IJCI V12(2025) 1-23

International Journal of Computers and Information (IJCI)

Available online at https://ijci.journals.ekb.eg/

A Speech-Based Hybrid Method for Parkinson's Detection using Pearson Correlation and Mutual Information

Mohamed Elkharadly [a,*], Khaled Amin [b], Osama M. Abo-Seida [c], Mina Ibrahim [d]

[a] *Information Technology Department, Faculty of Computers and Information, Kafr El-Sheikh University, Kafr El-Sheikh, 33511, Egypt*
[b] *Department of Information Technology, Faculty of Computers and Information, Menoufia University , Shebin El-Kom, 32511, Egypt*
[c] *Department of Computer Science, Faculty of Computers and Informaion, Kafr El-Sheikh University , Kafr El-Sheikh, 33511, Egypt*
[d] *Department of Machine Intelligence, Faculty of Artificial Intelligence, Menoufia University , Shebin El-Kom, 32511, Egypt*

* *mohamed.samy@fci.kfs.edu.eg , k.amin@ci.menofia.edu.eg, aboseida@yahoo.com, mina.ibrahim@ci.menofia.edu.eg*

Abstract

Parkinson's disease (PD) is a chronic and progressive neurodegenerative disorder that affects movement. Studies have shown that speech difficulties can appear early in PD, suggesting their potential use as an early diagnostic indicator. Our proposed method investigated a hybrid approach for Parkinson's detection based on Pearson Correlation (PC) and Mutual Information (MI). The approach combines PC and MI to identify the relevant features in the speech signals, utilizing these features for training five machine learning models, namely XGBoost, GBoost, CatBoost, AdaBoost, and LightGBM. Two datasets obtained from UCI repository were utilized for evaluation. To overcome the challenge of imbalanced classes in the datasets, synthetic minority oversampling technique (SMOTE) was implemented to achieve a more balanced representation. The proposed PCMI approach selects 10 features from dataset1 and 55 features from dataset2. The results show that CatBoost with SMOTE and PCMI achieved an accuracy of 97.3% using hold-out method 75:25 and 97.2% using 10-fold CV method for dataset1, while LightGBM with SMOTE and PCMI approach achieved an accuracy of 95.6% using hold-out method 60:40 and 97.6% using 10-fold CV method for dataset2.

Keywords: Parkinson's disease; speech signal; Pearson Correlation; Mutual Information; machine learning.; SMOTE

1. Introduction

Parkinson's disease (PD) presents significant challenges in early diagnosis, often leading to delayed treatment initiation and a consequent impact on patient outcomes [1]. The hallmark motor symptoms of PD, such as tremors [2], muscle stiffness [3], reduced movement [4], and instability while walking [5, 6], typically manifest after substantial neurodegeneration has occurred. This delay underscores the critical need for reliable early detection methods.

Speech difficulties [7-11], including changes in voice quality, articulation, and rhythm, are common among individuals with PD. Importantly, these speech impairments often emerge in the prodromal phase of the disease, preceding overt motor symptoms by several years. Therefore, analyzing speech signals holds promise as a non-invasive and potentially early indicator for PD.

Despite advancements in neuroimaging and clinical assessments, current diagnostic methods for PD primarily rely on subjective evaluations and late-stage symptom presentations. This limitation highlights a clear gap in diagnostic capabilities, particularly in detecting PD during its prodromal phase when neuroprotective interventions may be most effective.

This study aims to address this gap by developing a computer-aided diagnosis (CAD) system to assist in PD diagnosis, providing a supplementary opinion to physicians. This system aims to reduce diagnostic errors, streamline the diagnostic process, and improve efficiency.

Previous research in PD diagnosis has primarily focused on integrating effective classifiers with feature selection techniques to enhance diagnostic accuracy. The choice of both the feature selection method and classifier plays a crucial role in the outcomes of PD diagnosis. However, the previous research suffers from the following limitations:

- **Imbalanced Datasets**: affect model generalization ability.
- **Suboptimal feature selection**: Filter methods may occasionally overlook important features because the interconnection between features is not taken into account while ranking features.
- **Computation time**: Filter, Wrapper, and Hybrid methods may expand the training period and increase complexity.

In this study, an approach for Parkinson's disease detection is proposed utilizing a hybrid feature selection method combining Pearson Correlation (PC) and Mutual Information (MI). This method aims to identify and leverage the most relevant features extracted from speech signals, enhancing the accuracy of machine learning models used for classification. By integrating PC and MI, we prioritize features that exhibit both low correlation and high informativeness, thus improving the discriminatory power of the models. The experimental evaluation involved training five machine learning models—XGBoost, LightGBM, CatBoost, GBoost, and AdaBoost—using two distinct datasets obtained from the UCI repository. To address the challenge of imbalanced classes within these datasets, the Synthetic Minority Oversampling Technique (SMOTE) was applied to achieve a more balanced representation of the target classes. The proposed PCMI approach selects 10 features from dataset1 and 55 features from dataset2.

The results show that CatBoost with SMOTE and PCMI achieved an accuracy of 97.3% using hold-out method 75:25 and 97.2% using 10-fold CV method for dataset1, while LightGBM with SMOTE and PCMI approach achieved an accuracy of 95.6% using hold-out method 60:40 and 97.6% using 10-fold CV method for dataset2, highlighting the effectiveness of this approach across different validation strategies and datasets.

The core contributions of this study are:

- Firstly, the Parkinson's disease speech datasets used in this study are highly imbalanced. In the first dataset, 147 samples out of 195 are from Parkinson's patients. In the second dataset, 564 samples out of 756 are from Parkinson's patients. Therefore, SMOTE has been used to handle the class imbalance problem.

- A pioneering hybrid technique dubbed "PCMI" has been introduced, combining Pearson Correlation with Mutual Information to identify a subset of features that exhibit both low correlation and high informativeness.

- The performance of five classifiers namely XGBoost, GBoost, CatBoost, AdaBoost, and LightGBM was evaluated on reduced feature subset and the best classifier was found as CatBoost for PD diagnosis problem.

- The proposed method is better than the other methods with respect to computational cost since few number of speech features were used.

This study is structured as follows: Section 2 offers a brief overview of feature selection, feature scaling, and model selection, Section 3 covers related work, Section 4 outlines proposed approach, Section 5 presents experimental results, and Section 6 concludes proposed work.

2. **Background**

2.1 Feature Selection

Feature selection is a critical step in machine learning, involving the identification and extraction of the most relevant features from a dataset to improve model performance and interpretability. By selecting the most informative features, redundant or irrelevant ones are eliminated, reducing the dimensionality of the dataset and preventing overfitting [12].

2.1.1. Pearson Correlation (PC)

PC [13, 14], denoted by r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to +1, where +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}} \tag{1}$$

where X_i and Y_i are individual data points, $\overline{X}$ and $\overline{Y}$ are the means of X and Y, respectively.

2.1.2. Mutual Information (MI)

MI [13-17] quantifies the amount of information obtained about one random variable through the observation of another. It measures the degree of dependence between two variables by assessing how much knowing one variable reduces uncertainty about the other. Formally, mutual information between two random variables X and Y is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)\, p(y)} \tag{2}$$

where $p(x,y)$ is the joint probability distribution of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y, respectively.

2.2 Feature Scaling

Standard scaling, also known as z-score normalization, used to standardize the scale of features within a dataset. It involves transforming the data so that it has a mean of zero and a standard deviation of one. This process ensures that all features have the same scale, preventing features with larger magnitudes from dominating the analysis or model training process. The equation for standard scaling is [18]:

$$z = \frac{(x - \text{mean}(x))}{\text{stdev}(x)} \tag{3}$$

where x is the input feature, mean(x) is the mean of the input feature, stdev(x) is the standard deviation of the input feature, and z is the standardized feature.

2.3 Model Selection

We implemented five ML models – XGBoost, GBoost, CatBoost, AdaBoost, and LightGBM. These models were selected based on their strengths in handling complex datasets, their ability to handle both numerical and categorical features efficiently, and their performance in boosting weak learners to create robust classifiers.

2.3.1 XGBoost

XGBoost is tailored specifically for maximizing both model effectiveness and computational speed. It fully utilizes memory and hardware, offering advantages in algorithm enhancement, fine tuning, and deployment. the equation for objective function is as follows [19]:

$$\mathrm{Obj}(\beta) = \mathrm{Loss}(\beta) + \mathrm{Reg}(\beta) \tag{4}$$

where obj represents objective function, Loss stands for training loss that quantifies the error or disparity between predicted values and actual target values, and Reg denotes regularization term that address overfitting issues and facilitate the model's efficient generalization to unseen data.

2.3.2 GBoost

GBoost is a boosting algorithm that sequentially fits a classifier to the residual errors of the previous classifier. The equation for GBoost is as follows [20, 21]:

$$\mathrm{F}(\mathrm{x}) = \mathrm{F}_{\mathrm{T}-1}(\mathrm{x}) + \gamma_{\mathrm{T}}\mathrm{f}_{\mathrm{T}}(\mathrm{x}) \tag{5}$$

where F(x) represents prediction of classifer for input x, $\mathrm{F}_{\mathrm{T}-1}(\mathrm{X})$ represents output of previous T-1 classifier, γ_{T} represents learning rate for the T-th model, and $\mathrm{f}_{\mathrm{T}}(\mathrm{X})$ represents output of the T-th classifier for input x.

2.3.3 CatBoost

CatBoost is a boosting algorithm that uses a novel gradient boosting scheme known as "Ordered Boosting", contributing to mitigating overfitting and enhancing overall accuracy by incorporating categorical features directly into the model. The equation for CatBoost is as follows [22]:

$$\mathrm{F}(\mathrm{x}) = \sum_{\mathrm{t}=1}^{\mathrm{N}} \alpha_{\mathrm{t}}\, \mathrm{F}_{\mathrm{t}}(\mathrm{x}) \tag{6}$$

where F(x) represents overall prediction of the ensemble for input x. N represents total number of weak learners (trees) in the ensemble. α_t represents weight given to each tree. $F_t(x)$ represents prediction of t-th tree.

2.3.4 AdaBoost

AdaBoost is a boosting algorithm designed to create a robust classifier by combining multiple weak learners. The approach involves iterative training of weak learners on reweighted versions of data. During each iteration, weights are adjusted to give more emphasis to misclassified examples from the preceding iteration. The equation for AdaBoost is as follows [23]:

$$\mathrm{F}(\mathrm{x}) = \mathrm{sign}\left(\sum_{i=1}^{N} \alpha_{\mathrm{i}}\mathrm{h}_{\mathrm{i}}\,(\mathrm{x})\right) \tag{7}$$

where F(x) represents prediction of robust classifier for input x, N represents number of iterations, α_{i} represents weight given to each tree, and h_i (x) represents prediction of i-th tree. In each iteration, a tree is trained to minimize the weighted error on the training data, and its weight α_{i} is determined based on its performance. The robust classifier represents a sum of the trees, where the weights are determined by their individual performance and the number of iterations.

2.3.5 LightGBM

LightGBM is a boosting algorithm designed to be efficient and scalable, with the ability to handle large amounts of data [24, 25]. The algorithm works by building decision trees in a leaf-wise fashion, where each split is chosen to maximize the reduction in the loss function. The equation for LightGBM is as follows:

$$\mathrm{F}(\mathrm{x}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{w}_{\mathrm{i}}\mathrm{G}_{\mathrm{i}}(\mathrm{x}) \tag{8}$$

where F(x) represents final prediction for a given input x, N represents number of trees (weak learners) in the ensemble, w_{i} represents weight given to i-th tree, and $\mathrm{G}_{\mathrm{i}}(\mathrm{x})$ represents prediction of i-th tree for input x. w_{i} is determined based on its performance and the regularization parameters. LightGBM also handling missing values, categorical features, and supports parallel and GPU acceleration to further improve its efficiency.

3. Related Work

Prior research applied three techniques for selecting pertinent features, including Filter, Wrapper, and Hybrid methods [26].

3.1 **Filter Methods** focus on selecting or ranking features based on their inherent characteristics [27]. Yaman et al. [28] developed a statistical pooling method for PD detection with the help of vowels. The replicated acoustic features Parkinson's disease dataset was used in the study having 240 samples and 44 features vectors. The 44 features in the dataset were increased to 177 by the proposed method and then the Relief method was used for selecting the top-weighted features and 66 features were selected. For classification SVM and KNN algorithms were used and obtained 91.25% and 91.23% accuracy, respectively. Although they obtained high classification accuracies, they used 66 features. Therefore, their computational cost is high. Tuncer et al. [29] developed a method for the detection of PD using vowels. UCI PD classification dataset is used in this paper. The preprocessing of data was done by MAMA tree, then by using SVD, 122 features were extracted from total 754 features, and finally by relief-based feature selection method 50 most discriminative features were extracted and fed into eight different classifiers for classification of Parkinson's disease patients from healthy patients. The classification accuracy of 92.46 % is achieved by using KNN classifier and by doing post-processing the accuracy was increased to 96.83%. Bchir et al. [30] proposed a Gaussian mixture model (GMM) based classification strategy for Parkinson's disease diagnosis. This paper made use of Parkinson's speech data, which consisted of 756 voice measurements from 252 people. The mRMR feature selection technique chose the best 50 features out of 752. The performance measures were accuracy and MCC. When the suggested GMM model's performance was compared to that of other classifiers, the proposed model outperformed with an accuracy of 0.8912 and an MCC of 0.7060. Although 50 features are selected but the performance is limited. Ashour et al. [31] presented a two-sequential feature selection framework for Parkinson's diagnosis. This paper made use of Parkinson's speech data, which consists of 756 voice measures from 252 people. PCA and eigenvector centrality approaches were used to pick features. SVM-Cubic was used to accomplish the classification, and a reduced feature subset yielded 94% accuracy. Although they obtained high classification accuracy, they used 350 features. Therefore, their computational cost is high.

3.2 **Wrapper Methods** create different combinations of features and assess their effectiveness by feeding them into the model. Evaluation of the subsets is conducted through a performance measure calculated on the resulting model. Senturk [32] presented a machine learning algorithms-based diagnosis system for Parkinson's disease. Feature selection was done by RFE and feature importance methods. The dataset used consists of 23 features and 195 instances. The RFE feature selection technique chose the best 13 features out of 23. Regression tree, ANN, and SVM were used as classifiers. The combination of RFE with the SVM classifer shows 93.84% accuracy. Goyal et al. [33] demonstrate how a two-stage feature selection model outperforms existing methods. This work employed the UCI Parkinson's speech dataset, which has 19٥ cases and 2٣ features. An SVM classifier was employed to classify the data, and a three-fold cross-validation procedure was applied. Only nine features were chosen from the proposed GA+RFE technique, with the SVM classifier achieving the best accuracy of 88.71 %. Although 9 features are selected but the performance is limited. Lamba et al. [34] presented a Parkinson's disease detection system. Two speech datasets have been used in the design of this system: The first is an Italian Parkinson's Voice & Speech dataset, and the other is Mobile Device Voice Recordings at King's College London dataset. Seventeen acoustic features have been generated from the voice samples available in the datasets using Parselmouth library. In addition, based on the significance of features, the eight most significant features have been used in the design of the model. These features have been selected using genetic algorithm method. Four classifiers, k-nearest neighbors, XGBoost, random forest, and logistic regression, have been used during classification stage. The accuracy, sensitivity, f-measure, specificity, and precision parameters have been used for the analysis of the designed system. The combination of a genetic algorithm-based feature selection approach and logistic regression classifier has given 100% accuracy on Italian Parkinson's Voice & Speech dataset. The same feature extraction and classifier combination on the Mobile Device Voice Recordings at King's College London dataset have attained an accuracy level of 90%. Elshewey et al. [35] utilized Bayesian optimization (BO) to optimize the hyperparameters for six machine learning models, RF, SVM, NB, LR, RC, and DT to determine the categorization method that is both the most effective and precise for PD. The dataset used consists of 23 features and 195 instances. The experimental results demonstrated that the SVM model achieved the best outcomes when compared with various ML models before and after the process of hyperparameter tuning, with an accuracy 92.3% obtained using BO.

3.3 **Hybrid Methods** integrate two or more feature selection techniques to refine and optimize the selected feature set. Lamba et al. [36] presented a speech signal-based hybrid Parkinson's disease diagnosis system for early diagnosis. To achieve this, the authors have tested several combinations of feature selection approaches and classification algorithms and designed the model with the best combination. To formulate various combinations, three feature selection methods such as mutual information gain, extra tree, and genetic algorithm and three classifiers namely naive bayes, k-nearest-neighbors, and random forest have been used. To analyze the performance of different combinations, the dataset used consists of 23 features and 195 instances. The combination of genetic algorithm and random forest classifer has shown the best performance with 95.58% accuracy. Lamba et al. [37] presented a hybrid MIRFE feature selection approach based on mutual information gain and recursive feature elimination methods. A Parkinson's disease classification dataset consisting of 756 voice measures of 252 individuals was used in this study. The proposed feature selection approach is compared with the five standard feature selection methods by random forest and XGBoost classifier. The proposed MIRFE approach selects 40 features out of 754 features. MIRFE with XGBoost and RF achieved an accuracy of 93.88% and 92.72%, respectively. Abdel-fattah et al. [38] proposed a hybrid approach based on the Emperor Penguin Colony (EPC) swarm algorithm with Correlation-based Feature Selection (CFC), which is called CEPC. A Parkinson's disease classification dataset consisting of 756 voice measures was used in this study. Before using the proposed approach, five classification algorithms were used to compare accuracy results. Also, the ensemble classifier has been used in this paper. The CEPC proposed approach provides an improvement in the accuracy of results. An accuracy of 89.4% is obtained by the ensemble classifier. Chawla et al. [39] investigated methods for detecting Parkinson's disease using Nature Inspired Feature Selection (NIFS) with the Zebra Optimization Algorithm and Recursive Feature Elimination Cross Validation (RFECV). They utilized a vocal feature-based dataset for PD detection, reducing the number of features from 754 to 40. The classification results were obtained for two cases: a 70:30 train-test split and tenfold cross-validation, using 11 different classifiers. The Gaussian Process classifier showed the best accuracy, with values of 96% and 97.07% for the two cases, respectively. Al-Najjar et al. [40] presented a hybrid grey wolf and whale optimization for enhanced Parkinson's prediction based on machine learning models using biomedical sound. They utilized a vocal feature-based dataset for PD detection, reducing the number of features from 23 to 11. Six models, neural network, Quest, Chi-squared Automatic Interaction Detection, support vector machine, CR tree, and logistic regression, have been used during classification stage. The results showed that the CR tree performed better than all models, reaching 95 % for accuracy.

Although high classification rates were obtained in the literature for ML based diagnosis of PD, either they used many features (like [28, 31]) which increases computation time or the extraction of the features were hard even they use few features (like [39, 40]). Therefore, indirectly, the computation time is again high. In this paper, we aim to decrease computation time via less number of effective features.

4. Methodology

As shown in Fig. 1, the discussed approach for speech-based Parkinson's Disease (PD) detection consists of five stages: (1) Data Augmentation, (2) Feature Selection, (3) Feature Scaling, (4) Model Selection, and (5) Evaluation of Models' Performance. Two datasets obtained from the UCI repository are used in this work. To handle imbalanced data, the Synthetic Minority Over-sampling Technique (SMOTE) is applied. For feature selection, a hybrid FS method called "PCMI" has been introduced, combining Pearson Correlation with Mutual Information to identify a subset of features that exhibit both low correlation and high informativeness. The selected features are then standardized using standard scaling. These features are fed into different classifiers, including XGBoost, LightGBM, CatBoost, GBoost, and AdaBoost. Finally, the detection performance is evaluated using metrics such as accuracy, recall, precision, F-score, and AUC. This comprehensive approach aims to enhance the efficiency and accuracy of PD detection using speech-based features.

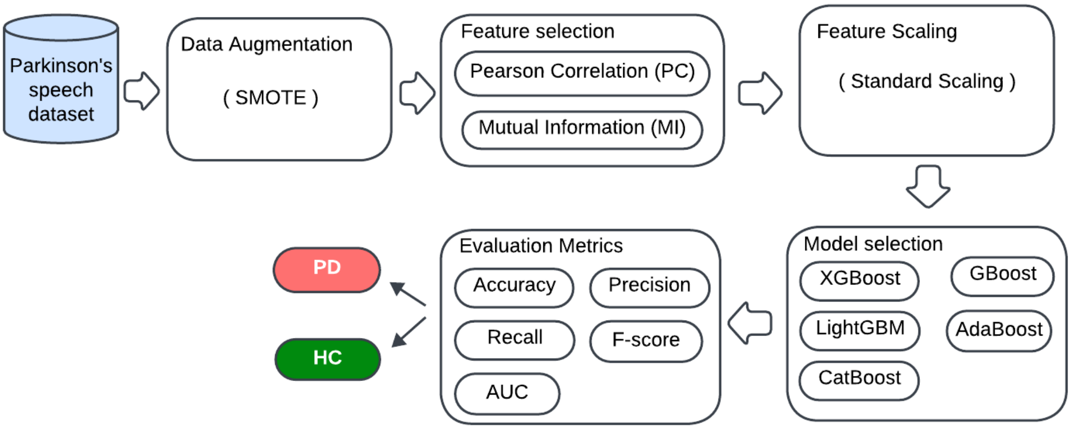

Fig. 1 Proposed diagnostic system

4.1 Data Augmentation

To overcome the challenge of imbalanced classes in the datasets, synthetic minority oversampling technique (SMOTE) was implemented to achieve a more balanced representation. This technique aids in mitigating the impact of imbalanced datasets on model training, particularly in scenarios where one class is significantly underrepresented. By creating synthetic examples that resemble the minority class instances, SMOTE contributes to better generalization, enhances overall model performance, and mitigates overfitting [41]. As shown in Fig. 2, the red-colored points are synthetic; it is noticed that all these points are lying between the boundaries of the original points, which gives a more accurate, reliable representation. Algorithm 1 shows SMOTE steps. Table 1 shows the datasets before and after the use of SMOTE. It is noted that the classes of the minority classes (healthy) became equal to the classes of the majority (Parkinson).

Table 1. SMOTE method for balancing the datasets

Datasets	**Dataset1**		**Dataset2**	
Classes	**Healthy**	**Parkinson**	**Healthy**	**Parkinson**
Before SMOTE	48	147	192	564
After SMOTE	147	147	564	564

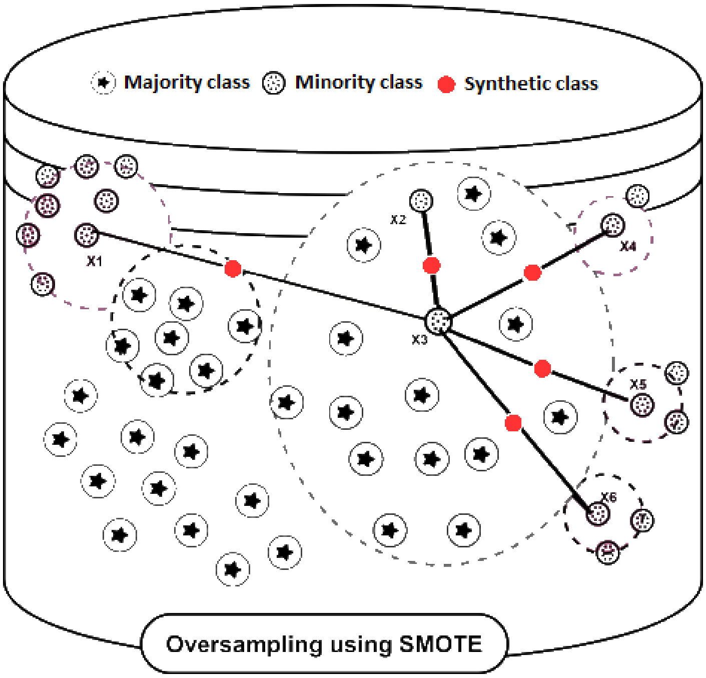

Fig. 2 Oversampling minority classes [42]

Algorithm 1

Inputs:

- Minority class samples (X_min)
- Number of synthetic samples to generate (N)
- Number of nearest neighbors to use (K)

Output:

- Augmented dataset with synthetic samples

```
Function SMOTE (X_min, N, K)
    1. Initialize an empty list for synthetic samples, S.
    2. For each minority sample x_i in X_min:
        a. Find the k nearest neighbors of x_i.
        b. For each neighbor x_j (where j = 1 to k):
            i. Randomly select a neighbor x_j.
            ii. Compute the difference vector: diff = x_j - x_i.
            iii. Multiply diff by a random number between 0 and 1: gap = random(0, 1).
            iv. Create the synthetic sample: synthetic_sample = x_i + gap * diff.
            v. Add synthetic_sample to the list S.
            vi. Repeat until N synthetic samples are generated.
    3. Return the original dataset augmented with the synthetic samples from S.
```

4.2 Proposed PCMI Method

The Pearson Correlation is calculated between all pairs of features in the dataset. If the correlation coefficient between two features is above a certain threshold, then the two features are considered to be correlated. The correlated features are then removed from the dataset by dropping the columns that contain them. This can solve the redundancy problem. Redundancy occurs when two or more features provide the same information. This can be a problem for machine learning models because it can lead to overfitting. By removing correlated features, redundancy in the dataset can be reduced and the performance of the model can be improved. Next, the Mutual Information of each feature is calculated. Mutual Information is a measure of how much information a feature provides about the target variable. The features with the highest Mutual Information are considered to be the most important features. The features with the highest Mutual Information are then selected. Algorithm 2 shows pseudo code of PCMI method.

Algorithm 2

```
Inputs:
    -   X_T: input feature matrix
    -   y_T: target variable
    -   corr_threshold: Pearson correlation threshold (float)
    -   mi_threshold: mutual information threshold (float)

Parameters:
    -   corr_matrix: Pearson correlation matrix
    -   corr_features: set of correlated features (set)
    -   mi_scores: mutual information scores
    -   mi_features: set of features selected by mutual information (set)
    -   selected_features: set of final selected features (set)

Output:
    -   selected_features: set of final selected features(set)

Steps:
    1.  Compute Pearson correlation matrix corr_matrix = X_T. corr ( )
    2.  Initialize empty set corr_features = set ( )
    3.  For each column i in corr_matrix:
            a. For each previous column j up to i-1:
                i. If abs (corr_matrix. iloc [i, j]) > corr_threshold:
                    1. Add the name of column i to corr_features
    4.  Drop the correlated features from X_T using X_T = X_T. drop (corr_features, axis=1)
    5.  Compute mutual information scores using mi_scores = mutual_info_classif (X_T, y_T)
    6.  Initialize empty set mi_features = set ( )
    7.  For each score i in mi_scores:
            a. If i > mi_threshold:
                1. Add the name of the corresponding column in X_T to mi_features
    8.  Set selected_features = mi_features
    9.  Return selected_features
```

4.3 Evaluation Metrics

The performance of models is evaluated by the following metrics:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{9}$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \tag{10}$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \tag{11}$$

$$\text{F} - \text{score} = 2 \times (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) \tag{12}$$

$$\text{AUC} = \sum_{\text{i}=1}^{\text{n}-1} \frac{1}{2} (\text{FPR}_{\text{i}+1} - \text{FPR}_{\text{i}}) \times (\text{TPR}_{\text{i}} + \text{TPR}_{\text{i}+1}) \tag{13}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

5. **Experimental Results**

5.1 Datasets

5.1.1 Dataset1

A dataset of Parkinson’s disease (PD) from the UCI Machine Learning Repository was created by Max Little [43]. It contains 195 rows, each corresponding to a voice measure from one of 31 individuals (23 with PD and 8 healthy). Each column represents a particular voice feature. Out of 195, 147 voice measures are from PD patients and the remaining are from healthy persons. The status column has two values, 0 for healthy individuals and 1 for PD patients. Table 2 details dataset features description.

Table2. Dataset1 features description

Feature Name	No.
Vocal fundamental frequency	3
Fundamental frequency variation measures	5
Amplitude variation measures	6
Noise to tonal measures	2
Recurrence and correlation measures	2
Detrended fluctuation analysis	1
Additional fundamental frequency variation measures	3

5.1.2 Dataset2

A Parkinson’s dataset available at [44] comprised 756 voice measures collected from 252 individuals. Among them, 188 individuals had Parkinson’s. Vowel /a/ was repeated thrice by each participant for data collection. From each voice measure, 754 distinctive features were extracted. Table 3 details dataset features description.

Table 3. Dataset2 features description

Feature Name	No.
Basal Features	21
Temporal Frequency Features	11
MFCCs	84
Vocal Cord Features	22
TQWT	615

5.2 Data Exploration

Figs. 3 and 4 show a heatmap of the dataset1 features and dataset2 features, respectively. In the context of this study, heatmap effectively used to identify highly correlated features in a dataset, which can then be removed to avoid redundancy. Resulting in a reduced dataset that more efficient to analyze and less prone to overfitting.

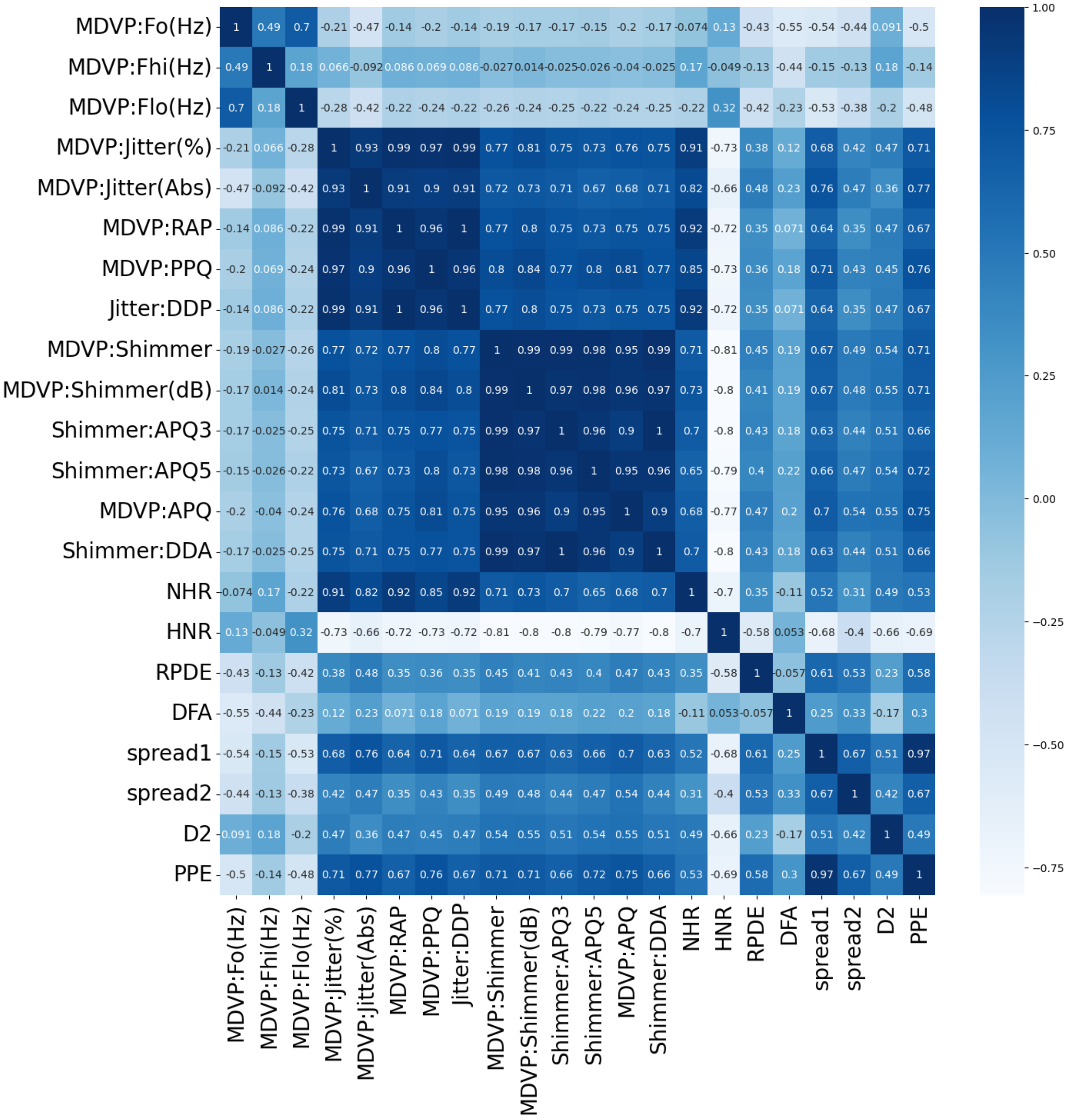

Fig. 3 Heatmap exploration of correlation between features in dataset1

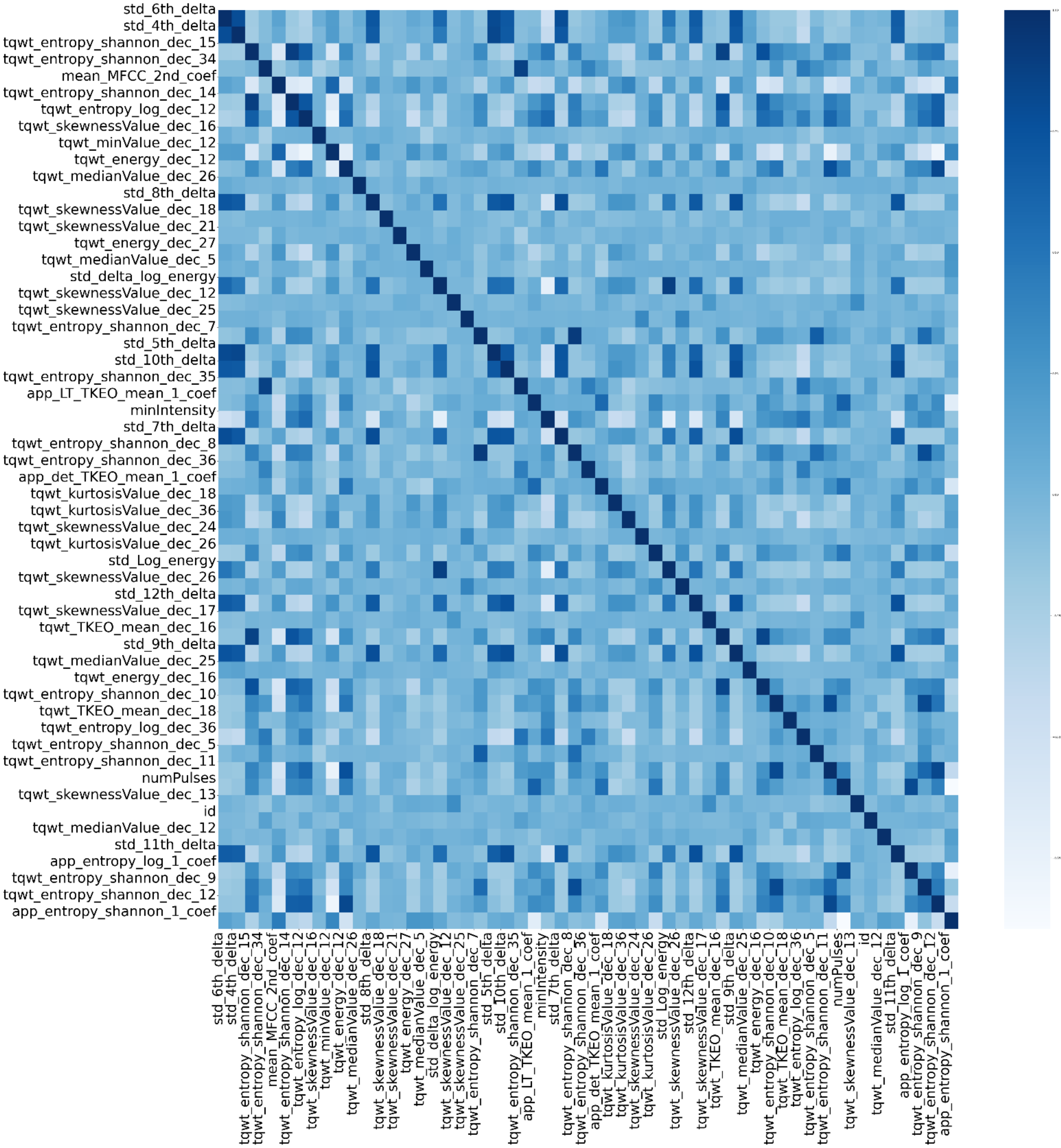

Fig. 4 Heatmap exploration of correlation between features in dataset2

5.3 Feature Importance

PCMI method selected 10 features out of 23 from dataset1 and 55 features out of 754 from dataset2. Figs. 5 and 6 illustrate the importance scores of features in dataset1 and dataset2, respectively.

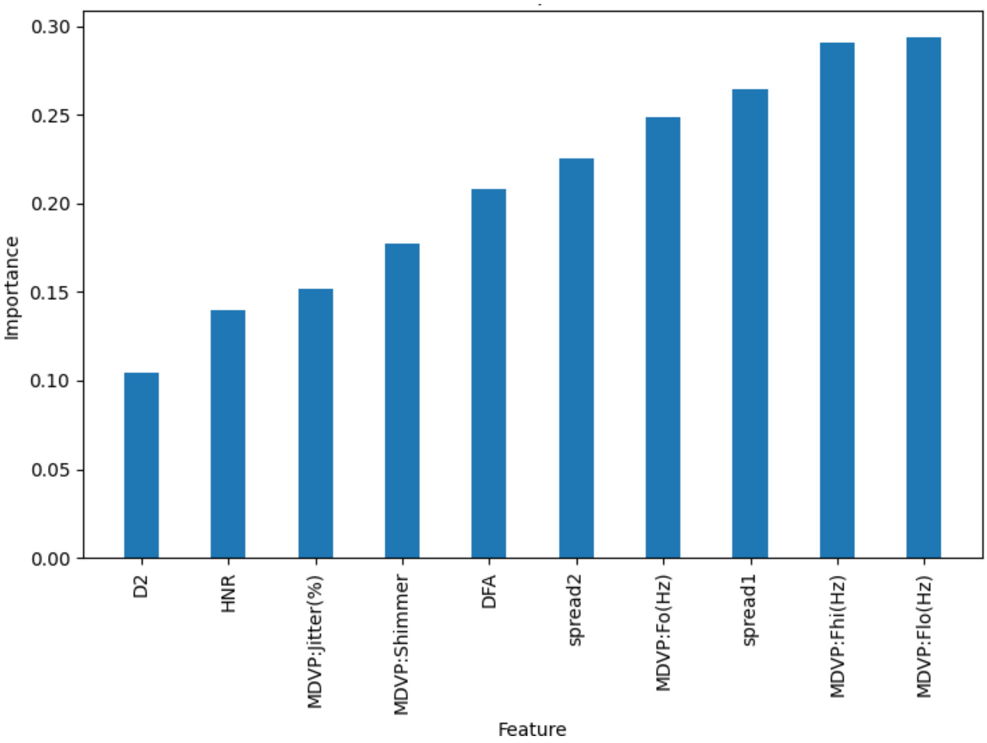

Fig. 5 Feature importance of dataset1

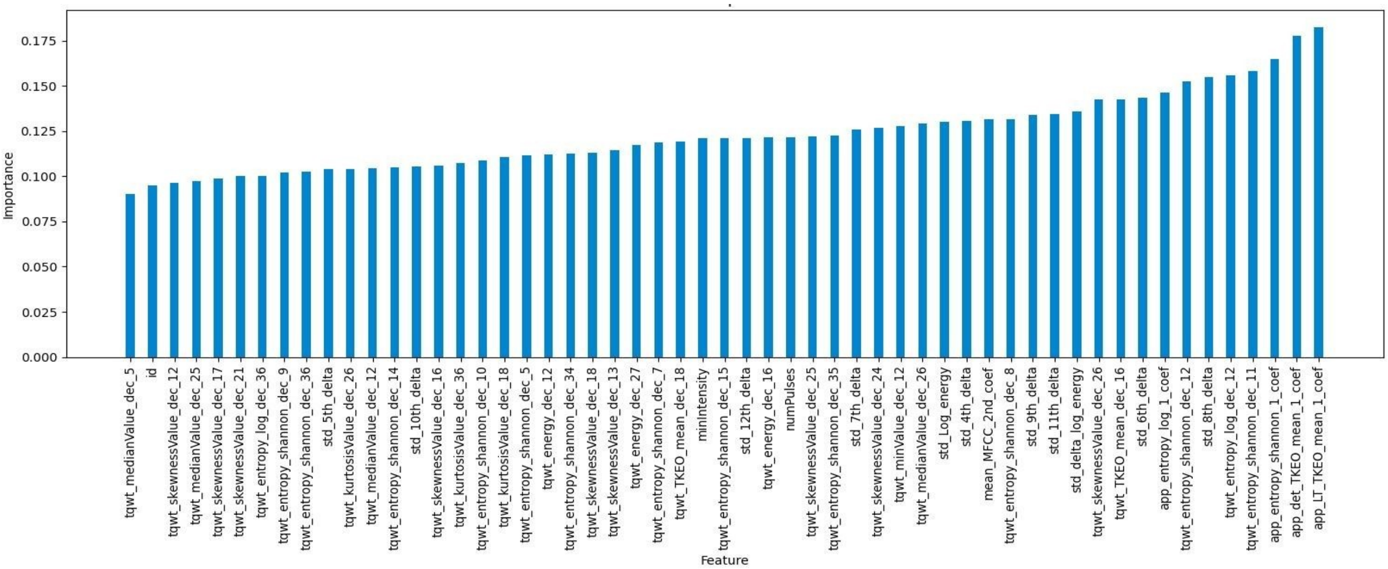

Fig. 6 Feature importance of dataset2

Figs. 7 and 8 illustrate histograms that analyze the distribution of important features in dataset1 and dataset2, respectively. A histogram visually represents data distribution, displaying frequencies within specified intervals. A balanced distribution ensures that the classifier is not biased towards the majority class.

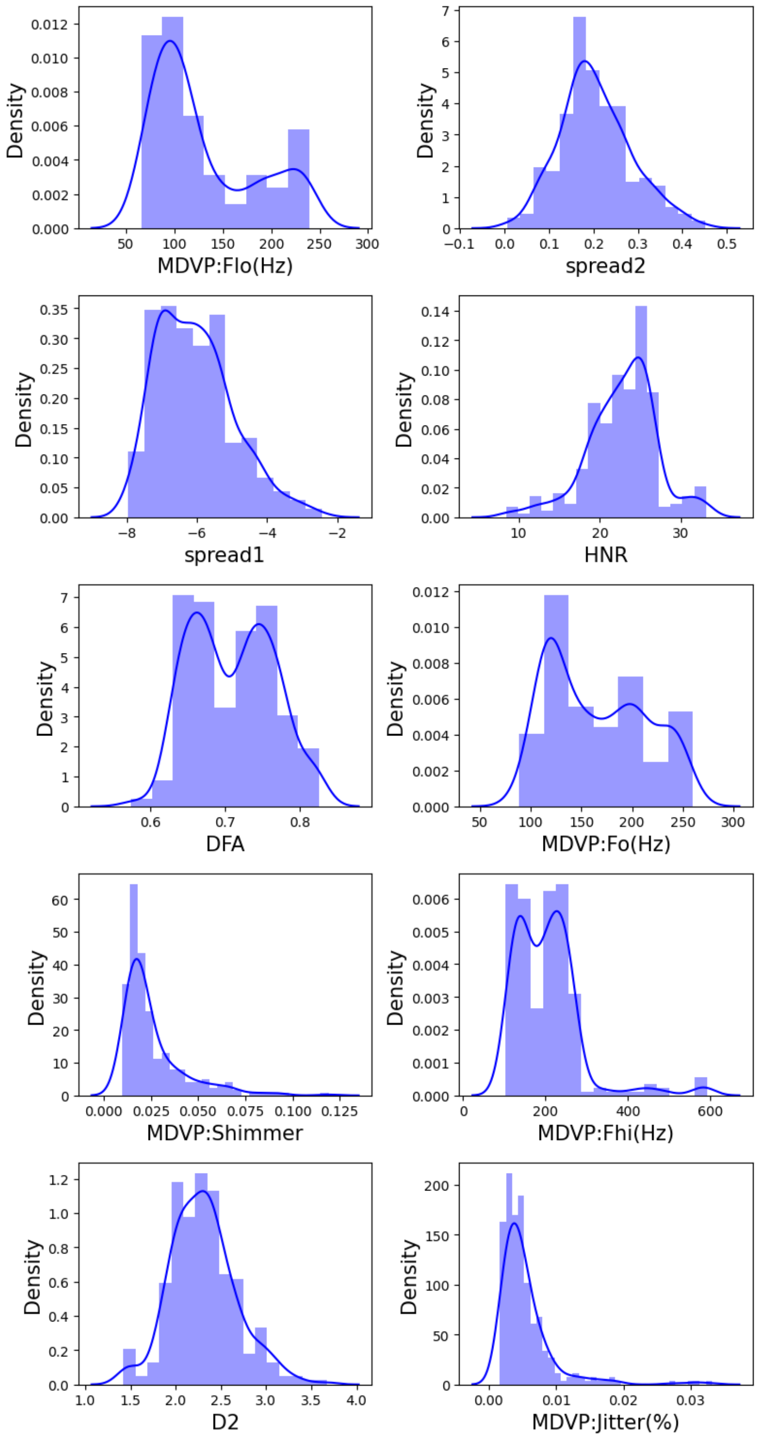

Fig. 7 Histogram exploration of features distribution in dataset1

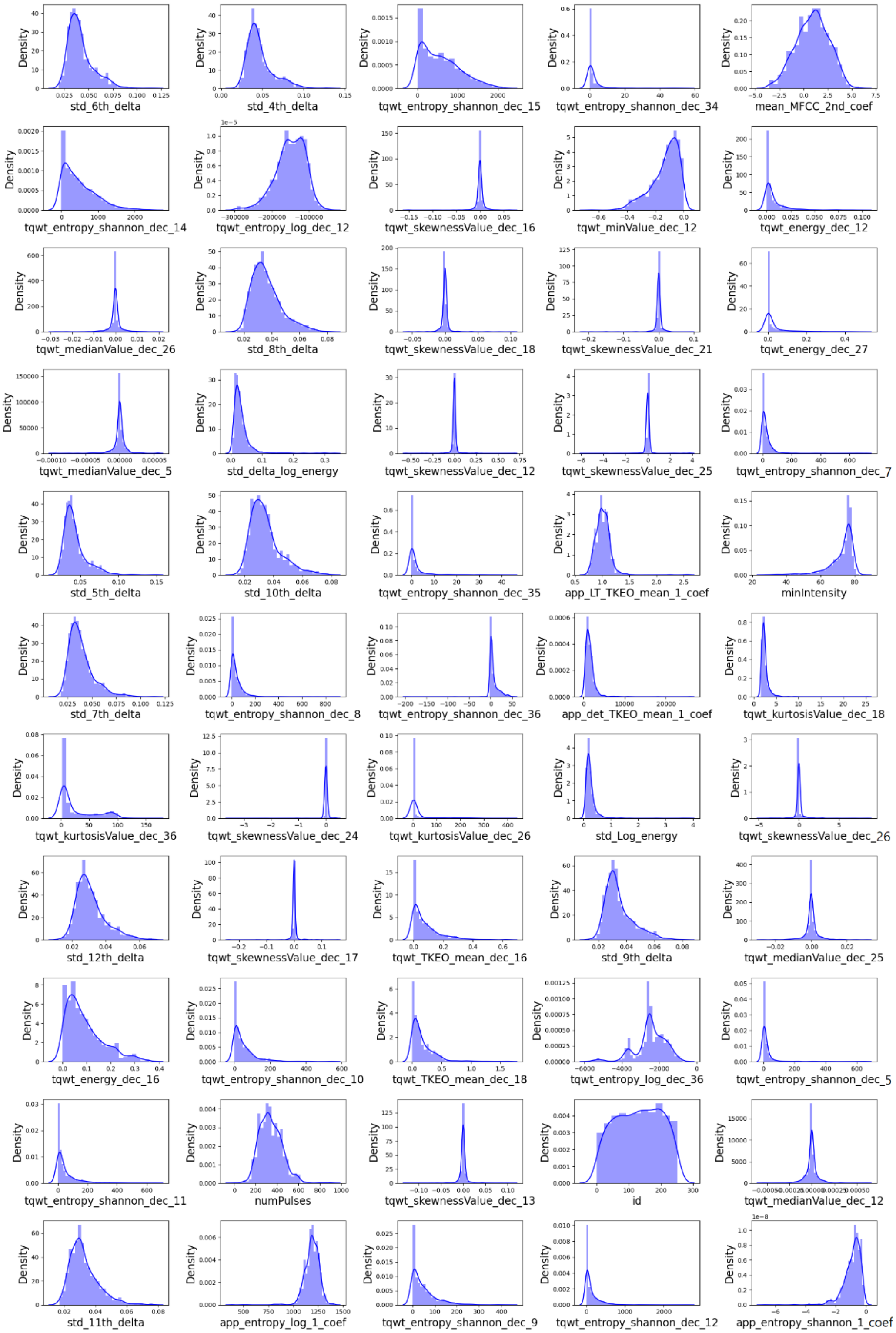

Fig. 8 Histogram exploration of features distribution in dataset2

5.4 Results and Discussion

The experiments were conducted on a system running Windows 10, powered by an Intel Core i5 G7 CPU and 8GB of RAM, and coded using python.

In our experiments, the proposed approach is tested on dataset1 and dataset2 published in UCI repository. Table 4 shows the performance of each model without SMOTE and PCMI approach for dataset1 in the term of accuracy. Table 5 shows the performance of each model without SMOTE and PCMI approach for dataset2 in the term of accuracy. Table 6 shows the outcomes achieved through the integration of SMOTE and PCMI approach, using the hold-out method 75-25% train-test partition for dataset1. Table 7 shows the outcomes achieved through the integration of SMOTE and PCMI approach, using the hold-out method 60-40% train-test partition for dataset2. Tables 8 and 9 show the outcomes achieved through the integration of SMOTE and PCMI approach with 10-fold CV for dataset1 and dataset2, respectively.

Tables 4 and 5 illustrate the performance of the classifiers without SMOTE and PCMI approach. In Dataset1, CatBoost achieved the highest accuracy of 93.8%, followed by XGBoost with 91.8%, GBoost with 89.8%, LightGBM with 86.3%, and AdaBoost with 84.7%. Conversely, Dataset2 witnessed LightGBM achieved the highest accuracy of 87.1% followed by XGBoost with 86.1%, CatBoost with 84.8%, GBoost with 84.2%, and AdaBoost with 83.8%. These results highlight the need for considering alternative techniques, such as SMOTE and PCMI, to enhance model generalization and robustness across diverse datasets.

Tables 6 and 7 illustrate the performance of the classifiers with SMOTE and PCMI approach, utilizing a hold-out method with 75:25 split for Dataset1 and 60:40 split for Dataset2, the results reveal that incorporating SMOTE and PCMI techniques can enhance the performance of classification models across both datasets. In Dataset1, CatBoost outperforms the other models with an accuracy of 97.3%, followed by XGBoost with 94.6%, GBoost with 93.2%, and LightGBM with 87.8%. AdaBoost has the lowest accuracy of 86.5%. The recall, precision, F-score, and AUC for CatBoost are also the highest, at 97.4%, 97.4%, 97.4%, and 0.998, respectively. This means that CatBoost with SMOTE and PCMI is the most accurate model for Dataset1. In Dataset2, LightGBM outperforms the other models with an accuracy of 95.6%, followed by CatBoost with 95.4%, XGBoost with 94.3%, and GBoost with 92.9%. AdaBoost has the lowest accuracy of 87.5%. The recall, precision, F-score, and AUC for LightGBM are also the highest, at 94.2%, 97.3%, 95.7%, and 0.993, respectively. This means that LightGBM with SMOTE and PCMI is the most accurate model for Dataset2. Further evaluation using confusion matrices provides additional insights into the model performance. For dataset1, CatBoost correctly identified 37 PD cases (TP) and incorrectly labeled 1 HC case (FP). Additionally, it correctly classified 35 HC cases (TN) and misclassified 1 PD case (FN). For dataset2, LightGBM accurately identified 389 PD cases (TP) while incorrectly classifying 11 HC cases (FP). Moreover, it correctly categorized 376 HC instances (TN) and mislabeled 24 PD cases (FN). The results indicated that CatBoost and LightGBM had the best performance for dataset1 and dataset2, respectively.

Tables 8 and 9 illustrate the performance of the classifiers with SMOTE and PCMI approach, utilizing a 10- fold CV for dataset1 and dataset2, respectively. In Dataset1, XGBoost demonstrated strong capabilities with an accuracy of 96.9%, recall at 95.2%, precision reaching 98.6%, F-score of 96.8%, and an AUC of 0.9897. GBoost closely followed, exhibiting robust performance across various metrics. However, the standout performer in Dataset1 was CatBoost, achieving an impressive accuracy of 97.2%, with a recall of 95.2%, precision of 99.3%, F-score of 97.1%, and an outstanding AUC of 0.9953. LightGBM and AdaBoost also yielded commendable results, contributing to the overall effectiveness of the hybrid approach. In Dataset2, XGBoost demonstrated exceptional performance with an accuracy of 97.2%, a recall at 95.6%, precision reaching 98.7%, F-score of 97.1%, and an AUC of 0.9963. Notably, LightGBM outshone other models with remarkable accuracy of 97.6%, along with a recall of 96.0%, precision of 99.2%, F-score of 97.6%, and an outstanding AUC of 0.9978. These results position LightGBM as the top-performing model for Dataset2. CatBoost also displayed commendable accuracy of 97.5%, a recall of 96.2%, precision of 98.8%, F-score of 97.5%, and an AUC of 0.9967, establishing itself as a robust classifier. GBoost and AdaBoost, while slightly trailing behind, contributed significantly to the overall efficacy of the hybrid approach with their respective performance metrics.

Table 4. Performance of each model without SMOTE and PCMI approach for dataset1 in the term of accuracy

Dataset	Model	Accuracy (%)
Dataset1	XGBoost	91.8
	LightGBM	86.3
	CatBoost	**93.8**
	GBoost	89.8
	AdaBoost	84.7

Table 5. Performance of each model without SMOTE and PCMI approach for dataset2 in the term of accuracy

Dataset	Model	Accuracy (%)
Dataset2	XGBoost	86.1
	LightGBM	**87.1**
	CatBoost	84.8
	GBoost	84.2
	AdaBoost	83.8

Table 6. Proposed hybrid approach results using hold-out method (75-25 % train–test partition) for dataset1

Dataset	Data Augmentation	FS Method	Model	Accuracy (%)	Recall (%)	Precision (%)	F-score (%)	AUC
Dataset1	SMOTE	PCMI	XGBoost	94.6	94.7	94.7	94.7	0.990
			LightGBM	87.8	86.8	89.2	88.0	0.980
			CatBoost	**97.3**	**97.4**	**97.4**	**97.4**	**0.998**
			GBoost	93.2	94.7	92.3	93.5	0.992
			AdaBoost	86.5	86.8	86.8	86.8	0.955

Table 7. Proposed hybrid approach results using hold-out method (60-40 % train–test partition) for dataset2

Dataset	Data Augmentation	FS Method	Model	Accuracy (%)	Recall (%)	Precision (%)	F-score (%)	AUC
Dataset2	SMOTE	PCMI	XGBoost	94.3	91.7	97.1	94.4	0.991
			LightGBM	**95.6**	**94.2**	**97.3**	**95.7**	**0.993**
			CatBoost	95.4	93.7	97.2	95.4	0.992
			GBoost	92.9	89.4	96.6	92.8	0.984
			AdaBoost	87.5	83.8	91.3	87.4	0.942

Table 8. Proposed hybrid approach results using 10-fold Cross Validation for dataset1

Dataset	Data Augmentation	FS Method	Model	Accuracy (%)	Recall (%)	Precision (%)	F-score (%)	AUC
Dataset1	SMOTE	PCMI	XGBoost	96.9	95.2	98.6	96.8	0.9897
			LightGBM	94.8	93.8	96.2	94.8	0.9891
			CatBoost	**97.2**	**95.2**	**99.3**	**97.1**	**0.9953**
			GBoost	95.5	93.8	97.4	95.4	0.9897
			AdaBoost	91.0	89.0	93.1	90.6	0.9516

Table 9. Proposed hybrid approach results using 10-fold Cross Validation for dataset2

Dataset2	Data Augmentation	FS Method	Model	Accuracy (%)	Recall (%)	Precision (%)	F-score (%)	AUC
Dataset2	SMOTE	PCMI	XGBoost	97.2	95.6	98.7	97.1	0.9963
			LightGBM	**97.6**	**96.0**	**99.2**	**97.6**	**0.9978**
			CatBoost	97.5	96.2	98.8	97.5	0.9967
			GBoost	94.4	92.4	96.3	94.3	0.9865
			AdaBoost	87.6	86.7	88.5	87.5	0.9532

The analysis results confirm that the proposed approach, employing a hybrid methodology with a strategic combination of SMOTE and PCMI, is highly effective in enhancing the efficacy of classification models for Parkinson’s detection. The initial experiments without employing the SMOTE and PCMI approach resulted in lower accuracy for the classification models in Parkinson’s detection. However, the subsequent integration of these techniques, along with a hold-out method, significantly improved models’ accuracy, recall, precision, F-score, and AUC metrics. In Dataset1, CatBoost emerged as the leading model with exceptional accuracy of 97.3% and well-balanced recall, precision, and F-score metrics. For Dataset2, LightGBM showcased outstanding accuracy of 95.6% and superior performance in recall, precision, F-score, and AUC. The comprehensive evaluation using confusion matrices provided deeper insights into the models’ abilities to accurately identify positive and negative cases. CatBoost and LightGBM consistently demonstrated superior performance in 10-fold CV, further substantiated the robustness of the hybrid approach. These findings underscore the effectiveness and adaptability of the proposed hybrid approach, positioning it as a reliable strategy for improving PD classification across diverse datasets and evaluation methodologies.

For comparison purpose, Tables 10 and 11 present the accuracies of previous PD diagnosis methods. Our PCMI-CatBoost and PCMI-LightGBM methods yield higher accuracy than all methods previously explored for dataset1 and dataset2, respectively.

For Dataset1, while [33] and [36] used fewer features, they often involved complex feature selection methods or algorithms with higher computational complexity. Moreover, [32] and [40] used more number of features compared to the proposed approach. For Dataset2, some studies such as [31] used a large number of features, which potentially increased computational burden, whereas studies like [29], [37], and [39] used fewer features but often involved more complex feature selection methods or algorithms. This study uses fewer features (10 for Dataset1, 55 for Dataset 2), reducing computational burden. In the proposed approach, computation time was decreased via less number of effective features, a lightweight feature extraction process and a classifier. The features were obtained from the speech signals and so to obtain these features is easier and less costly than the other methods in the literature.

Table10. Comparison of previous PD diagnosis methods with our method in term of accuracy for dataset1

Ref.	No. of features	Method	Accuracy (%)
Senturk [32]	13	RFE-SVM	92.84
Goyal et al. [33]	9	GA+RFE-SVM	88.71
Lamba et al. [36]	5	GA-RF	95.58
Al-Najjar et al. [40]	11	GWOWO-CR tree	95
This study	**10**	**PCMI-CatBoost**	**97.3 (hold-out)** **97.2 (avg. 10-fold CV)**

Table 11. Comparison of previous PD diagnosis methods with our method in term of accuracy for dataset2

Ref.	No. of features	Method	Accuracy (%)
Tuncer et al. [29]	50	MAMA tree +SVD+Relief-KNN	96.83
Ashour et al. [31]	350	PCA-SVM	94
Lamba et.al. [37]	40	MIRFE-XGBoost	93.88
Chawla et al. [39]	40	ZOARFE-GPC	97.07
This study	**55**	**PCMI-LightGBM**	**95.6 (hold-out)** **97.6 (avg. 10-fold CV)**

PD diagnosis system proposed in this study differs from the literature in terms of FS method, a lightweight feature extraction process and a classifier. A high enough classification performance has been achieved. Using speech features in the diagnosis of PD helped very much. Obtaining speech features are both easier and cheaper. In this study, CatBoost and LightGBM with PCMI gave the best classification accuracy. These findings suggest that using certain subset of speech features help researchers classify PD patients more accurately and less efforts can be made to extract features from speech signals of candidate PD patients. Besides, the classification can be realized by less computational cost.

Despite the promising results, several limitations and areas for improvement should be addressed: First, the performance variability across datasets underscores the need for further validation on larger and more diverse datasets to enhance model generalizability. Second, while the proposed system effectively diagnoses Parkinson's disease, it is unable to anticipate the severity of the illness. Lastly, due to the gradual progression of Parkinson's, the system currently lacks the capability to identify disease progression.

6. Conclusion

In this paper, the authors proposed a hybrid approach for Parkinson's detection based on PC and MI. The approach combines PC and MI to identify the relevant features in the speech signal. The identified features are subsequently utilized for training five machine learning models, namely XGBoost, GBoost, CatBoost, AdaBoost, and LightGBM. Two datasets obtained from UCI repository were utilized for evaluation. To overcome the challenge of imbalanced classes in the datasets, synthetic minority oversampling technique (SMOTE) was implemented to achieve a more balanced representation. The proposed PCMI approach selects 10 features from dataset1 and 55 features from dataset2. The results show that CatBoost with SMOTE and PCMI achieved an accuracy of 97.3% using hold-out method 75:25 and 97.2% using 10-fold CV method for dataset1, while LightGBM with SMOTE and PCMI approach achieved an accuracy of 95.6% using hold-out method 60:40 and 97.6% using 10-fold CV method for dataset2. Based on these results, CatBoost is suggested as the best classifier for Parkinson's detection due to its consistent performance in terms of accuracy across different evaluation methods. These findings bring the potential for improving the lives of people with PD by enabling earlier diagnosis and treatment. Developing more accurate and reliable methods for diagnosing PD can help to ensure that people receive the care they need as early as possible. Future work could focus on applying the proposed method on different datasets, having a large number of features, to diagnose numerous other diseases.

References

[1] J. Massano and K. P. Bhatia, "Clinical Approach to Parkinson's Disease: Features, Diagnosis, and Principles of Management," Cold Spring Harbor Perspectives in Medicine, vol. 2, no. 6, pp. a008870–a008870, 2012.

[2] O.-B. Tysnes and A. Storstein, "Epidemiology of Parkinson's disease," Journal of Neural Transmission, vol. 124, no. 8, pp. 901–905, 2017.

[3] L. Raiano, G. di Pino, L. di Biase, M. Tombini, N. L. Tagliamonte, and D. Formica, "PDMeter: A Wrist Wearable Device for an at-Home Assessment of the Parkinson's Disease Rigidity," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 6, pp. 1325–1333, 2020.

[4] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," Journal of Neurology, Neurosurgery & Psychiatry, vol. 79, no. 4, pp. 368–376, 2008.

[5] B. Palakurthi and S. P. Burugupally, "Postural Instability in Parkinson's Disease: A Review," Brain Sciences, vol. 9, no. 9, p. 239, 2019.

[6] C. Schlenstedt et al., "Quantitative assessment of posture in healthy controls and patients with Parkinson's disease," Parkinsonism & Related Disorders, vol. 76, pp. 85–90, 2020.

[7] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech Impairment in a Large Sample of Patients with Parkinson's Disease," Behavioural Neurology, vol. 11, no. 3, pp. 131–137, 1999.

[8] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova, "Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation," Journal of Neural Transmission, vol. 124, no. 3, pp. 303–334, 2017.

[9] H. Gunduz, “Deep Learning-Based Parkinson’s Disease Classification Using Vocal Feature Sets,” IEEE Access, vol. 7, pp. 115540–115551, 2019.

[10] S. Skodda, W. Grönheit, N. Mancinelli, and U. Schlegel, “Progression of Voice and Speech Impairment in the Course of Parkinson’s Disease: A Longitudinal Study,” Parkinson’s Disease, vol. 2013, pp. 1–8, 2013.

[11] B. E. Sakar et al., "Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings," IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 4, pp. 828-834, 2013.

[12] F. Ros and S. Guillaume, “From supervised instance and feature selection algorithms to dual selection: A review,” in *Sampling Techniques for Supervised or Unsupervised Tasks. Unsupervised and Semi-Supervised Learning* (F. Ros and S. Guillaume, eds.), pp. 83– 128, Cham: Springer, 2020.

[13] I. Guyon and A. Elisseeff, “An introduction to feature extraction,” in *Studies in Fuzziness and Soft Computing* (I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, eds.), vol. 207, pp. 1–25, Berlin, Heidelberg: Springer, 2006.

[14] C. Lazar et al., “A survey on filter techniques for feature selection in gene expression microarray analysis,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, pp. 1106–1119, 2012.

[15] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” IEEE Transactions on Neural Networks, vol. 5, pp. 537–550, 1994.

[16] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” Journal of Machine Learning Research, vol. 3, pp. 1289–1306, 2003.

[17] N. Kwak and C. H. Choi, “Input feature selection for classification problems,” IEEE Transactions on Neural Networks, vol. 13, pp. 143– 159, 2002.

[18] A. Eesa and W. Arabo, “A normalization method for backpropagation: A comparative study,” Science Journal of University of Zakho, vol. 5, p. 319, 2017.

[19] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[20] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” The Annals of Statistics, vol. 29, no. 5, 2001.

[21] J. H. Friedman, “Stochastic gradient boosting,” Computational Statistics & Data Analysis, vol. 38, no. 4, pp. 367–378, 2002.

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “Catboost: unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[23] R. E. Schapire, "Explaining AdaBoost," in *Empirical Inference* (B. Schölkopf, Z. Luo, and V. Vovk, eds.), Berlin, Heidelberg: Springer, 2013.

[24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[25] LightGBM GitHub Repository, “LightGBM.” Available:https:// github.com/microsoft/LightGBM, 2016. Accessed on 8 September 2023.

[26] R. Zhang, F. Nie, X. Li, and X. Wei, “Feature selection with multi-view data: A survey,” Information Fusion, vol. 50, pp. 158–167, 2019.

[27] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

[28] O. Yaman, F. Ertam, and T. Tuncer, “Automated Parkinson’s disease recognition based on statistical pooling method using acoustic features,” Medical Hypotheses, vol. 135, p. 109483, 2020.

[29] T. Tuncer, S. Dogan, and U. R. Acharya, “Automated detection of Parkinson’s disease using minimum average maximum tree and singular value decomposition method with vowels,” Biocybernetics and Biomedical Engineering, vol. 40, no. 1, pp. 211–220, 2020.

[30] O. Bchir, “Parkinson’s Disease Classification using Gaussian Mixture Models with Relevance Feature Weights on Vocal Feature Sets,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 4, 2020.

[31] A. S. Ashour, M. K. A. Nour, K. Polat, Y. Guo, W. Alsaggaf, and A. El-Attar, “A Novel Framework of Two Successive Feature Selection Levels Using Weight-Based Procedure for Voice-Loss Detection in Parkinson’s Disease,” IEEE Access, vol. 8, pp. 76193–76203, 2020.

[32] Z. Karapinar Senturk, “Early diagnosis of Parkinson’s disease using machine learning algorithms,” Medical Hypotheses, vol. 138, p. 109603, 2020.

[33] J. Goyal, P. Khandnor, and T. C. Aseri, “Analysis of Parkinson’s disease diagnosis using a combination of Genetic Algorithm and Recursive Feature Elimination,” *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 2020.

[34] R. Lamba, T. Gulati, A. Jain, and P. Rani, “A Speech-Based Hybrid Decision Support System for Early Detection of Parkinson’s Disease,” Arabian Journal for Science and Engineering, vol. 48, no. 2, pp. 2247–2260, 2023.

[35] A. M. Elshewey, M. Y. Shams, N. El-Rashidy, A. M. Elhady, S. M. Shohieb, and Z. Tarek, “Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification,” Sensors, vol. 23, no. 4, p. 2085, 2023.

[36] R. Lamba, T. Gulati, H. F. Alharbi, and A. Jain, “A hybrid system for Parkinson’s disease diagnosis using machine learning techniques,” International Journal of Speech Technology, vol. 25, no. 3, pp. 583–593, 2022.

[37] R. Lamba, T. Gulati, and A. Jain, “A Hybrid Feature Selection Approach for Parkinson’s Detection Based on Mutual Information Gain and Recursive Feature Elimination,” Arabian Journal for Science and Engineering, vol. 47, no. 8, pp. 10263–10276, 2022.

[38] M. A. Abdel-fattah, R. H. Eid, and A. E. Yakoub, "A hybrid approach for enhancing the classification of Parkinson’s disease using swarm optimization," Journal of Theoretical and Applied Information Technology, vol. 101, 2023.

[39] P. K. Chawla et al., “Parkinson’s disease classification using nature inspired feature selection and recursive feature elimination,” Multimedia Tools and Applications, vol. 83, no. 12, pp. 35197–35220, 2024.

[40] H. Al-Najjar, N. Al-Rousan, and D. Al-Najjar, “Hybrid grey wolf and whale optimization for enhanced Parkinson’s prediction based on machine learning models using biomedical sound,” Informatics in Medicine Unlocked, vol. 48, p. 101524, 2024.

[41] D. Elreedy, A. F. Atiya, and F. Kamalov, “A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning,” Machine Learning, 2023.

[42] N. Shehab, M. Badawy, and H. A. Ali, “Toward feature selection in big data preprocessing based on hybrid cloud-based model,” The Journal of Supercomputing, vol. 78, no. 3, pp. 3226–3265, 202٢.

[43] UCI Machine Learning Repository, “Parkinson’s DataSet.” Available: https://archive.ics.uci.edu/dataset/174/parkinsons, 2007. Accessed date: 1 August 2023.

[44] UCI Machine Learning Repository, “Parkinson’s DataSet.” Available: https://archive.ics.uci.edu/dataset/470/parkinson+s+disease+ classification, 2018. Accessed date: 1 August 2023.

المجلد (١٢) العدد (١) (٢٠٢٥)

المجلة الدولية للحاسبات والمعلومات

متاح على الإنترنت على الرابط : https://ijci.journals.ekb.eg/

Egyptian Knowledge Bank
بنك المعرفة المصري

طريقة هجينة تعتمد على الكلام للكشف عن مرض باركنسون باستخدام معامل الارتباط بيرسون والمعلومات المتبادلة

محمد الخرادلي[١]، خالد محمد أمين [٢] ، أسامة أبو سعدة[٣]، مينا إبراهيم [٤]

[١]قسم تكنولوجيا المعلومات – كلية الحاسبات والمعلومات – جامعة كفرالشيخ

[٢]قسم تكنولوجيا المعلومات – كلية الحاسبات والمعلومات – جامعة المنوفية

[٣]قسم علوم الحاسب – كلية الحاسبات والمعلومات – جامعة كفرالشيخ

[٤]قسم ذكاء الالة – كلية الذكاء الاصطناعي – جامعة المنوفية

الملخص:

مرض باركنسون (PD) هو اضطراب تنكسي عصبي مزمن ومتقدم يؤثر على الحركة. أظهرت الدراسات أن صعوبات الكلام قد تظهر في وقت مبكر من المرض، مما يشير إلى إمكانية استخدامها كمؤشر تشخيصي مبكر. تتناول طريقتنا المقترحة نهجاً هجينا للكشف عن مرض باركنسون يعتمد على الارتباط البيرسوني (PC) والمعلومات المتبادلة (MI) . يجمع النهج بين PC و MI لتحديد الميزات ذات الصلة في إشارات الكلام، ويتم استخدام هذه الميزات لتدريب خمسة نماذج من التعلم الآلي، وهي XGBoost ، GBoost، CatBoost، AdaBoost، LightGBM. تم استخدام مجموعتين من البيانات من مستودع التعلم الآلي بجامعة كاليفورنيا للتقييم. وللتغلب على تحدي الفئات غير المتوازنة في مجموعات البيانات، تم تطبيق تقنية التوليد الزائد للأقلية التركيبية (SMOTE) لتحقيق تمثيل أكثر توازناً. يختار نهج PCMI المقترح ١٠ ميزات من المجموعة الأولى و٥٥ ميزة من المجموعة الثانية. أظهرت النتائج أن نموذج CatBoost مع SMOTE ونهج PCMI حقق دقة تبلغ ٩٧.٣% باستخدام طريقة التحقق المتبقي ٧٥:٢٥ و٩٧.٢% باستخدام طريقة التحقق المتقاطع 10-fold لمجموعة البيانات الأولى، بينما حقق نموذج LightGBM مع SMOTE ونهج PCMI دقة تبلغ ٩٥.٦% باستخدام طريقة التحقق المتبقي ٦٠:٤٠ و٩٧.٦% باستخدام طريقة التحقق المتقاطع 10-fold لمجموعة البيانات الثانية.