INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Classification of Parkinson's Disease Data Using Traditional and Advanced Data Mining Techniques

Atish S Tangawade[1]*, **Aniket A Muley**[2]

1 Research Fellow, School of Mathematical Sciences, SRTM University, Nanded, Maharashtra, India
2 Associate Professor, School of Mathematical Sciences, SRTM University, Nanded, Maharashtra, India

OPEN ACCESS

Received: 02-12-2023

Accepted: 17-01-2024

Published: 05-03-2024

Citation: Tangawade AS, Muley AA (2024) Classification of Parkinson's Disease Data Using Traditional and Advanced Data Mining Techniques. Indian Journal of Science and Technology 17(11): 1043-1050. https://doi.org/10.17485/IJST/v17i11.3059

* **Corresponding author**.

tangawadeatish@gmail.com

Funding: None

Competing Interests: None

Copyright: © 2024 Tangawade & Muley. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: (1) To apply various traditional classification tools, (2) To check effectiveness of the classifiers to the Parkinson Dataset (3) To use boosting classification tools and (4) Compare performance of all used classification tools and find the best accuracy classifier algorithm. Thus, the main aim of the study is to discriminate healthy people from those with PD. **Methods:** The methodology of this study is categorised into three stages:(1) Preprocessing and feature selection; (2) Application of classifiers; (3) Comparative study. We have used secondary dataset of voice recordings originally collected by University of Oxford by Max Little. In first step, the voice data of PD patients is collected for analysis. Then the collected data is normalized using min-max normalization followed by feature extraction. Thus, uses classification Data Mining Techniques viz., KNN, Logistic Regression, Decision Tree, SVM, Random Forest and boosting algorithm etc. to predict whether the person is healthy or has Parkinson's disease. Finally, comparative analysis is made based on the accuracy provided by different data mining models. **Findings:** Results of our study reveals that GB algorithm is more accurate as compared with other models. It gives the highest accuracy, so that we recommend this algorithm to deal similar kind of studies in the future. These models are very useful in better and exact medical diagnosis and decision making. It is also found that, proposed methods are fully computerized and produce enhanced performance hence can be recommended for similar studies. Here, it is observed that Gradient Boost algorithm provide the best accuracy (100% for training and 92.02% for testing, 98.46% overall). **Novelty:** We have used boosting classification model for the classification of Parkinson's disease. Our proposed method is one such good example giving faster and more accurate results for the classification of Parkinson's disease patients with excellent accuracy. We have also compared the results with other existing approaches like linear discriminant analysis, support vector machine, K-nearest neighbour, decision tree, classification and regression trees, random forest, linear regression, logistic regression and Naive Bayes, but our proposed techniques were superior to existing studies in which gradient boost algorithm yielded an accuracy of 98.46%, so our method can be used as an effective means of computer-aided diagnosis of PD, and has important practical value.

Keywords: Data Mining; Parkinson's Disease; Classification; Boosting Algorithms; Feature Selection

1 Introduction

Nowadays, human life is busy and hectic. It is essential to maintain stress level, piece of mind and health consciousness. It is observed that, on an average after crossing the age of 60, there are numerous physical problems arise to human being. Parkinson's is one of the diseases in which most commonly found and it doesn't cure completely. Generally, due to unawareness among common people they are avoiding to visit to doctor and therefore it is difficult to trace at the earliest stage [1]. It is very interesting that, based on available collected dataset if classification is performed well then, it is very beneficial for future similar kind of studies by different researchers. It is really important that, if the appropriate algorithm is used for the corresponding disease dataset then in future it will be beneficial for user to apply the algorithm and classify the patients in the respective classes [2].

Different researchers classified the Parkinson's disease patients using various data mining tools. Ghorbani and Gousi have discussed the importance of data mining in the field of bioinformatics and various subfields of bioinformatics in which data mining has shown its great impact. They successfully used SVM method with random split of dataset to classify the instances into two classes, and hence can discriminate healthy people from those with Parkinson's disease [3]. Goyal and Rani conducted an extensive analysis using machine learning, ensemble learning, and deep learning models with different hyperparameters to develop accurate classification models for PD prediction. Their results showed that, the deep learning model has 91.33% training accuracy and 85.02% testing accuracy, suggesting that deep learning models perform comparably equivalent on small datasets compared to machine learning classifiers. [4] Baez et al. studied whether Parkinson's patient classification can be enhanced through the joint assessment of both extents using sentential incentives. Authors evaluated 31 early patients and 24 healthy controls via two syntactic measures and a verbal task tapping social emotions and compared their classification accuracy when analysed in isolation and in combination using one-way ANOVAs. Their study outcomes show the joint analysis improved the classification accuracy of patients and controls, regardless of their global cognitive and affective state [5]. Ricciardi et al. used SMOTE technique for balancing the data and then acquired gait analysis for each patient. Moreover, they differentiate PD patients at different stages through gait analysis [6]. Wingate et al. applied convolutional and recurrent neural network on medical images of Parkinson's disease patients' data. Their proposed unified model produced over different data sets can provide effective and transparent estimate of PD image data [7]. Salmanpour et al. employed Predictor algorithms and Feature Subset Selector algorithms (FSSAs) on two different datasets extracted from PPMI database. They used automated hyperparameter tuning for parameter optimization. Their results show that, applied algorithms can yield admirable forecast of intellectual outcome in PD patients [8].

Haq et. Al. conducted research study on comprehensive assessment of available surveys and DL-based diagnosis methodologies for PD recognition [9]. Temploton et al. used ML classification to assess the benefits and relevance of neurocognitive features both tablet based assessments and self-reported metrics, as they relate to Parkinson's disease (PD) and its stages. And their study depicts that perceived functionality of individuals with PD differed from sensor-based functionalities [10]. Tong et al. presents a method based on permutation-variable importance (PVI) and persistent entropy of topological imprints, and uses support vector machine (SVM) as a classifier to achieve the severity classification of PD patients. Their study achieved accuracy of 98.08% by 10-fold cross-validation[(11)]. Alalayah et al. used classification algorithms based on recursive feature elimination method for automatic and early detection of Parkinson's disease. Their experimental results yielded an accuracy of 97%, precision of 96.50%, recall of 94%, and F1-score of 95%[(12)]. Jing Zang used technique of mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of Parkinson's disease[(13)]. Lee et al. proposed a lightweight deep learning model to categorize resting-state EEG recorded from people with Parkinson's disease and healthy controls. Their results revealed that, CRNN model attains higher performance compared to baseline machine learning approaches and other newly proposed deep learning models[(14)]. Rasheed et al. proposed a technique for classification of PD using an improved version of ANN with adaptive momentum termed as BPVAM. In addition, BPVAM was also combined with PCA to further enhance the classification accuracy of the classifier. Both methods obtained high classification accuracy for PD identification. BPVAM-PCA has extra compensations in terms of processing time and reaching low error quickly[(15)]. Mittal and Sharma proposed a novel approach using data partitioning with feature selection algorithm principal component analysis (PCA) for Parkinson's disease classification. They used three different classifiers to classify all data partitions, including the weighted k-NN (nearest neighbour, wKNN), Logistic Regression (LR), and Medium Gaussian Kernel support vector machine (MGSVM). The classification accuracy of 74.2%, 85.0% and 82.1% achieved using Logistic algorithm, SVM with Gaussian, and weighted k-NN classifiers[(16)]. Ouhmida et al.performed a comparative analysis on machine learning (ML) techniques for PD identification based on voice disorders analysis. These ML methods included the Support Vector Machine (SVM), K-Nearest-Neighbors (KNN), and Decision Tree (DT) algorithms. In addition, two feature selection techniques mRMR and ReliefF were used to further improve the performance of the proposed classifiers. The efficiency of the developed model has been evaluated based on accuracy, sensitivity, specificity and AUC metrics, and it is higher than existing approaches. The simulation results show that the KNN algorithm yielded the best classifier performance in term of accuracy[(17)]. Priya et al. conducted study on early detection of Parkinson's disease using data mining techniques from multimodal clinical data[(18)]. Their results show that the SVM produces the most accuracy on the spiral imaging datasets, and the proposed ensemble method does the same for disease classification. Sharanyaa et al. evaluated the performance of various classification algorithms to detect Parkinson's disease. They used various performance metrics like precision, recall, F1 score are computed for all four machine learning techniques and the results show that nonparametric models using Random Forest and K-Nearest Neighbors produce higher classification accuracy of 87.2% and 90.2% compared to parametric models[(19)]. Ahmed et al. is classified PD patients' data with the help of human voice signals. Six different machine learning (ML) algorithms are used in the classification. Their results show that SGD-Classifier has 91% accuracy, XGB-Classifier has 95% accuracy, Logistic Regression has 91% accuracy, Random Forest shows 97% accuracy, KNN shows 95% accuracy, and Decision Tree has 95% accuracy[(20)].

1.1 Research Gap

It is observed that, the previous researchers dealt with the similar kind of work but the obtained accuracy for classification is sufficient? To get the more accuracy in classification this study is carried out and observed that, the results are able to get better results as compared with previous studies. So, in this paper we have used some advanced classifiers like random forest, adaboost, gboost, extra tree classifier. Hence, this paper proposes novel techniques to classification of PD patients from their early symptoms which is to help neurologists to make appropriate diagnostic decisions.

2 Methodology

The methodology of this study is categorised into three stages: (1) Preprocessing and feature selection; (2) Application of classifiers; (3) Comparative study. The proposed methodology for classification of the Parkinson's disease using data mining is outlined in Figure 1. In first step, the voice data of PD patients is collected for analysis. Then the collected data is normalized using min-max normalization followed by feature extraction. Then different data mining algorithms are employed for the classification task. Finally, comparative analysis is made based on the accuracy provided by different data mining models.

2.1 Data Collection

We have used secondary dataset of voice recordings originally collected by University of Oxford by Max Little. This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table (Table 1) is a particular voice measure, and each row corresponds to one of 195 voice recordings from these individuals ("name" column).

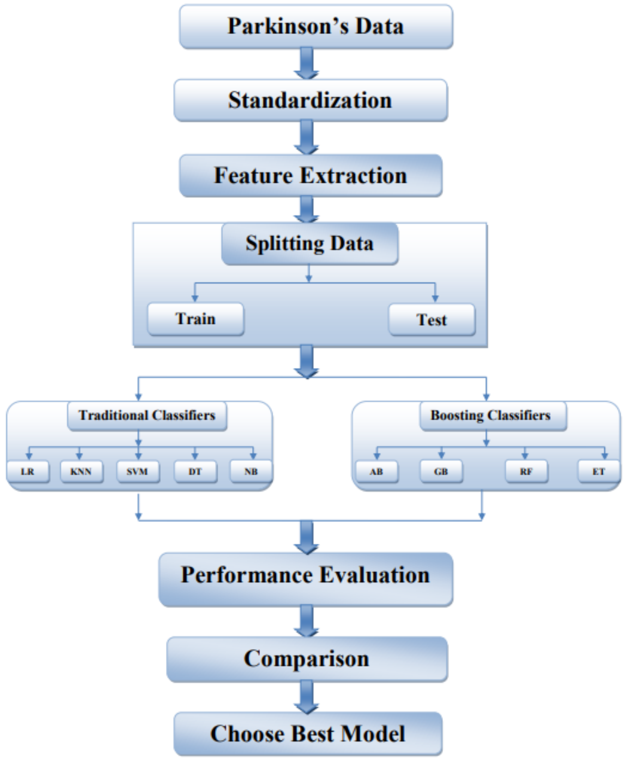

Fig 1. Flow chart

2.2 Attribute Information

Table 1. Information of Attributes in Dataset

Sr. No.	Name of Attributes	Type	Description of values
1.	Name	Categorical	ASCII subject name and recording number
2.	MDVP: Fo(Hz)	Numerical	Average vocal fundamental frequency
3.	MDVP:Fhi(Hz)	Numerical	Maximum vocal fundamental frequency
4.	MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP	Numerical	Several measures of variation in fundamental frequency
5.	MDVP: Shimmer, MDVP: Shimmer (dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ , Shimmer: DDA	Numerical	Several measures of variation in amplitude
6.	NHR, HNR	Numerical	Two measures of ratio of noise to tonal components in the voice
7.	Status	Numerical	Health status of the subject (one) - Parkinson's, (zero) – healthy
8.	RPDE, D2	Numerical	Two nonlinear dynamical complexity measures
9.	DFA	Numerical	Signal fractal scaling exponent
10.	Spread1, spread2, PPE - Three nonlinear measures of fundamental frequency variation"	Numerical	Three nonlinear measures of fundamental frequency variation"

For classification purpose, we have used Decision tree, Naïve Bayes, Support Vector Machine, Logistic regression and K-nearest neighbours (KNN) techniques from category of traditional classifiers. While, we also used some boosting classification techniques such as Adaptive boosting (Adaboost), Gradient boosting (G Boost), Random Forest and Extra Trees (Extremely Randomized trees).

3 Results and Discussion

In this section, we have analysed the data and discussed study outcomes. Here, we have performed correlation analysis of the study parameters and represented correlations coefficient values in the form of heat map. To perform further analysis, we have removed the less correlation coefficient columns that are related to irrelevant features which minimize the accuracy of an algorithm. It helps us to achieve good accuracy of classification. Based on correlation coefficient values, visualized top 10 coefficient values. Later we can decide which attribute have relevant and irrelevant features.

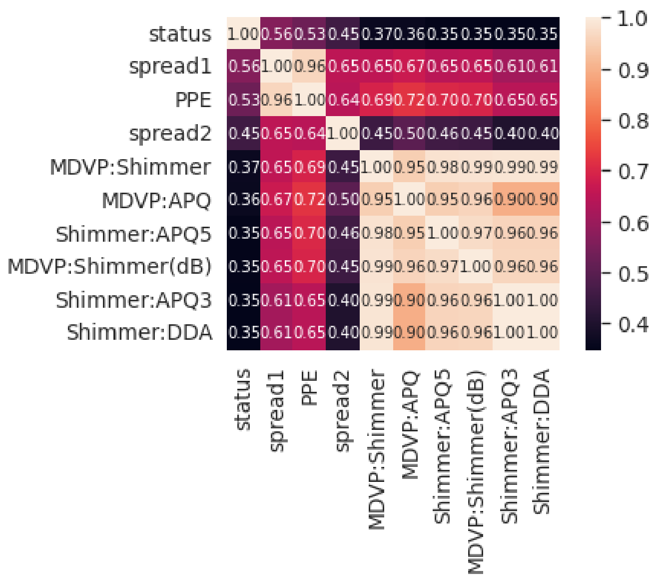

Fig 2. Correlation Plot

Table 2. Correlation with status target variable

Parameter	Correlation	Parameter	Correlation
spread1	0.5648	Shimmer:DDA	0.3476
PPE	0.531	D2	0.3402
spread2	0.4548	MDVP:Jitter(Abs)	0.3386
MDVP:Fo(Hz)	0.3835	RPDE	0.3085
MDVP:Flo(Hz)	0.3802	MDVP:PPQ	0.2886
MDVP:Shimmer	0.3674	MDVP:Jitter(%)	0.2782
MDVP:APQ	0.3643	MDVP:RAP	0.2666
HNR	0.3615	Jitter:DDP	0.2666
Shimmer:APQ5	0.3511	DFA	0.2317
MDVP:Shimmer(dB)	0.3506	NHR	0.1894
Shimmer:APQ3	0.3476	MDVP:Fhi(Hz)	0.1661

Table 2 represents the correlation values in descending order. To perform better analysis, we have dropped variables with the assumption that if we decrease the column count then accuracy will increase gradually, because we are not keeping the irrelevant features. So that, here we have dropped from MDVP:RAP to MDVP:Fhi(Hz) as it has less correlation with status column. Here, the counts of patients in terms of '0' and '1' as '0' indicates patient does not have Parkinson's disease whereas '1' represents patient is having Parkinson's disease. It is observed that, 147 patients are having Parkinson's disease and 48 patients do not have Parkinson's disease.

3.1 Classification Models with Feature Scale

Here, to perform classification of target variable following classification models were used with feature scales. The results are summarized in Table 3. It is observed that, decision tree classifier provides the best accuracy (0.8590) with minimum standard error (0.0433).

Table 3. Comparison of Classification Models with Feature Scale

Sr. No.	Model	Accuracy (Standard Error)
1	Logistic Regression	0.845766 (0.072339)
2	K nearest neighbor	0.845968 (0.096788)
3	Support Vector Machine	0.826613 (0.109452
4	Decision Tree	0.859073 (0.043356)
5	Naïve Bayes	0.730645 (0.052907)

3.2 Regularization Tuning for Top 2 Classification Algorithms

Based on Table 3, we have selected top 2 best performance algorithms and tuned them. After tuning the best algorithms observed are: Decision Tree Classification Algorithm (0.865524) and KNN Classification Algorithm (0.87157). Further we have used Ensemble and Boosting algorithm to improve performance and their results are summarized in Table 4.

Table 4. Comparison of Ensemble and Boosting Classification algorithms with feature scale

Classifier	Accuracy (Standard Error)
AdaBoostClassifier	0.878226(0.062488)
GradientBoostingClassifier	0.884476(0.072591)
RandomForestClassifier	0.846169(0.051526)
ExtraTreesClassifier	0.910282(0.024008)

Thereafter, top 2 Ensemble and Boosting Classification algorithms were tuned, after applying top 2 ensemble algorithms results summarized in Table 5.

Table 5. Accuracy of tuned Ensemble and Boosting Classification algorithms

Model	Accuracy
Gradient Boosting Classification Algorithm	0.9102
Extra Trees Classification Algorithm	0.9100

We finalized the Gradient Boost Classification Algorithm and evaluated the model for Detection Parkinson's disease.

3.3 Comparison of All 4 Tuned Algorithms and Selecting the Best Algorithm

Finally, all 4 tuned algorithms were compared for accurate classification of target variable. The results are summarized in Table 6.

Table 6. Comparison of all 4 tuned Classification algorithms

Sr. No.	Model	Accuracy
1	Decision Tree	0.8525
2	KNN Classification	0.8716
3	Gradient Boosting Classification	0.9103
4	Extra Trees Classification	0.9101

Gradient Boosting classification algorithm gives the best accuracy performance (0.9103) so we will use this ensemble algorithm to fit and predict our dataset.

3.4 Accuracy of an Optimal Algorithm

Table 7 explores the GB classification algorithm results for train and test dataset. We obtained training accuracy of the given trained dataset is100% and the test set is 92.3%, thus we achieved 98.46 % accuracy on overall dataset.

Table 7. Optimal results

		Predicted (Train set)		Predicted (Test set)	
		No	Yes	No	Yes
Actual	No	38	0	8	2
	Yes	0	118	1	28

3.5 Comparative Analysis with Existing Results

In this study, we used traditional classification techniques as well as boosting classification techniques. We observed that, boosting approaches are superior than that of traditional ones. We also compared our results with existing results. The results obtained through proposed study compared with previous studies [10–12,19–21], it is observed results are more superior to existing ones. Table 8 shows comparative results of the studies with the proposed study.

Table 8. Comparative results

Accuracy Results	Percentage
Goyal and Rani [4]	91.33
Templeton et al. [10]	92.6
Tong et al. [11]	98.08
Alalayah et al. [12]	97
Sharayanyaa et al. [19]	90.2
Ahmed et al. [20]	97
Jyotiyana et al. [21]	94.87
Proposed Approach	98.46

4 Conclusion

This study verified the effectiveness of the application of various classifiers to the Parkinson's dataset. The proposed classification system reveals that GB algorithm is more accurate as compared with other models. Proposed study reveals that our model gives the highest accuracy (98.46%), so that we recommend this algorithm to deal similar kind of studies in the future. This could lead to earlier analysis and treatment, which could improve the quality of life for people with Parkinson's disease. The systems optimize the early diagnosis of PD by evaluating selected features and hyperparameter tuning of ML algorithms for diagnosing PD based on voice disorders. The study found that the proposed techniques were superior to existing studies, indicating the superiority of the proposed techniques. These models are very useful in better and exact medical diagnosis and decision making. It is also found that, proposed methods are fully computerized and produce enhanced performance hence can be recommended for similar studies. In the future, interested researchers can increase the sample size to reach towards more accuracy level and it will be beneficial for exact classification of patients.

References

1) Raza C, Anjum R, Shakeel NUA. Parkinson's disease: Mechanisms, translational models and management strategies. *Life Sciences*. 2019;226:77–90. Available from: https://doi.org/10.1016/j.lfs.2019.03.057.
2) Bloem BR, Okun MS, Klein C. Parkinson's disease. *The Lancet*. 2021;397(10291):2284–2303. Available from: https://doi.org/10.1016/S0140-6736(21)00218-X.
3) Ghorbani R, Ghousi R. Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. *International Journal of Data and Network Science*. 2019;3(2):47–70. Available from: http://dx.doi.org/10.5267/j.ijdns.2019.1.003.
4) Goyal P, Rani R. Comparative Analysis of Machine Learning, Ensemble Learning and Deep Learning Classifiers for Parkinson's Disease Detection. *SN Computer Science*. 2023;5(1). Available from: https://doi.org/10.1007/s42979-023-02368-x.
5) Baez S, Herrera E, Trujillo C, Cardona JF, Diazgranados JA, Pino M, et al. Classifying Parkinson's Disease Patients With Syntactic and Socio-emotional Verbal Measures. *Frontiers in Aging Neuroscience*. 2020;12:1–11. Available from: https://doi.org/10.3389/fnagi.2020.586233.

6) Ricciardi C, Amboni M, De Santis C, Improta G, Volpe G, Iuppariello L, et al. Using gait analysis' parameters to classify Parkinsonism: A data mining approach. *Computer Methods and Programs in Biomedicine*. 2019;180:105033. Available from: https://doi.org/10.1016/j.cmpb.2019.105033.
7) Wingate J, Kollia I, Bidaut L, Kollias S. Unified deep learning approach for prediction of Parkinson's disease. *IET Image Processing*. 2020;14(10):1980–1989. Available from: https://doi.org/10.1049/iet-ipr.2019.1526.
8) Salmanpour MR, Shamsaei M, Saberi A, Setayeshi S, Klyuzhin IS, Sossi V, et al. Optimized machine learning methods for prediction of cognitive outcome in Parkinson's disease. *Computers in Biology and Medicine*. 2019;111:103347. Available from: https://doi.org/10.1016/j.compbiomed.2019.103347.
9) Haq AU, Li JP, Agbley BLY, Mawuli CB, Ali Z, Nazir S, et al. A survey of deep learning techniques based Parkinson's disease recognition methods employing clinical data. *Expert Systems with Applications*. 2022;208:118045. Available from: https://doi.org/10.1016/j.eswa.2022.118045.
10) Templeton JM, Poellabauer C, Schneider S. Classification of Parkinson's disease and its stages using machine learning. *Scientific Reports*. 2022;12(1):1–11. Available from: https://doi.org/10.1038/s41598-022-18015-z.
11) Tong J, Zhang J, Dong E, Du S. Severity Classification of Parkinson's Disease Based on Permutation-Variable Importance and Persistent Entropy. *Applied Sciences*. 2021;11(4):1–20. Available from: https://doi.org/10.3390/app11041834.
12) Alalayah KM, Senan EM, Atlam HF, Ahmed IA, Shatnawi HSA. Automatic and Early Detection of Parkinson's Disease by Analyzing Acoustic Signals Using Classification Algorithms Based on Recursive Feature Elimination Method. *Diagnostics*. 2023;13(11):1–24. Available from: https://doi.org/10.3390/diagnostics13111924.
13) Zhang J. Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of Parkinson's disease. *npj Parkinson's Disease*. 2022;8(1):1–15. Available from: https://doi.org/10.1038/s41531-021-00266-8.
14) Lee S, Hussein R, Ward R, Wang ZJ, Mckeown MJ. A convolutional-recurrent neural network approach to resting-state EEG classification in Parkinson's disease. *Journal of Neuroscience Methods*. 2021;361:109282. Available from: https://doi.org/10.1016/j.jneumeth.2021.109282.
15) Rasheed J, Hameed AA, Ajlouni N, Jamil A, Ozyavas A, Orman Z. Application of Adaptive Back-Propagation Neural Networks for Parkinson's Disease Prediction. In: 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI). IEEE. 2021;p. 1–5. Available from: https://doi.org/10.1109/ICDABI51230.2020.9325709.
16) Mittal V, Sharma RK. Machine learning approach for classification of Parkinson disease using acoustic features. *Journal of Reliable Intelligent Environments*. 2021;7(3):233–239. Available from: https://doi.org/10.1007/s40860-021-00141-6.
17) Ouhmida A, Raihani A, Cherradi B, Terrada O. A Novel Approach for Parkinson's Disease Detection Based on Voice Classification and Features Selection Techniques. *International Journal of Online and Biomedical Engineering (iJOE)*. 2021;17(10):111–130. Available from: https://doi.org/10.3991/ijoe.v17i10.24499.
18) Priya S, Priyatharshini R, Shruthi R, Pooja V, Swarna RS. Early detection of Parkinson's disease using data mining techniques from multimodal clinical data. In: Advanced Machine Vision Paradigms for Medical Image Analysis. Academic Press. 2021;p. 213–228. Available from: https://doi.org/10.1016/B978-0-12-819295-5.00008-1.
19) Sharanyaa S, Renjith PN, Ramesh K. Classification of Parkinson's Disease using Speech Attributes with Parametric and Nonparametric Machine Learning Techniques. In: 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). IEEE. 2021;p. 437–442. Available from: https://doi.org/10.1109/ICISS49785.2020.9316078.
20) Ahmed I, Aljahdali S, Khan MS, Kaddoura S. Classification of Parkinson Disease Based on Patient's Voice Signal Using Machine Learning. *Intelligent Automation & Soft Computing*. 2022;32(2):705–722. Available from: https://doi.org/10.32604/iasc.2022.022037.
21) Jyotiyana M, Kesswani N, Kumar M. A deep learning approach for classification and diagnosis of Parkinson's disease. *Soft Computing*. 2022;26(18):9155–9165. Available from: https://doi.org/10.1007/s00500-022-07275-6.