Predictive Modeling for Parkinson's Disease Diagnosis Using Machine Learning : CSE343/ECE363 Mid-Semester Project Report

Vansh Yadav
2022559
vansh22559@iiitd.ac.in

Utkarsh Dhilliwal
2022551
utkarsh22551@iiitd.ac.in

Shamik Sinha
2022468
shamik22468@iiitd.ac.in

Vaibhav Singh
2022555
vaibhav22555@iiitd.ac.in

Abstract

In this report, we undertake a comprehensive literature review focusing on machine learning applications for diagnosing Parkinson's Disease. We provide detailed insights into our dataset through Exploratory Data Analysis (EDA), highlighting key characteristics and trends. Additionally, we meticulously document our data preprocessing steps and feature engineering techniques aimed at optimizing model performance. Following the initial timeline outlined in our project proposal, we present our preliminary model training efforts.

1. Motivation

Early diagnosis of Parkinson's Disease (PD) is critical as significant neuronal damage occurs before clinical symptoms manifest, often compromising treatment efficacy. With over 60% of dopaminergic neurons potentially lost by symptom onset, timely intervention is essential to mitigate disease progression and enhance patient outcomes [2][3]. Moreover, three out of four team members are currently undertaking coursework in Cognition of Motor Movements (PSY308), deepening our understanding of PD aetiology and treatment challenges and further motivating our interest in advancing diagnostic methodologies.

2. Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disorder characterized by the loss of dopaminergic neurons in the brain. Dopamine plays a vital role in controlling, adapting, and ensuring the fluency of movements. When 60-80% of these neurons are compromised, the resultant deficit in dopamine production leads to the emergence of motor symptoms associated with Parkinson's Disease. Importantly, research suggests that the disease can initiate many years prior to the onset of noticeable motor-related symptoms, making early diagnosis critical for effective intervention and management[6].

Therefore our problem statement description: Developing machine learning techniques for the early detection of Parkinson's Disease in at-risk populations to improve diagnostic accuracy and patient outcomes, addressing the limitations of traditional clinical evaluations.

3. Literature Survey

Several studies have explored various features for distinguishing between early-stage Parkinson's disease (PD) and healthy subjects. For instance, [1] highlights the use of machine learning techniques in telemedicine to detect PD in its early stages. The authors conducted research on the MDVP audio data from 30 people with PD and healthy individuals while training four ML models: Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression. They found that the Random Forest classifier emerged as the most effective machine learning technique for PD detection, achieving a detection accuracy of 91.83% and sensitivity of 0.95. However, the dataset comprised only 31 participants, which raises concerns regarding the generalizability of their findings. Furthermore, the authors acknowledge that they utilized solely voice data, which may not capture the full spectrum of motor and non-motor symptoms associated with PD. Multimodal features could be used to enhance the robustness and accuracy of PD detection models.

On the other hand, [7] performed a successful diagnosis of Parkinson's disease (PD) using Principal Component Analysis (PCA) for dimensionality reduction, Fisher Discriminant Ratio (FDR) for feature selection, and Support Vector Machine (SVM) for classification. While their ap-

proach yielded very high classification accuracies, it involved using over 100 features extracted from brain MRI images, leading to high computational costs

Another noteworthy study is by [5], which used non-motor features such as REM sleep behaviour disorder (RBD) and olfactory loss, along with cerebrospinal fluid (CSF) and imaging markers from the PPMI database, to classify early PD in 401 subjects versus 183 healthy controls. The SVM classifier achieved the best performance with 96.40% accuracy, indicating that a combination of these multimodal features may enhance the preclinical diagnosis of PD.

4. Dataset

4.1. Dataset Description

The dataset for this study was sourced from the Parkinson's Progression Markers Initiative (PPMI) database (http://www.ppmi-info.org/data), a large-scale, international study aimed at identifying biomarkers for Parkinson's disease (PD) progression. Specifically, we used a curated version of the PPMI dataset, where data from multiple tables were merged into a single comprehensive table for ease of analysis. The dataset consists of over 13,000 records collected from 3,096 participants, which include 973 with sporadic PD, 763 with PD and major genetic factors, 1,018 prodromal cases (hyposmia/RBD), 279 healthy controls, and 63 participants from the SWEDD (Scan Without Evidence of Dopaminergic Deficit) cohort [4].

The dataset contains 158 features that capture both motor and non-motor symptoms. Unlike many previous studies that primarily focused on motor symptoms, this study emphasizes non-motor features present during the premotor stage of PD, such as REM sleep behavior disorder (RBD), olfactory loss, and cognitive and behavioral test results. These features hold potential for early detection of Parkinson's disease, even before the onset of classical motor symptoms.

4.2. Data Preprocessing & Feature Selection

To prepare the dataset for analysis, several preprocessing steps were applied. First, features with more than 50% missing values were removed, as they were unlikely to contribute valuable predictive power. Descriptive statistics were used to gain insight into the dataset's numerical features, including mean, standard deviation, and quartile values. Next, the dataset was inspected to identify which features were numerical and which were categorical. Missing values in numerical columns were imputed using the mean, while categorical columns were filled using the mode. Since all categorical features were already represented numerically, label encoding was not required.

For feature selection, Recursive Feature Elimination (RFE) was used to identify the most predictive features. Features such as handedness and education level, which did not significantly contribute to the model's performance, were removed, reducing the number of features from 158 to 44. The target variable initially contained four classes: Parkinson's Disease (PD), Prodromal (at-risk individuals), Healthy Controls, and SWEDD. To simplify the problem into a binary classification task, the SWEDD class was excluded, and the PD and Prodromal classes were combined. The data was then split into training and testing sets using an 80:20 ratio.

During visualization, an imbalance was observed between the PD/Prodromal cases and the Healthy Control group. To address this issue, SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the classes. Finally, standard scaling was performed to normalize the feature values, ensuring compatibility with models like logistic regression, which are sensitive to data scale.
Some of the features used are:

- **UPSIT (University of Pennsylvania Smell Identification Test)**: Olfactory dysfunction is a well-documented early sign of Parkinson's disease. A reduced sense of smell often appears before motor symptoms, making UPSIT a vital feature in early diagnosis.
- **REM Sleep Behavior Disorder (REM)**: REM sleep behavior disorder is a common non-motor symptom and strong early indicator of PD. This feature helps in identifying patients in the prodromal stage of PD, years before motor symptoms develop.
- **Abeta (Amyloid Beta)**: Monitoring levels of Abeta in cerebrospinal fluid (CSF) can help identify patients with overlapping pathologies, providing insights into the cognitive aspects of PD.
- **Tau (Total Tau Protein)**: Total tau protein levels are indicative of neuronal damage and degeneration, and their elevation in CSF has been observed in patients with PD.
- **Ptau (Phosphorylated Tau Protein)**: Elevated Ptau levels can suggest the progression of cognitive decline in PD and help distinguish PD from other neurodegenerative disorders.
- **MSEADLG (Modified Schwab & England ADL Score)**: The MSEADLG score assesses the impact of cognitive decline on daily activities, helping to gauge the severity of PD.
- **Age**: Age is a significant risk factor for developing Parkinson's disease, with prevalence increasing with advancing age.

- **Family History of Parkinson's Disease (FAMPD)**: A family history of Parkinson's disease increases the likelihood of developing the condition, providing valuable information regarding hereditary factors.

4.3. Visualisations

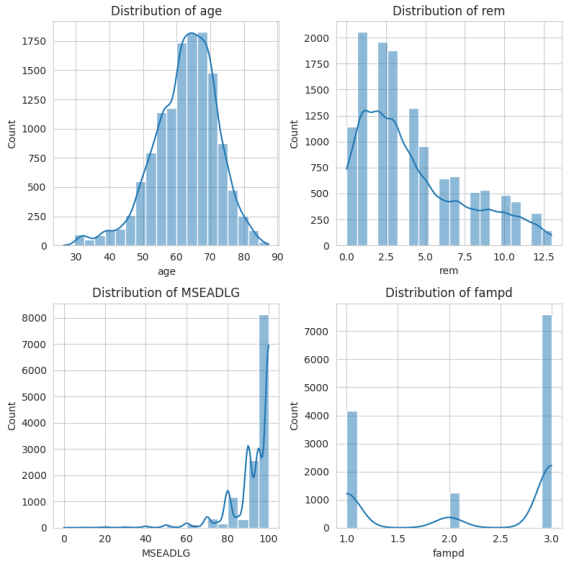

Figure 1. Density Distribution of Different Feautures

The Figure 1. shows the distribution of four features: age, rem , MSEADLG, and fampd. "rem" has right-skewed distributions, while "MSEADLG" and "fampd" have left-skewed distributions.

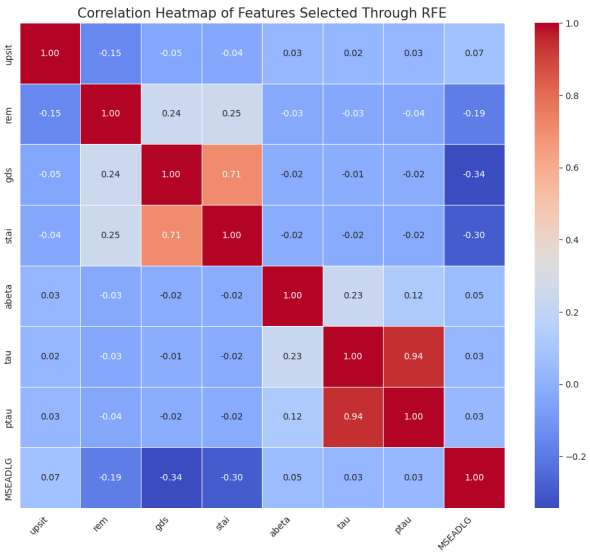

Figure 2. Correlation Heatmap of Features

The Figure 2. shows strong positive correlations between MSEADLG and tau/ptau, and negative correlations between MSEADLG and upsit/rem.

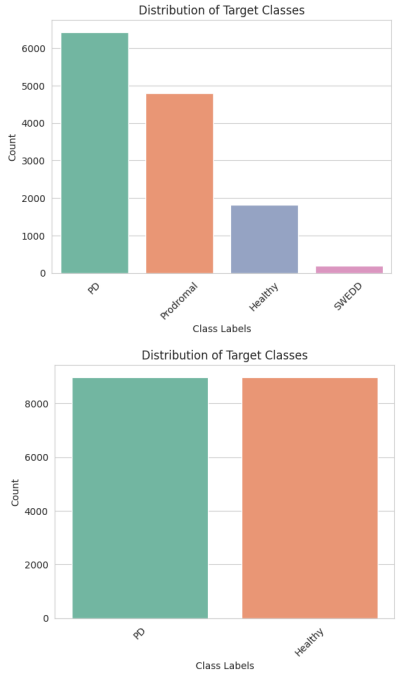

Figure 3. Target Class Distribution

The Figure 3. shows distribution of target classes before and after preprocessing. There is a reduction in class imbalances, with a more even distribution of samples across the classes.

5. Methodology and Models

The methodology followed in this project is described as follows: We utilized a dataset from the Parkinson's Progression Markers Initiative (PPMI), containing over 13,000 records and 158 features related to motor and non-motor symptoms of Parkinson's Disease (PD). After preprocessing, including handling missing values, performing feature selection with Recursive Feature Elimination (RFE), and applying SMOTE to balance classes, four machine learning models have been employed: Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine (SVM). These models have been trained and evaluated on the processed data, with Random Forest achieving the best overall performance in terms of accuracy (96.59%) and sensitivity (98.71%), though it has shown some signs of overfitting. The methodology has focused on early PD detection using multimodal data while balancing performance across key metrics.

A brief description of the models we utilised:

5.1. Logistic Regression

Logistic Regression is effective for binary classification, estimating the probability of class membership.

5.2. Naive Bayes

Naive Bayes assumes feature independence given the class label. Its simplicity allows for robust performance, especially with categorical data.

5.3. Random Forest

Random Forest builds multiple decision trees and outputs the majority vote of their predictions.

5.4. Support Vector Machine

SVM is effective for datasets with non-linear separation margins.

6. Results and 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

Accuracy: Among the models, Random Forest achieved the highest test accuracy (96.59%), followed by SVM (90.23%), Naive Bayes (85.97%), and Logistic Regression (82.10%). Random Forest generalized best to the testing data, while Logistic Regression exhibited the lowest test accuracy. The training accuracy for Random Forest was also perfect (100%), which suggests potential overfitting.

Sensitivity: Sensitivity, or the true positive rate, was highest for Random Forest on both the training (100%) and testing sets (98.71%). This makes Random Forest particularly effective at identifying positive cases in the dataset. SVM also showed strong sensitivity on the test set (90.78%), while Logistic Regression and Naive Bayes had similar performance, with test sensitivities of 81.17% and 84.42%, respectively.

Specificity: In contrast, Random Forest demonstrated lower specificity (83.47%) on the test set, despite having a perfect training specificity (100%). This trade-off between sensitivity and specificity suggests that Random Forest may have a slight bias toward positive predictions, which could explain the lower specificity. SVM and Naive Bayes performed reasonably well in terms of test specificity (86.78% and 95.59%, respectively), while Logistic Regression had the lowest test specificity (87.88

AUC: Random Forest achieved the highest test AUC (99.09%), followed by SVM (95.96%) and Naive Bayes (94.38%). Logistic Regression had the lowest test AUC (91.71%). Random Forest's high score suggests it strikes a good balance between sensitivity and specificity in distinguishing between classes.

Overfitting Considerations: Random Forest and Naive Bayes exhibited perfect or near-perfect performance on the training data, which suggests overfitting. On the other hand, Logistic Regression, may require further tuning to improve generalization.

Model Comparison: Random Forest stands out as the top performer, particularly in terms of accuracy, sensitivity, and AUC, making it the most reliable model for this dataset. SVM also shows strong performance across metrics, providing a balanced trade-off between sensitivity and specificity. Naive Bayes, while slightly lower in accuracy, demonstrates high specificity and AUC, making it suitable for applications where false positives are more critical to minimize. Logistic Regression, with lower test metrics across the board, may not be the best choice for this dataset without further adjustments.

7. Conclusion

7.1. Learnings and Future Work

This project has provided us valuable insights into machine learning models for predicting Parkinson's Disease. By comparing Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine (SVM), I gained a deeper understanding of key performance metrics such as accuracy, sensitivity, specificity, and AUC. A significant learning point was the importance of data preprocessing techniques, including feature selection, handling missing values, and scaling methods, particularly for Logistic Regression. We effectively addressed class imbalance using SMOTE on the training set, ensuring reliable results.

Moving forward, enhancing sensitivity and specificity will be a priority, along with optimizing the most promising model through **hyperparameter tuning**. Expanding the analysis to **multiclass classification** will add complexity and broaden the models' applicability. Additionally, **implementing one-hot encoding** and further feature engineering will strengthen model robustness. These efforts aim to improve the reliability of predictive tools for Parkinson's Disease.

7.2. Member Contribution

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

References

[1] R. Alshammri, G. Alharbi, E. Alharbi, and I. Almubark. Machine learning approaches to identify parkinson's disease using voice signal features. *Frontiers in Artificial Intelligence*, 6:1084001, 2023. 1

[2] J. L. Cummings, C. Henchcliffe, S. Schaier, T. Simuni, A. Waxman, and P. Kemp. The role of dopaminergic imaging in patients with symptoms of dopaminergic system neurodegeneration. *Brain*, 134(11):3146–3166, Nov 2011. 1

[3] S. Fahn. Description of Parkinson's disease as a clinical syndrome. *Annals of the New York Academy of Sciences*, 991(1):1–14, 2003. 1

[4] K. Marek et al. The parkinson progression marker initiative (ppmi). *Progress in Neurobiology*, 95(4):629–635, 2011. 2

[5] R. Prashanth, S. Dutta Roy, P.K. Mandal, and S. Ghosh. High-accuracy detection of early parkinson's disease through multimodal features and machine learning. *International Journal of Medical Informatics*, 90:13–21, 2016. 2

[6] Z. Karapinar Senturk. Early diagnosis of parkinson's disease using machine learning algorithms. *Medical Hypotheses*, 138:109603, May 2020. 1

[7] G. Singh, M. Vadera, L. Samavedham, and E.C.H. Lim. Machine learning-based framework for multi-class diagnosis of neurodegenerative diseases: a study on parkinson's disease. *IFAC-Papers OnLine*, 49(7):990–995, 2016. 1