RESEARCH ARTICLE

Bridging Predictive Modeling and Clinical Interpretability: An Explainable AI Approach to Parkinson's Disease Detection

Aamir Raza[1], **Aamir Ali**[2], **Aashesh Kumar**[1], **Nikhat Fatima**[3] and **Misbah Ali**[2,*]

[1] Department of IT and Management, Illinois Institute of Technology, Chicago, IL 60616, United States
[2] Department of Computer Science, COMSATS University Islamabad, Sahiwal Campus, Sahiwal 57000, Pakistan
[3] Concordia University Wisconsin, Mequon, WI 53097, United States

Abstract

Parkinson's disease (PD) is the second most common neurodegenerative disorder worldwide, predominantly affecting older adults. Early detection is crucial, as subtle motor and non-motor symptoms frequently overlap with other conditions, often resulting in delayed diagnosis. Many existing models rely on costly and less accessible imaging modalities such as MRI or PET scans, limiting their applicability in resource-constrained settings where only routine clinical data are available. This study develops interpretable AI models for early PD detection using structured clinical variables, incorporating feature selection techniques. Feature selection was conducted via Random Forest (RF) importance ranking combined with SelectKBest statistical scoring, retaining the most informative predictors for modeling. Five classifiers were implemented in parallel: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), RF, Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) network. Model performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The RF model achieved the highest accuracy of 92.72% (ROC-AUC: 0.968), with CNN and SVM showing competitive performance. LSTM exhibited balanced sensitivity and specificity, while KNN demonstrated relatively lower recall. To improve clinical interpretability, LIME was applied to each model to produce instance-level explanations, consistently highlighting tremor severity, motor impairment, cognitive scores, and age as key influential features. These results demonstrate that structured clinical variables alone can enable reliable PD detection without dependence on imaging. Integrating explainable artificial intelligence enhances transparency and supports responsible clinical adoption.

Keywords: Parkinson's disease, explainable artificial intelligence, machine learning, deep learning, feature selection, clinical decision support.

Submitted: 05 March 2026
Accepted: 09 March 2026
Published: 12 March 2026

Vol. 2, **No.** 1, 2026.
10.62762/BISH.2026.470997

***Corresponding author:**
✉ Misbah Ali
talktomisbah.ali@gmail.com

Citation
Raza, A., Ali, A., Kumar, A., Fatima, N., & Ali, M. (2026). Bridging Predictive Modeling and Clinical Interpretability: An Explainable AI Approach to Parkinson's Disease Detection. *Biomedical Informatics and Smart Healthcare*, 2(1), 20–37.

1 Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder globally, affecting over 10 million people, primarily older adults. It stands as the second most widespread neurological disorder globally, surpassed only by Alzheimer's Disease. Its growing prevalence underscores an urgent need for advanced research and improved treatments [1]. It is characterized by motor symptoms such as tremor, rigidity, and bradykinesia, along with non-motor symptoms like depression, sleep disorders, and cognitive impairment [2, 3]. Early and accurate diagnosis of PD is critical for effective clinical intervention, yet remains challenging due to the heterogeneity of symptoms and overlap with other neurological disorders. An overview of common motor and non-motor symptoms of PD relevant to clinical diagnosis and AI modeling is graphically shown in Figure 1.

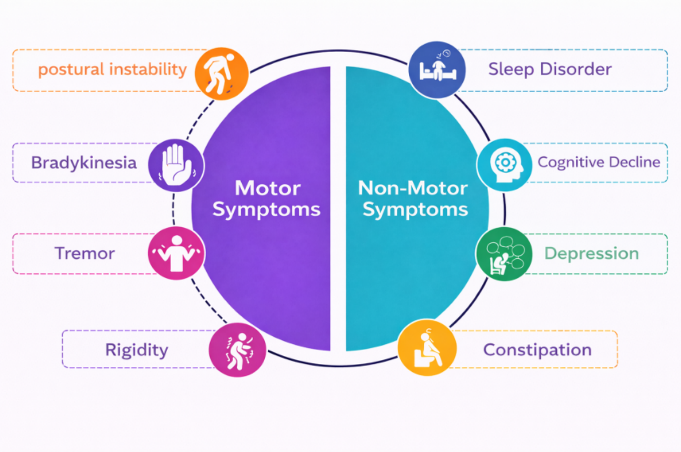


Figure 1. An overview of common motor and non-motor symptoms of PD relevant to clinical diagnosis and AI modeling.

Traditional diagnostic methods for PD primarily depend on clinical examination, patient history, and observation of symptom progression, often resulting in late-stage diagnosis when significant neuronal damage has already occurred [4]. This delay underscores the critical need for more objective, biomarker-driven approaches to enable earlier and more accurate detection. In this context, artificial intelligence (AI) has emerged as a transformative technology in healthcare, offering capabilities to analyze large-scale clinical datasets and detect subtle patterns indicative of disease onset [5, 6].

Several Machine Learning (ML) and Deep Learning (DL) models such as Support Vector Machines (SVM), Random Forest (RF), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks have shown promising performance in classifying PD based on clinical, demographic, and behavioral data [7–9]. However, most of these models operate as "black boxes," making their predictions difficult for clinicians to interpret and trust [10, 11].

This concern has led to increasing interest in Explainable AI (XAI), which aims to make AI predictions more transparent, interpretable, and clinically actionable [12]. Methods like SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations), Grad-CAM, and attention mechanisms have been employed to highlight important features contributing to model decisions, enhancing their acceptance in real-world clinical settings [13]. However, LIME provides clinically meaningful interpretations of ML models for PD by identifying the most influential features in individual predictions, such as motor symptom severity (UPDRS scores) or speech characteristics [14]. Its model-agnostic approach is particularly valuable for clinical datasets, as it can reveal potential biases, such as over-reliance on demographic factors rather than disease-specific markers, ensuring explanations align with medical knowledge [15].

Despite the growing body of work on artificial intelligence for PD detection, several limitations remain insufficiently addressed. Many existing studies rely heavily on imaging modalities or wearable sensor data. This can restrict applicability in resource constrained clinical settings where only structured clinical records are available. In addition, comparative evaluations across both traditional ML and DL models within a unified experimental pipeline are often limited. Interpretability is frequently treated as an auxiliary component rather than being systematically integrated into the modeling framework. Local explanation techniques are sometimes reported without consistent validation across different model types, reducing clinical coherence [16]. Furthermore, feature selection strategies and class balancing procedures are not always described within a reproducible and leakage controlled workflow. As a result, there remains a need for a clinically grounded, fully interpretable framework that evaluates multiple predictive models under a consistent preprocessing protocol using structured clinical variables alone [17, 18].

This study employed diverse ML models namely

RF, SVM, K-Nearest Neighbour (KNN) and DL architectures such as CNN, LSTM for PD detection, while applying feature selection technique.

Feature selection was performed using a two stage strategy that combined RF importance ranking with SelectKBest statistical scoring to identify the most informative clinical predictors. This procedure reduced redundancy within the feature space while retaining clinically meaningful variables. For traditional models, LIME effectively highlighted clinically meaningful features like motor symptoms, while for DL models it revealed less intuitive but predictive temporal patterns in complex data. The technique demonstrated particular value in making black-box DL outputs clinically interpretable by identifying relevant input segments, though its local approximations showed some instability with nonlinear architectures. The main contributions of the study are given below:

1. An explainable classification system was designed using structured demographic, lifestyle, neurological, symptom, and clinical history variables, enabling early PD detection without reliance on imaging modalities.
2. Feature selection strategy was implemented using Random Forest importance ranking followed by SelectKBest statistical scoring to identify and retain the most informative clinical predictors while reducing redundancy in the feature space.
3. Conventional ML models including RF, SVM, and K-Nearest Neighbors were developed alongside DL architectures comprising CNN and LSTM networks for comparative evaluation.
4. All classifiers were evaluated using diverse measures such as accuracy, precision, recall, F1 score, and ROC AUC.
5. Instance level explanations were generated to identify the contribution of key clinical features such as tremor severity, cognitive scores, and age, thereby supporting transparent and clinically interpretable decision making.

2 Literature Review

PD is a progressive neurodegenerative disorder that significantly affects the quality of life. Traditional diagnostic methods often rely on observable motor symptoms and clinical assessments, which may not detect PD at an early stage. With the advent of ML and AI, researchers have increasingly explored the use of clinical datasets for early diagnosis and prediction of PD. However, the black-box nature of many ML models has raised concerns in medical communities, highlighting the need for explainable AI (XAI) techniques.

Recent survey studies provide broader insight into the evolving role of artificial intelligence in Parkinson's disease diagnosis. Several review articles report that machine learning and deep learning techniques have been extensively applied to speech signals, wearable sensor data, handwriting analysis, and neuroimaging modalities. These studies collectively demonstrate that AI based models can achieve promising diagnostic performance across multiple data sources. However, survey findings also indicate persistent limitations in current research directions. Many predictive systems rely on specialized data acquisition procedures such as speech recordings or motion sensors, which may not always be available in routine clinical practice. In addition, interpretability is frequently treated as a secondary component rather than being incorporated within the core modeling framework. Systematic reviews also note that comparatively fewer studies investigate structured clinical variables as the primary data source for Parkinson's screening. Furthermore, unified experimental pipelines that integrate feature selection, multiple classification models, and explainability analysis within a consistent evaluation framework remain limited in the existing literature [16].

Quan et al. [19] introduced a DL approach for the diagnosis of PD from dynamic speech features with an accuracy of 84.29% using a Bidirectional LSTM model. The research indicated the importance of articulation transition features in the discrimination of PD patients and controls over conventional ML methods. Despite this, the small sample size limitation and the absence of heterogeneous language samples may impact the model's generalization to more general cases. Trabassi et al. [20] examined ML methods for diagnosing PD using IMU-based gait analysis and found SVM, RF, and DT to be high performers, achieving greater than 80% accuracy. This study highlighted the importance of feature selection in reducing overfitting and increasing model interpretability, particularly in clinical applications. Limitations included a small sample size and potential bias due to a skewed gender distribution, with broader validation recommended in subsequent studies.

Escobar-Grisales et al. [21] investigated DL and AI methods for speech and language biomarkers-based PD detection, reaching 88% accuracy with Wav2Vec 2.0 for speech and 77.9% using BETO-CNN for language modeling. The research identified speech as a discriminative modality compared to language for PD detection, albeit multimodal fusion strategies enhanced robustness. Limitations included dataset size restrictions and data loss during fusion, with recommendations for future work in fine-tuning pre-trained models and adjusting time-span alignment to maximize multimodal integration.

Um et al. [22] optimized PD classification with CNNs based on wearable sensor data, reaching 86.88% accuracy with rotation, permutation, and time-warping augmentations. The research considered limitations such as noisy labels and limited datasets, but could only be applied to 25 patients. Domain-specific data augmentation played a crucial role in enhancing the model's performance in healthcare use cases. Drotár et al. [23] differentiated PD patients from controls using handwriting kinematics and pressure, with 81.3% accuracy using SVM. Pressure features were found to perform better than kinematic ones (82.5% vs. 75.4%), illustrating their utility as diagnostic markers. The sample size was small, and no data from unmedicated patients were presented.

Gündüz [24] trained parallel CNN models for PD detection from vocal features with 86.9% accuracy on TQWT+MFCC+Concat features. The experiment proved the superiority of CNNs over SVMs but was constrained by possible data bias due to multiple patient recordings. Outcomes emphasized the worth of model-level feature fusion while calling for broader clinical verification. Yu et al. [25] created a DL model (CA_LSTM_FCN) for Parkinson's diagnosis from wearable sensors with 92.15% accuracy for PD vs. healthy controls and 85.38% for PD vs. comparable disorders. The model performed better than conventional approaches using temporal and spatial features with an attention mechanism. Short data collection durations and controlled laboratory settings are the limitations.

Shi et al. [26] applied ALFF-based radiomics and SVM to classify PD using rs-fMRI with 81.45% accuracy (primary) and 67.44% (validation). Biomarkers such as sensorimotor network high-order features were most important, although marked variability and small sample sizes preclude generalizability. The research emphasizes the tremendous promise of radiomics but requires multi-center, large validations. Martinez-Eguiluz et al. [27] employed ML to identify Parkinson's through non-motor signs, where SVM had 87.5% accuracy. RIPPER established important markers (autonomic/olfactory impairment) for explainable screening. The results of the study are encouraging but restricted by diversity in datasets.

Moon et al. [28]applied wearable sensors and ML to differentiate PD from essential tremor, with the best performance reported by neural networks (F1-score: 0.61). The research indicated the potential of gait/balance measures for differential diagnosis but was weakened by dataset imbalance and absence of prodromal cases. Future studies need to corroborate results on larger and more diverse cohorts. Fu et al. [29] employed radiomics from routine T2-FLAIR MRI to create ML models for the detection of Parkinson's, and MLP attained 0.85 AUC on external validation. The research showed the capacity of MRI for early screening of PD but was restricted by retrospective data and hand segmentation. Prospective validation and automated approaches are needed for future work to translate clinically. Summary of Key Studies on PD Detection Using ML and DL Approaches are shown in Table 1.

The reviewed experimental studies together with recent survey literature demonstrate the growing role of machine learning and deep learning in Parkinson's disease detection across diverse data modalities. However, variations in dataset composition, preprocessing protocols, feature selection strategies, and explainability integration limit direct comparability and clinical translation. A consistent, clinically grounded framework that evaluates multiple models under a unified pipeline with structured clinical variables remains insufficiently explored.

The remainder of this paper is organized as follows. Section 2 presents the literature review, where prior ML, DL, and XAI studies related to PD detection are critically examined. Section 3 describes the proposed methodology, including dataset characteristics, preprocessing procedures, feature selection strategy, model development, and interpretability integration. Section 4 reports the experimental results and comparative performance analysis across models. Section 5 outlines the limitations of the study and discusses implications for future research. Finally, Section 6 concludes the paper by summarizing the principal findings and their relevance to clinical

Table 1. Summary of Key Studies on PD Detection using ML and DL models.

Study	Dataset	Approach	Technique	Key Findings	Short Comings
[19]	268 voice samples, 45 subjects (15 HC, 30 PD)	DL	SVM, KNN, DT, CNN, Bidirectional LSTM, MLP, GNB	84.29% accuracy, outperforming traditional ML	Small regional dataset, complex architecture, no cross language validation
[20]	80 healthy and 81 Parkinson's patients	ML classification of IMU-derived gait features	SVM, DT, RF, ANN, KNN, Statistical, SBS, RF	Feature selection improved accuracy and interpretability	Gender imbalance, limited generalizability
[21]	165 Colombian Spanish speakers	Multimodel	1D-CNN, 2D-CNN, Wav2Vec 2.0, W2V, BERT, BETO, fusion	Speech 88%, language 77.9%, fusion improved stability	Small dataset, fusion information loss, no fine tuning
[22]	Wrist-worn accelerometer, 25 PD patients	DL with augmentation	Rotation, scaling, jittering, permutation, time-warping	Accuracy improved to 86.88% with augmentation	Noisy data, small sample, limited generalizability
[23]	38 healthy controls, 37 PD patients (handwriting task)	ML classification	SVM, KNN, AdaBoost, kinematic, and pressure feature analysis	Pressure features outperformed kinematic features	Small medicated cohort
[24]	Vocal recordings (UCI repository)	DL (Parallel CNN Framework)	MFCC, TQWT, LOPO_CV validation, SVM, and Wavelet features	CNN outperformed traditional models	Limited diversity, potential recording bias
[25]	PADS dataset (n=469)	DL for disease diagnosis using wearable sensors	Data Augmentation, Channel Attention Mechanism, LSTM, and FCN	92.15% PD vs HC, 85.38% PD vs PD	Short duration lab data, no real world validation
[27]	Biocruces (96 subjects), PPMI (687 subjects) dataset	Symptom-based PD classification using ML	RF, SVM, DT, NB, MLP, KNN, RIPPER, AdaBoost, Bagging	Non motor features improved classification	Focus on established PD, limited generalizability
[28]	567 patients (524 PD, 43 ET)	Wearable IMU sensors for balance/gait analysis during iSAW test	SVM, LR, RF, DT, kNN, Neural Network (NN), Gradient Boosting	NN effectively differentiated PD vs ET	Severe class imbalance, limited validation
[29]	Multicenter (5 cohorts)	Radiomics from T2W-FLAIR MRI	SVM, RF, KNN, AB, GNB, MLP	MLP achieved 0.85 external AUC	Manual segmentation, potential bias

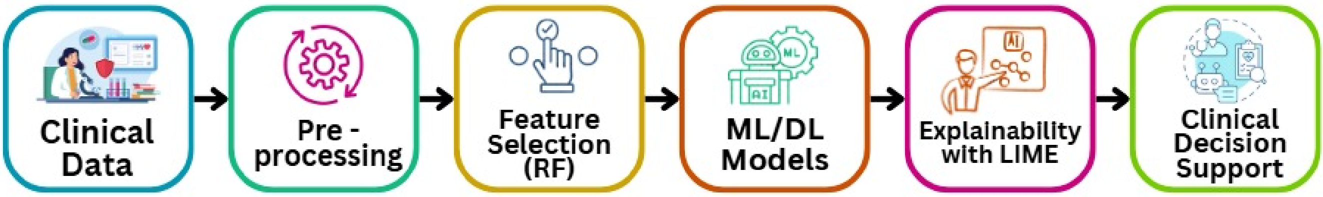


Figure 2. Overall framework of the explainable AI system.

decision support systems.

3 Proposed Methodology

The methodology of this research is designed to develop interpretable ML and DL models for the early detection of neurodegenerative diseases using only clinical data. It is comprised of the following steps. The overall framework for proposed XAI System is shown in Figure 2.

3.1 Data Collection and Preprocessing

This research study uses clinically structured dataset named Parkinson's Disease Classification for the early detection of Parkinson's Disease, obtained from the UCI ML Repository. It encompasses an elaborate set of features that fall under the categories of demographic, lifestyle, clinical, and neurological data. Demographic data includes age, gender, and Body Mass Index (BMI), which is basic patient background. Lifestyle data relevant to smoking, drinking, and exercising is tracked since these habits bear impacts on neurological health. Within the clinical domain, the dataset includes notable indicators such as the quality of sleep, presence of comorbidities including but not limited to hypertension, diabetes, depression, and stroke, as well as a family history of PD or history of brain injury which are known risk factors for PD. For neurological and cognitive evaluation, the UPDRS for assessing the severity of motor symptoms, the Montreal Cognitive Assessment (MoCA), and a score for functional living are three validated tools that are used. The Categories of clinical data features used for PD prediction in the proposed study are shown in Figure 3.

In addition, the dataset includes both motor symptoms such as tremor, rigidity, bradykinesia, postural instability as well as non-motor symptoms which include speech issues, sleep disturbances, and even constipation. The target variable of the dataset is the diagnosis label which classifies the patient as having PD (1) or not (0). In summary, the non-invasive dataset structure is well-suited for model development, as it is easy to explain and combine AI with human reasoning, making it appropriate for clinical settings and large-scale applications in

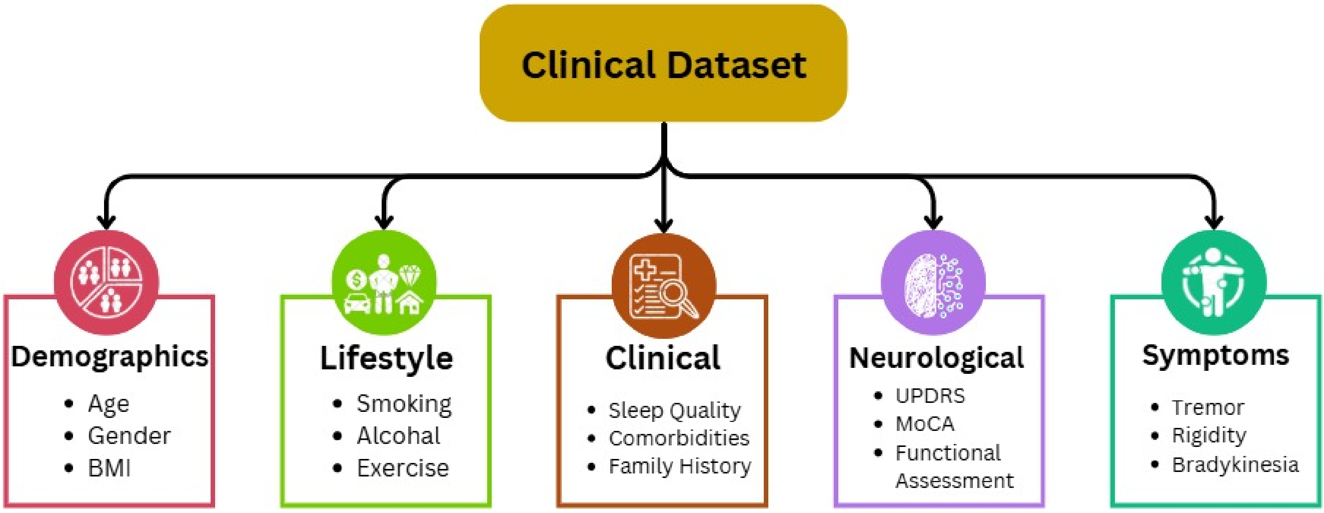


Figure 3. Categories of clinical data features used for PD prediction in the proposed study.

Table 2. Clinical dataset features and descriptions.

Category	Feature	Description
Demographics	Age	Patient's age in years
	Gender	Biological sex of the patient
	BMI	Body Mass Index (kg/m²)
Lifestyle Factors	Smoking	Indicates smoking status (Yes/No or frequency)
	Alcohol Consumption	Level or presence of alcohol intake
	Physical Activity	Frequency or intensity of physical exercise
Clinical Parameters	Sleep Quality	Self-reported or clinically assessed sleep score
	Comorbidities	Presence of hypertension, diabetes, depression, stroke
	Family History	Family history of PD or traumatic brain injury
Neurological Assessments	UPDRS Score	Unified PD Rating Scale score for motor evaluation
	MoCA Score	Montreal Cognitive Assessment for cognitive evaluation
	Functional assessment	Measures impact on daily life and independence
Symptoms	Tremor	Presence or severity of resting or action tremor
	Rigidity	Muscle stiffness typically seen in PD
	Bradykinesia	Slowness of movement
	Postural Instability	Difficulty in balance and posture control
	Speech Problems	Slurred or soft speech often seen in PD patients
	Sleep Disorders	Insomnia or REM sleep behavior disorder
	Constipation	Chronic difficulty in bowel movements
Target Variable	Diagnosis	Binary indicator: 1 for PD, 0 for non-PD

resource-limited environments, and Clinical Dataset Features and Descriptions are shown in Table 2.

The dataset contains a total of 756 instances derived from 252 individuals. Among them, 188 subjects were diagnosed with PD and 64 were healthy controls. The Parkinson cohort included 107 men and 81 women with ages ranging from 33 to 87 years, with a mean age of 65.1 ± 10.9 years. The control group consisted of 23 men and 41 women aged between 41 and 82 years, with a mean age of 61.1 ± 8.9 years. The dataset does not contain missing values. All records are anonymized and distributed for research use.

To ensure unbiased performance estimation, the dataset was divided using stratified sampling. 80 percent of the data were assigned to the training set, while 20 percent were reserved as an independent test set. Stratification preserved the original class distribution in both subsets. A fixed random seed was used to ensure reproducibility. All preprocessing procedures were performed exclusively on the training data. Feature scaling parameters were learned from the training set and subsequently applied to the test set.

3.2 Feature Selection

Feature selection was performed after stratified data partitioning. The dataset was first divided into training and testing subsets using an 80 to 20 stratified split to preserve class proportions. All subsequent preprocessing operations were restricted to the training subset. The independent test set remained untouched during feature selection and model construction. The training features were standardized using parameters estimated exclusively from the training data. These scaling parameters were later applied to the test subset without refitting. Class imbalance within the training data was then addressed using SMOTE oversampling. Synthetic minority instances were generated only within the training subset.

A two stage feature selection strategy was employed. In the first stage, a Random Forest classifier was trained on the balanced training data to compute feature importance scores based on mean decrease in impurity. Features were ranked according to their contribution to classification performance. Variables with negligible importance were discarded.

In the second stage, SelectKBest was applied to the reduced training feature space. Each remaining feature was evaluated statistically with respect to the class label. The top ranked predictors were retained according to the predefined selection parameter. This sequential approach reduced dimensionality while preserving clinically relevant variables. The complete two-stage feature selection workflow is illustrated in Figure 4.

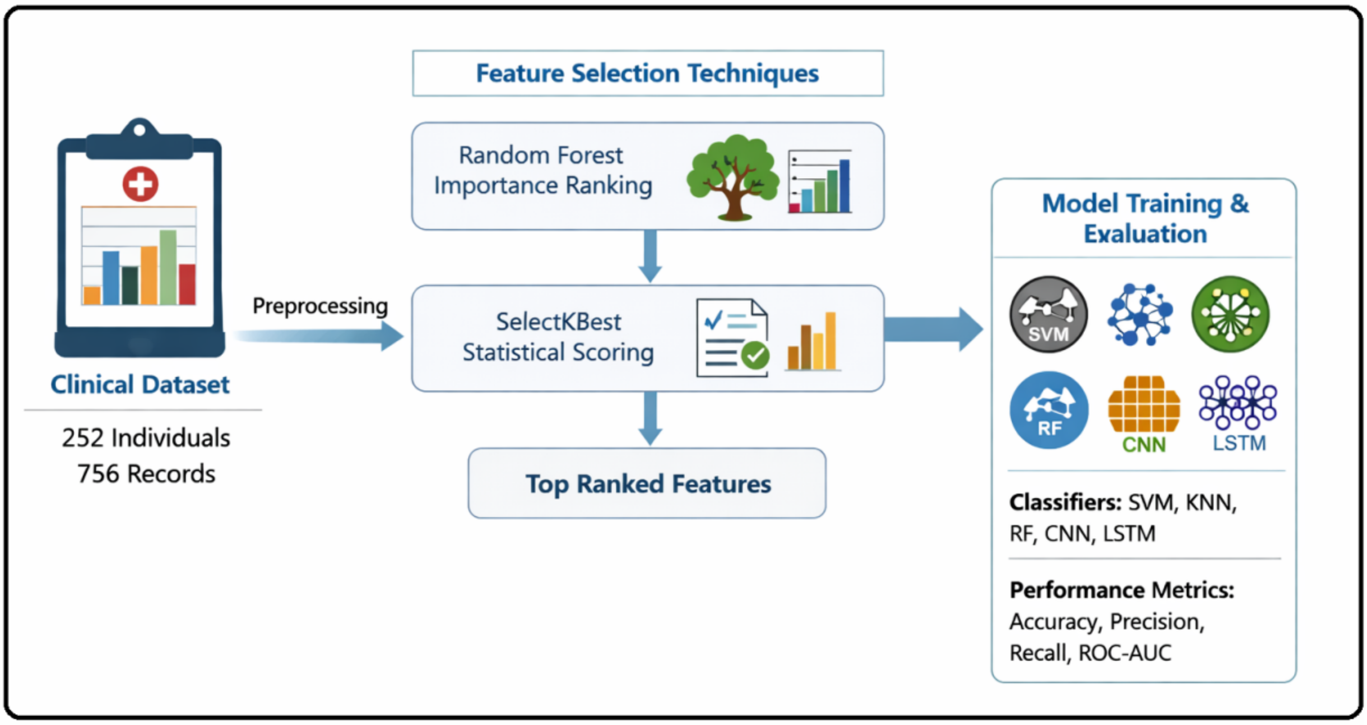

Feature Selection Techniques
Random Forest Importance Ranking
SelectKBest Statistical Scoring
Top Ranked Features
Clinical Dataset
252 Individuals
756 Records
Preprocessing
Model Training & Evaluation
SVM
RF
CNN
LSTM
Classifiers: SVM, KNN, RF, CNN, LSTM
Performance Metrics: Accuracy, Precision, Recall, ROC-AUC

Figure 4. Feature selection process using SelectKBest through RF.

The fitted selector was subsequently applied to the test subset using the transform operation. No fitting procedure involved the test data. The resulting reduced feature matrices were then used for downstream model training and evaluation. This controlled workflow ensured that performance assessment was conducted on data that remained fully independent from the feature selection process.

3.3 Model Development

The model development phase employs a hybrid set of algorithms combining traditional ML and DL. SVM and KNN are implemented due to their ability to handle high-dimensional spaces and offer strong performance in binary classification tasks using clinical tabular data. Random Forest (RF) is used for its high robustness, resistance to overfitting, and in-built feature ranking capabilities. For DL, Convolutional Neural Networks (CNN) and Long Short Term Memory Network (LSTM) have been adapted to operate on reshaped tabular clinical data to detect complex feature interactions. The Hyperparameter Configuration of Machine Learning Classifiers is given in Table 3.

Long Short Term Memory network was implemented to explore whether gated recurrent architectures can capture complex nonlinear dependencies among structured clinical variables within a single patient

Table 3. Hyperparameter configuration of machine learning classifiers.

Model	Hyperparameter	Value	Description
SVM	Kernel	RBF	Radial basis function kernel
	C	1.5	Regularization parameter
	Gamma	scale	Kernel coefficient
	Probability	TRUE	Enables probability estimation
	random_state	42	Reproducibility control
KNN	n_neighbors	5	Number of nearest neighbors
Random Forest	n_estimators	100	Number of decision trees
	max_depth	None	Nodes expanded until purity
	random_state	42	Reproducibility control

Table 4. The configuration summary of the one dimensional convolutional neural network.

Layer Type	Configuration	Output Dimension	Remarks
Input	n selected features	(n, 1)	Reshaped feature vector
Conv1D	32 filters, kernel size 3, ReLU	(n-2, 32)	Local feature interaction modeling
Max Pooling	Pool size 2	Reduced dimension	Dimensionality reduction
Dropout	Rate 0.3	Same as input	Regularization
Flatten	—	Flattened vector	Transition to dense layer
Dense	64 neurons, ReLU	64	Nonlinear transformation
Dropout	Rate 0.3	64	Regularization
Output	1 neuron, Sigmoid	1	Binary classification

record. In this configuration, the ordered feature vector of each sample was treated as a structured sequence. The ordering provides a consistent feature arrangement that enables recurrent computation across variables.

The Convolutional Neural Network was adapted for structured tabular input by reshaping the selected feature matrix into a compatible format. Convolutional layers were used to extract local feature representations, followed by fully connected layers for final classification.

Each clinical variable was mapped to a single timestep, resulting in a sequence length equal to the number of retained features after selection. The objective was to allow the LSTM to model inter feature relationships through its memory cell and gating mechanisms. Hence learning latent representations straightforwardly that may not be captured by purely feed forward architectures.

The input tensor was reshaped into three dimensions of size sequence length by 1 prior to being passed to the recurrent layer. A single LSTM layer with 64 hidden units was employed. Dropout with rate 0.3 was incorporated to reduce overfitting. The recurrent output was connected to a dense layer containing 32 neurons with Rectified Linear Unit activation, followed by a Sigmoid output neuron for binary classification. The model was optimized using Adam with binary cross entropy loss. Training was conducted for up to 50 epochs with early stopping based on validation loss.

A one dimensional convolutional neural network was implemented to model nonlinear interactions among selected clinical features. The input to the network consisted of the final subset of n selected features obtained from the feature selection stage. Each feature vector was reshaped to a two dimensional tensor of size n_features by 1 to enable convolutional operations. This reshaping was performed to facilitate parameter sharing and local interaction modeling rather than to impose spatial structure.

The architecture comprised a single Conv1D layer with 32 filters and a kernel size of 3. Rectified Linear Unit activation was applied to introduce nonlinearity. A max pooling layer with pool size 2 was then used to reduce dimensionality and control overfitting. Dropout with a rate of 0.3 was applied after pooling to improve generalization.

The convolutional output was flattened and passed to a fully connected dense layer containing 64 neurons with Rectified Linear Unit activation. An additional dropout layer with a rate of 0.3 was used prior to the final classification stage. The output layer consisted of a single neuron with Sigmoid activation to perform binary classification.

The model was trained using the Adam optimizer with binary cross entropy as the loss function. A batch size of 32 was employed, and training was conducted for up to 50 epochs. Early stopping based on validation loss was applied to prevent overfitting. To enhance architectural transparency and reproducibility, the configuration of the one dimensional convolutional neural network is summarized in Table 4. Table 4 outlines the sequence of layers, their structural parameters, and the corresponding functional roles within the network. This structured presentation facilitates clearer interpretation of the modeling design and supports methodological consistency.

Overall analytical framework illustrating the structured clinical dataset categories, followed by parallel modeling pathways is presented in Figure 5. The ML path comprises Support Vector Machine, K Nearest Neighbors, and Random Forest classifiers, while the DL path includes Convolutional Neural Network and Long Short Term Memory architectures. It presents the complete workflow from

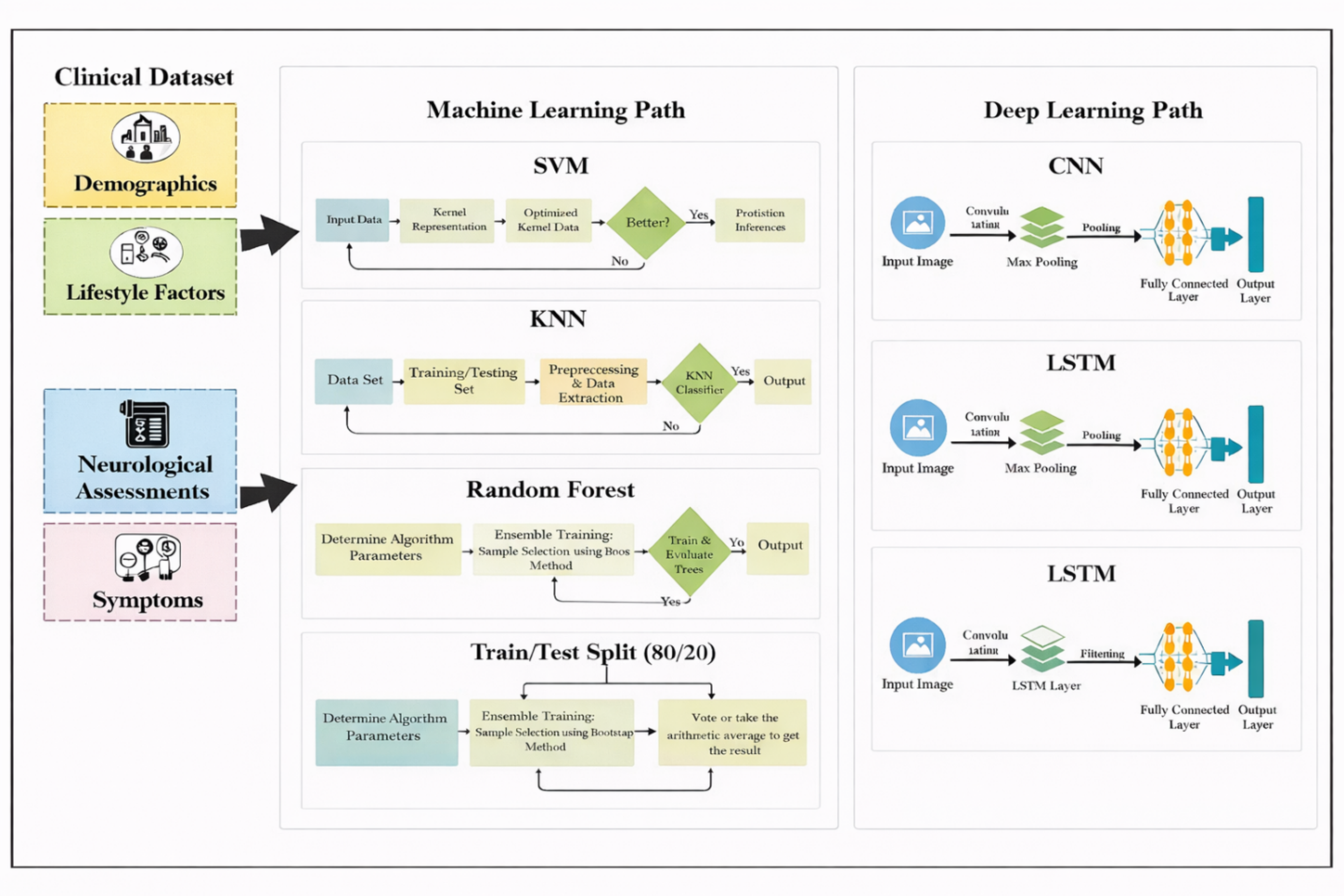


Figure 5. Analytical framework showing clinical dataset categories and parallel modeling pathways.

clinical inputs to model development and evaluation.

3.4 Explainable AI (XAI) Technique

To make the predictions interpretable and useful for clinical decision-making, Explainable AI (XAI) methods are integrated into the system. LIME (Local Interpretable Model-Agnostic Explanations) is used to generate human-interpretable surrogate models that approximate the predictions of complex models around specific instances. This method provides intuitive visualizations that display how specific clinical features such as tremor intensity, cognitive scores, or age impact model decisions, thereby enhancing trust and accountability.

4 Results and Discussion

This section presents the quantitative outcomes of the proposed classification framework. Performance was evaluated using accuracy, precision, recall, F1 score, and ROC AUC to provide a balanced assessment of diagnostic capability across models. The results achieved from both ML and DL models for PD classification are given in Table 5.

The SVM model reported good performance for Parkinson's classification with an accuracy of 90.04%, precision of 0.9, recall of 0.892, F1-score of 0.899, and ROC AUC of 0.949. It is seen that the model optimizes the trade-off between identifying true positives and reducing false positives. Its high precision indicates that the majority of the detected Parkinson's cases were

Table 5. Performance comparison of ML and DL models for PD classification.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
SVM	90.04	0.9	0.892	0.899	0.949
KNN	0.852	0.868	0.831	0.849	0.921
Random Forest	0.927	0.923	0.931	0.927	0.968
CNN	0.919	0.901	0.942	0.921	0.957

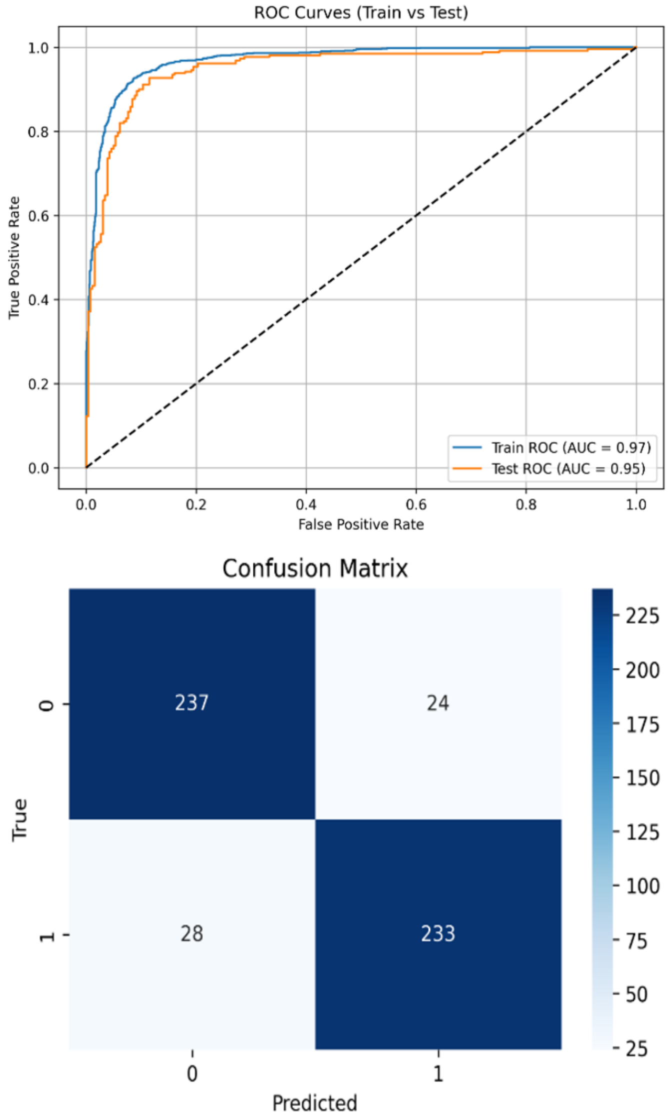


Figure 6. ROC curve and Confusion Matrix for PD using SVM.

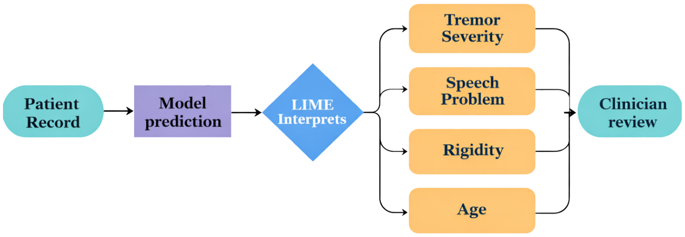


Figure 7. Instance level interpretability workflow using LIME for PD prediction based on structured clinical features.

well-classified, and the high recall represents the ability of the model in identifying true Parkinson patients.

The SVM ROC curve shows a high true positive rate with a low false positive rate consistently. The AUC of 0.9497 reflects the model's ability to distinguish between Parkinson's and non-Parkinson's conditions. The sharp curve in the upper-left corner of the plot reflects that the model has excellent performance for all the thresholds. The confusion matrix shows that the SVM model accurately labeled most Parkinson's and non-Parkinson's cases. Few false negatives and false positives were observed, proving the reliability of the model in a clinical setting. Its classification errors are minimal, where only a few Parkinson's patients are mislabeled. The ROC curve and confusion matrix are shown in Figure 6.

LIME explanations for SVM identify clinical features of tremor severity, motor function deterioration, rigidity, and speech disorders as major contributors to model predictions. Figure 7 illustrates the instance level explanation process implemented in the proposed framework. A patient record is first processed by the trained classification model to generate a prediction, after which LIME is applied to interpret the decision locally. The method identifies influential clinical variables such as tremor severity, speech problems, rigidity, and age, and presents their relative contributions to the predicted outcome. These interpretable outputs are then provided for clinician review to support transparent and informed decision making.

By extracting the most impactful features for each decision, LIME offers localized transparency that can be cross-checked by clinicians, hence justifying the use of the model in a clinical environment. This is shown in Figure 8.

The KNN classifier provided accuracy of 85.25%, precision of 0.868, recall of 0.831, F1-score of 0.849, and ROC AUC of 0.921. The results indicate slightly poorer performance than SVM and ensemble methods. The model is comparatively better at reducing false positives rather than detecting all true positives, as evident from the lower recall value.

The KNN confusion matrix shows that there is a significant number of false negatives, in that multiple true cases of Parkinson's were classified incorrectly as non-Parkinson's. This might decrease clinical reliability, particularly for early detection when the failure to detect positive cases is important. Nonetheless, it has acceptable true positive and true negative rates. ROC of KNN is reasonably steep and displays respectable classification ability. Although it falls short of CNN or Random Forest in curvature, an AUC of 0.9217 still reflects that the model possesses

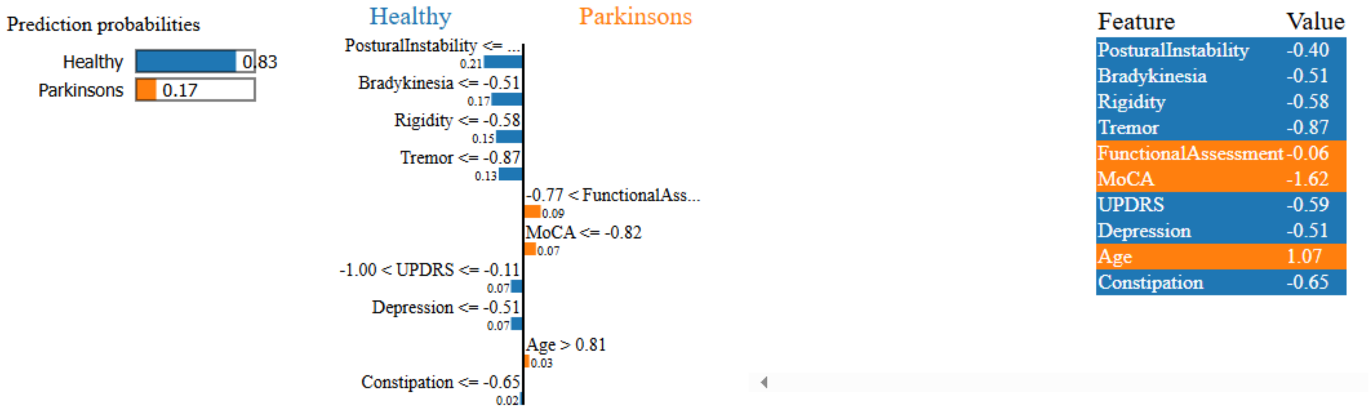


Figure 8. LIME explanation for PD after using SVM.

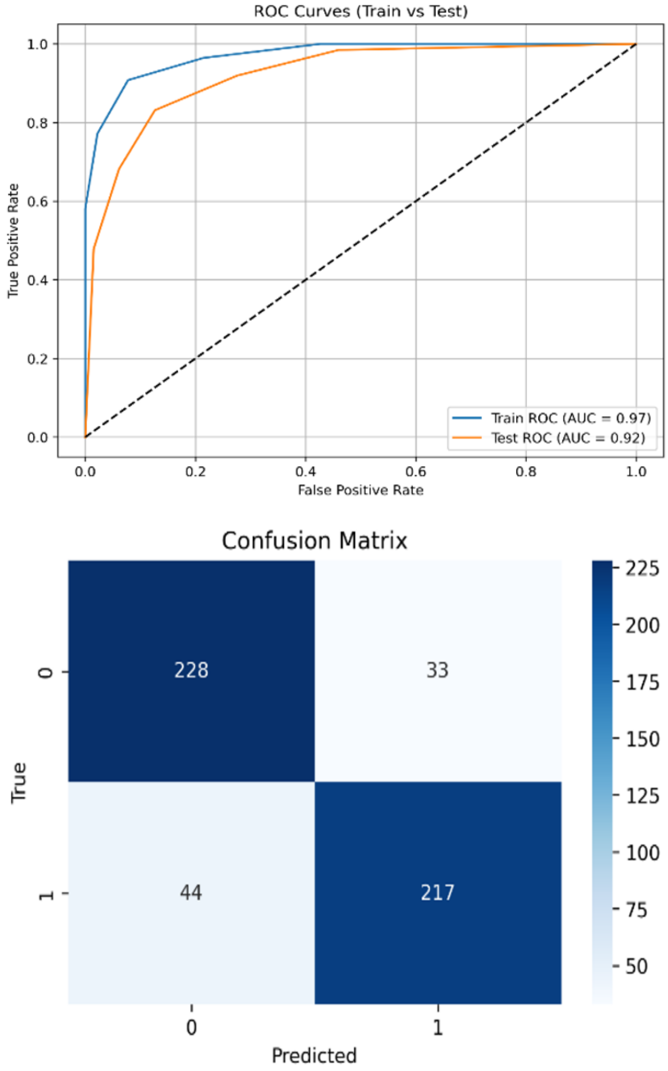


Figure 9. The ROC and Confusion Matrix for PD using KNN.

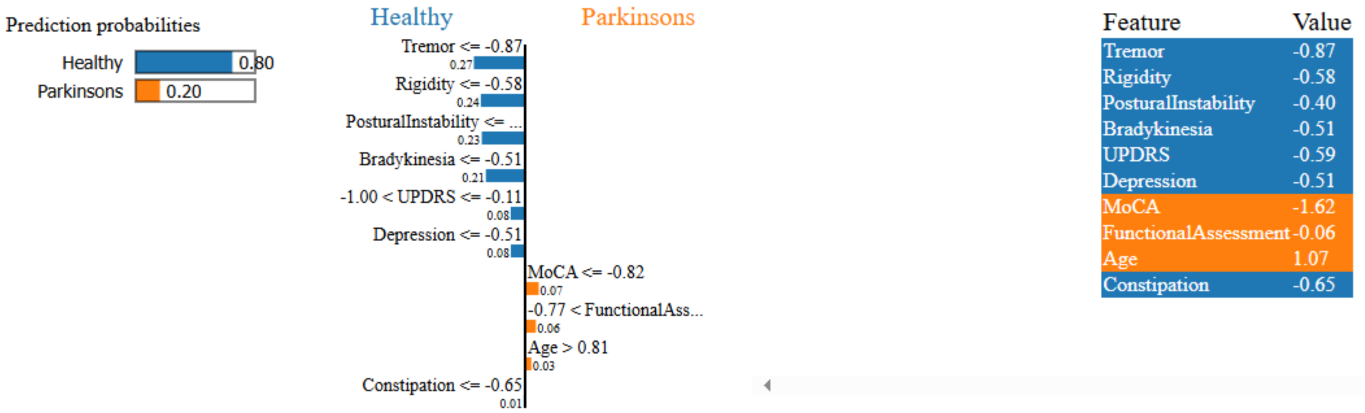
sufficient predictive power for different thresholds. The ROC and Confusion Matrix for PD using KNN are shown in Figure 9.

LIME explanations for KNN demonstrate the effect of local neighborhoods on the outcome. It places a strong focus on similarity-based reasoning, where the predictions are compelled by nearby instances having similar attributes. Although beneficial in explaining simple decisions, KNN's performance and explanation can be inconsistent in function based on the density and quality of training data in the feature space. This LIME explanation for PD after using KNN is shown in Figure 10.

Random Forest showed high accuracy in classification at 92.72%, precision at 0.923, recall at 0.931, F1-score at 0.927, and a stable ROC AUC of 0.968. It has a good performance for Parkinson's detection with a decent balance of detecting true positives and a minimal rate of false positives.

The ROC curve of Random Forest is close to 1, with a steep curve skirting close to the top-left boundary. Its AUC of 0.9682 confirms its remarkable performance at all thresholds, with high sensitivity and specificity. The confusion matrix yields extremely few misclassifications, with excellent capability to accurately label both Parkinson's and non-Parkinson's patients. Its consistent classification performance on the held out test data indicates its high reliability and low risk of diagnostic error, which makes it a viable candidate for real-world deployment. The ROC and Confusion Matrix for PD using RF are shown in Figure 11.

Random Forest LIME visualizations demonstrate that predictions are highly dominated by a mix of neurological attributes such as gait speed, limb stiffness, voice tremors, and muscle stiffness. These interpretations ascertain that Random Forest not only performs well but also makes its decisions based on medically interpretable parameters. This LIME explanation for PD after using RF are shown in



Feature	Value
Tremor	-0.87
Rigidity	-0.58
PosturalInstability	-0.40
Bradykinesia	-0.51
UPDRS	-0.59
Depression	-0.51
MoCA	-1.62
FunctionalAssessment	-0.06
Age	1.07
Constipation	-0.65

Figure 10. LIME explanation for PD after using KNN.

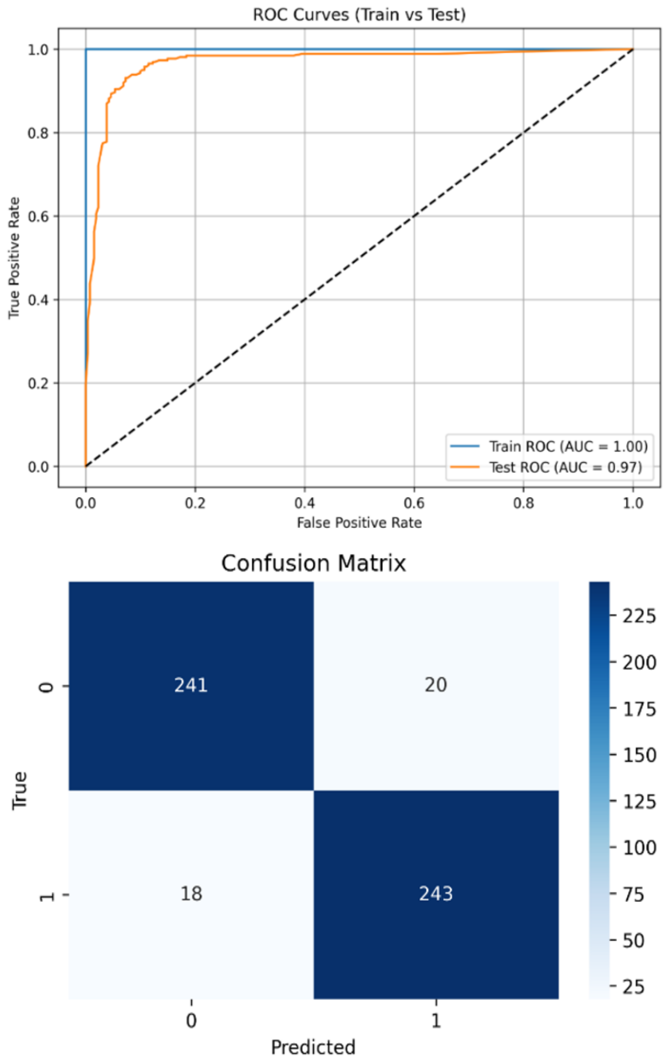


Figure 11. The ROC and Confusion Matrix for PD using RF.

Figure 12.

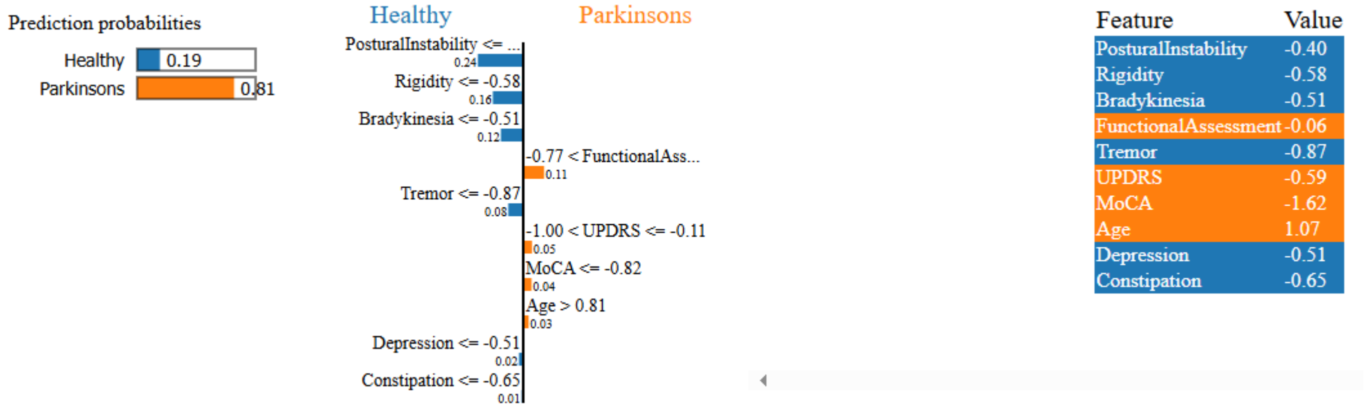
CNN gave robust performance with 91.95% accuracy, 0.901 precision, 0.942 recall, 0.921 F1-score, and 0.957 ROC AUC. The high recall rate means CNN is very sensitive and picks up most of the Parkinson's cases though it may lead to some extra false positives.

The ROC curve indicates that CNN can distinguish positive and negative classes effectively. AUC of 0.9578 assures good classification power. Its high recall is particularly useful for early-stage screening, where losing true cases is more expensive than a few false positives. The confusion matrix reveals that the model is very sensitive with a low false negative rate. Some false positives were also observed, i.e., some healthy people were falsely reported as having Parkinson's. The ROC and Confusion Matrix for PD using CNN are shown in Figure 13.

LIME was applied to explain CNN decisions by examining the input feature contribution at various layers of the network. Postural instability, severity of motor fluctuation, and tremor patterns were commonly determined as important. Local explanations provided by LIME reduce the black-box phenomenon of DL models, thereby increasing clinician trust. The LIME explanation for PD after using CNN is shown in Figure 14.

The LSTM model achieved 89.85% accuracy, precision of 0.896, recall of 0.900, F1-score of 0.898, and ROC AUC of 0.953. It is especially favorable for sequential clinical data, where onset of symptoms in time is significant for early diagnosis.

The ROC curve illustrates stable model performance at different thresholds, and the AUC of 0.9533 confirms its high discrimination power between classes in clinical sequences. LSTM's confusion matrix shows balanced classification with a comparably small number of false positives and false negatives. It has high sensitivity and specificity and so is very useful for diseases such as Parkinson's that show progression over time. ROC and Confusion Matrix for PD using LSTM are shown in Figure 15.



Feature	Value
PosturalInstability	-0.40
Rigidity	-0.58
Bradykinesia	-0.51
FunctionalAssessment	-0.06
Tremor	-0.87
UPDRS	-0.59
MoCA	-1.62
Age	1.07
Depression	-0.51
Constipation	-0.65

Figure 12. LIME explanation for PD after using RF.

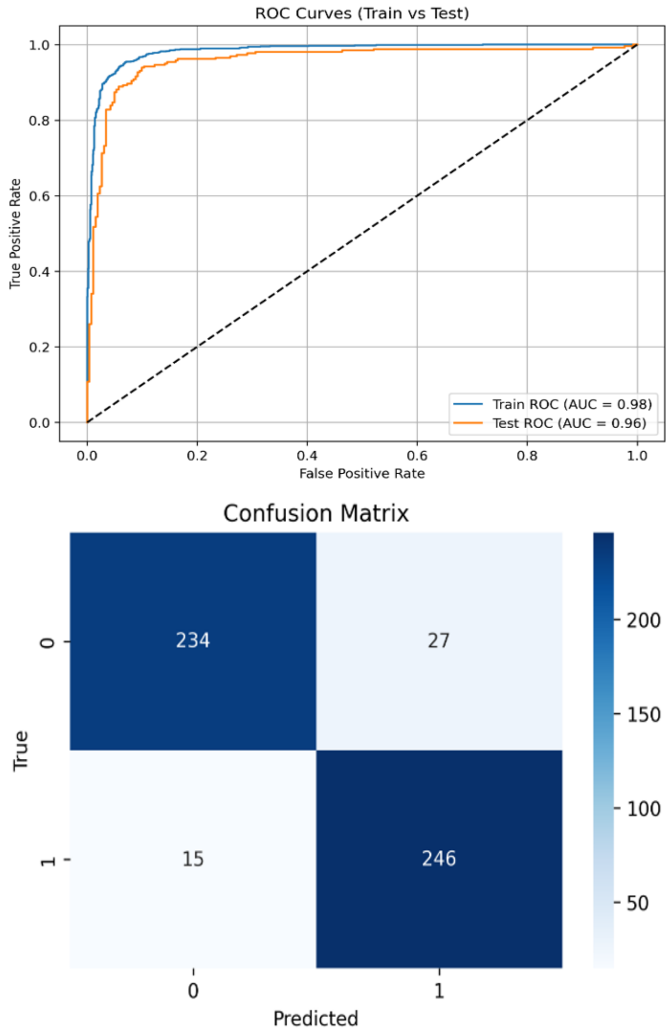


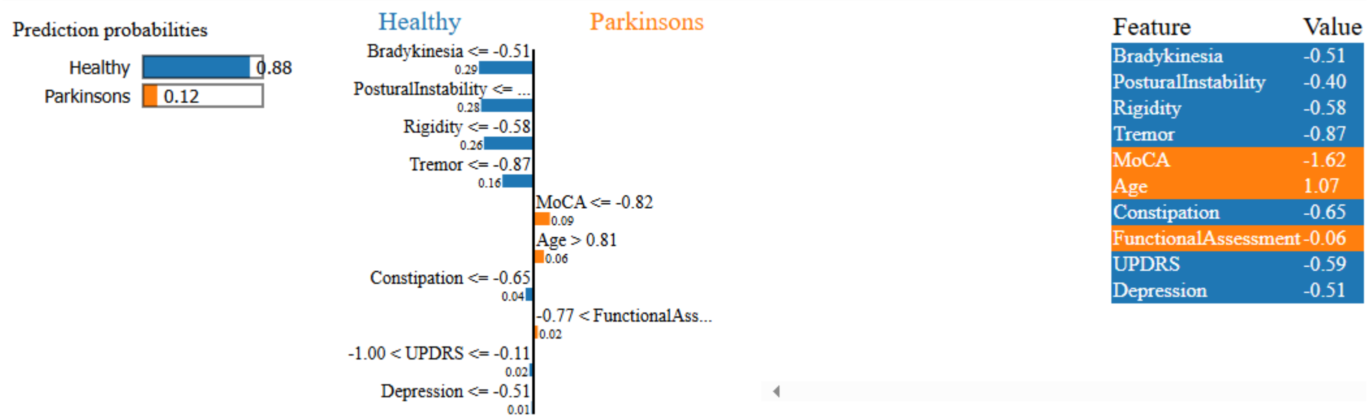
Figure 13. ROC and Confusion Matrix for PD using CNN.

LIME explanations for LSTM showed that the model is picking up on symptom trends and clinical history information, e.g., trend in tremors, facial expressions, and speech abnormalities, as being critical to predictions. This makes the model very interpretable with LIME, since clinicians can verify predictions based on familiar symptom patterns. This LIME explanation for PD; after using LSTM is shown in Figure 15.

Among all the models, RF stood out as the best performing model achieving the highest accuracy of 92.72%, thus reinforcing its position as the strongest traditional ML model. The DL model CNN also performed well, achieving an accuracy of 91.95% indicating its proficiency in identifying spatial correlations in the data. SVM and LSTM performed competitively as well, both achieving over 89% accuracy, while KNN lagged with an accuracy of 85.25%. The results support the use of latent AI in combination with clinical applications for reliable and comprehensible detection of PD devoid of imaging data.

4.1 Practical Feasibility and Clinical Deployment

Translation of predictive models into clinical environments requires consideration of operational feasibility. The framework proposed in this study relies exclusively on structured clinical variables rather than imaging or sensor based modalities. This characteristic improves its suitability for hospitals and community clinics where advanced diagnostic infrastructure is limited. Many of the variables used in the dataset such as UPDRS scores, demographic attributes, and cognitive assessments are already collected during routine neurological evaluations. Minimal additional data collection would therefore be required. Computational efficiency also supports practical implementation. The majority of models were trained using a reduced feature space generated through feature selection. Random Forest, Support Vector Machine, and K Nearest Neighbors operate effectively on tabular clinical data with moderate dimensionality. Prediction time is short. Inference can be executed within milliseconds on standard



Feature	Value
Bradykinesia	-0.51
PosturalInstability	-0.40
Rigidity	-0.58
Tremor	-0.87
MoCA	-1.62
Age	1.07
Constipation	-0.65
FunctionalAssessment	-0.06
UPDRS	-0.59
Depression	-0.51

Figure 14. LIME explanation for PD after using CNN.

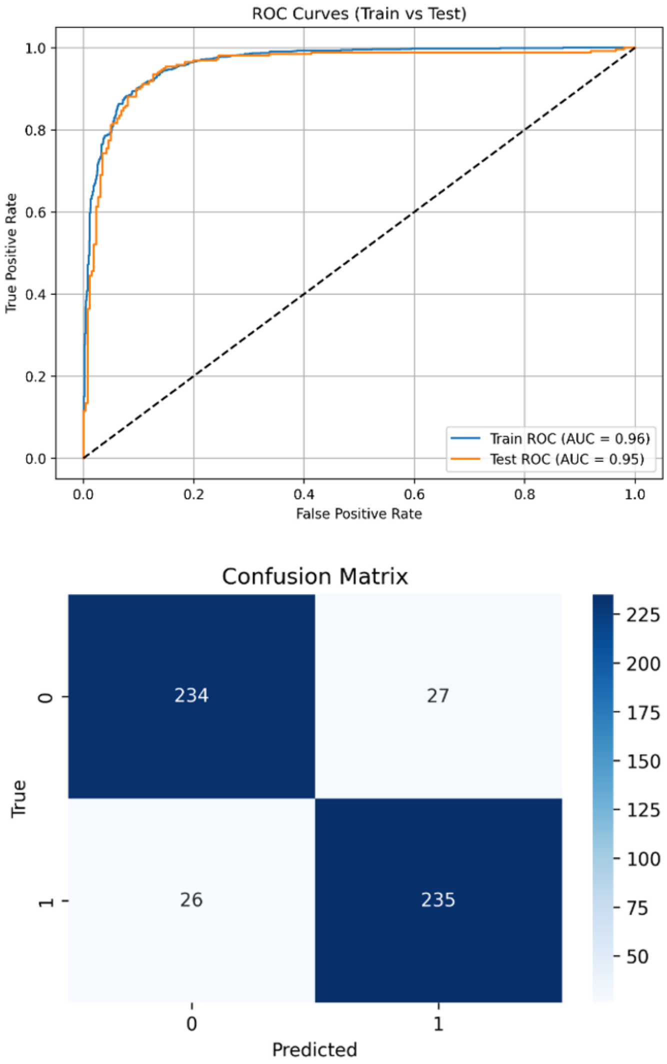


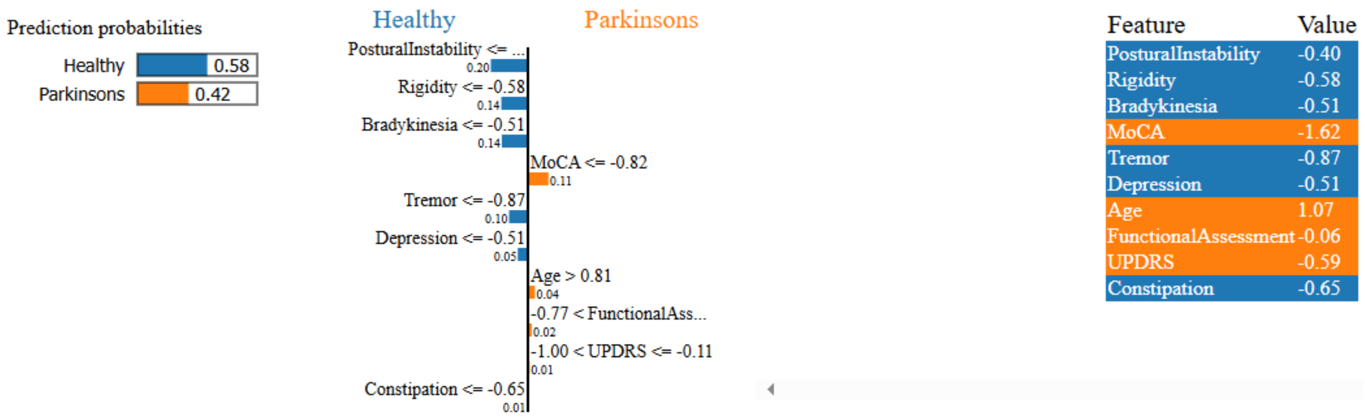
Figure 15. ROC and Confusion Matrix for PD using LSTM.

computing systems typically available in hospital environments. Real time screening assistance therefore becomes feasible during neurological consultations.

Interpretability remains an important requirement in clinical decision support systems. Many predictive algorithms generate outcomes that clinicians find difficult to interpret. In this framework, LIME explanations highlight the contribution of specific clinical variables such as tremor severity, rigidity, cognitive scores, and age. These explanations can be presented alongside prediction outputs. Clinicians may then compare the highlighted variables with observed patient symptoms before considering the system recommendation. Integration into clinical workflows can be implemented through electronic medical record platforms. A patient record containing demographic attributes and neurological assessment scores can be processed automatically by the trained model. A probability score indicating Parkinson related risk can then be generated. An accompanying explanation panel may display influential clinical factors. The system therefore functions as a supportive screening tool rather than a replacement for clinical judgment.

4.2 Limitations

Although the study provides useful insights, several limitations should be acknowledged. The analysis relies on a single publicly available dataset obtained from the UCI Machine Learning Repository. Population diversity remains limited within this dataset. Model performance may vary when applied to patient populations from different healthcare systems or demographic backgrounds. External validation was not conducted. Independent multicenter datasets were not available during the experimental stage. Generalization across hospitals and clinical environments therefore remains unverified. Additional evaluation using independent cohorts would strengthen reliability and clinical credibility. The dataset represents cross sectional clinical observations rather than longitudinal disease monitoring. Parkinson's disease develops gradually



Feature	Value
PosturalInstability	-0.40
Rigidity	-0.58
Bradykinesia	-0.51
MoCA	-1.62
Tremor	-0.87
Depression	-0.51
Age	1.07
FunctionalAssessment	-0.06
UPDRS	-0.59
Constipation	-0.65

Figure 16. LIME explanation for PD after using LSTM.

and symptom progression often unfolds over extended periods. Sequential modeling through LSTM relied on structured feature ordering rather than true temporal patient follow up. This constraint limits representation of long term disease dynamics.

Interpretability analysis was conducted using LIME as a local explanation method. The approach provides instance level insights for individual predictions. Global interpretability techniques were not incorporated in the present framework. Combining multiple explanation approaches may provide a broader understanding of feature influence and model behavior. Clinical deployment also requires additional validation stages. Regulatory approval, hospital system integration, and prospective clinical trials would be necessary before routine use in medical practice. These aspects remain beyond the scope of the present experimental investigation.

4.3 Conclusion

This study presented an explainable artificial intelligence framework for PD detection using structured clinical data. A unified pipeline was implemented that included preprocessing, feature selection, stratified data splitting, and parallel evaluation of machine learning and deep learning models. Random Forest achieved the highest overall discriminative performance, while CNN, SVM, and LSTM demonstrated competitive classification capability across evaluation metrics. The results indicate that reliable detection can be achieved without dependence on imaging or wearable sensor modalities. Interpretability was incorporated through the application of LIME to each trained model, and clinically relevant features such as tremor severity, cognitive assessment scores, and age were consistently identified as influential predictors. Transparent instance level explanations were generated to support clinical reasoning and decision support integration. Although external validation and broader dataset diversity remain necessary, the findings suggest that structured clinical variables can serve as a viable foundation for interpretable PD screening systems in resource constrained settings. In the future, we aim to include external validation using multicenter and longitudinal clinical cohorts. Additionally, multimodal integration with speech and gait data, along with evaluation in real clinical decision support settings, would also be incorporated.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

This study used a publicly available anonymized dataset from the UCI Machine Learning Repository, with no direct involvement of human subjects. Ethical approval and informed consent were not required.

References

[1] Velu, K., & Jaisankar, N. (2025). Design of an Early Prediction Model for Parkinson's Disease Using Machine Learning. *IEEE Access, 13*, 17457-17472. [CrossRef]

[2] Dentamaro, V., Impedovo, D., Musti, L., Pirlo, G., & Taurisano, P. (2024). Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: insights from the PPMI database. *Scientific Reports, 14*(1), 20941. [CrossRef]

[3] Priyadharshini, S., Ramkumar, K., Vairavasundaram, S., Narasimhan, K., Venkatesh, S., Amirtharajan, R., & Kotecha, K. (2024). A comprehensive framework for parkinson's disease diagnosis using explainable artificial intelligence empowered machine learning techniques. *Alexandria Engineering Journal, 107*, 568-582. [CrossRef]

[4] Tolosa, E., Garrido, A., Scholz, S. W., & Poewe, W. (2021). Challenges in the diagnosis of Parkinson's disease. *The Lancet Neurology, 20*(5), 385-397. [CrossRef]

[5] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine, 25*(1), 24-29. [CrossRef]

[6] Sharma, R., Sharma, K. D., & Bijalwan, A. (2025). HDSF: A Healthcare Decision Support Framework to Provide A Seamless and Adaptable Patient Experience. *Biomedical Informatics and Smart Healthcare, 1*(1), 1-8. [CrossRef]

[7] Sivaranjini, S., & Sujatha, C. M. (2020). Deep learning based diagnosis of Parkinson's disease using convolutional neural network. *Multimedia tools and applications, 79*(21), 15467-15479. [CrossRef]

[8] Tewari, Y., Parihar, N. S., Rautela, K., Kaundal, N., Diwakar, M., & Pandey, N. K. (2025). Diabetic retinopathy detection and analysis with convolutional neural networks and vision transformer. *Biomedical Informatics and Smart Healthcare, 1*(1), 18-26. [CrossRef]

[9] Srinivasan, S., Ramadass, P., Mathivanan, S. K., Panneer Selvam, K., Shivahare, B. D., & Shah, M. A. (2024). Detection of Parkinson disease using multiclass machine learning approach. *Scientific Reports, 14*(1), 13813. [CrossRef]

[10] Bauskar, S. (2021). An Analysis: Early Diagnosis and Classification of Parkinson's Disease Using Machine Learning Techniques. *International Journal of Computer Engineering and Technology, 12*(1), 54-66. [CrossRef]

[11] Farooq, W., Ali, A., Fatima, H. M., Rafiq, W., Zainab, N. E., & Ali, M. (2026). FusedCNN-LSTM: A Software-Oriented Multimodal Deep Learning Framework for Intelligent Hypertension Risk Prediction. *ICCK Journal of Software Engineering, 2*(1), 11-29. [CrossRef]

[12] Wiratsin, I.-O., & Ragkhitwetsagul, C. (2025). Effectiveness of Explainable Artificial Intelligence (XAI) Techniques for Improving Human Trust in Machine Learning Models: A Systematic Literature Review. *IEEE Access, 13*, 121326-121350. [CrossRef]

[13] Henninger, M., & Strobl, C. (2025). Interpreting machine learning predictions with LIME and Shapley values: theoretical insights, challenges, and meaningful interpretations. *Behaviormetrika, 52*(1), 45-75. [CrossRef]

[14] Adadi, A., & Berrada, M. (2020). Explainable AI for Healthcare: From Black Box to Interpretable Models. In *Proceedings of the International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering* (pp. 327-337). Springer. [CrossRef]

[15] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58*, 82-115. [CrossRef]

[16] Hossain, M. A., Traini, E., & Amenta, F. (2025). Machine learning applications for diagnosing parkinson's disease via speech, language, and voice changes: A systematic review. *Inventions, 10*(4), 48. [CrossRef]

[17] Shokrpour, S., MoghadamFarid, A., Bazzaz Abkenar, S., Haghi Kashani, M., Akbari, M., & Sarvizadeh, M. (2025). Machine learning for Parkinson's disease: a comprehensive review of datasets, algorithms, and challenges. *npj Parkinson's Disease, 11*(1), 187. [CrossRef]

[18] Nayan, N. M., Rana, A. M., Islam, M. M., Uddin, J., Yasmin, T., & Uddin, J. (2025). An interpretable and balanced machine learning framework for Parkinson's disease prediction using feature engineering and explainable AI. *PLOS One, 20*(10), e0333418. [CrossRef]

[19] Quan, C., Ren, K., & Luo, Z. (2021). A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech. *IEEE Access, 9*, 10239-10252. [CrossRef]

[20] Trabassi, D., Serrao, M., Varrecchia, T., Ranavolo, A., Coppola, G., De Icco, R., ... & Castiglia, S. F. (2022). Machine learning approach to support the detection of Parkinson's disease in IMU-based gait analysis. *Sensors, 22*(10), 3700. [CrossRef]

[21] Escobar-Grisales, D., Ríos-Urrego, C. D., & Orozco-Arroyave, J. R. (2023). Deep learning and artificial intelligence applied to model speech and language in Parkinson's disease. *Diagnostics, 13*(13), 2163. [CrossRef]

[22] Um, T. T., Pfister, F. M., Pichler, D., Endo, S., Lang, M., Hirche, S., ... & Kulić, D. (2017, November). Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM*

international conference on multimodal interaction (pp. 216-220). [CrossRef]

[23] Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., & Faundez-Zanuy, M. (2016). Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. *Artificial Intelligence in Medicine*, *67*, 39-46. [CrossRef]

[24] Gunduz, H. (2019). Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets. *IEEE Access*, *7*, 115540-115551. [CrossRef]

[25] Yu, J., Meng, K., Liang, T., Liu, H., & Wang, X. (2024). Improved deep learning for Parkinson's diagnosis based on wearable sensors. *Electronics*, *13*(23), 4638. [CrossRef]

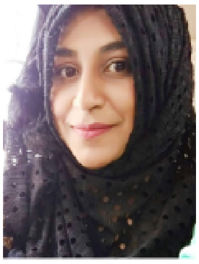
[26] Shi, D., Zhang, H., Wang, G., Wang, S., Yao, X., Li, Y., ... & Ren, K. (2022). Machine learning for detecting parkinson's disease by resting-state functional magnetic resonance imaging: A multicenter radiomics analysis. *Frontiers in aging neuroscience*, *14*, 806828. [CrossRef]

[27] Martinez-Eguiluz, M., Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perona, I., Murueta-Goyena, A., ... & Gabilondo, I. (2023). Diagnostic classification of Parkinson's disease based on non-motor manifestations and machine learning strategies. *Neural Computing and Applications*, *35*(8), 5603-5617. [CrossRef]

[28] Moon, S., Song, H. J., Sharma, V. D., Lyons, K. E., Pahwa, R., Akinwuntan, A. E., & Devos, H. (2020). Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach. *Journal of neuroengineering and rehabilitation*, *17*(1), 125. [CrossRef]

[29] Fu, J., Chen, H., Xu, C., Jia, Z., Lu, Q., Zhang, H., ... & Geng, D. (2025). Harnessing routine MRI for the early screening of Parkinson's disease: a multicenter machine learning study using T2-weighted FLAIR imaging. *Insights into Imaging*, *16*(1), 92-92. [CrossRef]

Aamir Ali is a dedicated researcher in artificial intelligence, healthcare analytics, and IoT environments. His expertise includes machine and deep learning applications in medical diagnosis, speech recognition, and facial emotion recognition. He has also worked on IoT-focused projects such as malware classification, anomaly detection, and cyberattack prediction. Aamir is passionate about developing AI-driven solutions for early disease detection and improving patient care. He actively contributes to interdisciplinary research across clinical and image data domains. (Email: amirali4436823@gmail.com)



Misbah Ali is a PhD scholar at COMSATS university Islamabad, with research interests in Machine Learning, Deep Learning, Generative Artificial Intelligence, and Software Engineering. Her work focuses on the development of secure, intelligent systems across domains such as healthcare, education, and industrial cyber-security. She has authored multiple peer-reviewed publications and presented her research at international conferences. She also contributes to the academic community as a reviewer for several reputed journals. (Email: talktomisbah.ali@gmail.com)