INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Machine Learning Approach for an Early Prediction of Parkinson's disease

Rajeshwar Prasad[1], **Amit Kumar Saxena**[2]*

1 Scholar, Department of Computer Science & Information Technology, Guru Ghasidas Vishwavidyalaya (C.U.), Koni, Bilaspur, (C.G.), 495009, Chhattisgarh, India
2 Professor, Department of Computer Science & Information Technology, Guru Ghasidas Vishwavidyalaya (C.U.), Koni, Bilaspur, (C.G.), 495009, Chhattisgarh, India

OPEN ACCESS

Received: 26-06-2024

Accepted: 28-07-2024

Published: 20-08-2024

Citation: Prasad R, Saxena AK (2024) A Machine Learning Approach for an Early Prediction of Parkinson's disease. Indian Journal of Science and Technology 17(33): 3410-3418. https://doi.org/10.17485/IJST/v17i33.2091

* **Corresponding author**.

amitsaxena65@rediffmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: Parkinson's disease (PD) is a critical disease and the early detection of PD is of high importance to start a proper treatment to avoid its emergence. An Optimized Filter-based Machine Learning Classifier (OFMLC) approach is used in this paper for prediction of PD. **Method**: The PD data is first balanced using random over-sampling, random down-sampling, and the Synthetic Minority Over-sampling Technique (SMOTE). Then, the SelectKBest approach is used to select K filter-based features. The dataset collected from the University of Oxford library has 195 items, 24 attributes (features), and two classes viz. healthy individuals and individuals with PD symptoms. The extensive experiments were performed with the 10 and 18 most relevant features of the dataset. The results are compared with some of the existing methods namely WFSK-NN, RFEANN, RFESVM, and ECL. The performance of the model is evaluated using the metrics: accuracy, precision, F1 scores, and recall. **Findings**: The proposed approach with 10 most important features of the PD dataset produced a maximum accuracy of 99.98% using Random Forest (RF), Gradient Boost (GB), Voting, Stacking, XG Boost (XG), LightGBM (LG) and CatBoost (CB) classifiers; while with 18 features, the Random Forest RF, Bagging, Stacking, LightGBM (LG) and CatBoost (CB) produced 99.98 % accuracy. **Novelty**: The proposed method applied firstly the data balancing method to balance the data for maintaining a proper class ratio among the patterns of the PD dataset to ensure unbiased data selection for experiments. The SelectKBest method has the advantage of improving the model's performance, making computations easier, and resulting in higher accuracy over other methods. The OFMLC approach is applied to the reduced feature subset which improves the accuracy of the model with different classifiers which justifies the strength of the proposed method.

Keywords: Machine Learning; Filter-based Feature Selection; Data Balancing; SelectKBest; Grid Search; Ensemble Classifiers

1 Introduction

Parkinson's Disease (PD) is a progressive neurological disorder characterized by physical and mental symptoms, including compromised postural stability, resting tremors, slowed movement, and stiffness. Several variables, such as environmental exposure, genetics, and ethnicity, contribute to the development of this disorder/disease. It affects several areas of the brain, including those associated to dopamine and others that are not related to dopamine. Early identification of PD is crucial to avoid and care in case it already started. The Unified PD Rating Scale (UPDRS) is an extensively investigated tool used to assess the severity and development of PD symptoms [1]. The use of Machine Learning (ML) techniques has provided helping tools for the detection of PD. These technologies analyze medical information to solve diagnostic issues and provide better-informed medical choices and treatments [2,3]. Feature selection is a process that reduces the number of features in the original set by removing redundant, noisy, or duplicate features. Grid search, used in the proposed model, is a key method in machine learning utilized to boost model performance, automate the tuning process, and ensure reliability and robust evaluation. This simplifies models, makes interpretation simpler, and lowers training durations [4–8].

Khamparia et al. [9] used a convolutional neural network (CNN) classifier with voice classification datasets and obtained a training phase accuracy of above 77%. In another study, Srinivasan et al. [10] applied FNN and KSVM models where FNN achieved an accuracy of 99.11%, with 98.78% recall, 99.96% precision, and a 99.23% F1-score. Similarly, the KSVM achieved an accuracy of 95.89%, recall of 96.88%, precision of 98.71%, and an F1-score of 97.62%. Zhang et al. [11] used a Naive Bayes (NB) classifier to identify significant features in speech data obtained from both individuals with PD and healthy individuals. Their approach achieved a detection accuracy of 69.24% and a precision of 96.02% using 22 voice features. Kadiri et al. [12], used support vector machines (SVM) to extract features from audio signals by using shifting delta cepstral (SDC) and single frequency filtering cepstral coefficients (SFFCC). Their approach led to a performance improvement of 9% compared to traditional features. The accuracy of the detection was 73.33%, while the F1 score was 73.32%. The decision forest models SysFor and ForestPA were used in a separate investigation [13]. The investigation revealed that the random forest classifier surpassed these models, with an average detection accuracy rate of 93.58% on incremental trees. In [14] by Govindu et al., four machine learning models were employed: RF, SVM, KNN, and LR. RF model attained an accuracy of 91.83% and a sensitivity of 0.95. A different research work cited by Rana et al. [15] used SVM, NB, KNN, and ANN. The SVM earned an accuracy of 87.17%, the NB achieved an accuracy of 74.11%, the ANN achieved an accuracy of 96.7%, and the KNN achieved an accuracy of 87.17%. Out of all these models, the ANN exhibited the highest level of accuracy, making it the most precise one. The methodology was examined in a research study proposed by Saeed et al. [16], which assessed the effectiveness of ML algorithms on both unaltered and modified datasets. Their findings demonstrated that the use of WFSK-NN significantly improved the accuracy of PD prediction by 88.33 %. Senturk et al [17] identified two diagnosis methods for PD named RFEANN and RFESVM. The RFESVM method outperformed the other methods in the testing, with an accuracy of 93.84%. In [18], Hussain et al. used the Ensemble of Classifiers (ECL) approach to categorize PD. ECL exhibited superior performance compared to individual pre-trained models.

Nevertheless, the early identification of PD remains challenging due to the complex and varied nature of the symptoms. Conventional diagnostic methods may rely heavily on the examiner's expertise and may fail to detect early indications. Traditional methods provide challenges in accurately identifying and investigating speech impairments that indicate the onset of PD. Moreover, the process of identifying suffers from challenges such as unbalanced datasets, irrelevant or duplicated data, and inadequate model performance. Prior studies mentioned earlier have shown encouraging outcomes by using various ML methodologies. However, there are significant deficiencies that need to be addressed, specifically the biased performance of algorithms caused by imbalanced datasets and inadequate feature selection. This work addresses these concerns by using data balancing techniques such as random over-sampling, random under-sampling, and the Synthetic Minority Over-sampling Technique (SMOTE) [19] to rule out biased performance. In addition, the SelectKBest method is used for filter-based feature selection, which enhances the inclusion of important features in OFMLC. Moreover, the use of Grid search to optimize the performance of models automates the process of adjusting hyperparameters and provides reliable and adaptable evaluation. This approach is anticipated to improve the reliability and accuracy of PD detection, offering a more dependable and unbiased method compared to those documented in the literature.

The paper is arranged as follows. Section 2 is a comprehensive and detailed overview of the methodology. Section 3 contains results and a discussion, Section 4 presents a conclusion of the findings with future scope, and Section 5 provides references.

2 Methodology

Figure 1 displays the flow of experimental actions.

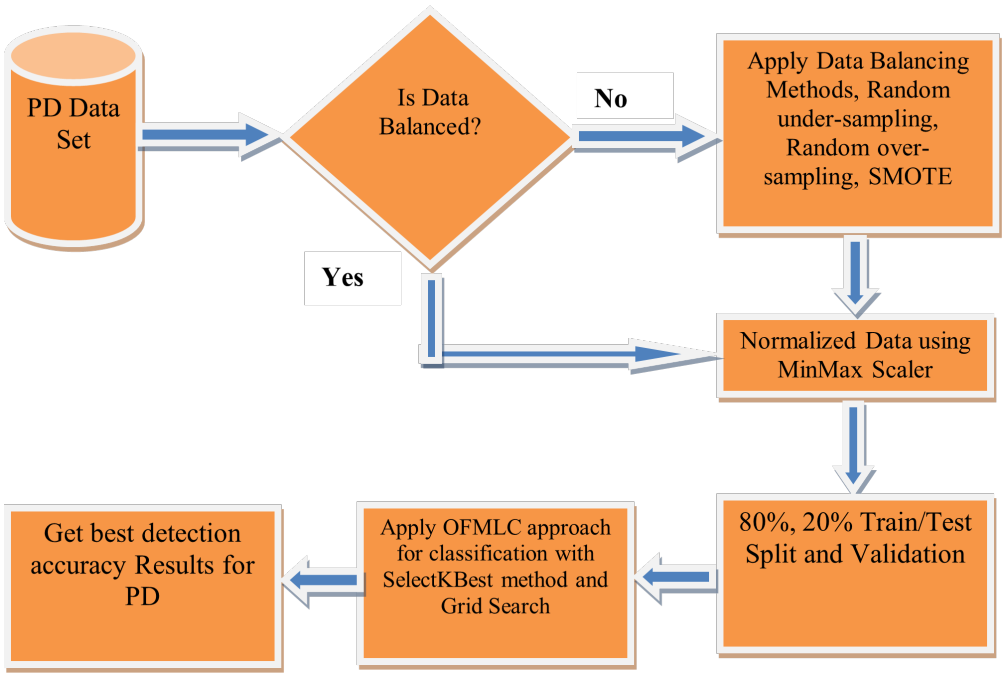

Fig 1. Work flow for proposed method

2.1 Dataset

This study makes use of the Parkinson's dataset, which is accessible online via Kaggle and the UCI repository. Little et al.[20] introduced this dataset and established the National Center for Voice to assist in retrieving material from the University of Oxford (UO) library. In 2008, little made the dataset available via the UCI ML Repository[20]. The dataset comprises data from 31 persons, of whom 23 had PD (16 men and 7 women) and 8 were healthy i.e. without PD (3 men and 5 women). The dataset's feature descriptions are shown in Table 1. The dataset comprises 195 objects distributed over 24 features. The 22 contestants had six records each, while the remaining 9 had seven. The individuals' ages varied from 46 to 85 years, with a mean of 65.8 years and a standard deviation of 9.8 years. The period since diagnosis varied between 0 and 28 years. Each row shows a 36-second speaking tape. Voice recordings were made in a soundproof room at an industrial acoustics business, using a microphone placed 8 cm from the lips and set up according to[20]. This dataset's status feature reflects 0 for healthy people and 1 for individuals with PD symptoms. In a typical classification language, this refers to it as a binary class, decision variable, or target.

Table 1. PD Data Features Description

Features name	Purpose
MDVP:Fo(Hz)	Fundamental frequency of pitch period.
MDVP:Fhi(Hz)	Upper limit of fundamental frequency or maximum threshold.
MDVP:Flo(Hz)	Lower limit or minimal vocal fundamental frequency.
MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP	These are various Kay Pentax's multi-dimensional voice program (MDVP) measures. MDVP is a traditional measure of frequency of vibrations in vocal folds at pitch period to vibrations at start of next cycle called pitch mark[20]
MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA	Measures of absolute difference between frequencies of each cycle, after normalizing the average.
NHR, HNR	Signal to noise and tonal ratio measures, that indicate robustness of environment to noise.
Status	0 indicates healthy person while 1 indicates PD. A decision variable or class
RPDE	Recurrence Period Density Entropy quantifies the extent to which signal is periodic.

Continued on next page

Table 1 continued

DFA	Detrended Fluctuation Analysis or DFA measures the extent of stochastic self-similarity of noise in speech signals.
spread1, spread2	Analysis of extent or range of variations in speech with respect to MDVP: Fo(Hz).
D2	Correlation dimension is used to identify dysphonia in speech using fractal objects. It is a nonlinear, dynamic attribute.
PPE	Pitch Period entropy is used to assess abnormal variations in speech on a logarithmic scale.

2.2 Preprocessing

In the preprocessing phase, duplicate and missing data were detected and removed. The dataset's status target feature exhibited an imbalanced distribution, with 25% of the instances being healthy controls (HC) and 75% representing PD. There were a total of 147 instances of PD and 48 cases of HC. In order to address the issue of imbalanced class labels, data balancing techniques namely random under-sampling, random over-sampling, and SMOTE were used [19]. Figure 2 displays the frequency of each class label in the dataset after it has been balanced. All the data, with the exception of the class designation, is given in numerical format. During the training and testing data phase, the data was normalized using the MinMax Scaler approach, which rescaled the values to a range of 0 to 1 [21]. The balanced data was divided into 80% for training and 20% for testing. Parameter values were optimized via grid search. The proposed OFMLC approach used the filter-based SelectKBest [22] method to choose the most important features. In the SelectKBest method, the variable "K" represents the desired number of important features to be selected. By choosing just the most important features from the initial set, the model's performance is enhanced, computations are simplified, and the risk of over-fitting is mitigated [23]. In this study, a total of 10 and 18 (K=10, K=18) features in the primary dataset were selected as relevant features from the dataset. A variety of Python libraries were used for preprocessing. The libraries include NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn (Sklearn) [21]. Python is often used for the purpose of importing and organizing data due to its robust data manipulation and analysis capabilities. Sklearn, the primary ML package in Python, is well recognized for its effectiveness and reliability. It offers a consistent interface and a wide range of methods for tasks like as classification, regression, clustering, and dimensionality reduction [21].

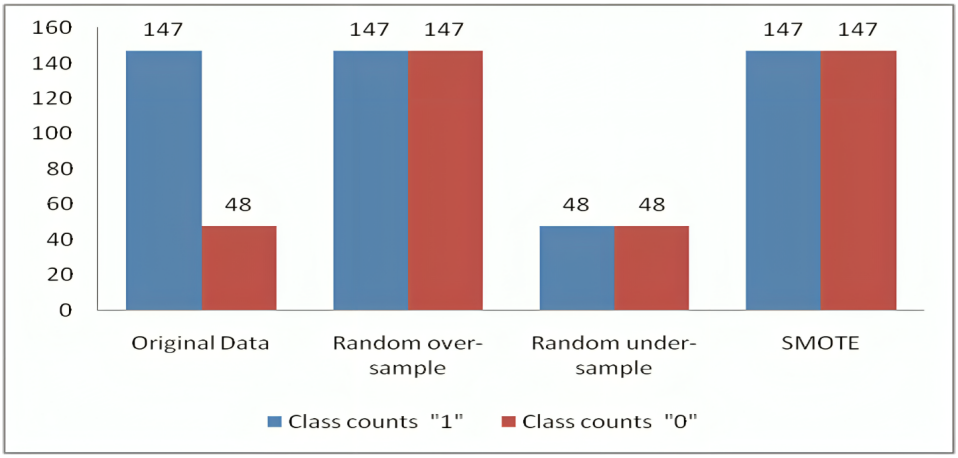

Fig 2. Different forms of data after applying Data Balancing Methods

2.3 Training / Testing Phase

In the training and testing phase, a balanced dataset is divided into an 80:20 ratio, and only the most relevant features are picked. For training purposes, a variety of eighteen ML classifiers are utilized, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), AdaBoost (AB), Voting, Bagging, Stacking, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Neural Network, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Gaussian Process (GP), XGBoost (XG), LightGBM (LG), and CatBoost (CB). Additionally, ensemble classifiers consisting of Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR) are employed with stacking, bagging, and voting approaches [24].

2.4 Experiments

The studies were conducted on an Intel i5-2nd generation CPU running on the Windows-10 Operating System. The studies were conducted using the Jupyter Notebook Python platform. The experimental findings of the suggested approach are shown in Tables 2 and 3.

Algorithm: Steps used in OFMLC

- **Input**: Load PD Dataset [20]
- Apply Data Balancing: Apply Data Balancing techniques random over-sampling, random under-sampling, and SMOTE [19] to Dataset
- Select Relevant Attributes: Apply feature selection techniques SelectKBest [22] to Balanced Dataset by SMOTE
- Train Classifiers: Train classifiers using the Balanced and preprocessed dataset (Selected Attributes)
- Assess Model Effectiveness: Build and evaluate the model using individual classifiers and ensemble approaches [24]. Evaluate metrics accuracy, precision, recall, and F1 score of the Trained Model
- Improve Model Performance: Utilize techniques of Grid search to optimize the parameters of the Trained Model
- **Outcome:** The model predicts/outputs strong PD identification/classification ability

Table 2. Classification Accuracies obtained using various 18 classifiers with OFMLC

Models	Accuracies in %					
	Normal PD Data	Random under-sample Data	Random over-sampled Data	SMOTE Data	OFMLC with 10 features	OFMLC with 18 features
LR	89. 74	95. 00	76. 27	84. 75	91. 53	91. 53
DT	89. 74	90. 00	94. 92	98. 31	98. 31	98. 31
RF	94. 87	95. 00	96. 61	99. 98	99. 98	99. 98
GB	94. 87	85. 00	96. 61	96. 61	99. 98	94. 92
AB	87. 18	85. 00	93. 22	96. 61	93. 22	96. 61
Voting	92. 31	90. 00	96. 61	98. 31	99. 98	93. 22
Bagging	94. 87	90. 00	96. 61	99. 98	98. 31	99. 98
Stacking	92. 31	90. 00	96. 61	99. 98	99. 98	99. 98
SVM	89. 74	90. 00	81. 36	94. 92	94. 92	94. 92
KNN	94. 87	90. 00	94. 92	96. 61	96. 61	96. 61
NB	71. 79	95. 00	76. 27	86. 44	86. 44	86. 44
ANN	87. 18	95. 00	77. 97	84. 75	88. 14	88. 14
LDA	87. 18	70. 00	77. 97	89. 83	88. 14	86. 44
QDA	89. 74	99. 98	98. 31	94. 92	94. 92	94. 92
GP	89. 74	95. 00	76. 27	88. 14	88. 14	88. 14
XG	94. 87	95. 00	96. 61	99. 98	99. 98	98. 31
LG	92. 31	95. 00	94. 92	99. 98	99. 98	99. 98
CB	94. 87	95. 00	96. 61	99. 98	99. 98	99. 98

Table 3. OFMLC approach with 10 important features on different performance evaluation metrics (All figures in %)

Models	Accuracy	Precision	Recall	F1 Score
LR	91. 53	92. 86	89. 66	91. 23
DT	98. 31	99. 98	96. 55	98. 25
RF	99. 98	99. 98	99. 98	99. 98
GB	99. 98	99. 98	99. 98	99. 98
AB	93. 22	96. 30	89. 66	92. 86
Voting	99. 98	99. 98	99. 98	99. 98
Bagging	98. 31	96. 67	99. 98	98. 31
Stacking	99. 98	99. 98	99. 98	99. 98
SVM	94. 92	99. 98	89. 66	94. 55

Continued on next page

Table 3 continued

KNN	96. 61	96. 55	96. 55	96. 55
NB	86. 44	99. 98	72. 41	84. 00
ANN	88. 14	86. 67	89. 66	88. 14
LDA	88. 14	89. 29	86. 21	87. 72
QDA	94. 92	93. 33	96. 55	94. 92
GP	88. 14	86. 67	89. 66	88. 14
XG	99. 98	99. 98	99. 98	99. 98
LG	99. 98	99. 98	99. 98	99. 98
CB	99. 98	99. 98	99. 98	99. 98

2.5 Performance Metrics:

Performance metrics assess the effectiveness of a classification model by comparing its predictions to the true labels in the dataset[23]. The metrics include accuracy, precision, recall, and F1 Score. Accuracy measures the ratio of successfully recognized instances to the total amount of the dataset as seen below:

- **Accuracy** is a quantitative measure that calculates the proportion of correctly identified instances to the total number of occurrences in a dataset. The calculation is as stated:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100$$

- **Precision** is a metric that quantifies the ratio of accurate positive predictions to all positive predictions generated by the model. The calculation is as follows:

$$Precision = \frac{TP}{TP + FP} \times 100$$

- **Recall**, also known as sensitivity, is the ratio of correctly predicted positive cases to the total number of positive instances in the dataset. The calculation is as follows:

$$Recall = \frac{TP}{FN + TP} \times 100$$

- The **F1 score** is calculated as the harmonic mean of accuracy and recall, offering a balanced evaluation of both measurements. The calculation is as follows:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100$$

3 Result and Discussion

Table 2 presents the accuracy results of eighteen ML classifiers employed in this study, which are based on experimental observations. The first column lists the classifiers/models used for experiments. The column numbers 2-7 show the accuracy values obtained for Normal PD Data, Random under-sample Data, Random over-sampled Data, SMOTE Data, and OFMLC approach with 10 and 18 important features, respectively. Table 3 displays the best results for the OFMLC method using the 10 significant features, as measured by the metrics precision, recall, and F1 scores. Table 2 shows that for LR model, the highest accuracy achieved is 95% with Random under-sample data for balancing. The OFMLC with 10 features generates an accuracy of 91.53% with 10 features and 18 features equally. The performance of OFMLC with 10 and 18 features with all 18 classifiers is shown in Table 2. It is observed that by employing the OFMLC approach with the 10 best features, the models namely RF, GB,

Voting approach, Stacking approach, XG, LG, and CB, obtained a maximum accuracy of 99.98% each. Another classifier, ANN achieved the highest accuracy of 95% with under-sample balancing of data whereas OFMLC with 10 and 18 features achieved an accuracy of 88.14%. When the performance of OFMLC with 10 and 18 features is considered at par for all models, it is noted that it achieves the highest accuracy for models DT(98.31%), RF(99.98%), Stacking(99.98%), SVM(94.92%), KNN(96.61%), LG(99.98%) and CB(99.98%). Further, the OFMLC with 10 features produces the highest accuracy 99.98% with the models GB, Voting, and XG. The OFMLC with 18 features produces the highest accuracies with the models' AB (96. 61%), and Bagging (99.98%). Table 2 can be referred to for a detailed explanation. It is deduced that the proposed approach OFMLC provides the highest accuracies in ten out of 18 models with only 10 features which is nearly 44% of the total number of features in the PD dataset. Further, it is noted that the OFMLC provides the highest accuracies in nine out of 18 models with only 18 features which is nearly 80% of the total number of features in the PD dataset.

Table 3 presents the performance of the OFMLCapproach with 10 important features with four common evaluation metrics namely: Accuracy, Precision, Recall, and F1 Scores. The OFMLC yields the highest values on different metrics. The proposed approach achieves the highest values in all four parameters for the models RF, GB, Voting, Stacking, XG, LG, and CB at 99.98% each.

The NB model obtained an accuracy of 86.44%, precision score of 99.98%, a recall score of 72.41%, and an F1 score of 84.00% which represents a weak performance. In general, if all four parameters achieve high values together, than the model is supposed to perform satisfactory. The remaining classifiers attained optimum accuracy and delivered satisfactory results. A prior study that used the WFSK-NN [16] approach on data related to PD reported an accuracy of 88.33%. In a separate study, Senturk [17] used the RFEANN and RFESVM techniques. The RFESVM [17] approach attained an accuracy of 93.84%. Using Parkinson's data, another research [18] obtained 96.6% accuracy with an ECL [18]. The comparative results of the proposed method with others reported are shown in Table 4. It is noted that the OFMLC succeeds in achieving the highest accuracy 99.98% with merely 10 features only with various models for classification.

Based on a comparative analysis, shown in Table 4 this study discovered that the suggested OFMLC approach, using the top 10 significant features, attained a much higher accuracy of 99.98% compared to earlier research investigations.

Table 4. Comparison of OFMLC with previous referenced researches

References/ Methods	Used Models	Accuracy in %
[16]	WFSK-NN	88.33
[17]	RFEANN, RFESVM	91.54, 93.84
[18]	VGG-16, ResNet-50, MobileNet, ECL	93.8, 93.3, 91.4, 96.6
Proposed work OFMLC approach with 10 important features	RF, GB, Voting, Stacking, XG, LG, CB	**99.98**

4 Conclusion

Detection of PD at an early stage is very crucial. This study examines the use of data and features in the categorization of PD to healthy and PD-grieved patients. Ensuring that data is balanced with evenly distributed class labels is crucial for ML classifiers to accurately identify patterns, as it reduces the risk of over-fitting and incorrect detections. This work applies data balancing approaches such as random over-sampling, random under-sampling, and SMOTE [19]. The judicious selection of relevant features may greatly enhance the accuracy of detection as the irrelevant/ redundant features might weaken the functioning of the classifier. The SelectKBest strategy was used, using a filter-based methodology that explicitly selects the K most important and essential attributes. The OFMLC approach employs Grid Search for parameter optimization, which further enhances classification accuracy. The previous studies mostly did not consider the data balancing issues while partitioning data into training and testing subparts. The SelectKBest method has the advantage of improving the model's performance, making computations easier, and resulting in higher accuracy over other methods. The model has been investigated for 10 and 18 important features to provide better classification with the minimum possible number of features. The OFMLC strategy yielded the highest accuracy (99.98%) when using merely the top 10 significant features using classifiers: RF, GB, Voting approach, Stacking approach, XG, LG, and CB. The proposed approach produces an accuracy of 99.98% as opposed to 96% as achieved in previous research. Moreover, the OFMLC utilizes only 43% of the features of the original PD dataset which reduces the

computational complexity, saves time, and improves accuracy due to the elimination of harmful redundant features. Thus, the use of the OFMLC technique has the potential to decrease treatment expenses by facilitating timely initial diagnosis. This technique may function as both an educational instrument for medical students and a preliminary diagnostic instrument for physicians. In future work, it would be advantageous to conduct a more extensive analysis of machine learning and deep learning methodologies by using a bigger dataset of disease data in combination with the OFMLC. Moreover, it would be advantageous to investigate alternative methodologies for feature selection to improve the performance of existing disease prediction methods.

Table 5. Abbreviations used in the paper

Abbreviation	Full Name	Abbreviation	Full Name	Abbreviation	Full Name
PD	Parkinson's Disease	SVM	Support Vector Machine	GP	Gaussian Process
LR	Logistic Regression	KNN	K Nearest Neighbour	XG	XGBoost
DT	Decision Tree	NB	Naive Bayes	LG	LightGBM
RF	Random Forest	ANN	Artificial Neural Network	CB	CatBoost
GB	Gradient Boosting	LDA	Linear Discriminant Analysis	PCA	Principal Components Analysis
AB	AdaBoost	QDA	Quadratic Discriminant Analysis	SMOTE	Synthetic Minority Over-sampling Technique
KSVM	Kernel Support Vector Machine	ML	Machine Learning	RBF	Radial Basis Function
ANN	Artificial Neural Network	LASSO	Least Absolute Shrinkage and Selection Operator	MLP	Multilayerd Perceptron
WFSK-NN	Wrapper- based Features Selection method with KNN	RFEANN	Recursive Feature Elimination methods with Artificial Neural Networks	RFESVM	Recursive Feature Elimination methods with Support Vector Machines
ECL	Ensemble Classifier	OFMLC	Optimized Filter based Machine Learning Classifiers	FNN	Feed Forward Neural network

References

1) Chen HL, Wang G, Ma C, Cai ZN, Liu WB, Wang SJ. An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson's disease. *Neurocomputing*. 2016;184:131–144. Available from: https://doi.org/10.1016/j.neucom.2015.08.139.
2) Brunak S. The bioinformatics: Machine learning approach. MIT Press. 2001. Available from: https://spygirl37.wordpress.com/wp-content/uploads/2018/04/bioinformatics-the-machine-learning-approach-second-edition-pierre-baldi-soren-brunak-1.pdf.
3) Naranjo L, Perez CJ, Campos-Roca Y, Martin J. Addressing voice recording replications for Parkinson's disease detection. *Expert Systems with Applications*. 2016;46:286–292. Available from: https://doi.org/10.1016/j.eswa.2015.10.022.
4) Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23:2507–2517. Available from: https://doi.org/10.1093/bioinformatics/btm344.
5) Hsu TW, Pare S, Meena MS, Jain DK, Li DL, Saxena A, et al. An Early Flame Detection System Based on Image Block Threshold Selection Using Knowledge of Local and Global Feature Analysis. *Sustainability*. 2020;12(21):8899–8899. Available from: https://doi.org/10.3390/su12218899.
6) Chugh D, Mittal H, Saxena A, Chauhan R, Yafi E, Prasad M. Augmentation of Densest Subgraph Finding Unsupervised Feature Selection Using Shared Nearest Neighbor Clustering. *Algorithms*. 2023;16:28–28. Available from: https://doi.org/10.3390/a16010028.
7) Saxena A, Chugh D, Mittal H, Sajid M, Chauhan R, Yafi E, et al. A Novel Unsupervised Feature Selection Approach Using Genetic Algorithm on Partitioned Data. *Adv Artif Intell Mach Learn*. 2022;2:500–515. Available from: https://www.oajaiml.com/uploads/archivepdf/82321134.pdf.
8) Saxena A, Wang J, Sintunavarat W. An Empirical Study on Initializing Centroid in K-Means Clustering for Feature Selection. *Int J Softw Sci Comput Intell (IJSSCI)*. 2021;13(1):1–16. Available from: https://www.igi-global.com/gateway/issue/254095.
9) Khamparia A, Gupta D, Nguyen NG, Khanna A, Pandey B, Tiwari P. Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access*. 2019;7:7717–7727. Available from: https://doi.org/10.1109/ACCESS.2018.2888286.
10) Srinivasan S, Ramadass P, Mathivanan S. Detection of Parkinson disease using multiclass machine learning approach. *Scientific Reports*. 2024;14:13813–13813. Available from: https://doi.org/10.1038/s41598-024-64004-9.
11) Zhang L, Qu Y, Jing JB, Gao L, Liang Z, Z. An intelligent mobile-enabled system for diagnosing Parkinson disease: Development and validation of a speech impairment detection system. *JMIR Public Health Surveill*. 2020;8:18689–18689. Available from: https://doi.org/10.2196/18689.
12) Kadiri SR, Kethireddy R, Alku P. Parkinson's disease detection from speech using single frequency filtering cepstral coefficients. In: Proceedings of the Interspeech 2020. 2020;p. 25–29. Available from: https://doi.org/10.21437/Interspeech.2020-2634.
13) Pramanik M, Pradhan R, Nandy P, Bhoi AK, Barsocchi P. Machine learning methods with decision forests for Parkinson's detection. *Appl Sci*. 2021;11:581–581. Available from: https://doi.org/10.3390/app11020581.

14) Govindu A, Palwe S. Early detection of Parkinson's disease using machine learning. *Procedia Computer Science*. 2023;218:249–261. Available from: https://doi.org/10.1016/j.procs.2023.01.007.
15) Rana A, Dumka A, Singh R, Rashid M, Ahmad N, Panda MK. An efficient machine learning approach for diagnosing Parkinson's disease by utilizing voice features. *Electronics*. 2022;11(22):3782–3782. Available from: https://doi.org/10.3390/electronics11223782.
16) Saeed F, Al-Sarem M, Al-Mohaimeed, Emara M, Boulila A, Alasli W, et al. Enhancing Parkinson's disease prediction using machine learning and feature selection methods. *Comput Mater Continua*. 2022;71(3):5639–5658. Available from: https://doi.org/10.32604/cmc.2022.023124.
17) Senturk ZK. Early diagnosis of Parkinson's disease using machine learning algorithms. *Med Hypotheses*. 2020;138:109603–109603. Available from: https://doi.org/10.1016/j.mehy.2020.109603.
18) Hussain MM, Weslin D, Kumari S, Umamaheswari S, Kamalakannan K. Enhancing Parkinson's disease identification using ensemble classifier and data augmentation techniques in machine learning. *Clin eHealth*. 2023;6:150–158. Available from: https://doi.org/10.1016/j.ceh.2023.11.002.
19) Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *JAIR*. 2002;16:321–357. Available from: https://doi.org/10.1613/jair.953.
20) Little M. UCI Machine Learning Repository. 2008. Available from: https://archive.ics.uci.edu/ml/datasets/parkinsons.
21) Desai R. Top 10 Python Libraries for Data Science. 2022. Available from: https://towardsdatascience.com/top-10-python-libraries-for-data-sciencecd82294ec266.
22) Kavya D. Optimizing Performance: SelectKBest for Efficient Feature Selection in Machine Learning. 2023. Available from: https://medium.com/@Kavya2099/optimizingperformance-selectkbest-for-efficient-feature-selection-in-machine-learning-3b635905ed48.
23) James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning;vol. 112. and others, editor;New York, USA. Springer. 2013. Available from: https://doi.org/10.1007/978-1-4614-7138-7.
24) Rahman MF, Basri H, Amin NS, Salleh H. A review of augmented reality classification methods. *International Journal of Computer Science and Network*. 2010;10(7):130–138. Available from: https://www.researchgate.net/publication/315701832_Classifying_different_types_of_augmented_reality_technology.