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. Link of GitHub Repo: ParkinsonX

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].

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 is as follows: 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 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 approach 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. Some of the features used were:

- **Age:** A major risk factor, with prevalence increasing with age.
- **Family History (FAMPD):** Indicates hereditary risk for Parkinson's disease.
- **UPSIT:** Olfactory dysfunction is an early marker of PD, detected via smell tests.
- **REM Sleep Disorder:** Strong early indicator, identifying prodromal PD years before motor symptoms.
- **Total Tau Protein (Tau):** Reflects neuronal degeneration, elevated in PD.
- **Phosphorylated Tau (Ptau):** Associated with cognitive decline, aids in distinguishing PD from other disorders.
- **MSEADLG:** Assesses daily activity impact, indicating PD severity.
- **MoCA:** Montreal Cognitive Assessment, a rapid screening tool for mild cognitive dysfunction.

4.2. Visualisations

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

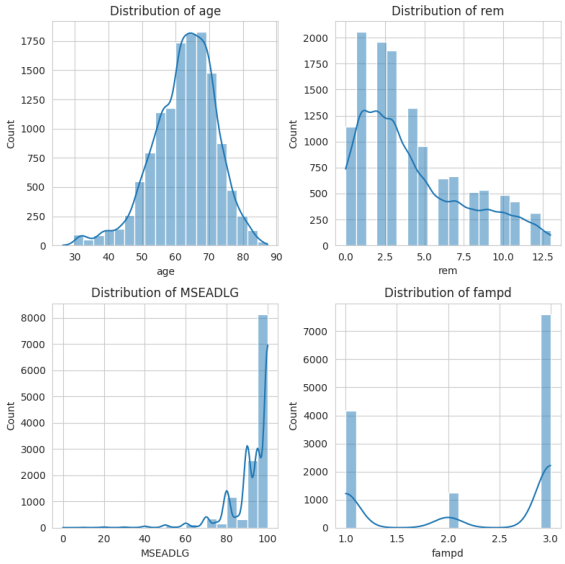

Figure 1. Density Distribution of Different Feautures

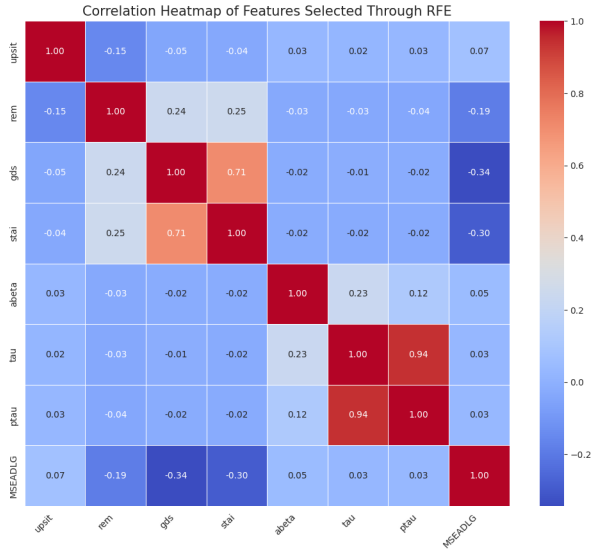

Figure 2. Correlation Heatmap of Features

distributions, while "MSEADLG" and "fampd" have left-skewed distributions.

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

The Figure 3. shows distribution of target classes before and after preprocessing.

4.3. Data Preprocessing & Feature Selection

Several preprocessing steps were applied to prepare the dataset for analysis. First, features with more than 50% missing values were removed, as they were unlikely to contribute valuable predictive power. Next, the dataset

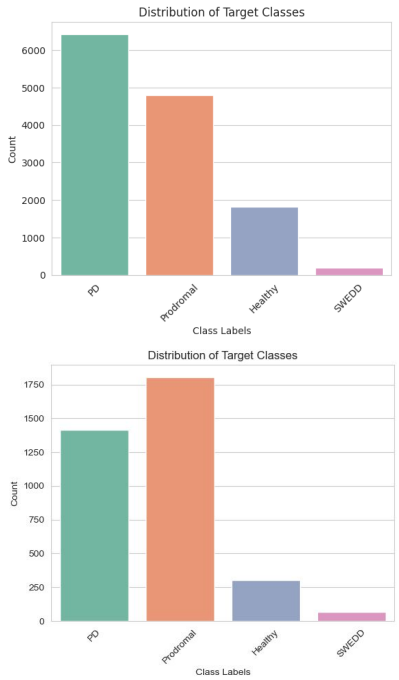

Figure 3. Target Class Distribution

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 38. 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 earlier, but since the number of healthy samples was quite less compared to the prodromal and PD classes, we randomly oversampled it to 400 and undersampled the other two to 400.

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 random oversampling and undersampling to balance classes, four machine learning models were employed initially: Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine (SVM). We then focused on enhancing initial results through hyperpaprameter tuning of SVM and RF models. Further, we also expanded to multiclass classification, and used random forests and ensemble of decision trees for the same. A brief description of the models we utilised and the results we got is given below.

Table 1. Performance metrics for different models without hyperparameter tuning. As can be

Metric	LR	NB	RF	SVM
Train Acc (%)	85.80	91.06	100.00	95.58
Test Acc (%)	82.10	85.97	96.59	90.23
Train Sens (%)	81.76	84.59	100.00	92.26
Test Sens (%)	81.17	84.42	98.71	90.78
Train Spec (%)	89.85	97.53	100.00	98.90
Test Spec (%)	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

6. Results and Analysis

6.1. Initial Results: Models Without Hyperparameter Tuning

The initial results from the models trained without hyperparameter tuning, including Support Vector Machine (SVM) and Random Forest (RF), exhibited issues of overfitting. The models performed well on the training set, but their performance on the testing set was significantly lower, indicating poor generalization.

For instance, the SVM model achieved an accuracy of 95.58% on the training data, but the test accuracy dropped to 90.23%. Similarly, the Random Forest model showed a training accuracy of 100%. These results highlighted the model's tendency to overfit, capturing noise and specific patterns from the training data rather than learning generalized features applicable to new, unseen data.

6.2. SVM and RF Hyperparameter Tuning: Combating Overfitting

To address overfitting and improve the generalization of the models, we applied hyperparameter tuning using GridSearchCV. This focused approach significantly enhanced the models' performance, especially on the test data.

For the SVM model, hyperparameter tuning revealed the Best Parameters: $\{$ C = 10, $\gamma = 0.001$, kernel = linear$\}$.

The tuning process resulted in increase in accuracy, with the best cross-validation accuracy reaching 89.84% and test accuracy improving to 92.50%. Moreover, the metrics for sensitivity, specificity, ROC AUC, and accuracy showed balanced performance across both training and testing data, suggesting that the model was no longer overfitting. The following table summarizes the performance metrics for the best SVM model:

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

The Random Forest (RF) model, when tuned, provided the best parameters: Best Parameters: $\{$ n_estimators = 500, max_depth = None, min_samples_split = 2, min_samples_leaf = 1, class_weight = balanced $\}$.

The optimized model achieved an accuracy of 95.46% on the training data, with a test accuracy of 92.50%. This demonstrated the model's improved ability to generalize and avoid overfitting.

These results show that hyperparameter tuning significantly improved the performance of the SVM model, while reducing overfitting and providing more balanced metrics across training and testing data for both.

6.3. Multiclass Classification Using XGBoost

For the multiclass classification problem, we utilized XGBoost with the best parameters obtained through GridSearchCV:

Best Parameters: $\{$ booster = gbtree, colsample_bytree = 0.6, learning_rate = 0.05, max_depth = 20, n_estimators = 100, objective = multi:softmax $\}$

Despite tuning, the XGBoost model achieved a lower test accuracy of 86.25% compared to the binary classification case. In multiclass classification, the complexity of distinguishing between multiple classes (3 in this case) generally requires more intricate models and often results in lower accuracy compared to binary classification. Complex decision boundaries, make it more challenging to achieve high accuracy.

The Random Forest model, when applied to the multiclass problem, achieved a test accuracy of 82.92%. XGBoost outperformed Random Forest in terms of test accuracy for the multiclass classification, both models showed a lower accuracy compared to the binary classification results.

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

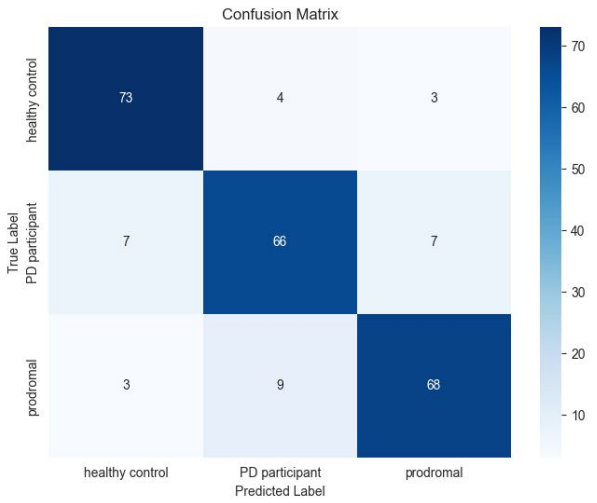

Figure 4. Confusion matrix for XGBoost on test data

7. Conclusion

From this project, we have been able to learn how diagnosis of Parkinson's disease can be done using machine learning. Reviewing different approaches and their parameters, we mentioned a potential problem of overfitting on data and the necessity to handle it. As much as there were some challenges that affect the machine learning models such as class imbalance, model generalization we had a great achievement towards improvement of the models. It added to our understanding of the issues in health care data on the predictive modeling and how preprocessing is crucial. In future work, the authors suggest that better tuning of these parameters and the use of more sophisticated algorithms might reveal better diagnostic results, which could aid in early diagnosis of Parkinson's disease.

7.1. Team Contributions

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. 1

[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