Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(8), 280-299
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Lightweight Interpretable Machine Learning Framework for Parkinson Disease Detection with Feature Selection Technique

Husne Farah[1,2*], Fahmida Islam[1,3], Mohammad Shorif Uddin[2]

[1]Department of Computer Science and Engineering, The People's University of Bangladesh, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh
[3]Department of Information and Communication Engineering, Islamic University, Kushtia, Bangladesh
Email: *husnefarahcse.pub@gmail.com, fahmidaislam.ice@gmail.com, shorifuddin@juniv.edu

How to cite this paper: Farah, H., Islam, F. and Uddin, M.S. (2025) A Lightweight Interpretable Machine Learning Framework for Parkinson Disease Detection with Feature Selection Technique. *Journal of Computer and Communications*, **13**, 280-299.
https://doi.org/10.4236/jcc.2025.138014

Received: July 26, 2025
Accepted: August 25, 2025
Published: August 28, 2025

Abstract

A degenerative neurological condition called Parkinson disease (PD) that evolves progressively, making detection difficult. A neurologist requires a clear healthcare history from the patients, as well as periodic scans, to make the diagnostic. In recent years, AI-based computer-aided diagnostic (CAD) programs have outperformed simpler approaches mainly because of their capacity to predict irregularities in healthcare data. Despite, the intricacy of AI models frequently leads to their employment as "black boxes" that may cause distrust among physicians owing to an absence of transparency regarding decision-making. This study introduces an interpretable machine learning approach to solve these difficulties. This approach offers both regional and worldwide insights for the auxiliary diagnostic of PD while preserving excellent prediction accuracy. This investigation used 894 healthcare instances contained several optimized characteristics. We used a two-stage data preparation strategy to manage extremes and equalize the data while preventing biased outcomes. We simulated multiple state-of-art ML models named boosting, voting and stacking with three features selectors such as mRMR, LDA, and PCA. Among these features selectors and models, the stacking + LDA approach provided the greatest accuracy of 100%. After that, two interpretable AI models named Local Interpretable Model-agnostic Explanations (LIME) and SHapely Adaptive Explanations (SHAP) are implemented for feature interpretability. This feature interpretability makes the proposed approach as a suitable candidate in medical sector.

Keywords

Parkinson Disease, Feature Selection, Interpretable AI, Machine Learning

1. Introduction

Parkinson disease (PD) frequently starts with modest and difficult-to-notice indications that worsen the prediction time. Slow motion, unsteady hands, rigid muscles, difficulty balancing, altered speaking patterns, and diminished facial emotions are typical symptoms. The abbreviation "BITMAP" is used to help recall these signs. Although, prompt identification is challenging since initial signs are sometimes too faint to notice. Parkinson's instances are thus rising quickly. The World Health Organization (WHO) estimates that the total number of individuals with Parkinson's disease increased by two times in 25 years, surpassing around 8.5 million (M) in 2019. In that year, Parkinson's disease caused 329,000 deaths and 5.5 M years of disability, which is twice as many as in 2000 [1]. The number of newly diagnosed PD cases in the US has increased to over 90,000 annually [2]. Women are less likely than men to get PD, which typically manifests around age 60, although it can begin earlier [3]. There is an accumulative demand for innovative techniques that employ machine learning (ML) to help diagnose PD because early indications are difficult to detect. ML can handle large datasets and supports medical personnel by enhancing patient safety, reducing healthcare costs, and improving the quality of treatment [4]. In addition to this, creating ML tools calls for highly qualified personnel and close coordination between technical and medical specialists. ML has demonstrated potential in detecting PD by analyzing data such as speech recordings, brain imaging, and clinical records [5]. Nevertheless, since non-technical individuals often lack an understanding of how these models work, there is public mistrust. This challenge is addressed using explainable AI (XAI) strategies like LIME and SHAP, which clarify framework decisions [6]. Given that speech symptoms are simple, low-cost, and non-invasive, the current study focuses on speech-based prediction of PD [7].

Recently, AI-driven ML approaches are applied in clinical data analysis like feature selection, data cleaning, testing, and classification. These strategies make it easier to find trends in medical information, especially after dealing with normalizing inputs, balancing the data, and outliers. The research contributions are as follows:

1) Two statistical data processing strategies like Winsor and z measure are employed for identifying data outlier and then balanced the data points applying SMOTE Tomek strategy.

2) Three novel feature selectors like mRMR, LDA, and PCA are employed to select top features and then the selected features are provided into three AI-driven ML models like boosting, voting, and stacking. After this experiment, stacking + LDA approach exhibited top results of accuracy 100%.

3) Two interpretable AI models like LIME and SHAP to interpret the predicted features.

The manuscript is separated into several parts: **Section 2** provided a summary of the prior work, **Section 3** demonstrated the methodology of the offered system, **Section 4** exhibited the outcomes with discussion and at last **Section 5** provided a conclusion.

2. Literature Review

The implementation of AI-driven interpretable ML in PD identification improves diagnosis precision and regulation while also allowing primary arbitration and personalized treatment methods. As study in this area involves, the use of interpretable method shows enormous potential for revolutionizing our knowledge and management of PD, eventually resulting to better the outcomes of patients. Several investigators have centered their studies on creating effective interpretable ML methods for PD analysis.

In order to improve categorization efficiency for healthcare and related datasets, a number of current research efforts have investigated several interpretable ML approaches. In article [8], Shastry *et al.* developed the Tree-based Nearest Neighbour (TNN) strategy, which outperformed individual regression models, but lack of statistical feature analysis. In paper [9], Mahesh *et al.* performed several ML approaches like XGBoost, Random Forest (RF), KNN and Support Vector Machine (SVM). Though their approach exhibited remarkable results, but their approach failed to explain clinical features. Another manuscript [10], authors recommended a three-level ML-based strategy. In first level, five base classifiers like KNN, Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT) and SVM were used, a stacked ML model used in second level, and four combined methods *i.e.*, Bagging, AdaBoost, RF, and gradient boosting (GB) used in third level. Among these three levels, the GB method obtained the best precision of 97.43% with low computational cost. Similarly, Oguri *et al.* [11] used four tree-based algorithms—RF, DT, LightGBM, and XGBoost—were employed for PD diagnosis. Though their methods achieved an excellent precision of 97.43%; but their research reflects the black-box nature. In article [12], Nissar *et al.* discovered many updated methods, including NB, LR, DT, SVM, KNN, RF, XGBoost, and MLP for PD diagnosis based on speech features. However, they improved their methods using RFE and mRMR feature optimizers. Nahar *et al.* [13] employed diverse ML models like RFE, bagging, Extra Tree, extreme GB, GB, and RF for PD prediction. However, their method fails to provide feature interpretability. Another study [14]. Saleh *et al.* implemented a hybrid approach using Artificial Neural Network (ANN) and ML classifiers for PD identification. They also applied an ensemble voting classifier with cross-validation that given greatest performance, but they conducted their work on a small dataset. In [15], Asmae *et al.* suggested a novel ML model named stacking that integrates multiple ML models for improved PD prediction. Though they proposed improved model, but their architecture was very complex and time consuming. In [16], authors examined both conventional and combined ML strategies (DT, RF, SVM, LR, bagging, GB, and stacking) for PD diagnosis. Their stacking-based SVM + GB + LR combination demonstrated an excellent precision of 96.05%, but it needed a feature selection technique. Finally, Bukhari *et al.* [17] implemented the AdaBoost classifier with PCA feature selector and SMOTE data balancing method. Although it lacks feature comprehensibility the approach's effectiveness was improved using grid search and cross-validation.

The above mentioned articles have significantly improved healthcare practitioners' capacity to recognize PD in its initial stages. Early recognition of PD is crucial to preventing serious consequences. In addition to this, it is possible to enhance the outcomes and lower the memory cost of a network by putting feature optimization approaches into operation. Moreover, the results produced by these simulations are difficult for healthcare providers to comprehend. In order to guarantee expert understanding of the approach's results, this study used interpretable ML (IML) procedures.

3. Materials and Method

This research offers a reliable and interpretable framework to diagnosis PD using ML and XAI techniques with optimized feature selector method. The conceptual diagram of the offered system is depicted in Figure 1. The suggested system consists of multiple phases: 1) working dataset; 2) data preprocessing; 3) top feature selection; 4) model training; 5) trained model and 6) result evaluation. In data preprocessing phase data outliers are removed using Winsor and z-score approaches. For addressing the data unbalancing situation, the SMOTE-Tomek approach is used. Feature selection is then carried out using three feature selectors, which provided the maximum precision in every combination. Three ML strategy are evaluated on the chosen characteristics to choose the best appropriate algorithm. The ultimate prediction is the technique that performs most accurately. Then, two IAI models named SHAP and LIME are applied to guarantee the method is transparent and reliable.

3.1. Dataset

The dataset was collected from the UCI ML data source which contains 195 speech instances with 24 features [18]. This dataset was generated in 2008 from the University of Oxford, Irvine. The working dataset consists of biomedical speech recordings from persons of various ages, where 147 positive (PD) cases and 48 healthy instances. After that, the "status" features was set into "1" for identifying positive cases and "0" for negative cases. Table 1 summarizes the features of the working dataset. The working dataset was divided into training set (80%) and testing set (20%).

Table 1. Summary of the features of PD dataset.

No. of features	Description
Name	ASCII subject name and recording number
MDVP: Fo (Hz)	Average vocal fundamental frequency (VFF)
MDVP: Fhi (Hz)	Maximum VFF
MDVP: Flo (Hz)	Minimum VFF
MDVP: Jitter (%), Jitter: DDP, MDVP: Jitter (Abs), MDVP: PPQ, MDVP: RAP	Several methods of variation in fundamental frequency/ Multiple indicators of fundamental frequency fluctuation

Continued

Shimmer: DDA, MDVP: Shimmer, MDVP: APQ, MDVP: Shimmer(dB), Shimmer: APQ5, Shimmer: APQ3	Several methods of variation in amplitude/ Multiple amplitude variation measurements
HNR, NHR	Two measures of ratio of noise to tonal components in the voice
Status	Health status: Parkinson's (1) and healthy (0)
D2, RPDE	Two nonlinear dynamical complexity measures
DFA	Signal fractal scaling exponent
Spread1, PPE, spread2	Three nonlinear measures of fundamental frequency variation

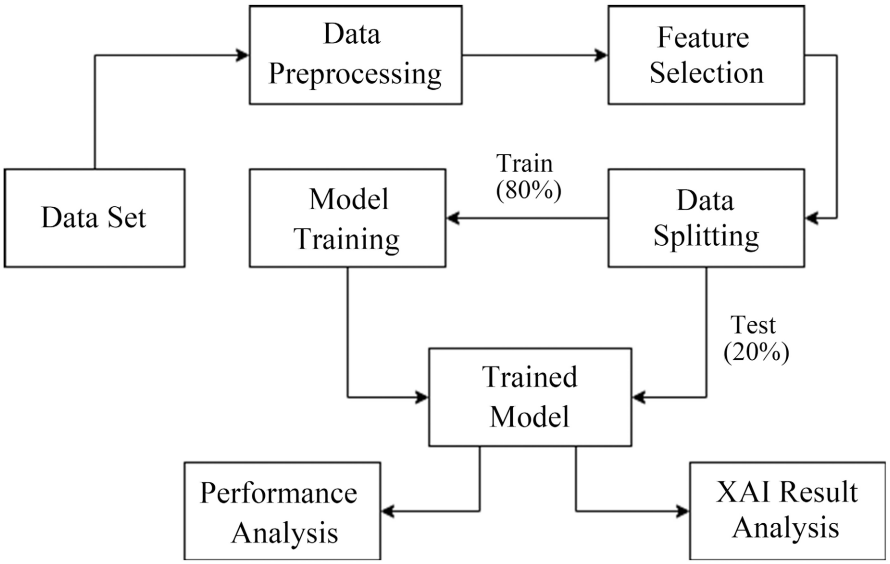

Figure 1. Conceptual diagram of the suggested system.

3.2. Data Preprocessing

The working dataset was processed through the following steps: outlier identification, data augmentation, and balancing. Identifying outliers is a vital phase in preparing data because unusual data layers could have a negative influence on model performance and dependability [19]. Two methods named Winsor and z-score techniques are applied to identify outliers in data layers. The Z-score measures the standard deviation of a data layer using the average value for easily identifying outliers [20] Winsorization or Winsor method [21] is applied for identifying outliers that involves substituting very high or low values of data with fewer severe ones. Instead of eliminating outliers directly, this approach substitutes high values with the nearest values within a certain percentile range. According to this approach, large numbers might occur as a result of measurement mistakes or random oscillations, rather than being outliers.

Data balancing is the crucial part in ML approach which may distort the findings of the test. To overcome this issue, our experiment used a challenging data

balancing strategy known as SMOTE-Tomek. It balances the unbalanced class by integrating SMOTE over-sampling and Tomek under-sampling strategies. It generates simulated specimens in feature set. It chooses a minimal feature set to calculate its k-nearest neighbors [22].

Data augmentation is a critical operation that enriches datasets by using various methods to increase their quantity and variety [23]. Several augmentation strategies have been utilized in this case. When the Tomek connection is established between two scenarios, then almost every feature from the feature set is deleted. This mechanism is useful when a large number of dataset is needed. Applying revised steps of the working data increases the framework's resilience. Table 2 shows the data preprocessing stages for processing PD dataset. Figure 2 shows the data distribution chart. The above two techniques (SMOTE-Tomek and data augmentation) are applied after the train/test split to prevent information leakage.

Table 2. Data preprocessing stages for processing PD dataset.

Step	Method	Description
Outlier detection	Winsor and z-score	Measures standard deviation of the dataset
Data balancing	SMOTE-Tomek	Generates simulated specimens in the feature space
Data augmentation	Traditional technique	Randomly added number of data

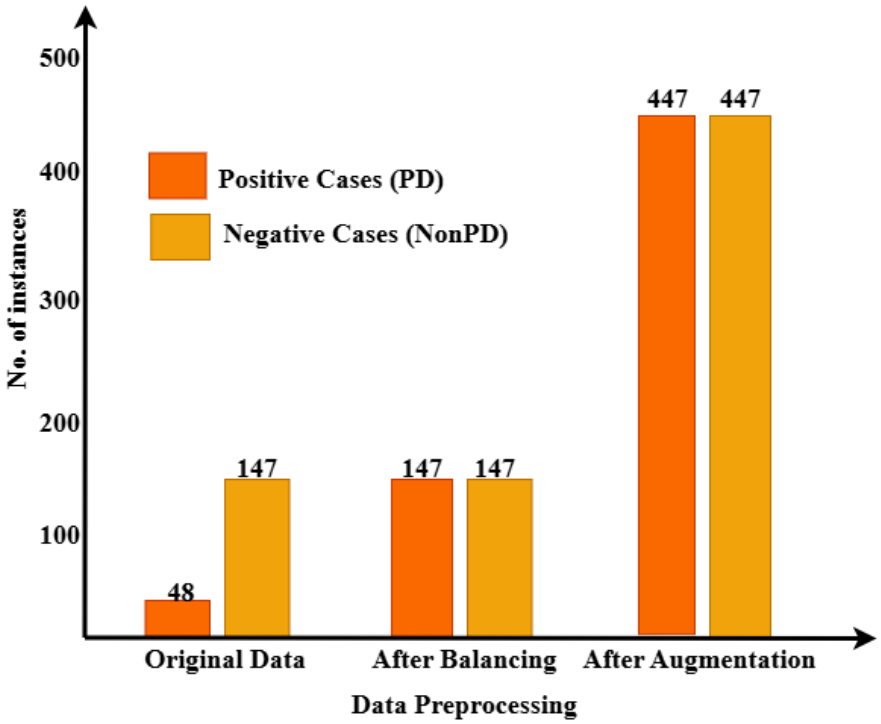

Figure 2. Data distribution chart.

3.3. Feature Selection Approach

Feature selection approach selects high impactful features from the final data set to boost the efficacy of the suggested method. Several feature selection techniques *i.e.*, mRMR for feature selection, LDA and PCA for feature reduction are used in feature engineering field. Among these techniques, LDA exhibited outstanding performance. The working principle of this technique is described in below.

Linear Discriminant Analysis (LDA)

The dimensionality reduction approach named LDA is widely utilized in feature reduction and selection problems. It aims to identify the linear combination of best features (LCBF) from the entire dataset. It transforms the higher dimensional-feature space (DFS) into a lower DFS based on the LCBF [24]. It improves the ratio of class variance, resulting in the optimum routes in the LCBF for future differentiation [25]. It boosts class reparability, integrates classified data for supervised training, and enhances the effectiveness of classification, especially in cases when classes are well divided [26].

Let S be the dataset with f features and n instances that is partitioned into m classes. The output of the feature set F is defined by Equation (1).

$$F_j = \frac{1}{I_j} \sum_{j:u_j=m} v_j \tag{1}$$

Where, I_j represents the instances for feature m, u_j indicates class label for instance j, and v_j indicates feature map for instance j. The class scatters inside (L_{in}) and outside (L_{out}) the feature map are defined by Equation (2) and Equation (3), respectively.

$$L_{in} = \sum_{j=1}^{n} I_j \left(F_j - F\right)\left(F_j - F\right)^t \tag{2}$$

$$L_{out} = \sum_{j=1}^{n} \sum_{j:u_j=m} \left(u_j - F_m\right)\left(u_j - F_m\right)^t \tag{3}$$

3.4. Machine Learning Model

Three ensemble ML models named boosting, voting, and stacking are applied in this work to predict PD. Each ensemble ML model is the combination of multiple ML classifier which boosts the prediction accuracy by leveraging the power of multiple ML models [27]. By leveraging the variety of the basic classifiers, each model can capture many data attributes while limiting the risk of overfitting [28]. Among these models, stacking provides best results. The working mechanism of this model is described below.

Stacking

Stacking is an ensemble ML model that combines several base classifiers (BCL) [29] and trained them to produce unique prediction. Then these unique predictions are stacked and transmitted into a meta-classifier (MTC) [30] to produce

final prediction. The final prediction of this model is defined by Equation (4). The main concept underlying stacking is to train the base predictors on a single dataset concurrently. In working dataset, the training samples are represented by S and the predicted outcome is p_k; where k indicates the number of classifiers. Lastly, the predicted results are represented by Equation (4).

$$\hat{\boldsymbol{p}} = \sum_{\boldsymbol{k}=1}^{\boldsymbol{S}} \boldsymbol{p_k}(\boldsymbol{q}) \tag{4}$$

These predictions are used as input for the MTC. The output of the MTC for data point q is calculated by vector P^(q)and defined by Equation (5). The desired result p^is obtained by applying the MTC named LR [31] on the BCL output P^(q). The final outcomes are calculated by Equation (6).

$$\hat{P}(q) = \left[\hat{p}_1(q), \hat{p}_1(q), \cdots, \hat{p}_s(q)\right] \tag{5}$$

$$\hat{p} = f\left(\hat{P}(q)\right) \tag{6}$$

3.5. Proposed Framework for Parkinson Disease Prediction

Algorithm 1 Step by step working procedure of the suggested system.

start

input:

Training instances, $D_{u,v} = \left[(u_1, v_1), (u_2, v_2), \cdots, (u_t, v_t)\right]$;

L = Number of base learners;

Output:

Final predicted result from stacking classifier, C_{stack};

Phase 1: Train base classifiers

for l = 1 to L:

$C_{base}^{(l)}$ = train base classifier l on $D_{u,v}$;

end for

Phase 2: Generate meta level dataset

for each training instance u_i, where i = 1 to k, do:

Obtain predictions from all base classifiers;

$\hat{u}_i = \left[C_{base}^{(1)}(u_1), C_{base}^{(2)}(u_2), \cdots, C_{base}^{(L)} u_i\right]$;

Create new dataset $\hat{D}_{(u,v)} = (\hat{u}_i, v_i)_{i=1}^{k}$;

for end

Phase 3: Train meta classifier for final prediction

Train the meta classifier $\hat{M}$ on dataset $\hat{D}_{(u,\hat{v})}$;

return $C_{stack}(u) = \hat{C}\left(C_{base}^{(1)}(u), C_{base}^{(2)}(u), \cdots, C_{base}^{(L)}(u)\right)$;

end

The current investigation averages the expected likelihoods of each category and chooses the class with the greatest mean possibility as the end result. A unique technique is utilized to evaluate a MTC on the BCL projections in order to reach the ultimate choice, with the goal of learning how to optimally integrate the assumptions. **Figure 3** depicts the entire architecture of the suggested framework

concept used in this article. Algorithm (1) illustrates how the offered frame work produces final result step by step. The incorporation of numerous models into the suggested model may raise computing demands and make the system more difficult to administer.

To deal with this drawback, four feature analysis techniques are utilized to minimize the complexity of the model by lowering the dimensionality of the feature space and simplifying the framework. LDA determines the optimum linear combinations to distinguish classes, PCA reduces the initial attributes to make an optimal feature set and mRMR chooses the greatest number of significant and minimal redundant characteristics. By lowering the number of characteristics, these strategies result in quicker computing, a decreased risk of excessive fitting, and more comprehensible scenarios, eventually making the algorithm more effective and manageable.

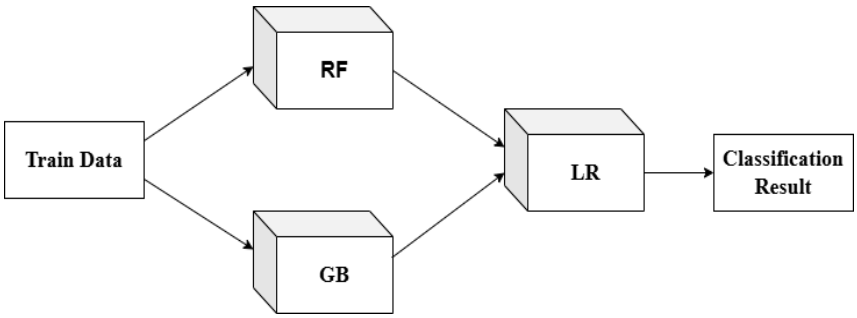

Figure 3. Internal structure of the stacking model. Here RF and GB are the base learners, and LR is the meta learner.

3.6. Interpretable Artificial Intelligence Approach

To develop a dependable and transparent system for PD diagnosis, it is critical to graphically display and explain how ML models make judgments [6]. The use of interpretable AI approach is crucial for verifying the final decision of the ML classifier [32] [33]. This research reflects two interpretable methods' named LIME and SHAP that increase the accessibility and interpretation of the final results.

3.6.1. SHapley Additive exPlanations (SHAP)

SHAP is a collaborative theory-based strategy to interpret the final outcomes of the ML classifier [34]. In collaborative strategy, the Shapley result is a way for evenly distributing the "payout" among individuals on the basis of marginal contribution. In this strategy, each feature in the prediction is treated as a player, and the SHAP score measures its individual contribution by comparing the model's output with and without that feature, relative to the average prediction [35]. SHAP evaluates all conceivable feature subsets (coalitions) and estimates each feature's contribution to the probability through evaluating the modification in prognosis when the attribute is added to the coalition.

The Shapley value δ_m for m^{th} feature quantifies the average feature contribution of the model's prediction and considers all probable subsets of features that

is defined by Equation (7). In this context, *y(w)* denotes the feature map for feature *w* and it can be calculated as the difference between the final estimation and the baseline estimation of the proposed model. The possible feature set is represented by 2^k for all *k*. To compute a Shapley value, SHAP evaluates the marginal contribution of feature m across these subsets and then averages the results over all possible permutations. As shown in Equation (8), the model's outcome for a specific input p is expressed as the sum of a baseline value δ_0 and the individual contributions of all attributes. The final outcome from the SHAP algorithm is denoted by $\hat{g}(p)$ and calculated by Equation (8).

$$\delta_m = \sum_{w \subseteq k-\{m\}} \frac{|w|!.(|k|-|w|-1)!}{|k|!}\left(y(w \cup \{m\}) - y(w)\right) \tag{7}$$

$$\hat{g}(p) = \delta_0 + \sum_{m=1}^{k} \delta_m \tag{8}$$

3.6.2. Local Interpretable Model-Agnostic Explanations (LIME)

This mechanism emphasizes the projected outcomes for particular scenarios above providing a thorough understanding of the system across the entire dataset [36]. Applying the LIME methodology offers useful knowledge into how many aspects affect PD, distinguishing those that participate positively from those that have a negative effect. LIME aims to interpret the selected features (represented by Z) by analyzing the local vicinity of that feature. To achieve this, it generates a set perturbed samples z', which are slight disparity of z. These perturbed samples form a neighborhood around z. Each sample z' is allocated a proximity weight ϖ(Z') which reflects its similarity to z. From this set, a subset z'ϵ Z of perturbed samples is chosen, and each is weighted based on its closeness to z, as calculated in Equation (9). Then an explainable network M_L is examined on these weighted samples locally to find out the estimated predictions of the local model M_L. These local predictions are calculated by Equation (10).

$$\omega(Z') = \frac{\pi(z')}{\sum_{z'_k \in Z'_N} \pi(z'_k)} \tag{9}$$

$$M_I(z) = \arg\min_z \sum_{z'_k \in Z'_N} \omega(z'_k).L_f(M_L(z), M_L(z'_k) + \delta(M_I) \tag{10}$$

In Equation (10), L_f is the loss function that is measured using the difference between M_L and M_I over the local neighborhood, and $\delta(M_L)$ is the regularization parameter of the interpretable model M_I.

4. Experimental Result Analysis

This part of the study compares different feature selection techniques applied to various ML algorithms, emphasizing the interpretability of the selected features.

The instance from the test set are collected using two-step preprocessing phases: outlier removal and data balancing with augmentations. This study applies four

statistical feature selectors to collect best features from the entire characteristics. Three ML models are applied to predict PD from normal cases based on the selected best features. In this work, all ML models were simulated on a platform with specifications: Intel Core i9-14900KS CPU, NVIDIA® GeForce RTX™ 5090 GPU 1 TB of disk space, 128 MB of cache, and 64 GB of RAM.

In this experiment, four-evaluation matrix such as, Precision (Pr), Accuracy (Ac), F1-measure (Fm), and Recall (Rc) are used to evaluate the proposed system that are formulated in Table 3. In Table 3, TN indicates True Negative, TP indicates True Positive, FN indicates False Negative and FP indicates False Positive.

Table 3. Performance evaluation matrices with equation.

Metrics	Formula
Accuracy (Ac)	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision (Pr)	$\frac{TP}{TP+FP}$
Recall (Rc)	$\frac{TP}{TP+FN}$
F1-measure (Fm)	$2\times\frac{Pr\times Rc}{Pr+Rc}$

4.1. Result Analysis with ML Method

Firstly, we evaluate three ML models without feature selection technique that is shown in Table 4. Table 4 indicates that the stacking network provided high classification rate. We also show that the accuracy of the voting model is lower than the stacking network. However, we trained these ML models with three feature selection algorithms to increase the classification rate. The experimental results of all ML models with three feature selection techniques are demonstrated in Table 5. Table 5 reflects that the stacking EML model with LDA feature selector provides best results. The voting ML model with mRMR shows the lowest accuracy than other models. On the other hand, boosting and stacking models are provided same and highest precision of 100% using LDA feature selector. In Table 5, the boosting model with PCA provides lowest recall value of 85.26% and the voting model with mRMR provides lowest F1 value of 85.72%.

Table 4. Experimental results of all ML models without feature selection techniques.

ML Model	Ac	Fm	Pr	Rc
Voting	0.7765	0.8701	0.971	0.7882
Stacking	0.9274	0.9257	0.9101	0.9419
Boosting	0.8436	0.8264	0.7667	0.8961

Table 5. Experimental results of all ML models with feature selection techniques.

ML Model	Feature Selector Technique	Ac	Pr	Rc	Fm
Voting	mRMR	0.8659	0.809	0.9114	0.8572
	LDA	0.9664	0.9775	0.956	0.9666
	PCA	0.8939	0.809	0.9351	0.8675
Stacking	mRMR	0.9441	0.8989	0.9877	0.9412
	LDA	0.1	0.1	0.1	0.1
	PCA	0.9832	0.9775	0.9886	0.983
Boosting	mRMR	0.8771	0.809	0.9351	0.8675
	LDA	0.9944	0.1	0.9888	0.9943
	PCA	0.8771	0.9101	0.8526	0.8804

In this experiment, we implement k-fold (k = 5) cross-validation on the PD dataset to assess the model's robustness and generalizability. The PD dataset is fold into 5 parts and then trained each part using the best Stacking model. Table 6 reflects the experimental results of the Stacking model with 5-fold cross validation.

Table 6. Experimental results of the Stacking model with 5-fold cross validation.

No. of fold	Ac	Pr	Rc	Fm
K = 1	0.9744	0.9677	1.0000	0.9836
K = 2	0.9231	0.9355	0.9667	0.9508
K = 3	0.9231	0.9333	0.9655	0.9492
K = 4	0.8718	0.8529	1.0000	0.9206
K = 5	0.8205	0.8235	0.9655	0.8889

In this experiment, diverse statistical metrics like classification report, ROC curve, confusion metric, and AUC-ROC curve for all ML models are shown to evaluate the proposed stacking + LDA model. Figure 4 represents different evaluation metrics of the proposed model.

The LDA was fitted on the training portion of the dataset instead of refitting on the full dataset before testing to avoid information leakage issue. This was ensured by placing the model within the cross-validation loop. At no point was the model trained on or exposed to the test data prior to evaluation.

Recently, many authors have developed PD disease prediction system applying four ML models. Though their proposed system was provided outstanding results, but their proposed system has some pitfalls like data imbalances and irregularly optimum characteristics set. That's why three feature selection methods are employed to mitigate these pitfalls. We compare our proposed work with the previous works in Table 7.

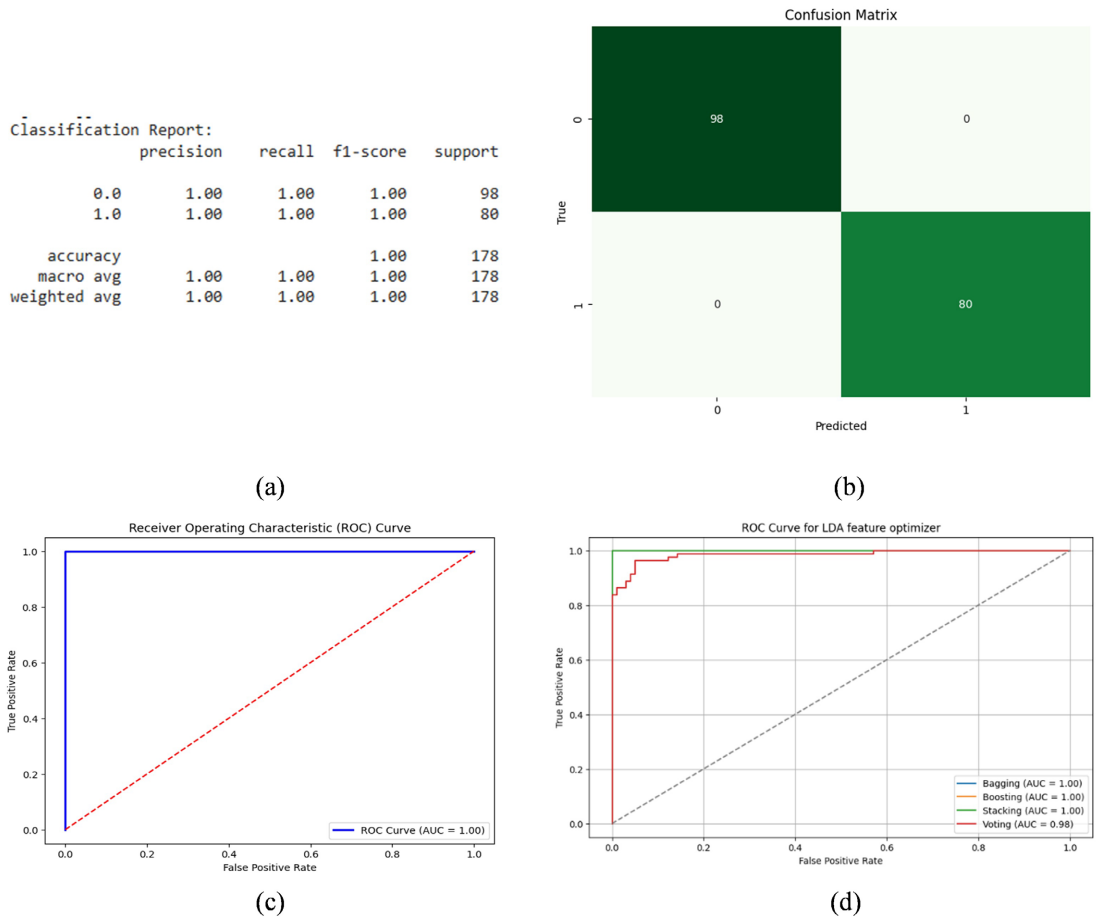

Figure 4. Evaluation metrics for stacking + LDA model: (a) classification report, (b) confusion matrix, (c) ROC curve and (d) ROC curve for all ML models.

Table 7. Result analysis of the current study and the previous studies.

Ref./Year	Dataset	Approach	Performance (%)
[37]/2020	Max Little	PCA + BPVAM	Ac = 97.5
[38]/2022	Max Little	RF + Genetic Algorithm + SMOTE	Ac = 95.58
[39]/2023	Max Little	GridSearchCV + SMOTE + MLP	Ac = 98.31, Pr = 100, Rc = 98, Fm = 99
[40]/2023	Max Little	LSTM + Hybrid GRU	Ac=98
[17]/2024	Sakar	AdaBoost + PCA + SMOTE	Ac = 96, Pr = 98, Rc = 93, Fm = 95
[41]/2024	Max Little	LightGBM	Rc = 100, Ac = 95, Fm = 90, Pr = 93.3
[16]/2024	Max Little + Sakar	Stacking	Ac = 96
[42]/2024	Max Little	SMOTE + RF + XGBoost	Ac = 98, Pr = 97.24, Rc = 97.56, Fm = 97.40
This work	Max Little	Stacking + LDA + SMOTE-Tomek	Ac = **100**, Pr = **100**, Rc = **100**, Fm = **100**

The proposed stacking model demonstrates competitive performance compared to state-of-the-art deep learning (DL) approaches for PD detection, particularly when evaluated using k-fold cross-validation to ensure robustness and generalizability. Unlike deep models such as CNNs, LSTMs, or Transformers, the stacking model is lightweight, easy to deploy, and computationally efficient. DL models require large datasets, significant computational resources and complex implementation, where our proposed system is suitable for real-time applications. While DL models may achieve slightly higher accuracy on large datasets, the stacking model maintains high predictive accuracy on smaller datasets with lower training time and inference cost. Moreover, it offers better model interpretability through XAI tools, which is crucial for clinical decision support. This makes the proposed approach not only effective but also practical for healthcare deployment scenarios where transparency, speed, and cost-efficiency are essential.

4.2. Result Analysis with XAI Method

4.2.1. SHAP Result Analysis

The output of the SHAP algorithm for the PD features are shown in **Figure 5.** In Figure 5, x-axis shows the predicted SHAP scores that impact on framework output and y-axis indicates the overall importance of individual characteristic. Each dot represents feature contributions for specific data instances, with their position along the x-axis showing the extent of the influence. The color of each dot reflects the actual value of the characteristic—red for max values and blue for min values—highlighting how different value ranges influence the estimation. This visualization offers a strong interpretable overview of which features drive the method's decisions.

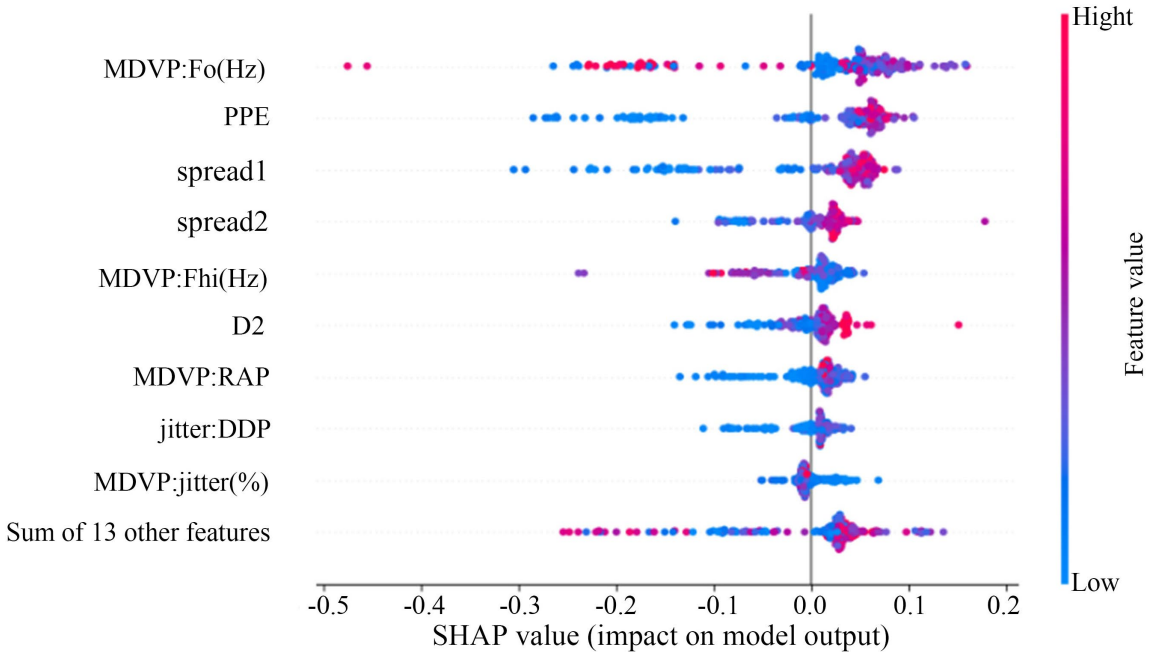

Figure 5. Beeswarm output of the SHAP algorithm.

Figure 6 reflects the bar chart output of the SHAP algorithm. In Figure 6, x-axis shows the mean absolute SHAP value and y-axis lists the individual features displaying the mean SHAP values of different features. From Figure 6, we see that three features named MDVP: Fo (Hz), PPE, and spread1 have the highest SHAP value where the MDVP: jitter (%) has the lowest SHAP value (+0.01). The bar labeled "Sum of 13 other features" reflects the combined contribution of less influential features. This plot provides a clear summary of which features are most significant in driving the model's decisions.

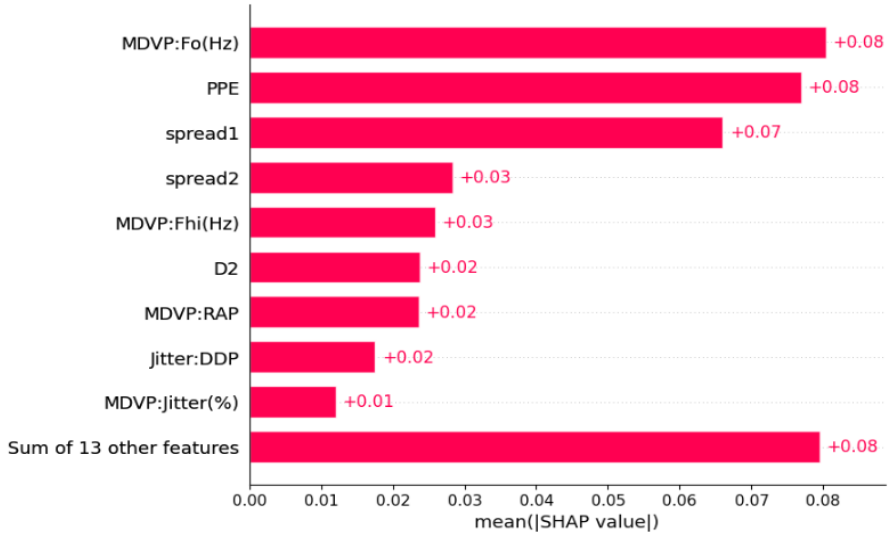

Figure 6. Bar chart output of the SHAP algorithm.

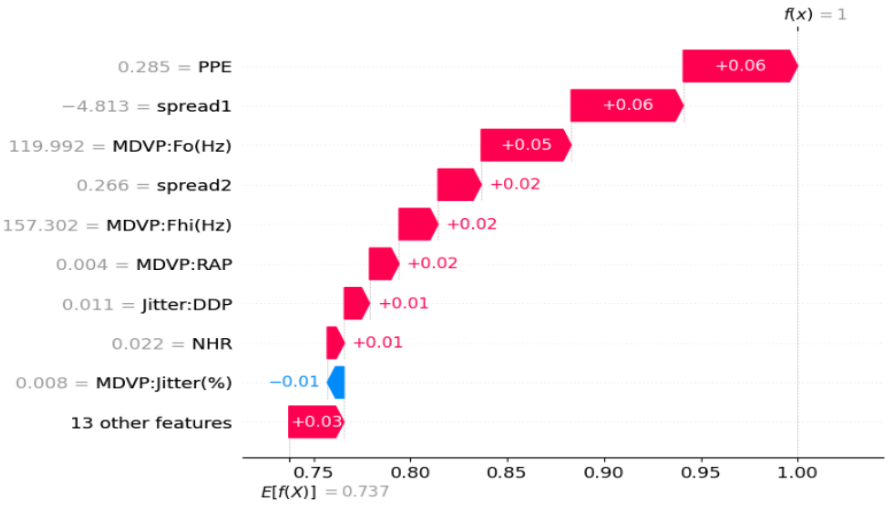

Figure 7. Waterfall output of the SHAP algorithm.

Figure 7 reflects the waterfall output of the SHAP algorithm. The base value (E[f(x)] = 0.737E [f(x)] = 0.737E [f(x)] = 0.737) represents the average model output across all data. In Figure 7, the red arrow indicates that the prediction of the features are increased, while blue arrow represents that the prediction of the fea-

tures are decreased. For example, features like "PPE", "spread1", and "MDVP: Fo(Hz)" pushed the prediction higher, while "MDVP: Jitter(%)" slightly decreased it. This plot provides an interpretable breakdown of how specific feature values influenced the final prediction for one particular data point.

4.2.2. LIME Result Analysis

Figure 8 reflects a decision plot of a ML model using LIME algorithm. In Figure 8, the bar chart at the top left shows the final predicted probabilities—0.08 for PD and 0.92 for Non-PD. The center portion contains a tree-based breakdown of decision paths, where each branch shows how specific feature thresholds influence the prediction toward either class. Each decision node includes the feature, its condition, and the corresponding contribution to the framework's prediction. On the right, a summary table lists the actual results of the best characteristic used in the estimation. Most feature contributions in this example push the prediction toward the Non-PD category, aligning with the high Non-PD probability score. This visualization provides a step-by-step explanation of how individual features and their thresholds contributed to the final classification decision.

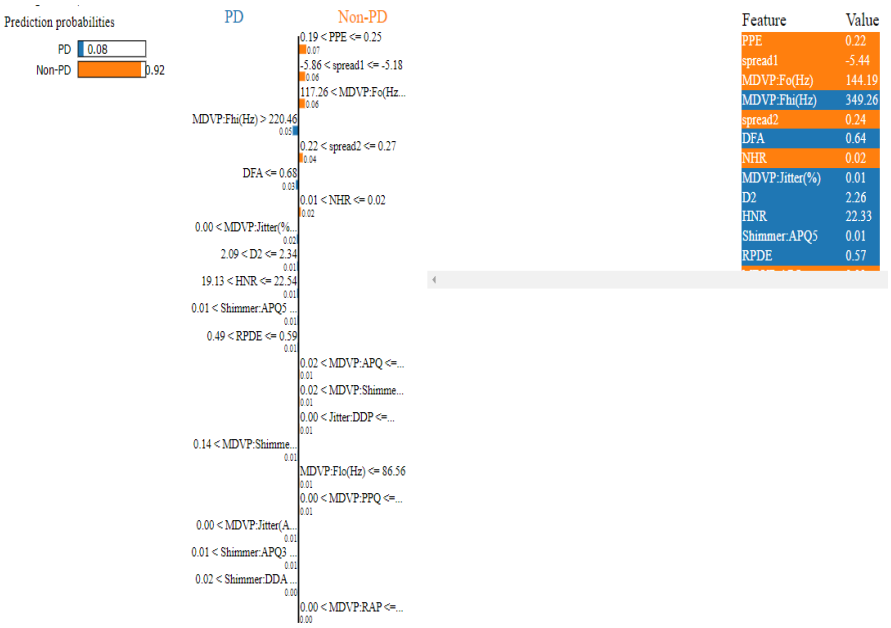

Figure 8. Graph for investigating prediction probabilities using LIME algorithm.

5. Conclusion and Future Scope

Efficient identification of Parkinson Disease (PD) is critical for appropriate therapy. This study offers a smart scheme for by leveraging ML and explainable AI with optimized feature. In this work, at first, three data preprocessing phases: data misbalancing, augmentation to enhance dataset amount, and outlier identification are applied to perform PD diagnosis. Then, three feature optimization strategies (PCA, mRMR, and LDA) are implemented on three ML models to retrieve the important characteristics. Although feature selection strategy increases the ef-

fectiveness of this system, choosing the appropriate approach is critical. This study revealed that not every feature optimization strategy increased model efficiency; a few even scored poorly than using without optimization. The simulated results reflects that the LDA feature optimizer produced a best precision of 100% when applied the stacking model. In testing phase, the stacking strategy beat all other approaches in given feature set. To assure the framework's receptiveness and validity, we used two XAI methods: LIME and SHAP. In future we have a plan to increase the dataset and apply federated ML approach to increase the privacy preserving in healthcare domain.

Data Availability Statement

The working dataset can be donloaded via the link: https://archive.ics.uci.edu/dataset/174/parkinsons.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Schiess, N., Cataldi, R., Okun, M.S., Fothergill-Misbah, N., Dorsey, E.R., Bloem, B.R., *et al.* (2022) Six Action Steps to Address Global Disparities in Parkinson Disease. *JAMA Neurology*, **79**, 929-936. https://doi.org/10.1001/jamaneurol.2022.1783

[2] Reddy, A., Reddy, R.P., Roghani, A.K., Garcia, R.I., Khemka, S., Pattoor, V., *et al.* (2024) Artificial Intelligence in Parkinson's Disease: Early Detection and Diagnostic Advancements. *Ageing Research Reviews*, **99**, Article 102410. https://doi.org/10.1016/j.arr.2024.102410

[3] Mawe, G.M., Browning, K.N., Manfredsson, F.P., Camilleri, M., Hamilton, F.A., Hollander, J.A., *et al.* (2022) 2021 Workshop: Neurodegenerative Diseases in the Gut-Brain Axis—Parkinson'S Disease. *Gastroenterology*, **162**, 1574-1582. https://doi.org/10.1053/j.gastro.2022.02.004

[4] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., *et al.* (2017) Artificial Intelligence in Healthcare: Past, Present and Future. *Stroke and Vascular Neurology*, **2**, 230-243. https://doi.org/10.1136/svn-2017-000101

[5] Rana, A., Dumka, A., Singh, R., Rashid, M., Ahmad, N. and Panda, M.K. (2022) An Efficient Machine Learning Approach for Diagnosing Parkinson's Disease by Utilizing Voice Features. *Electronics*, **11**, Article 3782. https://doi.org/10.3390/electronics11223782

[6] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, **30**, 4768-4777.

[7] Sakar, B.E., Isenkul, M.E., Sakar, C.O., Sertbas, A., Gurgen, F., Delil, S., *et al.* (2013) Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings. *IEEE Journal of Biomedical and Health Informatics*, **17**, 828-834. https://doi.org/10.1109/jbhi.2013.2245674

[8] Shastry, K.A. (2023) Ensemble Machine Learning Regression Model Based Predictive Framework for Parkinson's UPDRS Motor Score Prediction from Speech Data. *International Journal of Speech Technology*, **26**, 433-457. https://doi.org/10.1007/s10772-023-10026-z

[9] T.R., M., V., V.K., Bhardwaj, R., Khan, S.B., Alkhaldi, N.A., Victor, N., *et al.* (2024) An Artificial Intelligence-Based Decision Support System for Early and Accurate Diagnosis of Parkinson's Disease. *Decision Analytics Journal*, **10**, Article 100381. https://doi.org/10.1016/j.dajour.2023.100381

[10] Chaurasia, V. and Chaurasia, A. (2023) Detection of Parkinson's Disease by Using Machine Learning Stacking and Ensemble Method. *Biomedical Materials & Devices*, **1**, 966-978. https://doi.org/10.1007/s44174-023-00079-8

[11] Oguri, V.S.B., Poda, S., Satya, A.K. and NK Prasanna, P. (2023) Parkinson's Disease Detection Using Tree Based Machine Learning Algorithms. *Current Trends in Biotechnology and Pharmacy*, **17**, 808-818. https://doi.org/10.5530/ctbp.2023.2.19

[12] Nissar, I., Raza Rizvi, D., Masood, S. and Nazir Mir, A. (2019) Voice-Based Detection of Parkinson's Disease through Ensemble Machine Learning Approach: A Performance Study. *EAI Endorsed Transactions on Pervasive Health and Technology*, **5**, e2. https://doi.org/10.4108/eai.13-7-2018.162806

[13] Nahar, N., Ara, F., Neloy, M.A.I., Biswas, A., Hossain, M.S. and Andersson, K. (2021) Feature Selection Based Machine Learning to Improve Prediction of Parkinson Disease. In: Mahmud, M., Kaiser, M.S., Vassanelli, S., Dai, Q. and Zhong, N., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 496-508. https://doi.org/10.1007/978-3-030-86993-9_44

[14] Saleh, S., Cherradi, B., El Gannour, O., Hamida, S. and Bouattane, O. (2024) Predicting Patients with Parkinson's Disease Using Machine Learning and Ensemble Voting Technique. *Multimedia Tools and Applications*, **83**, 33207-33234. https://doi.org/10.1007/s11042-023-16881-x

[15] Asmae, O., Saleh, S., Abdelhadi, R. and Bachir, B. (2024) Enhancing Parkinson's Disease Diagnosis: A Stacking Ensemble Approach Leveraging Machine Learning Techniques. 2024 4*th International Conference on Innovative Research in Applied Science*, *Engineering and Technology* (*IRASET*), FEZ, 16-17 May 2024, 1-7. https://doi.org/10.1109/iraset60544.2024.10549375

[16] Al-Tam, R.M., Hashim, F.A., Maqsood, S., Abualigah, L. and Alwhaibi, R.M. (2024) Enhancing Parkinson's Disease Diagnosis through Stacking Ensemble-Based Machine Learning Approach. *IEEE Access*, **12**, 79549-79567. https://doi.org/10.1109/access.2024.3408680

[17] Bukhari, S.N.H. and Ogudo, K.A. (2024) Ensemble Machine Learning Approach for Parkinson's Disease Detection Using Speech Signals. *Mathematics*, **12**, Article 1575. https://doi.org/10.3390/math12101575

[18] Dataset Link. https://archive.ics.uci.edu/dataset/174/parkinsons

[19] Boukerche, A., Zheng, L. and Alfandi, O. (2020) Outlier Detection: Methods, Models, and Classification. *ACM Computing Surveys*, **53**, 1-37. https://doi.org/10.1145/3381028

[20] Martinez-Millana, A., Hulst, J.M., Boon, M., Witters, P., Fernandez-Llatas, C., Asseiceira, I., *et al.* (2018) Optimisation of Children Z-Score Calculation Based on New Statistical Techniques. *PLOS ONE*, **13**, e0208362. https://doi.org/10.1371/journal.pone.0208362

[21] Hoo, K.A., Tvarlapati, K.J., Piovoso, M.J. and Hajare, R. (2002) A Method of Robust Multivariate Outlier Replacement. *Computers & Chemical Engineering*, **26**, 17-39. https://doi.org/10.1016/s0098-1354(01)00734-7

[22] Fernandez, A., Garcia, S., Herrera, F. and Chawla, N.V. (2018) SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary. *Journal of Artificial Intelligence Research*, **61**, 863-905.

https://doi.org/10.1613/jair.1.11192

[23] Park, D.S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E.D., *et al.* (2019) SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *INTERSPEECH* 2019, Graz, 15-19 September 2019, 2613-2617. https://doi.org/10.21437/interspeech.2019-2680

[24] Mostafiz, R., Rahman, M.M., Kumar, P.K.M. and Islam, M.A. (2018) Speckle Noise Reduction for 3D Ultrasound Images by Optimum Threshold Parameter Estimation of Bi-Dimensional Empirical Mode Decomposition Using Fisher Discriminant Analysis. *International Journal of Signal and Imaging Systems Engineering*, **11**, 93-101. https://doi.org/10.1504/ijsise.2018.091886

[25] Amin, S. and Singhal, A. (2017) Identification and Classification of Neuro-Degenerative Diseases Using Feature Selection through PCA-LD. 2017 4*th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics* (*UPCON*), Mathura, 26-28 October 2017, 578-586. https://doi.org/10.1109/upcon.2017.8251114

[26] Mostafiz, R., Rahman, M.M., Mithun Kumar, P.K. and Islam, M.A. (2017) Speckle Noise Reduction for 3-D Ultrasound Images by Optimum Threshold Parameter Estimation of Wavelet Coefficients Using Fisher Discriminant Analysis. *International Journal of Imaging and Robotics*, **17**, 73-88.

[27] Polikar, R. (2012) Ensemble Learning. In: Zhang, C. and Ma, Y., Eds., *Ensemble Machine Learning*, Springer, 1-34. https://doi.org/10.1007/978-1-4419-9326-7_1

[28] Bind, S., *et al.* (2015) A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction. *International Journal of Computer Science and Information Technol Technology*, **6**, 1648-1655.

[29] Wolpert, D.H. (1992) Stacked Generalization. *Neural Networks*, **5**, 241-259. https://doi.org/10.1016/s0893-6080(05)80023-1

[30] Liang, M., Chang, T., An, B., Duan, X., Du, L., Wang, X., *et al.* (2021) A Stacking Ensemble Learning Framework for Genomic Prediction. *Frontiers in Genetics*, **12**, Article 600040. https://doi.org/10.3389/fgene.2021.600040

[31] Dreiseitl, S. and Ohno-Machado, L. (2002) Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review. *Journal of Biomedical Informatics*, **35**, 352-359. https://doi.org/10.1016/s1532-0464(03)00034-0

[32] Biswas, S., Mostafiz, R., Paul, B.K., Uddin, K.M.M., Hadi, M.A. and Khanom, F. (2024) DFU_XAI: A Deep Learning-Based Approach to Diabetic Foot Ulcer Detection Using Feature Explainability. *Biomedical Materials & Devices*, **2**, 1225-1245. https://doi.org/10.1007/s44174-024-00165-5

[33] Biswas, S., Mostafiz, R., Uddin, M.S. and Paul, B.K. (2024) XAI-Fusionnet: Diabetic Foot Ulcer Detection Based on Multi-Scale Feature Fusion with Explainable Artificial Intelligence. *Heliyon*, **10**, e31228. https://doi.org/10.1016/j.heliyon.2024.e31228

[34] Alotaibi, A., Alnajrani, L., Alsheikh, N., Alanazy, A., Alshammasi, S., Almusairii, M., *et al.* (2023) Explainable Ensemble-Based Machine Learning Models for Detecting the Presence of Cirrhosis in Hepatitis C Patients. *Computation*, **11**, Article 104. https://doi.org/10.3390/computation11060104

[35] Haneesha Samudrala, S.S., Thambi, J., Vadluri, S.R., Mahalingam, A. and Pati, P.B. (2024) Enhancing Parkinson's Disease Diagnosis Using Speech Analysis: A Feature Subset Selection Approach with LIME and SHAP. 2024 3*rd International Conference for Innovation in Technology* (*INOCON*), Bangalore, 1-3 March 2024, 1-5. https://doi.org/10.1109/inocon60754.2024.10511805

[36] Tiwari, U., Jahanve, P.R., Karna, S., M, A., Pati, P.B. and KN, B.P. (2024) Parkinson's

Disease Severity Assessment: A Comparative Study & Interpretability Analysis. 2024 5*th International Conference for Emerging Technology* (*INCET*), Belgaum, 24-26 May 2024, 1-5. https://doi.org/10.1109/incet61516.2024.10593180

[37] Rasheed, J., Hameed, A.A., Ajlouni, N., Jamil, A., Ozyavas, A. and Orman, Z. (2020) Application of Adaptive Back-Propagation Neural Networks for Parkinson's Disease Prediction. 2020 *International Conference on Data Analytics for Business and Industry*: *Way Towards a Sustainable Economy* (*ICDABI*), Sakheer, 26-27 October 2020, 1-5. https://doi.org/10.1109/icdabi51230.2020.9325709

[38] Lamba, R., Gulati, T., Alharbi, H.F. and Jain, A. (2022) A Hybrid System for Parkinson's Disease Diagnosis Using Machine Learning Techniques. *International Journal of Speech Technology*, **8**, 1-11.

[39] Alshammri, R., Alharbi, G., Alharbi, E. and Almubark, I. (2023) Machine Learning Approaches to Identify Parkinson's Disease Using Voice Signal Features. *Frontiers in Artificial Intelligence*, **6**, Article ID: 1084001. https://doi.org/10.3389/frai.2023.1084001

[40] Rehman, A., Saba, T., Mujahid, M., Alamri, F.S. and ElHakim, N. (2023) Parkinson's Disease Detection Using Hybrid LSTM-GRU Deep Learning Model. *Electronics*, **12**, Article 2856. https://doi.org/10.3390/electronics12132856

[41] Karapinar Senturk, Z. (2020) Early Diagnosis of Parkinson's Disease Using Machine Learning Algorithms. *Medical Hypotheses*, **138**, Article 109603. https://doi.org/10.1016/j.mehy.2020.109603

[42] Mahesh, T.R., Bhardwaj, R., Khan, S.B., Alkhaldi, N.A., Victor, N., *et al.* (2024) An Artificial Intelligence-Based Decision Support System for Early and Accurate Diagnosis of Parkinson's Disease. *Decision Analytics Journal*, **10**, Article 100381. https://doi.org/10.1016/j.dajour.2023.100381