



# Detection of Parkinson's Disease by Using Machine Learning Stacking and Ensemble Method

Vikas Chaurasia<sup>1</sup> · Aparna Chaurasia<sup>2</sup>

Received: 15 November 2022 / Accepted: 30 March 2023 / Published online: 27 April 2023  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC 2023

## Abstract

Using speech data, it is difficult to learn through machine learning how to diagnose Parkinson's disease (PD) and evaluate the effects of treatment. For this issue, the article has developed a three-stage PD discovery method. The base classifiers used in the initial stage are logistic regression (LR), K-nearest neighbor (KNN), naive bayes (NB), support vector machine (SVC), and decision tree (DT). The second stage, or stack model, is a meta-model that combines all of the classifiers mentioned earlier. The third stage ensemble model consists of Bagging, AdaBoost, Random Forest (RF), and Gradient Boosting (GBC) components. The RF and GBC classifiers are utilized to estimate the most important features from the PD dataset. The models' validation has been evaluated using the confusion matrix and validation metrics like precision, recall, and F1 score. Out of all the ensemble models, the GBC—the third model—had the highest accuracy with testing data—97.43%. KNN from the base model and stacking from the meta-model, on the other hand, had the highest accuracy, with 94.87% each. Out of all the models mentioned in this manuscript, the GBC is the only ensemble model classifier with the highest accuracy. The proposed classifier appears to be an extremely useful model for the discovery of Parkinson's disease, as demonstrated by the exploratory findings and factual analyses.

**Keywords** Parkinson's disease · Machine learning · Stacking model · Ensemble model · Meta-learner · Confusion matrix

## Introduction

PD is the second most common neurodegenerative disorder worldwide, after Alzheimer's. PD is a common, moderate neurodegenerative condition primarily affecting the central sensory system and the body's structure [1]. Shaking, firmness, difficulty talking and walking, and other common body side effects are included. The dopamine-producing neurons in a particular part of the human brain are profoundly affected by PD. A significant number of the side effects are caused by the body's lack of dopamine. The majority of Parkinson's disease patients over the age of 60 first experience symptoms, but 5 to 10% of cases begin earlier [2]. Most of the time, the side effects start off slowly and get worsen over time. Non-body side effects, such as mental and behavioral

issues, memory issues, sleep issues, discouragement, fatigue, and so on, may arise as the disease progresses.

As previously stated, PD diagnosis has been challenging due to the absence of reliable tests. However, PD patients exhibit voice and speech impairments, according to recent research. However, medical professionals are unable to identify these voice defects in clinics. As a result, in order to identify PD in its earliest stages and capture these voice impairments, automated signal processing tools are required [3]. Recent studies demonstrate that automated risk factor extraction and classification using machine learning and signal processing algorithms is effective for disease detection. There have been more AI-based PD treatment trials in recent years [4]. Even though previous studies looked at how AI could be used to find and evaluate PD, they only looked at motor side effects, kinematics, and data from wearable sensors. As a result, the application of AI to non-clinical and clinical information from a variety of modalities has frequently resulted in high symptomatic exactness in human members. This may make it possible to use novel biomarkers and AI calculations in clinical settings to support more precise and informed direction. A number of

✉ Vikas Chaurasia  
chaurasia.vikas@gmail.com

<sup>1</sup> Department of Computer Applications, MHPGC, VBS  
Purvanchal University, Jaunpur, India

<sup>2</sup> Research Enthusiastic, Jaunpur, India

artificial intelligence methods have recently been utilized to identify Parkinson's disease indicators [5]. Among them, stacking models that combine at least two independent classifiers with a meta-model to improve accuracy can reduce the propensity accuracy of a single AI model. In addition, it has been demonstrated to predict the outcome factors with greater precision. It achieves better expected execution than individual computation demonstrations by employing stacking ensemble computations [6].

In this paper, we also attempt to develop a PD detection strategy based on machine learning and signal processing algorithms, inspired by these studies. Using signal processing algorithms for feature extraction from voice signals and machine learning algorithms for classification, we developed an automated model to detect Parkinson's disease (PD). This was inspired by the automated disease detection methods discussed above.

The literature review and background are presented in "Literature Review" and "Background" sections of the remaining manuscript. "Experimental Methodology" section is discussion of the experimental approach. The experimental setup is shown in "Experimental Setup" Section. The proposed model's results are presented in "Results" section. "Discussion" section is representation of the discussion and "Conclusion" section is conclusion.

## Literature Review

Many analysts already do their part in forecasting, more importantly, the investigation of Parkinson's disease. Partial exploration of this article analyzes the work.

The model is proposed by Joshi et al. [7] There are 12 ML-based models, including NB, KNN, LR, MLP, DT, SVM (linear, Poly, and RBF), and RF classifiers, as they can even find relapse signs of altered ranges to identify Parkinson's Infect individuals and able-bodied individuals. The proposed model is a bunch of RF, SVM, and KNN and ET and Extreme GBC; the model is performed with mean square error (MSE) and mean absolute error (MAE). Overall, the proposed model achieves (90–91) % testing accuracy and (98–100) % training accuracy, and the proposed model likewise utilizes PCA and LDA for better accuracy. Behroozi and Sami [8] obtained the dataset from the UCI AI store. Using the voice feature set, there are two design of CNNs are proposed to PDs. They used CNN. While mainly organized methods have many capabilities before using them to a nine-layer contour CNN as a data source, the subsequent structure confers feature sets on similar data layers closely related to convolutional layers. ANN and SVM are standard computations in this PD permutation KNN when preparing RF and SVM. CNNs want to secure an MLP for CNN. The central framework is a nine-layer convolutional neural

network consisting of a data layer, six convolutional layer contour expansion and max pooling layers, a fully correlated layer edge, and a yield layer contour. The structure consists of adjoint layers: a data layer with  $n$  feature sets, an equivalent layer with  $n$  branches, a solidification layer, and four medium convolutional layer contours. The accuracy is 82%, the F score is 88%, and the MCC is approx 50%. An article is presented by Lucijano et al. about the remote monitoring of PD [9]. TQWT was used to investigate the vocal signs in PD patients. For merge extraction, a configurable q-factor waveform offset is coordinated with the Pd's victim sound sign. When the MFCC and TQWT coefficients are added to the PD request question, they provide reciprocal information that further informs the accuracy of the plan. Once again, TQWT has steady repeat greetings. The calculated limits are enforced for the adjustable q-factor channel to more accurately delineate the perceptibility ends. TQWT guarantees short-lived confinement of important subcomponents in the process. TQWT method by cultivating an important field of remote monitoring engineering, it can be used to determine the true joint PD review standard level for PD patients. Using quantitative methods, the maximum accuracy of supplies picked with minimal redundancy to the classifier was 0.86, with an F1 score of 0.84 and an MCC of 0.59. Sarkar et al. [10] presents an exploratory article on "Parkinson Dataset with Reproduced Acoustic Elements Informational index." Specifically, 45 sound modules corresponding to the assessment were set, including Parkinson's disease patients and ordinary subjects. Scientists have found evidence that leverages several scientific methods, such as Xgboost, RF, SVM, KNN, and basically backwards. At that time, a variable importance examination was also performed to understand the key elements delineating PD patients. A total of 80 people are expected: 40 people with Parkinson's disease and 40 able-bodied people. LGB was used by the best performing model, achieving an AUC of 95.1% in fourfold CV with only seven audible parts, with a 95% certainty range of 0.946–0.955. They achieved 84% accuracy. Gunduz [11] utilized a speech test dataset to discriminate PD patients. Scientists hope to use strait & recursive recurrence records to dissect the localization of merged Parkinson's. Efforts have been made to gain a thorough understanding and connection between these two components and ensure that Pds scores are calculated using global Insights due to the diligence followed by change testing and processing of data into SVM classifiers with selected features. Most notable Relevant minimal repeats. To avoid trends, the analyst model's cross-validity system has been eliminated. The scientists' program achieved 92.75% precision. They used KNN, SVM, SVM with RBF, and NB. The accuracy was 88%, the sensitivity was 90%, and the specificity was 85%. Based on the highlights extracted from 26 different speech tests, Akshay and Kiran [12] developed PD diagnosis in the study subjects.

The optimal sequential method for Pd analysis without a method for selecting elements has been shown to be several ANNs. The test accuracy reached 86.47% after calibrating the neuronal tissue. A 1D ANN is used to achieve 100% preparation accuracy and inspection accuracy. In addition to using additional hidden layers, additional neurons are added to the existing layers. It has been demonstrated that the appropriate ANN response depends on the ANN design, as the ability to alter the outcome has been established. The correlation coefficient component selection method outperformed PCA, although their detection percentages were similar. Sztah and Hemmerling [13] used speech signals as biosignals to identify PD patients and serve as a reliable baseline set, represented by a set of provisions taken from temporal, recursive, and cepstral spaces, and applied to PCA and nonlinear SVM. The results guarantee an ensemble accuracy of 93.43%. They utilized various AI algorithms such as DNN, SVM, Naive Bayes, DT, Regression, and RF to identify PD. The study discusses how to use parallel distributed NN classifiers with backpropagation to learn computation and majority voting. Studies have shown that the search accuracy of PD can reach 90%. During the investigation, the following descriptions were used: Regression, NN, DT, and DM Neural Computation. The neural network had the highest representation rate at 92.9%. SVM produced the best results with 89.3% accuracy. For PD, the accuracy is 95% and the accuracy is 93%. Karabayir et al. [14] combine network research, ensemble learning, deep neural networks, ANN, DL, C-SVM back, NN, and DNN. ANN and KNN are used together with PSO algorithm to evaluate PD speech. After normalizing speech and applying FFT computation, speech data are cleaned and separated from unpleasant speech in surrounding area. The PSOANN model is particularly effective when easily predicting PD at population 300. The accuracy of the PSOANN model is 93.25%. Dasari et al. [15] leveraged ensemble learning, autoencoders, and SVMs. In addition, many computations, such as SVM, RF, and ANN, are frequently used in various applications, including the characterization of PD by counting LR and Voting with a computational accuracy of 97.22%. Such studies show that example representation models outperform other deep learning strategies. By combining LR and voting the researchers asserted that Parkinson's patients could be accurately identified using the regression model combined with the output of the classifier, GBC, MLP, and RNN classification techniques. In the end, the researchers asserted that Parkinson's patients could be accurately identified using the regression model combined with the output of the classifier, GBC, MLP, and RNN classification techniques. These findings were confirmed by a min–max normalized fivefold CV technique. In the early stages, they used ML classifiers such as SVM, XGBoost, and MLP to evaluate PD from speech features. Afterward, they instruct the auto-encoder to transfer

useful components to the classifier, in this case, an SVM or a single sigmoid neuron. Jankovic [16] was used to differentiate between people with PD and those who did not use the test ( $N=195$ ). The MLP classification accuracy for information collection was 93.22%, while the RBF classification accuracy was 86.44%. Harding et al. [17] proved Genetic Algorithm-RF, Genetic Algorithm-SVM, SVM, and RF. Accuracy (69–94) %, influence ability (60–92) %, and specificity (70–95) % all pass the GASVM classifier. Hughes et al. [18] found the earthquake onset motion is correctly predicted by RBFNN, PCA, RBFNN, PCA, PSO, and RBFNN based on PSO and PCA using local field potential data obtained through incitement terminal acquisition, respectively. By combining a similarity classifier with entropy-based composition determination, Goetz et al. [19] improved information metrics and reduced computation time. Informative metrics were created using various voice assessments from healthy and PD individuals. Using Parkinson's information, the mean alignment accuracy was 85.03%. Warrior et al. [20] suggested the use of wavelet analysis to extract component vectors from speech tests and contributed to three-layer feed-forward multilayer neural networks.

## Background

There are a number of options for analyzing PD in the field of research on information technology. In this section, we talk about how some of the methods that the majority of researchers use are taken into account by some of the methods that are already in use. This work analyzes PD data using existing methods such as LR, KNN, NB, SVM, and DT [21]. This study proposes an ensemble approach of RF, GBC, Bagging, and AdaBoost algorithms, inspired by previous ensemble work. In this study, the stacked model (LR + KNN + NB + SVM + DT) is used to improve the prediction accuracy.

## Classifiers

The process of identifying, understanding, and organizing concepts and things into predetermined categories or "subgroups" is called classification. Pre-classified training datasets are used by various machine learning algorithms to classify the subsequent datasets [22]. Classification algorithms in machine learning predict whether subsequent data will fall into a predetermined set of categories by using the input training data. Using classification to classify Parkinson's disease into "with PD" and "non-PD" is one of the most common uses. To put it simply, classification is a type of "pattern recognition" in which a classification algorithm is used on training data to find patterns that are similar to those

in subsequent datasets [23]. Five of the most common classification algorithms are used in this article to distinguish between patients with and without Parkinson's disease, as described above in the Background section.

## Ensemble Method

In machine learning, ensemble methods combine insights from multiple learning models to make better and more accurate decisions. Noise, variance, and bias are the majority of errors in learning models. The accuracy and stability of machine learning algorithms are guaranteed by ensembles of methods used to reduce these error-causing factors. Predictions about each data point are made using multiple models in an ensemble technique of machine learning. Therefore, most of the model's predictions are considered final predictions. When using the mean ensemble technique to make the final forecast, the data analyst considers the average forecast made by all the models. By the using of the weighted average ensemble method to make predictions, giving each model a different weight that tells how important it is.

## Stacked model

In model stacking, we don't make predictions with just one model; rather, we make predictions with multiple models and use those predictions as features for a meta-model at a higher level. It can be particularly effective with a variety of lower-level learners, each of whom brings unique strengths to the meta-model [24]. There is no one "correct" way to use stacking, and there are many different ways to build model stacks. With multiple levels, weights, averaging, and other features, it can be made more complicated than the example presented here. We will use a series of intermediary models to make non-leaky predictions on our train data in our stack. These models will then be used as features alongside the original training features on a meta-model.

## k-Fold Model

This model is a statistical technique for estimating the skill of a ML model [25]. Because it is easy to understand, easy to implement, and produces skill estimates that are generally less biased than other methods, learning is often used in applied machines to compare and select models for specific predictive modeling problems. The single parameter  $k$  in the procedure specifies the number of groups into which a given data sample will be divided. Therefore, this method is often called  $k$ -fold cross-validation. When choosing a specific value of  $k$ , replace  $k$  in the model reference, such as  $k=10$  for tenfold cross-validation. When making predictions using data not used in model training, we use small samples to estimate the overall performance of the model. It

is important to note that each observation in the data sample is placed into a specific group and remains in that group throughout. This shows that the model is trained  $k$  times with each sample and once with the holdout set.

## Experimental Methodology

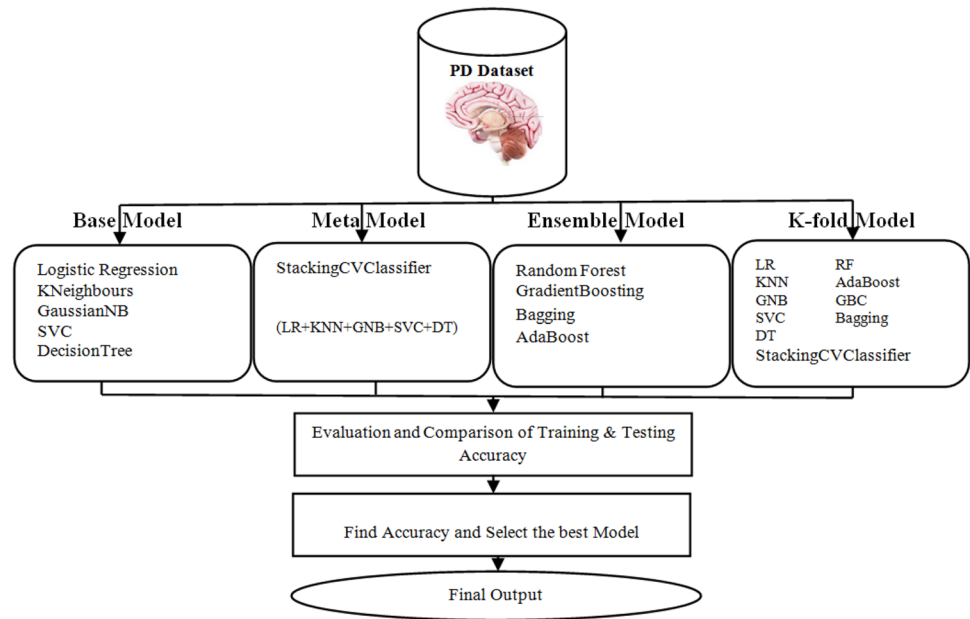
The authors of this paper conducted experiments aimed at creating a targeted diagnostic system that can use machine learning algorithms to create objective classifications of patient voices. During a machine learning examination, such a system would enable physicians to initiate treatment and monitor its efficacy, providing technical support for their subjective assessment of PD speech impairment. First, this paper demonstrates how multiple base classifiers can be used to identify voices in PD patients. The ability of base classifiers to solve numerous recognition problems without using experts to locate relevant features is the main reason for their application [26]. This approach may be significant without sufficient domain knowledge. The meta-model, ensemble model, and  $k$  folded model are then discussed in the study for comparison. Over the past few years, there has been an incorporate to interest in ensemble models and meta-models. To the best of our knowledge, this architecture is a novel problem in speech-based Parkinson's disease diagnosis. However, there is not enough room for comparison between the PD recognition results of the duet approach and traditional feature engineering.

Further feature engineering requires locating a set of parameters that can be used as a basis for generating feature vectors for modeling the patient's voice, including those derived from the physician's expertise. This study evaluates nonlinear speech signal analysis while also taking into account the perceptual frequency scale. The classification scheme utilizing speech signals is based on traditional RF and GBC methods, utilizing vectors of selected speech descriptors [27]. However, it is still difficult to compare the PD identification results of various models. The comparison in this article will help readers better understand the advantages and disadvantages of each strategy. The experimental method of the proposed model is shown in Fig. 1.

## Experimental Setup

Biomedical speech measurements of 195 objects, 147 of whom had Parkinson's disease, constituted the PD dataset used in this study. The person's name is the first of 24 attributes in the dataset, and the person's status is the 18th. This dataset is taken from UCI machine learning repository [28]. Biomedical speech measurements that characterize recorded speech are found in the remaining 22 attributes. Each person

**Fig. 1** Proposed methodology for diagnosing the PD Patients



**Table 1** Description of attributes

Features name	Description
Name	Subject name and recoding number
MDVP:Fo(Hz)	Average vocal fundamental frequency
MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
MDVP:Flo (Hz)	Minimum vocal fundamental frequency
MDVP:Jitter (%)	Measures of variation in fundamental frequency
MDVP:Jitter (Abs)	Measures of variation in fundamental frequency
MDVP:RAP	Measures of variation in fundamental frequency
MDVP:PPQ	Measures of variation in fundamental frequency
Jitter:DDP	Measures of variation in fundamental frequency
MDVP:Shimmer	Measures of variation in amplitude
MDVP:Shimmer(dB)	Measures of variation in amplitude
Shimmer: APQ3	Measures of variation in amplitude
Shimmer: APQ5	Measures of variation in amplitude
MDVP:APQ	Measures of variation in amplitude
Shimmer: DDA	Measures of variation in amplitude
NHR	Measures of ratio of noise to tonal components in the voice
HNR	Measures of ratio of noise to tonal components in the voice
RPDE	Nonlinear dynamical complexity measures
DFA	Signal fractal scaling exponent
spread1	Nonlinear measures of fundamental frequency variation
spread2	Nonlinear measures of fundamental frequency variation
D2	Two nonlinear dynamical complexity measures
PPE	Nonlinear measures of fundamental frequency variation
Status	Health status of the subject: 1. Parkinson’s disease 0. Healthy

has 6 recordings made of their voice. Jitter, low light, ratio of noise to tonal components, scaling components, and various fundamental frequency measurements are included. The actual classification will be based on measurements of the sound. The dataset does not contain any missing values. Table 1 gives a description of each attribute.

### Results

Four models have been proposed in this study to categorize PD in the selected dataset. Base Model, Meta-Model, Ensemble Model, and K-fold Model are the names given to these models. One of these four models produces the best outcomes, which can be visualized for disease prediction.

Table 2 and Fig. 2 displays the results from the first three models—the Base model, the Meta-model, and the Ensemble model—as well as their architecture. The obtained results are analyzed in light of these models' evaluation criteria.

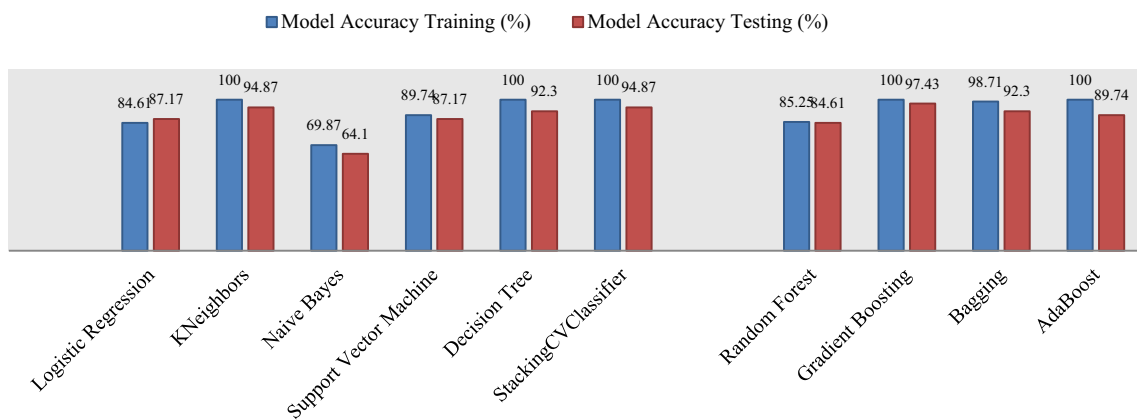
Estimates for the model's training and testing accuracy, as well as its validation scores—Precision, Recall, and F1 score—are provided in Table 2 [29]. The following formulas are used to calculate these scores:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

**Table 2** Accuracies acquired by different models

Results from based-mode								
Model name	Model accuracy		Validation score					
	Training (%)	Testing (%)	Precision (%)		Recall (%)		F1 score (%)	
			0	1	0	1	0	1
Logistic regression	84.61	87.17	86	88	60	97	71	92
K neighbors	100	94.87	100	94	80	100	89	97
Naive bayes	69.87	64.10	41	94	90	55	56	70
Support vector machine	89.74	87.17	86	88	60	97	71	92
Decision tree	100	92.30	82	96	90	93	86	95
Results from meta-model								
Stacking CV classifier (LR + KNN + NB + SVM + DT)	100	94.87	100	94	80	100	89	97
Results from ensemble model								
Random forest	85.25	84.61	83	85	50	97	62	90
Gradient boosting	100	97.43	100	97	90	100	95	98
Bagging	98.71	92.30	89	93	80	97	84	95
AdaBoost	100	89.74	75	96	90	90	82	93



**Fig. 2** Graphical representations of model's score

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Accurate classification is described by TP (true positive) and TN (true negative), while inaccurate classification is defined by FP (false positive) and FN (false negative). The number of cases correctly identified as having Parkinson's disease (PD) is referred to as TP, while the number of healthy cases is referred to as TN. The number of healthy patients who were mistakenly identified as having Parkinson's disease (PD) is shown in FP, while the number of PD cases is shown in FN [30]. The accuracy of a test is determined by its capacity to accurately differentiate between healthy and sick individuals. The ability of a test to correctly identify the healthy instances and the sensitivity of the test to correctly identify the patient instances are two different terms.

In the base model, the KNN classifier has the highest test accuracy of 94.87%; in the meta-model, the stacking CV classifier has the highest accuracy of 94.87%; and in the ensemble model, the Gradient Boosting (GBC) classifier has the highest accuracy of all three models with 97.43%.

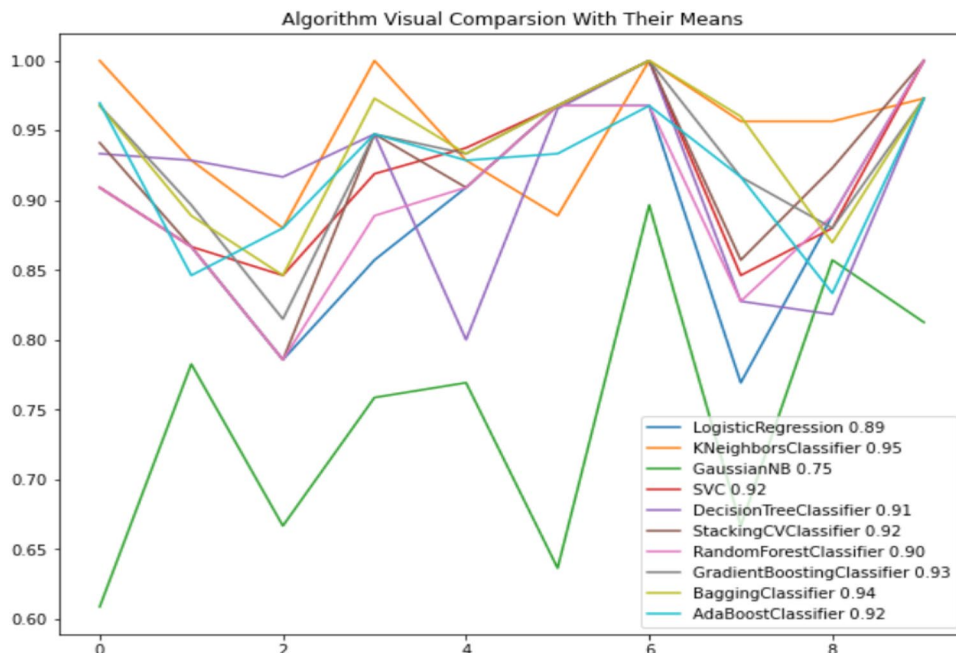
In contrast, in the fourth model, the k-fold is depicted in Fig. 3, where cross-validation employs a tenfold scale. Again, KNN has the highest accuracy, 95% higher than the base model. The mean values of classifiers as a graphical representation are depicted in Fig. 3.

In machine learning, features are distinct, independent variables that serve as system inputs. In fact, these features are utilized by models when making predictions. In ML, new features can also be created from existing ones by employing the process of feature engineering. On the PD dataset, RF and GBC classifiers are used to determine the importance of features. There are 22 features in the feature set, 16 of which are (MDVP:MDVP, Fo(Hz):MDVP, Fhi(Hz):MDVP, Flo(Hz):MDVP, Jitter (Abs):MDVP, RAP:PPQ, Thrill:MDVP, DDP:MDVP, shimmering:(dB) Shimmer, Shimmer:APQ3, Glare:MDVP, APQ5:Shimmer:APQDDA, spread1, spread2, and PPE) have a significant impact when the RF classifier finds a value greater than 1. Similarly, 14 important features of the GBC classifier are (MDVP:MDVP, Fo(Hz):MDVP, Fhi(Hz):MDVP, Jitter (Abs):RAP, Thrill:MDVP, DDP:(dB) Shimmer, Shimmer:MDVP, APQ5:spread1, spread2, D2, and PPE) depicted in Fig. 4.

The performance of ML classifiers whose output can be two or more classes is measured using the Confusion Matrix. There are four distinct combinations of predicted and actual values in this table. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the four divisions of the combinations table in the confusion matrix. The accuracy or falsehood of the predictions made by the classifiers regarding PD patients is measured in Table 3.

A classification model's true positive rate (TPR) and false positive rate (FPR) performance are depicted on a graph called a receiver operating characteristic curve (ROC). Here, recall is referred to as the TPR [31]. Figure 5

Fig. 3 Maximum accuracy of classifier in k-fold model



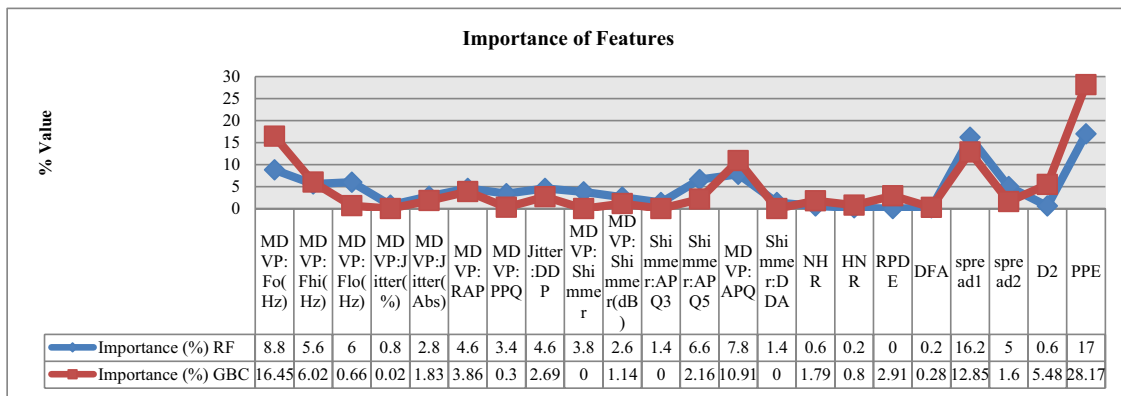


Fig. 4 Importance of features with RF and GBC

shows that GBC has the highest accuracy (AUC) of 95%, followed by DT with 92%. Figure 5 demonstrates that GBC serves as an important ensemble classifier in the diagnosis of Parkinson's disease patients.

### Discussion

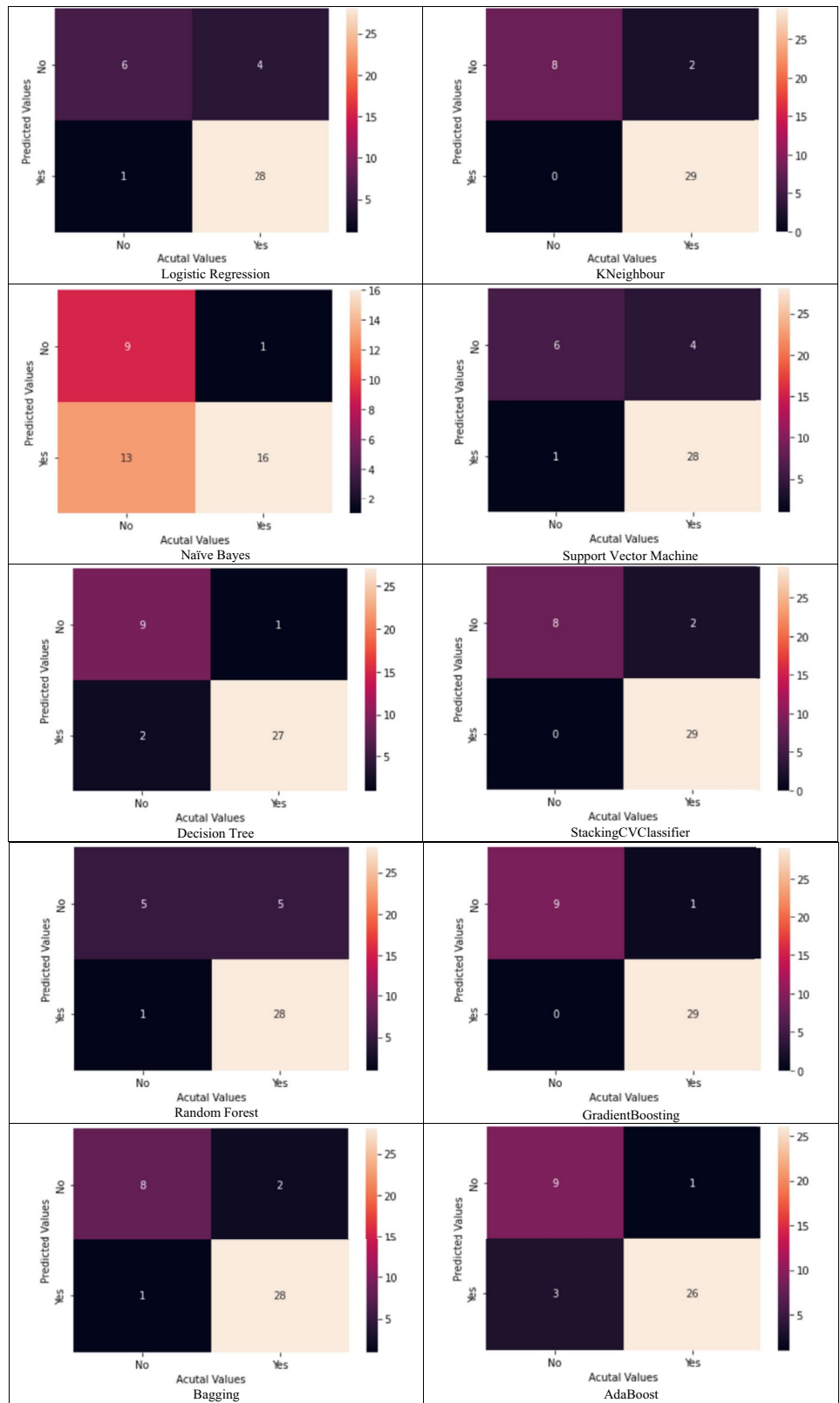
Python software was used to apply classification, stacking, ensemble, and k-fold model (considering mean value of accuracy) techniques to the dataset in this investigation. Each classifier model was trained with eighty percent of the dataset and tested with twenty percent. Four distinct models for anticipating Parkinson's infection have been created for this study. Using these four models, we can make predictions about PD patients. The confusion matrix as well as the ROC curve has been created for these models' validation. The experiment's scientific proof can be found in the validation result. The PD dataset was categorized by the first and most fundamental model, which used LR, KNN, NB, SVM, and DT classifiers as its base model. We have evaluated each classifier's Validation score (precision, recall, and F1 score) and model accuracy (Training and Testing) in Table 2. K-nearest neighbors outperformed LR, NB, SVM, and DT in terms of training and testing accuracy, scoring 100% in training and 94.87% in testing. Therefore, considering the PD conclusion in the field of ML for a specific dataset, we might consider KNN to be a significant classifier. Stacking is an ensemble machine learning algorithm in model 2 that uses meta-learning to learn how to combine predictions from two or more basic machine learning algorithms. The META model, which combines the predictions of two or more base models, is included in stacking model architecture. The accuracy of the model and the validation score of the stacking model (LR + KNN + NB + SVM + DT) are both shown in Table 2. In term of accuracy, for instance, training and testing which score 100% and 94.87% independently. Precision, recall,

and the F1 score all reached their usual maximum scores in the validation score. The RF, GBC, Bagging, and Adaboost are taken as ensemble model is the third model. The goal of using this ensemble strategy was primarily to lower the error rate. An ensemble can make better predictions and deliver better results than a single contributing model. In terms of testing accuracy, the GBC performed better than any other classifier (97.43%), like RF, Bagging, and Adaboost. As can be seen in Table 2, its accuracy has increased in additional metrics like precision, recall, and the F1 score. In reality, ensemble model is outstanding model in all models in general. Finally, in the fourth model, the accuracy of the classifiers at  $k = 10$  scale is estimated using k-fold cross-validation. The KNN classifier is the most accurate of all classifiers, with 95% accuracy. In the end, each of the above-mentioned four models generates testing accuracy and a variety of training. With a prediction accuracy of 97.43%, the ensemble model (Gradient Boosting) outperforms all of those trainings. After that, the K-Neighbors K-fold cross-validation model has the highest prediction accuracy—95%. In addition, the accuracy of the base model—K-Neighbors—and the meta-model—LR + KNN + NB + SVM + DT—is tie at the end.

Last but not least, when it comes to validation metrics like the confusion matrix and the ROC curve, GBC is the only classifier as model 3 (Ensemble model) that validates our experimental findings. Overall, we can conclude that the base model (Model 1) and the k-fold model (Model 4) both use KNN as a highly effective classifier. The proposed ensemble method is contrasted with other existing approaches for performance evaluation, demonstrating its superior implementation. Based on the data presented in Table 4, it is abundantly clear from the results of this experiment that the proposed approach, particularly GBC ensemble model, outperforms all of the other methods in terms of performance metrics.



**Table 3** Validation score of different classifiers through confusion matrix



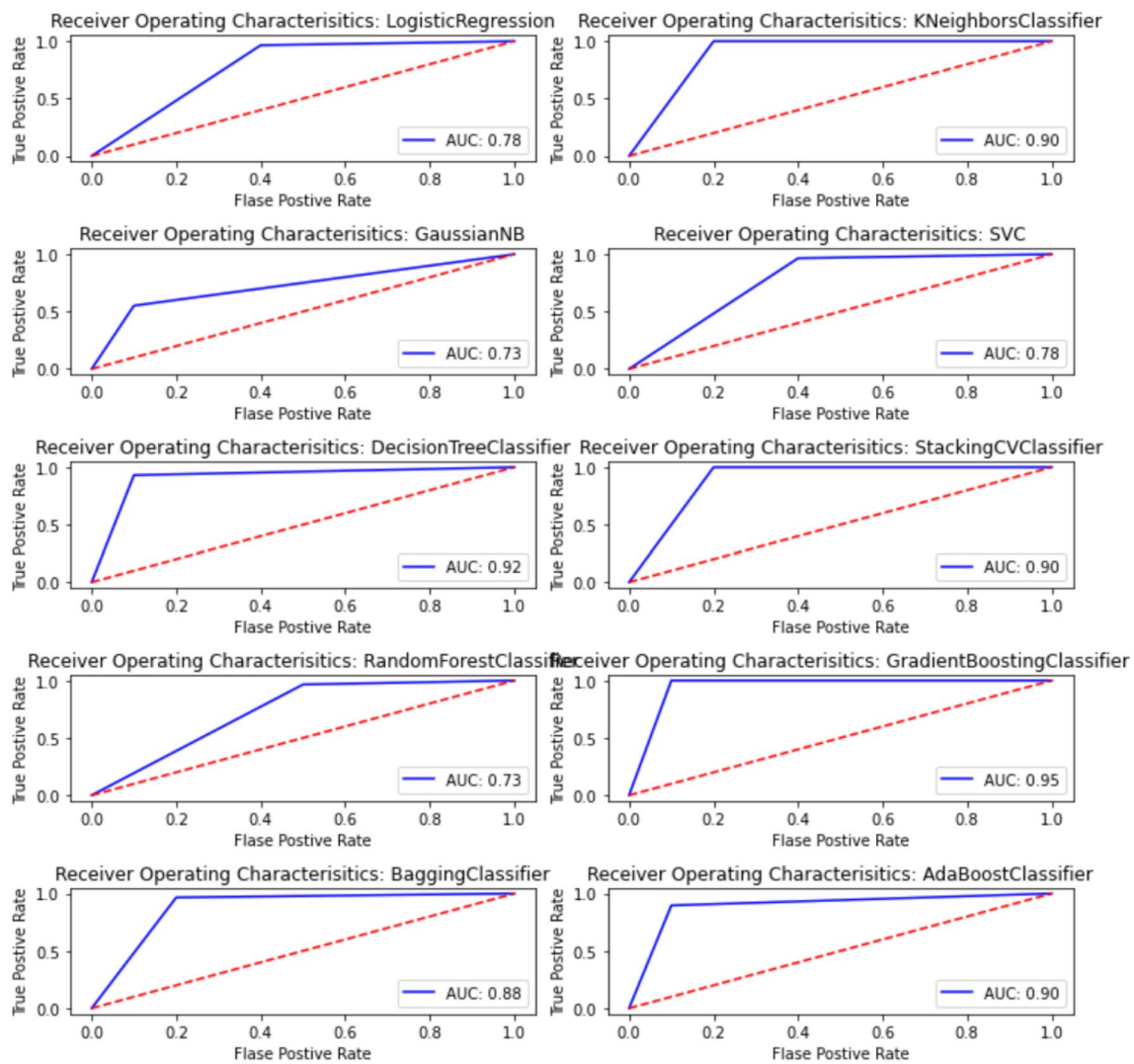


Fig. 5 ROC curves of different classifiers

## Conclusion

Today, medical data analysis employs a variety of methods for disease diagnosis. The majority of current strategies combine different methods. One approach suggested in this study is to predict and classify Parkinson's disease based on the publicly available dataset. Only voice-based parameters were used for disease prediction. Although it has been demonstrated that voice-based Parkinson's disease classification is effective, the capability to evaluate speech samples, which is necessary for effectively improving Parkinson's disease classification, is lacking in the methods currently in use. It is essential to propose a

simple, cost-effective method for quickly and accurately diagnosing Parkinson's disease due to its prevalence and difficulty in medical diagnosis. Comparing the frequency of voice in controlled environments among people with Parkinson's disease is a reliable method for diagnosing the condition. In this study, four model frameworks were utilized. The ensemble model outperforms the most recent cutting-edge work by a significant margin. According to the other metrics, the model was able to identify the diseases of patients with 97.43% accuracy. In future, our work on using imaging biomarkers to train a model using deep learning techniques will improve this method, which could be used to find additional diseases.

**Table 4** Comparison of proposed work with existing works

Study	Applied Methods	Used Dataset	Availability	Accuracy (%)
Cai et al. (2018) [32]	CBFO-FKNN	Oxford Parkinson's disease dataset Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University	<a href="https://archive.ics.uci.edu/ml/datasets/parkinsons">https://archive.ics.uci.edu/ml/datasets/parkinsons</a> <a href="https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+Multiple+Types+of+Sound+Recordings">https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+Multiple+Types+of+Sound+Recordings</a>	83.68
Khan et al. (2018) [33]	NN ensemble	Oxford Parkinson's disease dataset Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University	<a href="https://archive.ics.uci.edu/ml/datasets/parkinsons">https://archive.ics.uci.edu/ml/datasets/parkinsons</a> <a href="https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+Multiple+Types+of+Sound+Recordings">https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+Multiple+Types+of+Sound+Recordings</a>	90
Berus et al. (2019) [34]	ANNs ensemble	Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University	<a href="https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+Multiple+Types+of+Sound+Recordings">https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+Multiple+Types+of+Sound+Recordings</a>	86.47
Nissar et al. (2019)[35]	SMOTE and Random Forest	Parkinson's Disease Classification Dataset	<a href="http://archive.ics.uci.edu/ml/datasets/Parkinson's+Disease+Classification">http://archive.ics.uci.edu/ml/datasets/Parkinson's+Disease+Classification</a>	94.89
Castro et al. (2020) [36]	mRMR and XGBoost	Multiple datasets analyzed	<a href="https://www.sciencedirect.com/science/article/pii/S2590005621000278#bib11">https://www.sciencedirect.com/science/article/pii/S2590005621000278#bib11</a>	95.39
Rehman et al. (2021) [37]	MFCC, LDA, SVM	Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University	<a href="https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+Multiple+Types+of+Sound+Recordings">https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+Multiple+Types+of+Sound+Recordings</a>	77.5
Velmurugan and Dhinaran (2022) [38]	RF, XGBoost, Logistic regression and MLP	Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University	<a href="https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+Multiple+Types+of+Sound+Recordings">https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+Multiple+Types+of+Sound+Recordings</a>	90.13
Proposed method	GBC ensemble	Oxford Parkinson's disease dataset	<a href="https://archive.ics.uci.edu/ml/datasets/Parkinsons">https://archive.ics.uci.edu/ml/datasets/Parkinsons</a>	97.43

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- R.A. Hickman, S.A. O’Shea, M.F. Mehler, W.K. Chung, Neurogenetic disorders across the lifespan: from aberrant development to degeneration. *Nat. Rev. Neurol.* **18**(2), 117–124 (2022)
- P. Limousin, T. Foltynie, Long-term outcomes of deep brain stimulation in Parkinson disease. *Nat. Rev. Neurol.* **15**(4), 234–242 (2019)
- S. Rewar, A systematic review on Parkinson’s disease (PD). *Indian J. Res. Pharm. Biotechnol.* **3**(2), 176 (2015)
- D.V. Dimitrov, Medical internet of things and big data in healthcare. *Healthcare Inform. Res.* **22**(3), 156–163 (2016)
- S. Paul, M. Maindarkar, S. Saxena, L. Saba, M. Turk, M. Kalra, P.R. Krishnan, J.S. Suri, Bias investigation in artificial intelligence systems for early detection of parkinson’s disease: a narrative review. *Diagnostics* **12**(1), 166 (2022)
- N. Singh, P. Singh, A hybrid ensemble-filter wrapper feature selection approach for medical data classification. *Chemom. Intell. Lab. Syst.* **217**, 104396 (2021)
- D.D. Joshi, H.H. Joshi, B.Y. Panchal, P. Goel, A. Ganatra (2022) A Parkinson disease classification using stacking ensemble machine learning methodology. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1335–1341). IEEE
- M. Behroozi, A. Sami, A multiple-classifier framework for Parkinson’s disease detection based on various vocal tests. *Int. J. Telemed. Appl.* **2016**, 6837498 (2016). <https://doi.org/10.1155/2016/6837498>
- L. Berus, S. Klancnik, M. Brezocnik, M. Ficko, Novel discourse signal preparing calculations for high accuracy grouping of Parkinson’s illness. *Biomed. Design. IEEE Trans.* **59**(5), 1264–1271 (2018)
- C. OkanSakar, G. Serbes, A. Gunduz, H.C. Tunc, H. Nizam, B.E. Sakar, M. Tutuncu, M. Tarkan Aydin, E. Isenkul, H. Apaydin, Component Significance Analysis and Classification of Parkinson Disease Tele- Observing Data Through Data Mining Techniques. *Global Diary Adv. Res. Comput. Sci. Softw. Design.* **2**(3), 15 (2018)
- H. Gunduz (2019). Determination of Parkinson’s illness utilizing head part examination and boosting advisory group machines.
- S. Akshay, K. Vincent (2019). Component determination in Parkinson’s illness: A harsh sets approach, In: Software engineering and Information Technology, 2009. IMCSIT’09. Global Multiconference on (pp. 425–428). IEEE.
- D. Sztah, D. Hemmerling, Nonlinear discourse investigation calculations planned to a standard measurement accomplish clinically valuable measurement of normal Parkinson’s sickness manifestation seriousness. *Diary R. Soc. Interface* **8**(59), 842–855 (2019)
- I. Karabayir, S.M. Goldman, S. Pappu, O. Akbilgic, Gradient boosting for Parkinson’s disease diagnosis from voice recordings. *BMC Med. Inform. Decis. Mak.* **20**(1), 228 (2020). <https://doi.org/10.1186/s12911-020-01250-7>
- S.B. Dasari, P.R. Vital, T.V. K. Gangu (2020). Programmed Recognition of Parkinson’s sickness by means of artificial
- J. Jankovic, Parkinson’s disease: clinical features and diagnosis. *J. Neurol. Neurosurg. Psychiatry* **79**(4), 368–376 (2008). <https://doi.org/10.1136/jnnp.2007.131045>
- A.J. Harding, E. Stimson, J.M. Henderson, G.M. Halliday, Clinical correlates of selective pathology in the amygdala of patients with Parkinson’s disease. *Brain* **125**(Pt 11), 2431–2445 (2002). <https://doi.org/10.1093/brain/awf251>
- A.J. Hughes, Y. Ben-Shlomo, S.E. Daniel, A.J. Lees, What features improve the accuracy of clinical diagnosis in Parkinson’s disease: a clinicopathologic study. *Neurology* **42**(6), 1142–1146 (1992). <https://doi.org/10.1212/wnl.42.6.1142>
- C.G. Goetz, G.T. Stebbins, B. Ouyang, Visual plus nonvisual hallucinations in Parkinson’s disease: development and evolution over 10 years. *Mov Disord.* **26**(12), 2196–2200 (2011). <https://doi.org/10.1002/mds.23835>
- D. Roberts-Warrior, A. Overby, J. Jankovic, S. Olson, E.C. Lai, J.K. Krauss, R. Grossman, Postural control in Parkinson’s disease after unilateral posteroventral pallidotomy. *Brain* **123**(10), 2141–2149 (2000)
- V. Chaurasia, S. Pal, Ensemble technique to predict breast cancer on multiple datasets. *Comput. J.* **65**(10), 2730–2740 (2022)
- S. Messaoud, A. Bradai, S.H.R. Bukhari, P.T.A. Quang, O.B. Ahmed, M. Atri, A survey on machine learning in internet of things: algorithms, strategies, and applications. *Internet Things* **12**, 100314 (2020)
- J. Carreira, A. Zisserman (2017) Quo vadis, action recognition? A new model and the kinetics dataset. In: proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6299–6308)
- A. Vázquez-Ingelmo, A. García-Holgado, F.J. García-Peñalvo, R. Therón, A meta-model integration for supporting knowledge discovery in specific domains: a case study in healthcare. *Sensors* **20**(15), 4072 (2020)
- H. Tabrizchi, M.M. Javidi, V. Amirzadeh, Estimates of residential building energy consumption using a multi-verse optimizer-based support vector machine with k-fold cross-validation. *Evol. Syst.* **12**(3), 755–767 (2021)
- O. Sagi, L. Rokach, Ensemble learning: a survey. *Wiley Interdiscip. Rev.* **8**(4), e1249 (2018)
- J. Nayak, B. Naik, P.B. Dash, A. Souri, V. Shanmuganathan, Hyper-parameter tuned light gradient boosting machine using memetic firefly algorithm for hand gesture recognition. *Appl. Soft Comput.* **107**, 107478 (2021)
- UCI machine learning repository dataset accessed on 10th October 2022. (<https://archive.ics.uci.edu/ml/datasets/Parkinsons>)
- V. Chaurasia, A. Chaurasia, Novel method of characterization of heart disease prediction using sequential feature selection-based ensemble technique. *Biomed. Mater. Devices* **20**, 1–10 (2023)
- M. Rumman, A.N. Tasneem, S. Farzana, M.I. Pavel, M.A. Alam (2018) Early detection of Parkinson’s disease using image processing and artificial neural network. In: 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR) (pp. 256–261). IEEE.
- D.M. Powers (2020) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint [arXiv:2010.16061](https://arxiv.org/abs/2010.16061).
- Z. Cai, J. Gu, C. Wen, D. Zhao, C. Huang, H. Huang, C. Tong, J. Li, H. Chen, An intelligent parkinson’s disease diagnostic system based on a chaotic bacterial foraging optimization enhanced fuzzy knn approach. *Comput. Math. Methods Med.* (2018). <https://doi.org/10.1155/2018/2396952>
- M.M. Khan, A. Mendes, S.K. Chalup, Evolutionary wavelet neural network ensembles for breast cancer and Parkinson’s disease prediction. *PLoS ONE* **13**(2), e0192192 (2018)

34. L. Berus, S. Klancnik, M. Brezocnik, M. Ficko, Classifying parkinson's disease based on acoustic measures using artificial neural networks. *Sensors* **19**(1), 16 (2019)
35. I. Nissar, D.R. Rizvi, S. Masood, A.N. Mir, Voice-based detection of Parkinson's disease through ensemble machine learning approach: a Performance study. *EAI Endorsed Trans Pervasive Health Technol* **5**(19), e2–e2 (2019)
36. C. Castro, E. Vargas-Viveros, A. Sánchez, E. Gutiérrez-López, D.L. Flores (2020) Parkinson's disease classification using artificial neural networks. In: VIII Latin American Conference on Biomedical Engineering and XLII National Conference on Biomedical Engineering: Proceedings of CLAIB-CNIB 2019, October 2–5, 2019, Cancún, (pp. 1060–1065). Springer International Publishing.
37. A. Rahman, S.S. Rizvi, A. Khan, A. AfzaalAbbasi, S.U. Khan, T.S. Chung, Parkinson's disease diagnosis in cepstral domain using MFCC and dimensionality reduction with SVM classifier. *Mobile Inform. Syst.* (2021). <https://doi.org/10.1155/2021/8822069>
38. T. Velmurugan, J. Dhinakaran. (2022) A novel ensemble stacking learning algorithm for Parkinson's disease prediction. *Math. Probl. Eng.* 2022.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.