Explainable AI for Parkinson’s Disease Prediction: A Machine Learning Approach with Interpretable Models

Adebimpe O. Esan , David B. Olawade , Afeez A. Soladoye , Bolaji A. Omodunbi , Ibrahim A. Adeyanju , Nicholas Aderinto

PII:	S2452-3186(25)00050-9
DOI:	https://doi.org/10.1016/j.retram.2025.103541
Reference:	RETRAM 103541
To appear in:	*Current Research in Translational Medicine*
Received date:	27 March 2025
Accepted date:	2 September 2025

Please cite this article as: Adebimpe O. Esan , David B. Olawade , Afeez A. Soladoye , Bolaji A. Omodunbi , Ibrahim A. Adeyanju , Nicholas Aderinto , Explainable AI for Parkinson’s Disease Prediction: A Machine Learning Approach with Interpretable Models, *Current Research in Translational Medicine* (2025), doi: https://doi.org/10.1016/j.retram.2025.103541

HIGHLIGHT

- Explainable AI enhances clinical trust in Parkinson’s predictive models.
- Random Forest achieves best predictive performance for Parkinson’s Disease.
- SHAP clarifies global importance of PD features like UPDRS and cognition.
- LIME provides clear, patient-specific explanations of model predictions.
- Functional assessment scores strongly influence PD model predictions.

Explainable AI for Parkinson's Disease Prediction: A Machine Learning Approach with Interpretable Models

Adebimpe O. Esan[a], David B. Olawade [b,c,d,e*], Afeez A. Soladoye[a], Bolaji A. Omodunbi[a], Ibrahim A. Adeyanju[a], Nicholas Aderinto[f]

[a]Department of Computer Engineering, Federal University, Oye-Ekiti, Nigeria.

[b]Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom.

[c]Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham ME7 5NY, United Kingdom

[d]Department of Public Health, York St John University, London, United Kingdom.

[e]School of Health and Care Management, Arden University, Arden House, Middlemarch Park, Coventry, CV3 4FJ, United Kingdom

[f]Department of Medicine and Surgery, Ladoke Akintola University of Technology, Ogbomoso, Nigeria

[*]Corresponding author:

Email address: d.olawade@uel.ac.uk (David B. Olawade)

Abstract

Background: Parkinson's Disease (PD) is a chronic, progressive neurological disorder with significant clinical and economic impacts globally. Early and accurate prediction remains challenging with traditional diagnostic methods due to subjectivity, delayed diagnosis, and variability. Machine Learning (ML) approaches offer potential solutions, yet their clinical adoption is hindered by limited interpretability. This study aimed to develop an interpretable ML model for early and accurate PD prediction using comprehensive multimodal datasets and Explainable Artificial Intelligence (XAI) techniques.

Methods: The study applied five ML algorithms: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), XGBoost, and a stacked ensemble method to a publicly available dataset (n=2105) from Kaggle. Data encompassed demographic, medical history, lifestyle, clinical symptoms, cognitive, and functional assessments with specific inclusion/exclusion criteria applied. Preprocessing involved normalization, Synthetic Minority Oversampling Technique (SMOTE), and Sequential Backward Elimination (SBE) for feature selection. Model performance was evaluated via accuracy, precision, recall, F1-score, and Area Under Curve (AUC). The best-performing model (RF with feature selection) was interpreted using SHAP and LIME methods.

Results: Random Forest combined with Backward Elimination Feature Selection achieved the highest predictive accuracy (93%), precision (93%), recall (93%), F1-score (93%), and AUC (0.97). SHAP and LIME analyses indicated UPDRS scores, cognitive impairment, functional assessment, and motor symptoms as primary predictors, enhancing clinical interpretability.

Conclusion: The study demonstrated the effectiveness of an interpretable RF model for accurate PD prediction. Integration of ML and XAI significantly improves clinical decision-making, diagnosis timing, and personalized patient care.

Keywords: Parkinson’s Disease; Machine Learning; Explainable Artificial Intelligence; Predictive Modeling; Clinical Decision-Making

1.0 Introduction

Parkinson's Disease (PD) is a chronic and progressive neurodegenerative disorder affecting millions globally, characterized by motor symptoms such as tremors, rigidity, bradykinesia (slowness of movement), and postural instability, alongside non-motor symptoms including cognitive impairment, mood disorders, and sleep disturbances [1]. This condition significantly impairs quality of life and places substantial burdens on healthcare systems and caregivers [2]. Early and accurate prediction of PD is critical for effective management and timely intervention, yet current diagnostic practices face significant challenges.

Despite advances in neurological assessment, current PD diagnostic methods remain suboptimal, with diagnostic accuracy as low as 70–80%, particularly in early stages [3]. Traditional diagnosis relies on clinical evaluations, patient history, and neurological examinations by specialists [4]. These methods are subjective, time-consuming, and dependent on clinician expertise, resulting in variability and risk of misdiagnosis [3]. Moreover, diagnoses often occur after significant symptom onset, limiting opportunities for early intervention [4]. These shortcomings have driven interest in machine learning (ML) as an objective, scalable diagnostic tool [5]. ML excels at analysing complex datasets, identifying patterns and biomarkers overlooked by traditional methods [6]. By integrating multimodal data, such as demographic details, medical histories, lifestyle factors, clinical assessments, cognitive tests, and symptomatology, ML offers a comprehensive approach to predict PD and track its progression [6]. This data-driven strategy supports earlier detection and personalised treatment plans.

However, many ML studies for PD prediction rely on single-modality datasets, limiting their scope [7]. For example, Grover et al. used voice recordings to predict PD accurately but excluded medical history and cognitive data [7]. Similarly, Pereira et al. leveraged gait analysis effectively but omitted lifestyle and demographic factors [8]. Afonso et al. employed wearable sensors to monitor motor symptoms, yet their small sample size and limited data diversity reduced generalizability [9]. These studies highlight ML's potential but underscore the need for broader, more inclusive datasets. Another limitation of ML models is their "black-box" nature, which obscures interpretability and hinders clinical adoption where transparency is essential [10]. Explainable AI (XAI) techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), address this by clarifying how predictions are made [11, 12]. These methods reveal the features driving outcomes, fostering trust and enhancing clinical decision-making [10].

The rationale for this study is rooted in addressing the critical limitations associated with traditional diagnostic methods for Parkinson's Disease, including subjectivity, diagnostic delays, and variability, which significantly limit early intervention opportunities and effective disease management. The novelty of this research lies in adopting a multimodal machine learning approach that integrates comprehensive datasets, encompassing demographic information, medical histories, lifestyle factors, clinical assessments, cognitive tests, and symptom profiles, offering a more holistic, accurate, and interpretable predictive model compared to previous single-modality studies.

The central research question guiding this investigation is: Can machine learning algorithms combined with explainable AI techniques develop a clinically interpretable and accurate predictive model for early Parkinson's Disease detection using multimodal patient data? Our primary hypothesis posits that machine learning models, particularly ensemble methods like Random Forest, when combined with comprehensive multimodal datasets and explainable AI techniques (SHAP and LIME), can achieve superior predictive accuracy (>90%) compared to traditional diagnostic methods whilst maintaining clinical interpretability for early PD prediction. We propose several secondary hypotheses to support our investigation. We hypothesise that multimodal datasets incorporating demographic, clinical, cognitive, and functional assessments will yield better predictive performance than single-modality approaches. Additionally, we expect that UPDRS scores and cognitive assessments will emerge as primary predictive features in explainable AI analysis. Furthermore, we anticipate that ensemble methods will outperform traditional ML algorithms for PD prediction.

The primary objective of this study is to develop an accurate, interpretable, and clinically relevant machine learning model capable of predicting PD at an early stage using diverse data sources. The specific objectives of this study encompass several key areas of investigation. We aim to collect and integrate multimodal data for PD prediction to establish a comprehensive dataset foundation. Our research will evaluate and compare predictive performances of different machine learning algorithms to identify the most effective approach for PD classification. The investigation will apply explainable AI (XAI) techniques like SHAP and LIME to enhance transparency and interpretability of the selected models, ensuring clinical applicability. Finally, we will validate the model using appropriate statistical and clinical methods to demonstrate its effectiveness and practical clinical applicability in real-world healthcare settings.

2.0 Methodology

The methodology of this study targets the development of a comprehensive and systematic framework for predicting Parkinson's Disease (PD) using machine learning (ML) approaches, integrated with Explainable AI (XAI) techniques for interpretability. The methodology is structured into five significant phases: data acquisition, preprocessing, prediction, evaluation, and explanation, each carefully designed to ensure robustness, precision, and transparency throughout the predictive modelling process as shown in Figure 1 below.

2.1 Data Acquisition and Sample Characteristics

2.1.1 Dataset Selection Rationale

The data for this study was sourced from Kaggle, a prominent platform for sharing datasets for data science and ML projects. The Kaggle dataset was chosen over the Parkinson's Progression Markers Initiative (PPMI) database for several practical and methodological reasons:

a. Comprehensive Multimodal Coverage: The Kaggle dataset provides integrated multimodal data including demographic, clinical, cognitive, lifestyle, and functional assessments in a preprocessed format, facilitating direct ML implementation.
b. Balanced Representation: Unlike PPMI which focuses primarily on early-stage PD patients, the Kaggle dataset includes both PD and control participants across various disease stages.
c. Accessibility: Open access availability enables reproducibility and comparison with other studies.
d. Feature Completeness: The dataset contains standardised clinical assessments (UPDRS, MoCA) essential for interpretable AI analysis.

2.1.2 Inclusion and Exclusion Criteria

Inclusion Criteria:

- Complete demographic information (age, gender, ethnicity)
- Available clinical assessments including UPDRS and MoCA scores
- Documented medical history including family history of PD
- Lifestyle factors (diet, exercise, smoking status)
- Complete motor and non-motor symptom profiles
- Participants aged 18-85 years

Exclusion Criteria:

- Incomplete demographic or clinical data (>20% missing values)
- Participants with other neurodegenerative disorders (Alzheimer's, multiple sclerosis)
- Insufficient symptom documentation
- Secondary parkinsonism due to medications or other causes
- Participants with severe cognitive impairment preventing reliable assessment

Final Sample: After applying inclusion/exclusion criteria, 2,105 records were retained for analysis, comprising 1,052 PD patients and 1,053 healthy controls.

2.2 Data Acquisition

The data for this study was sourced from Kaggle, a prominent platform known for sharing and retrieving datasets for data science and ML projects. This publicly available dataset, optimised specifically for PD prediction, consists of 2,105 records featuring comprehensive details, including demographic factors (age, gender, ethnicity), medical histories (family history of PD, comorbidities, medication use), lifestyle parameters (diet, exercise, smoking status), clinical measurements (motor and non-motor symptoms), cognitive and functional assessments (MoCA and UPDRS scores), and symptoms reported by both clinicians and patients. The dataset is compiled from multiple sources, such as clinical trials, wearable sensors, patient surveys, and clinician evaluations, providing a holistic representation of PD factors. Data anonymisation and open-access licensing were implemented to maintain patient confidentiality and facilitate broad research usage.

2.3 Data Preprocessing

The dataset underwent extensive preprocessing to ensure suitability for ML modelling. Variables including age, BMI, alcohol consumption, clinical parameters (e.g., Diastolic BP, Cholesterol levels), and cognitive assessments (e.g., UPDRS, MoCA scores) were normalised using Min-Max scaling to achieve uniformity within a 0-1 range. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was employed, generating synthetic data points for the minority class, thus balancing the dataset and mitigating bias. Additionally, Sequential Backward Elimination (SBE) was used for feature selection, systematically removing less significant features until the optimal subset was identified. These preprocessing steps enhanced the dataset's balance, cleanliness, and readiness for accurate and efficient ML modelling.

2.4 Machine Learning Algorithms for PD Prediction

Five ML algorithms were employed in this study: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), and XGBoost, supplemented by a stacked ensemble to leverage their combined predictive strengths. SVM was selected for its capacity to manage high-dimensional data and complex relationships using kernel functions. KNN was chosen for its simplicity and effectiveness in capturing local data patterns without assuming specific distributions. LR provided interpretability and baseline performance in binary classification tasks with primarily linear relationships. RF was adopted due to its ensemble nature, enhancing predictive accuracy and reducing overfitting, while simultaneously providing intrinsic feature importance measures. XGBoost, known for its gradient-boosting mechanism, was included for its robustness with imbalanced datasets and exceptional predictive power. A stacked ensemble integrated these models, capitalising on individual strengths to deliver superior overall performance through reduced bias and variance.

2.5 Explainable Artificial Intelligence (XAI) with SHAP and LIME

Explainable AI methods, specifically SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), were utilised to enhance interpretability of predictions, particularly from the Random Forest (RF) model, identified as the most accurate predictor among the tested algorithms. SHAP, employing a game-theoretic approach, quantified each feature's contribution globally, enabling clinicians to identify crucial predictive factors and gain deeper insights into PD mechanisms. Conversely, LIME provided local interpretability by approximating predictions for individual instances, offering clinicians detailed explanations for specific patient predictions. This combination provided comprehensive transparency, building clinical trust and facilitating model adoption in practical healthcare settings.

2.6 Performance Evaluation

The Hold-out evaluation method was employed, partitioning the dataset into a 70-30 training-testing split while ensuring class proportions were stratified for balanced representation. Model performance was quantitatively evaluated using standard metrics, including accuracy, precision, recall, and F1-score, ensuring thorough and reliable assessment of the predictive models developed in this study.

3.0 Results

This section presents the results obtained by applying various machine learning algorithms: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), XGBoost, and Stacked Ensemble, to predict Parkinson's Disease (PD). These algorithms were evaluated using metrics including accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC). Among these models, Random Forest (RF) combined with Backward Elimination Feature Selection (BEFS) outperformed all others, achieving the highest accuracy and overall predictive performance. To enhance the interpretability of the RF model's predictions, Explainable AI (XAI) techniques, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), were utilized. The results are described in detail, with interpretations through SHAP and LIME clearly outlined to emphasize clinical implications.

3.1 Experimental Results of Machine Learning Models

Table 1 summarizes the performance metrics of all evaluated models. The Random Forest (RF) model achieved the highest accuracy of 93%, with precision, recall, and F1-score also each at 93%, and an AUC of 0.97. The Stacked Ensemble and XGBoost models performed similarly, each attaining an accuracy of 92% and an AUC of 0.96. SVM and Logistic Regression demonstrated moderate performances, achieving accuracies of 84% and 83%, respectively. KNN had the lowest accuracy at 79%. RF's superior performance is largely due to its ensemble structure, combining multiple decision trees to minimize overfitting and enhance generalization. Its inherent ability to rank feature importance complements feature selection techniques, further boosting performance. Conversely, KNN's weaker performance is likely due to limitations associated with handling high-dimensional datasets and sensitivity to noise.

3.2 Interpretation of ML Predictions for Parkinson's Disease

Two prominent XAI techniques, SHAP and LIME, were employed to interpret the predictions of the best-performing RF model, enhancing transparency and clinical usability.

3.2.1 SHAP Interpretation of Random Forest Predictions

SHAP analysis provided a global understanding of how each feature influenced the RF model's predictions. The SHAP waterfall plot (Figure 2) highlighted the importance of cognitive impairment (MoCA), functional assessments, and hypertension as key predictors for PD. Notably, traditional motor symptoms such as tremor, rigidity, bradykinesia, and postural instability showed lower contributions in this analysis. Additionally, SHAP summary plots (Figures 3 and 4) identified UPDRS scores and functional assessments as the most influential features across the dataset, aligning closely with clinical knowledge about PD severity indicators. Features such as BMI, diet quality, physical activity, and comorbid conditions had comparatively lower impact, suggesting these lifestyle-related factors were secondary in prediction.

3.2.2 LIME Interpretation of Random Forest Predictions

LIME provided detailed, instance-level explanations for model predictions, clarifying the decision-making process for individual cases. Figure 5a presents the LIME interpretation for an instance correctly predicted as a PD patient, with a prediction confidence of 93%. Here, UPDRS, tremor, functional assessments, and rigidity were critical in influencing the positive prediction. Conversely, higher diet quality and physical activity slightly reduced PD probability, highlighting their minor protective influence. Figure 5b demonstrates the interpretation for an instance correctly classified as a non-PD patient, where low UPDRS scores and absence of significant tremor or rigidity strongly contributed to a negative prediction.

3.2.3 Comparison of SHAP and LIME Interpretations

Both SHAP and LIME provided complementary insights into the model's decision-making. SHAP offered a broad, dataset-wide interpretation, emphasizing the global importance of UPDRS scores and functional assessment. In contrast, LIME delivered localized explanations, confirming the importance of specific motor and cognitive symptoms for individual predictions. The convergence of findings from SHAP and LIME underscores the model's interpretability and clinical relevance, highlighting consistent utilization of clinically significant features. However, both analyses indicated limited roles for lifestyle factors, suggesting potential benefits from incorporating more detailed lifestyle or genetic data to enhance future predictive accuracy and clinical applicability.

4.0 Discussion

4.1 Performance Comparison with Existing Studies

Our Random Forest model achieved 93% accuracy, which positions it competitively within the existing literature whilst offering unique advantages. Several studies have reported varying performance levels that merit careful comparison when considering methodological differences and dataset characteristics [13].

Grover et al. achieved 94.2% accuracy using speech features alone [7], while Pereira et al. reported 95.1% accuracy with gait analysis [8]. However, direct comparison requires careful consideration of several key factors. First, dataset characteristics differ significantly - Grover et al. used a smaller, homogeneous sample (n=195) focused solely on voice recordings, whilst our study employed a larger, multimodal dataset (n=2,105) providing more comprehensive patient representation. Second, regarding feature scope, unlike single-modality approaches achieving higher accuracy on specific features, our model integrates diverse data types (demographic, clinical, cognitive, lifestyle), potentially trading peak performance for clinical applicability and interpretability.

Recent studies using multimodal deep learning approaches have demonstrated the potential for enhanced early detection capabilities. Dentamaro et al. [14] achieved 96.6% accuracy using DenseNet combined with an Excitation Network on PPMI data (n=90), focusing specifically on prodromal stage detection using multimodal deep learning with 3D MRI scans and clinical data. Their explainable AI analysis using SHAP and LIME revealed that UPDRS scores, cognitive impairment measures, and functional assessments were primary predictors, which aligns with our findings regarding the importance of clinical assessment features. Their use of joint co-learning for multimodal fusion enabled end-to-end training and learning of complementary information from both imaging and clinical modalities, demonstrating superior performance compared to single-modality approaches.

Priyadharshini et al. [15] developed a comprehensive framework using T2-weighted 3D MRI datasets (n=500) and achieved 96.8% accuracy with Gradient Boosting combined with SMOTE for data balancing. Their study extracted 107 radiomics features from subcortical regions and used a systematic feature selection approach to identify the top 20 most significant features. The integration of multiple XAI techniques (SHAP, LIME, and SHAPASH) provided both global and local explanations, with UPDRS scores and cognitive assessments emerging as primary predictive features. This finding corroborates our results regarding the critical importance of clinical assessment scales in PD prediction.

Zhang et al. [16] conducted a systematic comparison of eight machine learning algorithms using PPMI data (n=747) and achieved optimal performance with penalized logistic regression (AUC=0.94) and XGBoost (AUC=0.92). Their study demonstrated that models incorporating demographic variables, clinical assessments, and polygenic risk scores (PRS) achieved the best prediction performance without requiring invasive biomarkers. Their SHAP analysis consistently identified olfactory function (UPSIT) and polygenic risk scores as the most important predictors across different ML methods, emphasising the value of non-invasive assessment tools.

The generalisability versus specificity trade-off represents a crucial consideration when evaluating these performance differences. Studies reporting >95% accuracy often focus on specific patient populations or controlled settings, whilst our multimodal approach prioritises real-world clinical applicability across diverse patient presentations. The interpretability trade-off is equally important - higher-performing deep learning models often sacrifice interpretability, whereas our 93% accuracy comes with comprehensive XAI analysis, crucial for clinical adoption.

Furthermore, the modest performance difference (1-2%) compared to some studies is offset by significant advantages in clinical interpretability, broader applicability, and comprehensive feature integration. The convergence of findings across these studies regarding the importance of UPDRS scores, cognitive assessments, and olfactory function validates the clinical relevance of these features for early PD detection. The consistent success of ensemble methods (Random Forest, XGBoost, Gradient Boosting) across multiple studies suggests their superior suitability for PD prediction compared to traditional algorithms.

These comparative findings indicate that our approach provides a more suitable framework for practical healthcare implementation, balancing accuracy with interpretability and clinical applicability. The integration of explainable AI techniques across all compared studies demonstrates the critical importance of model transparency in medical applications, enabling clinicians to understand and trust AI-driven diagnostic decisions.

4.2 Clinical Significance and Feature Interpretation

The current study investigated the efficacy of multiple machine learning algorithms for predicting Parkinson's Disease. The Montreal Cognitive Assessment (MoCA) emerged as a significant predictor, consistent with evidence that cognitive impairment is a prevalent early non-motor symptom in PD, often detectable before motor deficits become pronounced [17]. This aligns with longitudinal studies showing cognitive decline as a marker of disease onset and progression [18].

SHAP analysis identified UPDRS and functional assessment scores as the most influential predictors, resonating with clinical consensus on UPDRS as a gold-standard measure of PD severity and progression [19]. The integration of Explainable AI techniques addressed the "black-box" challenge that often hampers clinical adoption [10]. UPDRS's dominance likely reflects its comprehensive evaluation of both motor and non-motor symptoms, offering a holistic view of patient status [1].

4.3 Algorithm Performance Analysis

Comparatively, KNN yielded the weakest results, likely due to the curse of dimensionality, where distance-based metrics become less meaningful as feature count increases [20]. LR exhibited moderate performance, possibly reflecting its reliance on linear relationships between predictors and outcomes [21]. SVM showed intermediate results, which could be due to difficulties in selecting optimal kernels and tuning hyperparameters [22]. RF's success can be attributed to its ensemble structure, which aggregates predictions from multiple decision trees, reducing overfitting and enhancing robustness against noisy or incomplete data [23]. These findings suggest that ensemble methods, particularly RF with feature selection, are better suited for PD prediction than traditional algorithms [24].

4.4 Clinical Implementation and Future Directions

From a clinical perspective, the RF model's high performance and interpretability offer substantial utility. The model's emphasis on non-motor symptoms, including cognitive dysfunction and comorbidities like hypertension, supports recent literature recognising these features as critical in early PD diagnosis [25].

Future research should prioritise several key areas to enhance the clinical applicability and accuracy of PD prediction models. Integrating genetic markers such as GBA and SNCA variants would help capture hereditary risk factors that significantly influence PD development [26]. Incorporating longitudinal data would improve progression modelling capabilities, allowing for better understanding of disease trajectory over time [27]. Expanding non-motor features including sleep patterns and autonomic dysfunction would provide a more comprehensive assessment framework aligned with current clinical understanding [25]. Finally, deploying the model in clinical trials would enable assessment of its impact on diagnostic accuracy and patient care outcomes in real-world settings [28].

The clinical translation pathway for implementing this predictive framework involves a systematic four-phase approach. Phase 1 focuses on integration with electronic health records for automated risk assessment, enabling seamless incorporation into existing clinical workflows. Phase 2 involves the development of clinician decision support tools that incorporate SHAP and LIME explanations, providing transparent and interpretable guidance for healthcare providers. Phase 3 encompasses implementation in specialist neurology clinics for early detection screening, allowing for targeted application in high-risk populations. Phase 4 involves validation across diverse healthcare settings and populations, ensuring the model's robustness and generalisability across different clinical environments and patient demographics.

5.0 Study Limitations

Despite promising results, several limitations must be acknowledged:

5.1 Dataset Limitations

Firstly, the data utilised was obtained from a publicly available source (Kaggle), potentially limiting the study due to constraints in dataset size, diversity, and feature completeness. The dataset lacks several critical elements:

a. Longitudinal Data: The cross-sectional design limits our ability to track disease progression over time, which is crucial for understanding PD trajectory and validating long-term model performance. Unlike PPMI which provides longitudinal follow-up, our dataset represents a single time-point assessment.

b. Genetic Biomarkers: The absence of genetic markers (LRRK2, GBA, SNCA mutations) limits prediction accuracy, as genetic predisposition significantly influences PD risk and progression. Modern precision medicine approaches increasingly rely on genetic profiling for personalised risk assessment.
c. Neuroimaging Data: The lack of MRI, DaTscan, or other neuroimaging biomarkers represents a significant limitation, as these provide objective measures of neurodegeneration and are increasingly used in clinical practice for PD diagnosis.
d. Environmental Factors: Limited data on environmental exposures (pesticides, heavy metals, head trauma) that contribute to PD risk may impact model comprehensiveness.

5.2 Methodological Limitations

Secondly, whilst the Random Forest model exhibited superior predictive performance, the model's reliance on backward elimination feature selection (BEFS) may have inadvertently excluded relevant predictive features due to automated selection criteria, potentially affecting comprehensiveness and interpretability.

Thirdly, the reliance on synthetic oversampling techniques (SMOTE) to address class imbalance may introduce artificial patterns or biases, potentially impacting model generalisability to real-world clinical scenarios where natural class distributions differ.

5.3 Interpretability Limitations

Additionally, although SHAP and LIME methodologies substantially enhanced interpretability, these methods provide post-hoc interpretations that are inherently limited by their approximation mechanisms. Complete transparency in understanding underlying decision-making processes may remain partially constrained.

5.4 External Validation

Finally, the absence of external validation datasets limits the ability to fully assess robustness and real-world applicability. Further validation using independent datasets from diverse clinical settings would be beneficial to confirm the model's predictive accuracy and clinical utility.

5.5 Generalisability Concerns

The dataset's demographic composition may not represent global PD populations, potentially limiting generalisability across different ethnicities, healthcare systems, and socioeconomic backgrounds. The study's focus on English-speaking populations may not translate to other linguistic and cultural contexts.

Addressing these limitations in future research would enhance model accuracy, interpretability, and applicability to clinical practice.

6.0 Conclusion and Recommendations

This study aimed to develop and evaluate predictive models for Parkinson's Disease using machine learning techniques, enhanced by Explainable Artificial Intelligence methods for improved interpretability. The primary hypothesis that machine learning models, particularly ensemble methods like Random Forest, could achieve superior predictive accuracy (>90%) whilst maintaining clinical interpretability was confirmed, with our RF model achieving 93% accuracy combined with comprehensive XAI analysis.

Key findings supporting our hypotheses include: Multimodal datasets incorporating demographic, clinical, cognitive, and functional assessments yielded superior performance compared to reported single-modality approaches in terms of clinical applicability; UPDRS scores and cognitive assessments (MoCA) emerged as primary predictive features, confirming their clinical significance; Ensemble methods (RF, XGBoost, Stacked Ensemble) consistently outperformed traditional ML algorithms.

Among the evaluated algorithms, KNN, SVM, LR, XGBoost, Stacked Ensemble, and Random Forest, the Random Forest model combined with Backward Elimination Feature Selection demonstrated the highest predictive accuracy and robustness. The utilisation of SHAP and LIME provided essential insights into model decision-making processes, highlighting the importance of clinically relevant features such as UPDRS scores, cognitive impairment, and functional assessments.

6.1 Clinical Implications

The study's findings have several important clinical implications:

a. Early Detection Potential: The 93% accuracy achieved suggests the model could serve as a valuable screening tool for early PD detection, potentially identifying at-risk patients before significant symptom onset.
b. Clinical Decision Support: The interpretable nature of the RF model, enhanced by SHAP and LIME explanations, makes it suitable for integration into clinical workflows, providing clinicians with transparent, evidence-based diagnostic support.
c. Personalised Care: The model's ability to identify individual risk factors for specific patients enables personalised treatment planning and targeted interventions.
d. Resource Optimisation: Automated preliminary screening could help prioritise patients for specialist evaluation, optimising healthcare resource allocation.

6.2 Recommendations for Future Research

Based on the findings and limitations identified, several recommendations are suggested for future research and practice:

a. Data Enhancement: Incorporate longitudinal datasets to track disease progression and validate long-term predictive accuracy; Include comprehensive genetic markers (LRRK2, GBA, SNCA variants) to capture hereditary risk factors; Integrate neuroimaging biomarkers (MRI, DaTscan) for objective neurodegeneration assessment; Expand environmental and lifestyle data collection for comprehensive risk profiling.

b. Methodological Improvements: Conduct external validation using independent datasets from varied clinical settings to enhance robustness and generalisability; Explore alternative feature selection methods to complement BEFS and potentially capture additional relevant features; Investigate ensemble methods combining traditional ML with deep learning approaches; Develop real-time learning algorithms that can adapt to new patient data.
c. Clinical Translation: Phase I: Pilot testing in specialist neurology clinics to assess clinical workflow integration; Phase II: Multi-centre validation trials across diverse healthcare settings; Phase III: Health economic evaluation to assess cost-effectiveness and clinical impact; Phase IV: Long-term implementation studies with healthcare provider training programmes.
d. Technology Development: Create user-friendly clinical interfaces displaying SHAP/LIME explanations for clinician interpretation; Develop mobile applications for point-of-care screening in primary healthcare settings; Establish continuous model updating mechanisms based on new clinical data; Implement robust data security and privacy protection measures for clinical deployment.

6.3 Broader Impact and Future Directions

This research contributes to the growing field of explainable AI in healthcare, demonstrating that high-performance machine learning models can maintain clinical interpretability. The successful integration of multimodal data with XAI techniques provides a template for similar applications in other neurodegenerative diseases.

The study's emphasis on clinical interpretability addresses a critical barrier to AI adoption in healthcare, potentially accelerating the translation of machine learning advances into routine clinical practice. Future work should focus on expanding this approach to other neurological conditions and developing comprehensive diagnostic support systems.

Ultimately, this research represents a step towards precision medicine in neurology, where data-driven approaches complement clinical expertise to improve patient outcomes through earlier detection, personalised treatment strategies, and more efficient healthcare delivery.

References

[1] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, and A. E. Lang, "Parkinson disease," Nature Reviews Disease Primers, vol. 3, p. 17013, Mar. 2017. https://doi.org/10.1038/nrdp.2017.13

[2] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," Journal of Neurology, Neurosurgery & Psychiatry, vol. 79, no. 4, pp. 368–376, Apr. 2008. https://doi.org/10.1136/jnnp.2007.131045

[3] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino, "Accuracy of clinical diagnosis of Parkinson disease: A systematic review and Bayesian meta-analysis," Neurology, vol. 86, no. 6, pp. 566–576, Feb. 2016. https://doi.org/10.1212/WNL.0000000000002350

[4] R. B. Postuma, D. Berg, M. Stern, W. Poewe, C. W. Olanow, W. Oertel, J. Obeso, K. Marek, I. Litvan, A. E. Lang, G. Halliday, C. G. Goetz, T. Gasser, B. Dubois, P. Chan, B. R. Bloem, C. H. Adler, and G. Deuschl, "MDS clinical diagnostic criteria for Parkinson's disease," Movement Disorders, vol. 30, no. 12, pp. 1591–1601, Oct. 2015. https://doi.org/10.1002/mds.26424

[5] R. Prashanth, S. D. 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, vol. 90, pp. 13–21, Jun. 2016. https://doi.org/10.1016/j.ijmedinf.2016.03.001

[6] M. D. Hssayeni, J. L. Adams, B. Ghoraani, and M. A. Burack, "Automatic assessment of medication states of patients with Parkinson's disease using wearable sensors and machine learning," Sensors, vol. 19, no. 19, p. 4122, Sep. 2019. https://doi.org/10.3390/s19194122

[7] S. Grover, S. Bhartia, A. Yadav, and K. R. Seeja, "Predicting severity of Parkinson's disease using deep learning," Procedia Computer Science, vol. 132, pp. 1788–1794, 2018. https://doi.org/10.1016/j.procs.2018.05.154

[8] C. R. Pereira, D. R. Pereira, F. A. Silva, J. P. Masieiro, S. A. T. Weber, C. Hook, and J. P. Papa, "A new computer vision-based approach to aid the diagnosis of Parkinson's disease," Computer Methods and Programs in Biomedicine, vol. 169, pp. 49–57, Feb. 2019. https://doi.org/10.1016/j.cmpb.2018.12.002

[9] L. C. Afonso, G. H. Rosa, C. R. Pereira, S. A. T. Weber, C. Hook, V. H. C. Albuquerque, and J. P. Papa, "A recurrence plot-based approach for Parkinson's disease identification," Future Generation Computer Systems, vol. 94, pp. 282–292, May 2019. https://doi.org/10.1016/j.future.2018.11.033

[10] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, Jun. 2020. https://doi.org/10.1016/j.inffus.2019.12.012

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

[13] Omodunbi BA, Olawade DB, Awe OF, Soladoye AA, Aderinto N, Ovsepian SV, Boussios S. Stacked Ensemble Learning for Classification of Parkinson's Disease Using Telemonitoring Vocal Features. Diagnostics. 2025 Jun 9;15(12):1467. https://10.3390/diagnostics15121467

[14] V. Dentamaro, D. Impedovo, L. Musti, G. Pirlo, and P. Taurisano, "Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: insights from the PPMI database," Scientific Reports, vol. 14, no. 1, p. 20941, Sep. 2024. https://doi.org/10.1038/s41598-024-70165-4

[15] S. Priyadharshini, K. Ramkumar, S. Vairavasundaram, K. Narasimhan, S. Venkatesh, R. Amirtharajan, and K. Kotecha, "A Comprehensive framework for Parkinson's disease diagnosis using explainable artificial intelligence empowered machine learning techniques," Alexandria Engineering Journal, vol. 107, pp. 568–582, Jul. 2024. https://doi.org/10.1016/j.aej.2024.07.106

[16] J. Zhang, W. Zhou, H. Yu, T. Wang, X. Wang, L. Liu, and Y. Wen, "Prediction of Parkinson's Disease Using Machine Learning Methods," Biomolecules, vol. 13, no. 12, p. 1761, Dec. 2023. https://doi.org/10.3390/biom13121761

[17] D. Aarsland, K. Bronnick, C. Williams-Gray, et al., "Mild cognitive impairment in Parkinson disease: A multicenter pooled analysis," Neurology, vol. 75, no. 12, pp. 1062–1069, Sep. 2010. https://doi.org/10.1212/WNL.0b013e3181f39d0e

[18] D. Weintraub, T. Simuni, C. Caspell-Garcia, et al., "Cognitive performance and neuropsychiatric symptoms in early, untreated Parkinson's disease," Movement Disorders, vol. 30, no. 7, pp. 919–927, Jun. 2015. https://doi.org/10.1002/mds.26170

[19] C. G. Goetz, B. C. Tilley, S. R. Shaftman, et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," Movement Disorders, vol. 23, no. 15, pp. 2129–2170, Nov. 2008. https://doi.org/10.1002/mds.22340

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, New York, NY: Springer, 2013, pp. 103–105.

[21] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed., Hoboken, NJ: Wiley, 2013, pp. 35–40.

[22] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, Sep. 1995. https://doi.org/10.1007/BF00994018

[23] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001. https://doi.org/10.1023/A:1010933404324

[24] S. M. McKinney, M. Sieniek, V. Godbole, et al., "International evaluation of an AI system for breast cancer screening," Nature, vol. 577, no. 7788, pp. 89–94, Jan. 2020. https://doi.org/10.1038/s41586-019-1799-6

[25] K. R. Chaudhuri, D. G. Healy, and A. H. Schapira, "Non-motor symptoms of Parkinson's disease: Diagnosis and management," Lancet Neurology, vol. 5, no. 3, pp. 235–245, Mar. 2006. https://doi.org/10.1016/S1474-4422(06)70373-8

[26] M. A. Nalls, N. Pankratz, C. M. Lill, et al., "Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease," Nature Genetics, vol. 46, no. 9, pp. 989–993, Sep. 2014. https://doi.org/10.1038/ng.3043

[27] M. B. Makarious, H. L. Leonard, D. Vitale, et al., "Multi-modality machine learning predicting Parkinson’s disease" npj Parkinson's Disease, vol. 8, no. 1, p. 35, 2022. https://doi.org/10.1038/s41531-022-00288-w

[28] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, Mar. 2003. http://jmlr.org/papers/v3/guyon03a.html

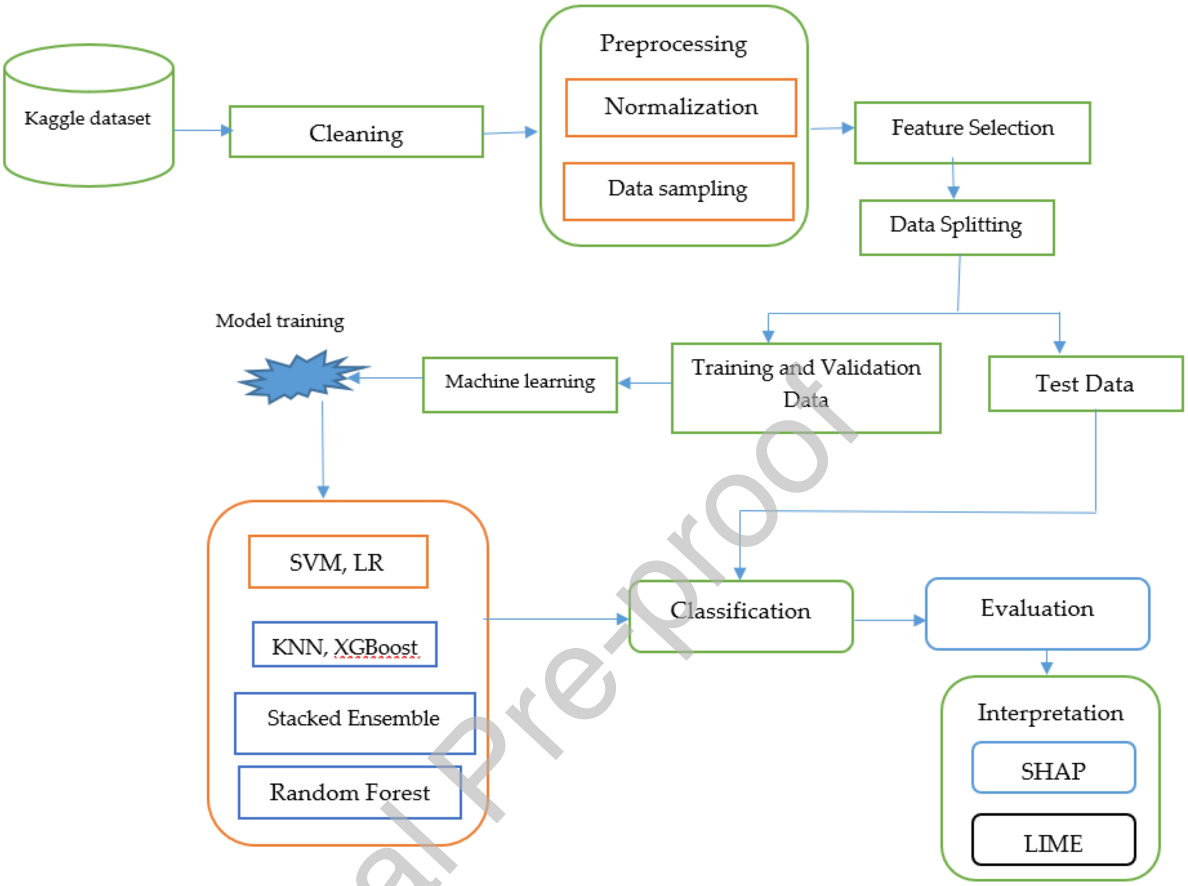


Figure 1: Research workflow for prediction of PD with XAI

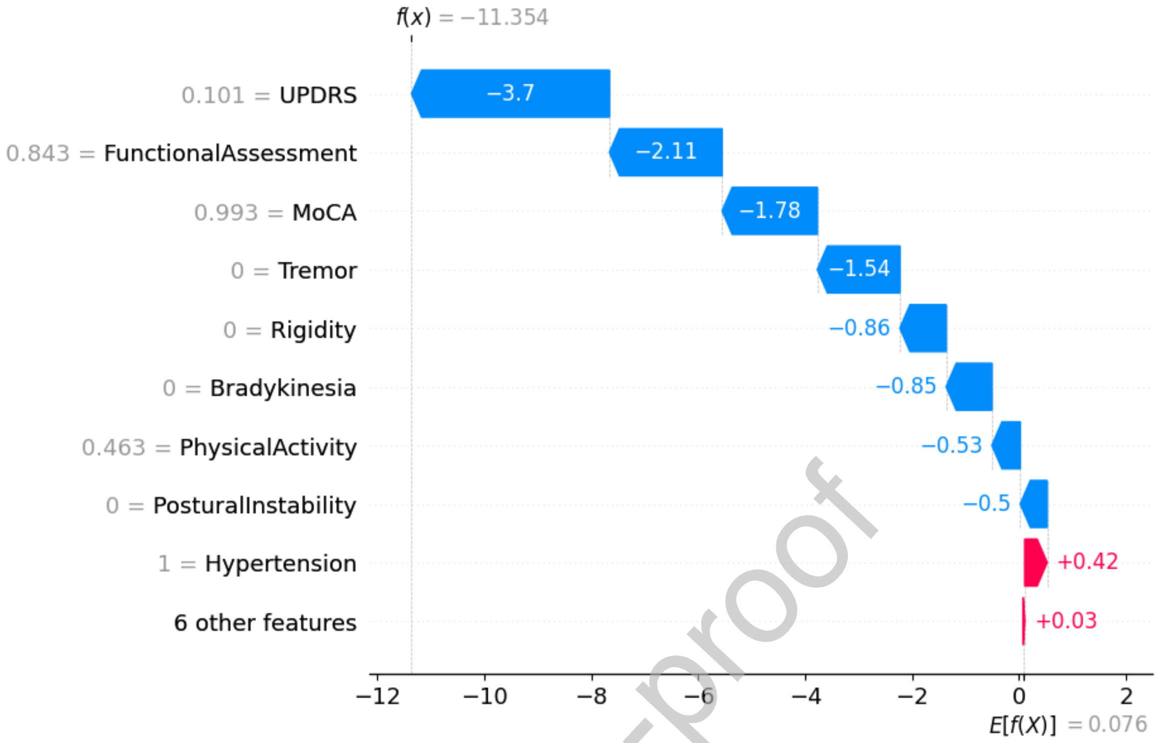


Figure 2: SHAP's waterfall plot for Interpretation of PD

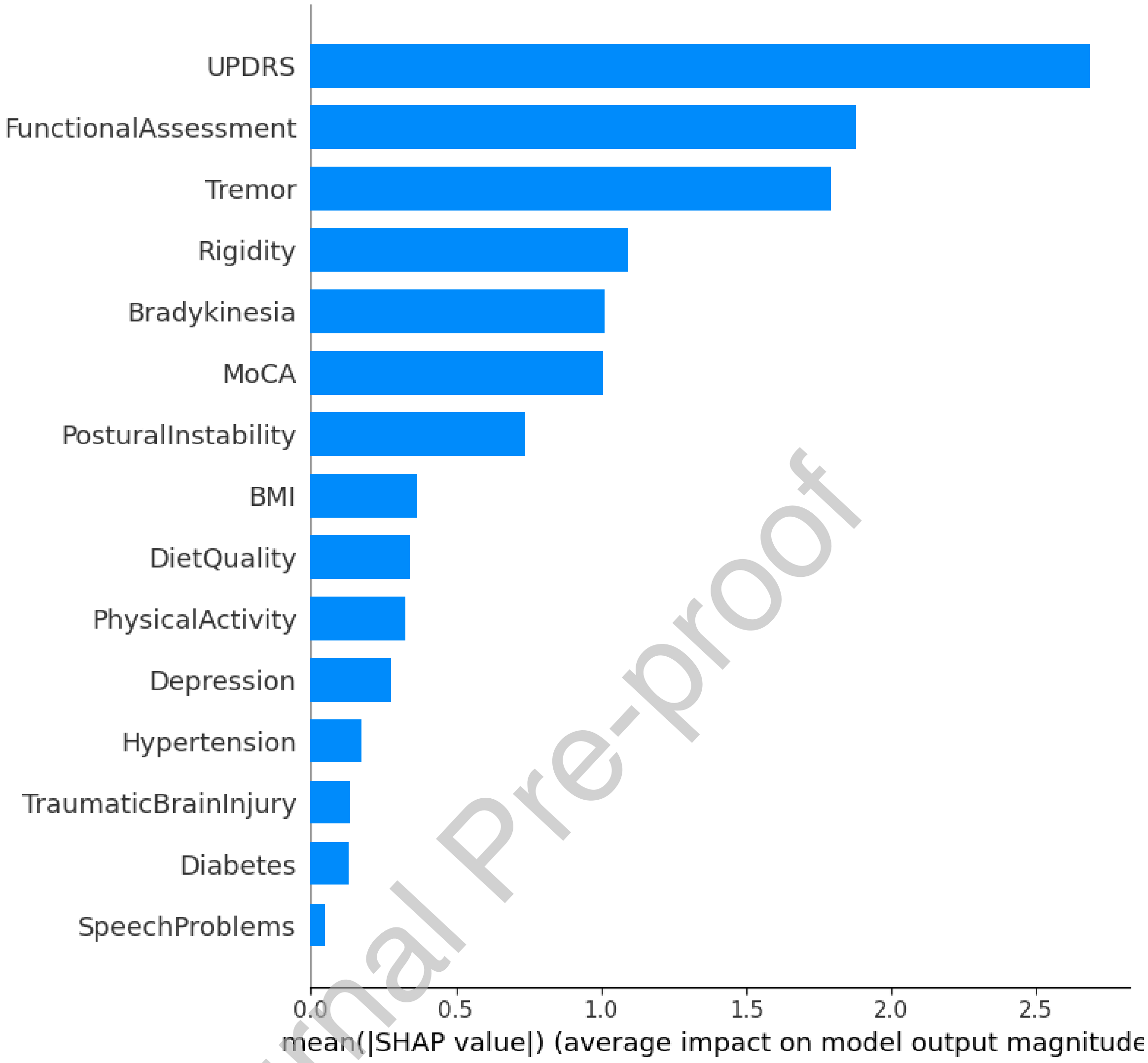


Figure 3: SHAP Summary on the training and testing set for interpretation of prediction of PD

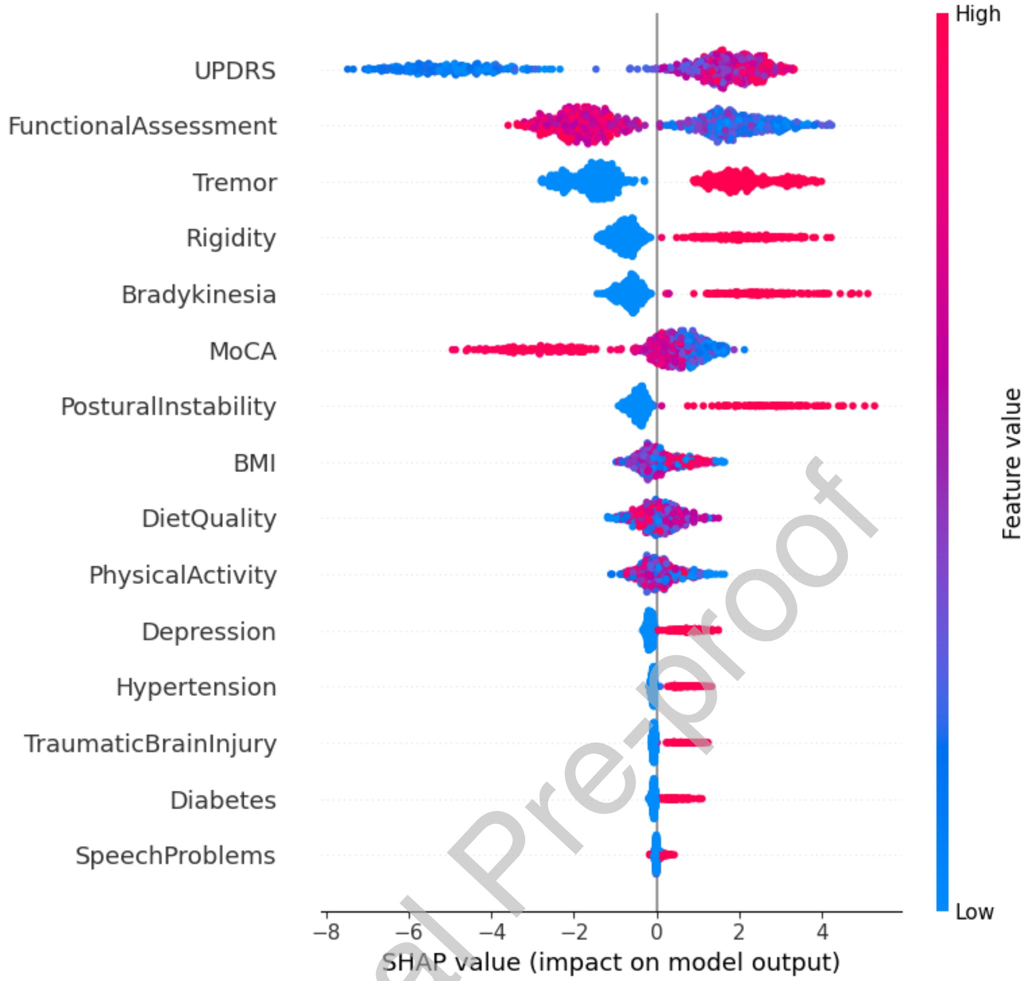
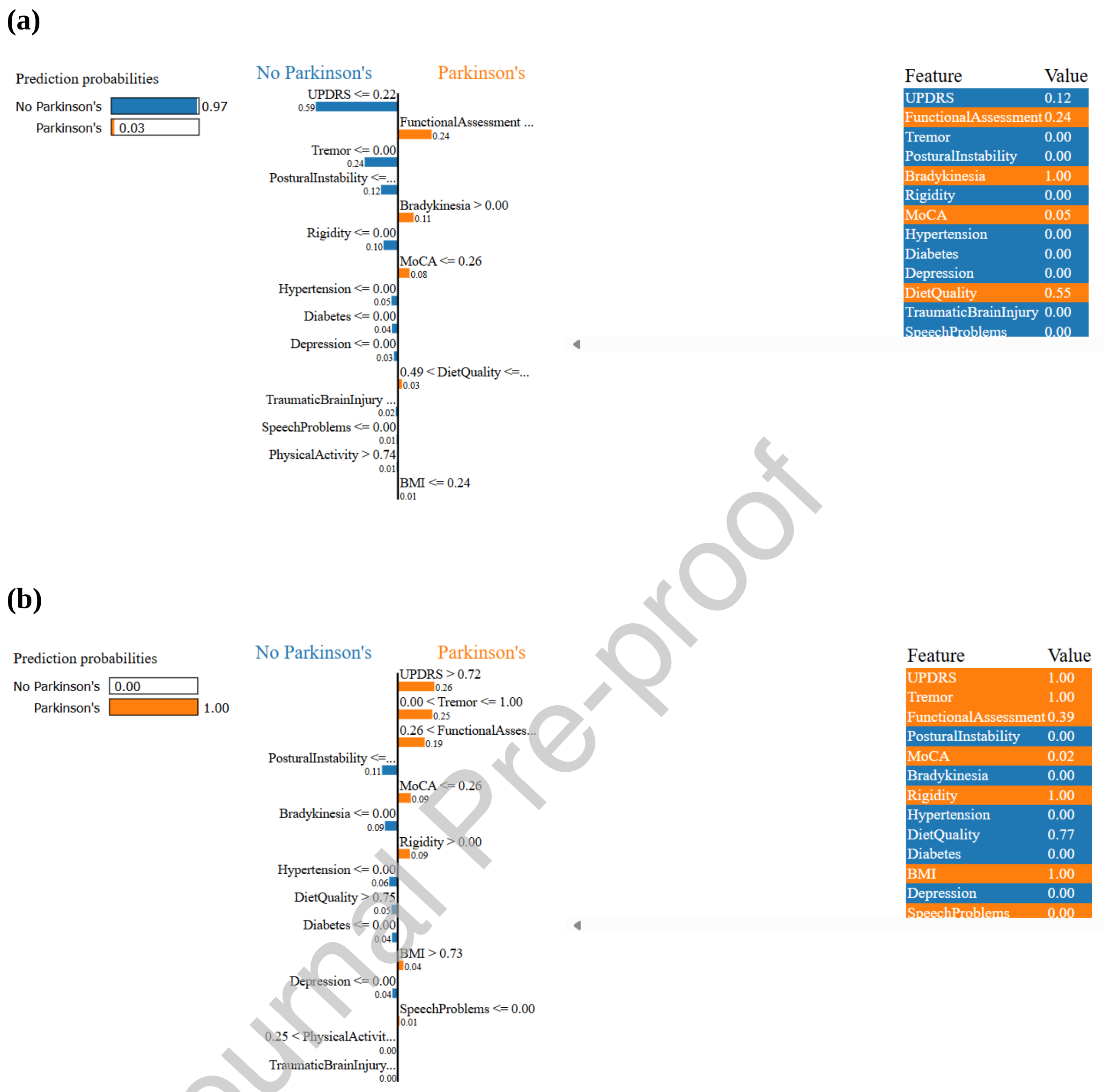


Figure 4: SHAP's Summary of the Random Forest for Prediction of PD



Feature	Value
UPDRS	0.12
FunctionalAssessment	0.24
Tremor	0.00
PosturalInstability	0.00
Bradykinesia	1.00
Rigidity	0.00
MoCA	0.05
Hypertension	0.00
Diabetes	0.00
Depression	0.00
DietQuality	0.55
TraumaticBrainInjury	0.00
SpeechProblems	0.00

Feature	Value
UPDRS	1.00
Tremor	1.00
FunctionalAssessment	0.39
PosturalInstability	0.00
MoCA	0.02
Bradykinesia	0.00
Rigidity	1.00
Hypertension	0.00
DietQuality	0.77
Diabetes	0.00
BMI	1.00
Depression	0.00
SpeechProblems	0.00

Figure 5: (a) LIME for instances predicted to have Parkinson's disease, (b) LIME for instances predicted not to have Parkinson's disease

Table 1: Performance Comparison of Machine Learning Algorithms in Predicting Parkinson's Disease (PD)

S/N	Algorithm	Avg. accuracy	Avg. precision	Avg. recall	Avg. f1-score	AUC
1	KNN	0.79	0.79	0.79	0.79	0.84
2	SVM	0.84	0.84	0.84	0.84	0.90
3	LR	0.83	0.83	0.83	0.83	0.90
4	XGBoost	0.92	0.92	0.92	0.92	0.96
5	Stacked ensemble	0.92	0.92	0.92	0.92	0.96
6	BEFS+ AACOA$_{hp}$+RF	0.93	0.93	0.93	0.93	0.97