Accuracy | 93.44% | 92.50% |

Table 2. Metrics for the Best SVM Model after Hyperparameter Tuning

# Results Analysis Slide 2: Multiclass Classification Performance

- **Objective**: Classify data into three categories for Parkinson's diagnosis.
- **XGBoost**:
  - **Best Parameters**: booster=gbtree,learning rate=0.05,max depth=20booster = \text{gbtree}, \text{learning rate} = 0.05, \text{max depth} = 20booster=gbtree,learning rate=0.05,max depth=20
  - **Performance**:
    - **Training Set**: Accuracy = 97.29%, F1-Score = 98.20%, Sensitivity = 98.20%
    - **Testing Set**: Accuracy = 86.25%, F1-Score = 96.50%, Sensitivity = 96.70%
  - **Observation**: Lower test accuracy due to the complexity of multiclass classification.
- **Random Forest**:
  - **Performance**:
    - Testing Accuracy = 82.92%
  - **Observation**: XGBoost outperformed Random Forest in test accuracy for multiclass classification.

| Metric | Training Set | Testing Set |
|---|---|---|
| Sensitivity | 98.20% | 96.70% |
| Specificity | 97.90% | 94.20% |
| F1-Score | 98.20% | 96.50% |
| Accuracy | 97.29% | 86.25% |

Table 3. Metrics for the Best XGBoost Model for Multiclass Classification

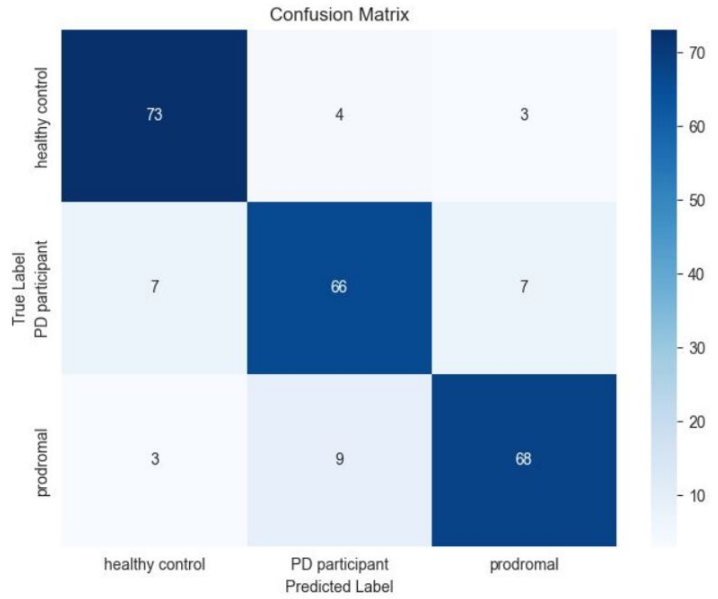



Figure 4. Confusion matrix for XGBoost on test data

# Conclusion & Learnings

- **Hyperparameter Tuning**: GridSearchCV significantly improved generalization and mitigated overfitting in SVM and RF models.
- **Binary Classification**:
  - SVM and RF models achieved comparable and high accuracy, with balanced metrics across training and testing.
- **Multiclass Classification**:
  - XGBoost outperformed RF but showed lower accuracy compared to binary classification due to complexity.

**Challenges**:

- Class imbalance impacted model performance.
- Generalization remains a critical challenge in predictive modeling for healthcare data.

**Lessons Learned**:

- Preprocessing is vital in addressing class imbalance and improving model performance.
- Hyperparameter tuning enhances model robustness.

# Contributions

- Vaibhav Singh: Data Collection, Literature Survey
- Vansh Yadav: Feature Selection, Model Training
- Utkarsh Dhilliwal: Model Training, Data Collection
- Shamik Sinha: Feature Selection, Literature Survey

# Thank You!

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
**DELHI**