Multimedia Tools and Applications (2024) 83:33207–33234
https://doi.org/10.1007/s11042-023-16881-x

Predicting patients with Parkinson's disease using Machine Learning and ensemble voting technique

Shawki Saleh[1] · Bouchaib Cherradi[1,2] · Oussama El Gannour[1] · Soufiane Hamida[1,3] · Omar Bouattane[1]

Received: 25 March 2023 / Revised: 25 July 2023 / Accepted: 4 September 2023 /
Published online: 15 September 2023

Abstract
Parkinson's disease is the second most common neurological disorder that causes significant physical disabilities, decreases the quality of life, and does not have a cure. Because it is a nervous system disorder, it impacts people in different ways, affecting movement and speech and causing muscle stiffness. Approximately, 90% of people with Parkinson's disease have speech disorders. Machine Learning (ML) algorithms can mostly be employed for the early detection of diseases to enhance the lifespan and improve the lifestyle of people with chronic diseases such as Parkinson's disease. In this paper, we have employed an Artificial Neurons Network (ANN) and nineteen ML algorithms to predict people with Parkinson's disease using two different acoustic datasets. Contrary to the train-test split approach, this work aims to utilize the cross-validation technique to estimate the performance. The objective is to ensure that each sample in these small and unbalanced acoustic datasets contributes to both the training and testing processes to provide accurate estimations for the performance of the classifiers on unseen dataset, and to provide a clear insight into the effectiveness of ML algorithms in diagnosing Parkinson's disease via voice disorder. To enhance the performance of the prediction, we employed several techniques such as Optimal Hyperparameters Tuning and Cross-Validation to obtain the best performance and results, and we have provided a detailed explanation of these algorithms' performance and the Optimal Hyperparameters used for each of them. Based on the results and performance, the best classifiers have been selected to build two independent ensemble voting classifiers for the two different datasets. We calculated and represented the accuracy, sensitivity, specificity, precision and AUC. They reached 96.41% and 97.35% of accuracy, respectively.

Keywords Healthcare · Parkinson's disease prediction · Hyperparameters · Cross-validation · Ensemble voting classifier · Disease diagnosis · Machine Learning

✉ Bouchaib Cherradi
bouchaib.cherradi@gmail.com

[1] EEIS Laboratory, ENSET of Mohammedia, Hassan II University of Casablanca, Mohammedia, Morocco

[2] STIE Team, CRMEF Casablanca-Settat, Provincial Section of El Jadida, 24000 El Jadida, Morocco

[3] GENIUS Laboratory, SupMTI of Rabat, Rabat, Morocco

1 Introduction

Parkinson's disease (PD) is a degenerative nervous system illness marked by tremors, muscle stiffness, inaccurate movement, and speech difficulty and disorders. It especially affects middle-aged and elderly people [1, 2]. It is associated with degeneration of the brain's basal ganglia and a deficiency of the neurotransmitter dopamine. Parkinson's disease indicators and symptoms may vary from one patient to another. Early symptoms of the disease may be mild and unnoticed [3–5].

Historically, the diagnosis of maladies has depended on the experience of doctors to identify the nature of an illness or other problem by examination of the symptoms. Recently, Artificial intelligence (AI) and the internet of things (IoT) have provided multiple opportunities for researchers and the scientific community to produce effective decision-support systems that can be used locally and remotely to diagnose various diseases in healthcare centers [6], and medical fields [7], [8]. Nowadays, decision-support systems are available in several medical fields and provide an essential role in diagnosing many diseases such as Breast cancer [9], Atherosclerosis diseases [10–12], Cervical cancer diagnosis [13], Diabetes diseases [14, 15], Cardiovascular diagnosis [16–18], Arrhythmia [19], Brain tumours [20, 21], Pneumonia [22], COVID-19 [23–29] and its consequences that impact mental health such as depression [30], Parkinson's disease [31–36], etc.

Supervised learning is a popular category of machine learning algorithms where the model learns patterns and relationships from labelled training datasets. It is widely used in various domains, including image classification, text sentiment analysis, fraud detection, and healthcare diagnosis. It is a powerful approach that leverages labelled data to enable the prediction and classification of future, unseen instances. On one hand, we noticed that the train-test split technique is widely used for training and evaluating ML algorithms or models without taking into account the size of the data set, availability of data and specific research objectives, whereas this technique is commonly used when the dataset is large. On the other hand, cross-validation is preferred when the dataset is relatively small and unbalanced or when a more robust performance estimation is required; especially in medical field. While training and evaluation process using train-test splits, the performance of the classifier increases and decreases based on three parameters which are train-test splits' size, Stratify, and Random State (shuffle). Also, as shown in the related work section, some literature studies have proposed systems achieving nearly 100% accuracy in distinguishing healthy individuals from those who have PD using an unbalance or limited datasets. The main questions are: Is the performance obtained accurate as well as trustworthy**?** How much confidence can we have in these classifiers to support decision-makers in predicting Parkinson's Disease**?**

In this study, we focused on diagnosing Parkinson's disease via speech disorders using Machine Learning (ML) algorithms and Artificial Neural Networks (ANN) [37–39]. Thus, we have employed an ANN and nineteen ML algorithms to perform this mission using two different acoustic datasets. Actually, Some of ML algorithms have good performance on some datasets, but they are unable to achieve the desired results with other datasets. Moreover, the performance can be improved or increased by adjusting the hyperparameters of these algorithms, because they control the behaviour of the algorithms during training processes and can significantly impact the performance and generalization ability of the models [40]. Therefore, selecting the values of hyperparameters is critical. Any incorrect selection of these hyperparameters delays learning and confuses the algorithm and then offers an inappropriate or poor performance [40,

41]. Two small and unbalanced datasets have been used for training and evaluating the performance of our classifiers. For this reason, we used two types of cross-validation (CV) techniques which are Stratified k-fold CV and Leave-one-out CV to estimate the performance of our classifiers [42]. These techniques allowed us to use the entire dataset for training and evaluating the models, rather than just a single train/test split. These techniques can provide a more accurate estimation of the model's performance, and its principal goal is to estimate the performance of classifiers on unseen data. Moreover, CV techniques are often used for model selection, hyperparameter tuning, performance evaluation, etc. Furthermore, we presented a detailed comparison between Stratified k-fold CV and Leave-one-out CV, including their working principle and the results obtained during the training and evaluation of classifiers for providing insight into the impact of the number of training and testing samples in small and unbalanced datasets on performance.

To improve the generalizability and robustness of the classifiers, as well as reduce overfitting, we employed an ensemble voting technique which is a machine-learning model that combines the predictions of multiple individual classifiers and uses a voting system to make a final prediction [43, 44]. The idea behind ensemble models is that multiple models working together can achieve better performance than any individual model alone. During the prediction phase, each individual classifier makes a prediction, and the ensemble model aggregates the predictions and makes a final prediction based on a majority vote or some other voting scheme such as soft voting that calculates the average of probabilities for each class. to find the best combination of hyperparameter. As we mentioned earlier, we have employed an ANN and nineteen ML algorithms for predicting PD. To maximize the performance of these algorithms on unseen data, we used several techniques of search including Grid Search and Manual search to find the best combination and to obtain the optimal hyperparameters which offer the most favourable distinction (or classification) of people with Parkinson's disease from those who are healthy. Due to the need for a robust performance estimation for the classifiers in the medical field and to provide a clear insight into the effectiveness of ML algorithms in diagnosing Parkinson's disease via voice disorder, these classifiers have been evaluated using two techniques of CV to estimate the classifiers' performance and to provide a clear visualization of the models' behaviour on unseen data. And finally, we selected the best classifiers to build two Ensemble Voting classifiers that predict people with Parkinson's disease using two different datasets in order to enhance the classification performance due to its critical importance in the medical field. We can summarize the main contributions of this research as follows:

- Selecting and implementing twenty ML algorithms to predict Parkinson's disease via speech disorders using two different acoustic datasets.
- Tuning the hyperparameters of these classifiers to obtain the optimal hyperparameters that provide maximum performance.
- Training and evaluating the classifiers using different techniques such as stratified K-fold, LOOCV and AUC to estimate the performance of the classifiers on unseen data.
- Development and implementation of ensemble voting classifiers to enhance the prediction performance of Parkinson's disease by combining the predictions of multiple models.
- To provide a comprehensive comparison concerning the training, evaluation and techniques used in this work.

This paper has been structured as follows: Section 2 appears previous literature related to this work, and Section 3 provides an overview concerning methodology, techniques and materials used to evaluate the classifiers and establish their hyperparameters. Then, the discussion and the results are presented in Section 4. Finally, Section 5 concludes this paper.

2 Related work

In Parkinson's disease, changes in voice quality may be the first symptom because this symptom is common in people with Parkinson's disease patients. In fact, there are many studies which use voice vocal to diagnose the illness.

The authors in study [45] aim to understand various classifiers which predict Parkinson's disease using vocal datasets and compare their performance. The five algorithms had been used are SCFW with KELM, ACO, PSO, FCM and the algorithm ABO.

In [8], the authors had compared the performance between SVM, Least Square Support Vector Machines (LS-SVM), General Regression Neural Network (GRNN), and Multilayer Perceptron Neural Network (MLPNN). These algorithms had been used for tracking Parkinson's disease progression remotely, and they found that the algorithm LS-SVM offers the best accuracy and performance.

The authors in study [46] used fuzzy k-NN for predicting Parkinson's disease. And They have compared it with SVM. They mentioned that their methodology proposed was offered appropriate results comparing with the other methods in previous literatures, and they reached accuracy of (96.07%).

The authors in paper [47] had used UCI ML repository dataset and four algorithms of classification. They found that the accuracies of SVM is (81.0%), Logistic regression is (75.0%), Extra tree classifier is (77.0%) and Decision tree classifier is (73.0%).

In the paper [48], The authors had employed a Random Forest Classifier (RFC) and ANN model to predict Parkinson's disease using UCI dataset. Then, they compared the accuracy of both RF Classifier and ANN model with the Principal Component Analysis (PCA). They captured a visible difference because the accuracy of ANN model increases from 79.47% to 97.354 when they use PCA. On the other side, RFC reduces its accuracy from 89.534 to 76.65 when they were applying the feature reduction technique. Also, the purpose of the study [32] is to compare ML approaches for tracking and monitoring of Parkinson's disease progression remotely.

In the study [35], the authors have focused on the analysis, evaluation and comparison of nine ML Algorithms using an acoustic Parkinson's disease dataset. They reached accuracy rate equal 97.22% which belongs to KNN classifier. Also in [34], they had employed both ANN and CNN model to classify patients with Parkinson's disease from healthy people based on acoustic features. They reached an accuracy of 93.10 % when they were utilizing the CNN model.

The study [49] examines variables extracted from voice analysis, presenting a non-intrusive technique. The paper proposes a deep learning method with two objectives. The first, determining severe or non-severe Parkinson's disease presence. The second, estimating the disease's progression in a specific patient using regression techniques. The Unified Parkinson's Disease Rating Scale (UPDRS) is utilized, considering motor and total labels. Results indicate that a mixed Multi-Layer Perceptron (MLP) combined with an autoencoder for feature selection achieves a 99.15% success rate in predicting severe or non-severe Parkinson's disease.

The presented Comparative Table 1 of Related Works provides a comprehensive overview and comparison of various studies. It highlights the algorithms used, datasets utilized, train-test split ratios, as well as feature extraction and selection techniques employed in each study. Additionally, it showcases the achieved accuracy for each approach, offering valuable insights into the performance of different methods and their suitability for specific tasks.

3 Materials and methods

3.1 Parkinson's disease diagnosis

Parkinson's disease signs and symptoms may vary from one person to another. Early signs of disease may be mild and unnoticed. In fact, there are many methods for detecting these symptoms to diagnose Parkinson's using AI such as Detecting Parkinson's disease from speech disorders [50], detecting Parkinson's disease from handwriting using deep learning [36], detecting Parkinson's disease from the sketching of the spiral and wave [51, 52], detecting Parkinson's disease using EEG signals [53], detecting Parkinson's disease from Hand tremor [54] and detecting Parkinson's disease using nocturnal breathing signals [31].

3.2 Predicting system global overview

There are many solutions to predict Parkinson's disease, and many techniques have been proposed in previous literature to improve the prediction of this disease. Since one of the earliest symptoms of this disease is a speech disorder, this paper provides an overview and a comprehensive comparison of the effectiveness of diagnosing the disease through voice using ML algorithms. As shown in Fig. 1, the process began by recording the voices of some volunteers with Parkinson's disease and healthy people. These recordings have been processed and features extracted from these voices, and this dataset became available in UC Irvine ML Repository.

In this paper, we contributed by the selection of twenty classifiers and two different databases of Parkinson's disease, and to provide an accurate assessment of the performance of the classifiers using these datasets, which we suspect about their performances when they are evaluated by the common technique of Train-Test split; because these datasets are small and unbalanced. We started by processing and analyzing the datasets and then choosing the classifiers that will be used for this study. The speed of classification and the good performance of these classifiers depends mainly on tuning their hyperparameters which we tuned using Grid Search. After selecting the optimal hyperparameters, the classifiers have been trained and evaluated using Stratified K-Fold and LOO CV to estimate their real performance on unseen data. Based on the performance indicators that we obtained, the best-performing classifiers have been selected to build two ensemble voting classifiers in order to enhance the quality of classification due to its importance in the medical field.

3.3 Dataset I description

The first dataset used in this study was created by Max Little [55, 56] and can be found in the UCI (University of California Irvine) ML Repository [57]. This dataset was recorded at the National Centre for Voice and Speech in Denver, Colorado, and consists of 195 voice

Table 1 comparison between related works and feature extraction approach

Study	Datasets	Algorithm	Train-test approach	Feature extraction	Accuracy
[45]	UCI Repository I 195 records x 23 features	ABO	Train-test split	Relief Feature Selection (RFS): is a supervised feature selection method that aims to rank features to maximize the performance of a specified learning algorithm.	97.50
		SCFW+KELM			93.82
		FCM			91.33
		PSO			87.02
		ACO			83.17
[48]	UCI Repository II 756 records x 753	ANN	Train-test split	Principal Component Analysis (PCA): It is a technique in data ana-lysis and ML to reduce dimensionnality in a dataset while presserving the most important information.	79.47
		ANN+PCA			97.35
		RF			89.53
		RF+PCA			76.65
[35]	UCI Repository III 240 records x 45 features	SVM	Train-test split 85% and 15%	Default features of dataset	91.67
		DA	Train-test split 85% and 15%	Default features of dataset	88.89
		KNN	Train-test split 85% and 15%	Default features of dataset	97.22
		DT	Train-test split 85% and 15%	Default features of dataset	97.22
		FR	Train-test split 85% and 15%	Default features of dataset	91.67
		BT	Train-test split 85% and 15%	Default features of dataset	94.44
		NB	Train-test split 85% and 15%	Default features of dataset	91.67
		AB	Train-test split 85% and 15%	Default features of dataset	94.44
[34]	UCI Repository I 195 records x 23	ANN	Train-test split 85% and 15%	For ANN: Default features	82.76
		CNN	Train-test split 85% and 15%	For CNN: Convolution Layers + Normalization Layer + Activation Layer (ReLU)	93.10
	UCI Repository III 240 records x 45	ANN	Train-test split 85% and 15%		72.22
		CNN	Train-test split 85% and 15%		88.89
[49]	UCI Repository I 195 records x 23	ANN: MLP	Train-test split 80% and 20%	Autoencoders: self-supervised algorithm which are used to reduce the size of input data by recreating it.	99.00

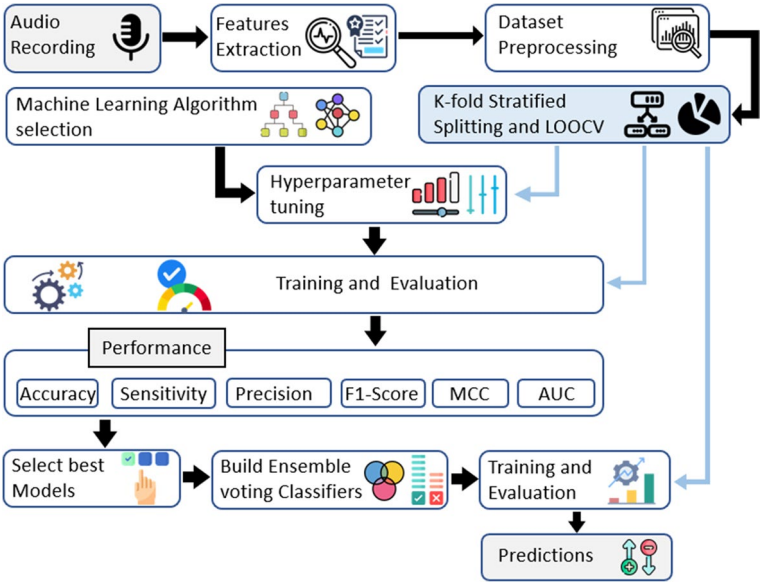

Fig. 1 Flowchart of the system to predict Parkinson's disease via speech disorder

recordings (samples) of 31 individuals (23 persons have Parkinson's disease and 8 are healthy). Moreover, this dataset has 23 features including a number of the record, some essential features of audio and the 'status' that indicates whether the individual has Parkinson's disease (value of 1) or is healthy (value of 0) [34]. As shown in Table 2, the voice measures in this dataset include various features related to vocal fundamental frequency (average, maximum, and minimum), fundamental frequency perturbation (jitter), absolute jitter in microseconds, relative amplitude perturbation (RAP), period perturbation quotient (PPQ), local amplitude perturbation (shimmer), and amplitude perturbation quotient (APQ). Other features comprise noise-to-harmonics ratio (NHR), harmonics-to-noise ratio

Table 2 Dataset I of Parkinson's disease (Feature Description)

Features	Description
MDVP: Fo (Hz), Fhi (Hz) & Flo (Hz)	Vocal Fundamental Frequencies: Average, maximum and minimum
MDVP: Shimmer, Shimmer (dB), APQ3, APQ5, APQ & DDA	Amplitude Perturbation (Amplitude Variation Measurements)
MDVP: Jitter (%), Jitter (Abs), RAP, PPQ & DDP	Fundamental Frequency Variation Measurements
HNR and NHR	Harmonics_to_Noise Ratios and Noise_to_Harmonics Ratio
Spread 1, Spread2, & PPE	Fundamental Frequency Variation (Non-linear Measurements)
RPDE & D2	Measures of non-linear dynamical complexity
DFA	Detrended Fluctuation Analysis
Status	*Labels (0: Healthy and 1: Parkinson's Disease)*

(HNR), recurrence period density entropy (RPDE), correlation dimension (D2), detrended fluctuation analysis (DFA), fundamental frequency variation (Spread1 and Spread2), pitch period entropy (PPE), etc.

3.4 Dataset II description

The second dataset used in our study was collected at Istanbul University (Department of Neurology in Cerrahpaşa Faculty of Medicine). This dataset consists of 188 records gathered from 81 females and 107 males who were diagnosed with Parkinson's disease and their ages range between 33 and 87 (65.1Y±10.9) years old. In addition, it consists of 64 records gathered from healthy individuals (41 females and 23 males) whose ages range between 41 and 82 (61.1Y±8.9). This dataset was being collected by a microphone set on 44.1 kHz while participants pronounce prolonged the vowel "a" with three repetitions for each person. The total dataset contains 756 samples and 753 features including the gender, TWQT features, Time-Frequency, Mel-Frequency-Cepstral Coefficients, Vocal Fold Features and Wavelet Transform-based Features [58–60]. In this dataset of Parkinson's Disease, various acoustic features have been extracted to analyze and diagnose vocal abnormalities. The baseline features encompass Jitter and Shimmer variants, capturing instabilities in the oscillating pattern of vocal folds by quantifying changes in fundamental frequency and amplitude. Fundamental frequency parameters provide information about vocal fold vibration, including mean, median, standard deviation, and minimum and maximum values. Harmonicity parameters assess the effects of incomplete vocal fold closure, measuring the ratio of signal information to noise. Recurrence Period Density Entropy (RPDE) evaluates the ability of vocal folds to sustain stable oscillations, while Detrended Fluctuation Analysis (DFA) quantifies the self-similarity of turbulent noise. Pitch Period Entropy (PPE) measures impaired control of fundamental frequency using a logarithmic scale. Time frequency features include Intensity Parameters, Formant Frequencies, and Bandwidth. Mel Frequency Cepstral Coefficients (MFCCs) capture Parkinson's Disease effects in the vocal tract. Wavelet Transform-based Features analyze deviations in fundamental frequency. Vocal fold features consist of Glottis Quotient (GQ), Glottal to Noise Excitation (GNE), Vocal Fold Excitation Ratio (VFER), and Empirical Mode Decomposition (EMD) for quantifying glottis movements, turbulent noise, and pathological vocal fold vibration. The tunable Q-factor wavelet transform (TQWT) [58]. These features provide valuable insights for Parkinson's Disease diagnosis and assessment of vocal pathology.

3.5 Dataset preprocessing and normalization

Dataset preprocessing in this study is used for handling missing values by imputing or removing them, removing duplicates to avoid bias, and converting categorical data into numerical representations. Data normalization, also known as feature scaling or standardization, is a crucial preprocessing technique in ML. It is typically applied to features with different scales or units of measurement. This technique aims to allow ML algorithms to treat all features equally, preventing one feature from dominating the learning process due to its larger scale. In fact, there are many methods of normalization such as Z-Score Normalization, Constant Factor Normalization, Decimal Scaling, Min-Max Scaling, etc. In this work, we used the Min-Max Scaling technique to scale all numerical features of both datasets I and II to a specific range between [Min and Max: -1 and 1] according to Eqs. 1 and 2.

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

$$\widehat{x} = y^{*}(\text{Max} - \text{Min}) + \text{Min} \tag{2}$$

Where, X is the original, $\widehat{X}$ is the target, Max and Min are the desired specific range, Xmax and Xmin are the max and the min numbers in the dataset. To perform this process, we used a pre-developed algorithm available in preprocessing package of scikit-learn using Python Language.

3.6 Finding appropriate hyperparameters for ML models

As a ML engineer building a model, you define and specify the values of the hyperparameters that your classifier will use before training begins. In fact, there are two kinds of parameters, the first is Hyperparameters and the second is Parameters. The parameters which are chosen for the model while creating are Hyperparameter such as Sigmoid, ReLU, Tanh, Adam optimizer, K-Neighbors in KNN, etc. In addition, the Parameters which are estimated or evaluated during training such as the mapping of weights and biases between the input features and targets.

Controlling the behavior of a ML model requires hyperparameter adjustment. If the hyperparameters are not properly tuned, the parameters of model (or classifier) yield unsatisfactory or unacceptable results. This means the model makes more errors. Moreover, the accuracy, sensitivity and the other indicators in confusion matrix will be worse. We can determine these optimal hyperparameters values using manual or automated methods. In fact, there are various automated methods to search for the optimal hyperparameters values such as Grid search, Random Search, Bayesian optimization, etc. In this study, we have used Grid search with two acoustical datasets and different Hyperparameters with 20 ML Algorithms; and the results of these searches enable us to choose the best hyperparameters for the models that provide the best performance and minimum error. Flowchart in Fig. 2 represents the behavior of the grid search, and how the dataset is split into k-fold in each split, every split is used individually for training and validation of the model. And in each

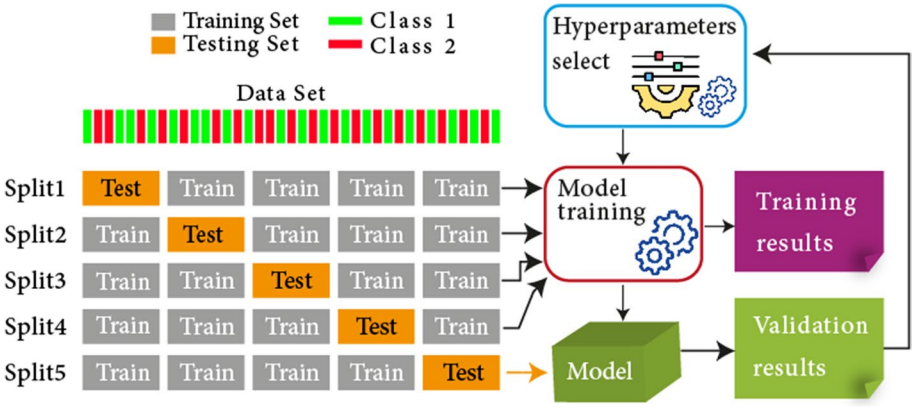

Fig. 2 Flowchart of Grid search used for Hyperparameters tuning

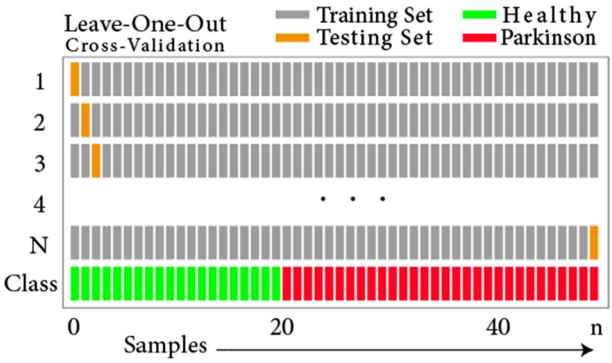

Fig. 3 Leave-one-Out CV

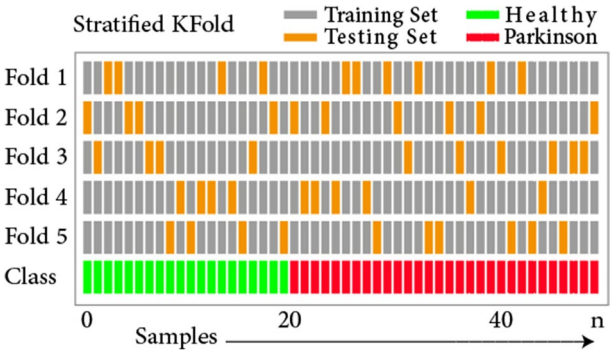

Fig. 4 Stratified K-Fold CV

iteration, it uses the same hyperparameter and returns the average of the performance of these iterations and the group (combinations) of the hyperparameter used.

3.7 Algorithms evaluation

Most of the times, we split the dataset into train/test and use confusion matrix to evaluate the performance of the models. However, this is insufficient always to truly judge the performance of these models. Actually, there are several techniques to evaluate the models' performance using CV [61–64] such as K-fold, Hold-out, Stratified k-fold, Leave-p-out, Leave-one-out (LOO), Monte Carlo (shuffle-split), Time series (rolling cross-validation), etc. In this study we used LOOCV and Stratified k-fold CV.

- **LOOCV**: It is the most computationally expensive variant of k-fold CV because only one sample point is utilized as a validation set; whilst the remaining n-1 samples are using for training as shown in Fig. 3. Also, it is a more robust estimate of model performance that means each sample in dataset has an opportunity to play the role as entirety test set. It is useful with a small and unbalanced dataset or when estimated model performance is critical.
- **Stratified k-fold CV**: It is a strategy that splits a dataset into k equal folds. As shown in Fig. 4, it is equivalent to k-fold except that each fold has the same ratio of instances of target variables. This technique enables us to evaluate the performance of our models perfectly using unbalanced acoustic datasets.

To calculate the performance of the models in the context of binary classification of Parkinson's disease within the LOOCV or Stratified k-fold CV framework, we designed an algorithm for this mission according to the following pseudocode (Algorithm 1).

```
Input: Dataset with features X and corresponding labels y, CV type, Number of folds k
Output: Evaluation metrics (e.g., accuracy, recall, precision)
if LOOCV:
1.      | Split the dataset into k folds equal to the number of dataset samples.
Else If Stratified k-fold Cross Validation:
1.      | Split dataset into k-folds while preserving the class distribution.
2.  Initialize an empty array.
3.  For each fold i in k:
    a.  | Set fold i as the validation set.
    b.  | Combine the remaining k-1 folds to create the training set.
    c.  | Train the machine learning model using the training set.
    d.  | predicted labels ← Make predictions on the validation set.
        | If LOOCV:
    e.  |    Store the predicted label of the validation set fold i in the array[i].
        | Else If Stratified k-fold Cross Validation:
    e.  |    Calculate the evaluation metrics using the true and predicted labels.
    f.  |    Store the evaluation metrics for fold i in the array[i].
    End
    If LOOCV:
4.      | Calculate the evaluation metrics using the original and predicted labels.
5.      | Return the evaluation metrics.
    Else If Stratified k-fold Cross Validation:
4.      | Calculate the average performance across all k folds in the array.
5.      | Return the average performance of Evaluation metrics.
```

Algorithm 1 Leave One Out and Stratified k-fold Cross Validation

In LOOCV, the model is trained and evaluated for each data point in the dataset. The prediction results are stored for each LOOCV iteration, and the performance is calculated according to the predicted labels and the original labels of the dataset, while in Stratified K-fold the model is trained on the training set consisting of the remaining folds and evaluated on the current fold. The performance metrics are stored for each stratified CV fold, and the average performance metric is calculated.

3.8 Ensemble voting methodology

Ensemble voting is a ML technique that combines the predictions of multiple models to create a more accurate and robust prediction [43, 44]. One way to achieve ensemble learning is through voting classifiers, which involves training several different models and having each model make a prediction for a given input. The final prediction is then made based on the majority vote of the individual models. This can be used for both classification and regression tasks. One of the main advantages of using voting classifiers is that it can often improve the performance of the model by reducing overfitting and increasing the generalization ability of the model. This is because each individual model may have its own strengths and weaknesses, and by combining the predictions of multiple models, the ensemble can take advantage of the strengths of each model while mitigating the weaknesses. Additionally, ensemble methods can be more robust to noise and variations in the data, which can further improve the performance of the model. The working principal of this technique is represented in Fig. 5.

3.9 Performance evaluation

Performance evaluation using the confusion matrix provides detailed insights into the predictive accuracy of a classifier. The confusion matrix is a table that summarizes the predictions

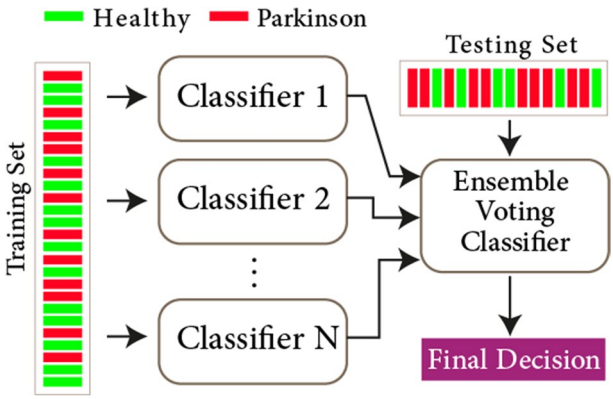

Fig. 5 Working principal of Ensemble voting classifier

made by the model against the actual values in the dataset. As shown in Table 3, it consists of four key indicators which are true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Based on the confusion matrix, the performance of all classifiers used in this paper was evaluated using multiple evaluation criteria, including accuracy, specificity, sensitivity, precision and f1-score.

TP represents the number of instances where the model correctly predicted Parkinson's disease (PD) when the actual value was PD. TN represents the number of instances where the model correctly predicted Healthy when the actual value was Healthy. FP represents the number of instances where the model incorrectly predicted as PD when the actual value was Healthy. FN represents the number of instances where the model incorrectly predicted as Healthy when the actual value was PD.

- **Accuracy:** It indicates the right predictions of the model from the total forecasts and is calculated using Eq. 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{3}$$

- **Sensitivity (*Recall*):** Also known as true positive rate, it focuses on the model's ability to avoid FN. It is particularly useful in scenarios where the identification of positive instances is critical, such as detecting diseases, fraud, or any situation where FN has severe consequences. It measures the proportion of actual instances of PD that are correctly identified by the model and can be calculated by Eq. 4.

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{4}$$

Table 3 Confusion Matrix

	As Healthy	As PD *	
Healthy person	TN	FP	***Actual***
PD * patient	FN	TP	
	Predicted		

* Parkinson's Disease

- **Specificity:** The number of correct negative predictions divided by the total of numbers that the algorithm correctly predicted as healthy and the people who are healthy and the algorithm forecasted them with Parkinson. According to Eq. 5, Specificity demonstrates the model's ability to correctly identify individuals who are healthy.

$$Specificity = \frac{TN}{TN + FP} \times 100 \tag{5}$$

- **Precision:** It represents the number of correct positive predictions divided by the total of numbers that the algorithm predicted as Parkinson correctly and the people who are healthy and the algorithm forecasted them with Parkinson incorrectly Eq. 6. A higher precision value indicates that the model has a lower tendency to label negative instances as positive and is better at accurately identifying positive instances. A low precision value would suggest a higher rate of FP; it indicates that the model may incorrectly classify healthy individuals (class 0) as having Parkinson's disease.

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{6}$$

- **F1-score**: The indicator F1-score is calculated as a weighted average of the sensitivity (recall) and the precision of the model as determined in Eq. 7.

$$F1 - score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \times 100 \tag{7}$$

4 Results and discussion

4.1 Machine Learning algorithms

Each classifier has a nature of decision boundaries and a different performance comparing with the others on the same dataset. For that, we have selected 20 predeveloped ML algorithms (or in other term classifiers) to compare their performance for detecting Parkinson's disease using two different acoustic datasets. These algorithms are presented in Table 4. These classifiers have been selected from Scikit-Learn packages [65]. Concerning the Integrated Development Environment (IDE), we used the Standard Google Collaboratory which is provided by Google as a cloud-based IDE. It allows us to write and execute Python code on external servers that come with the computing resources which are an Intel Xeon(R) CPU @ 2.20GHz with 2 vCPUs (2 virtual CPUs) and 13 GB RAM.

We used the Grid search approach to find the best combination of hyperparameters for these classifiers that provide good performance for detecting Parkinson's disease using dataset I and dataset II. The optimal combination that outcome from this process is illustrated in Table 5.

In fact, each classifier in our study comes with a set of pre-defined initial parameters provided by scikit-learn (version 1.2.2). The search process focused on exploring the hyperparameters of each classifier. It is important to note that any hyperparameter not explicitly mentioned in this work takes the default value available in the scikit-learn library as an optimal value.

Table 4 Machine Learning Algorithms

Algorithm and Reference	Abbreviation
Logistic Regression [66]	**LR**
Ridge Classifier [68]	**RidC**
SGD Classifier [71]	**SGDC**
Passive Aggressive Classifier [74]	**PAC**
K-Neighbors Classifier [77]	**KNN**
Decision Tree Classifier [79]	**DTC**
Extra Tree Classifier [81]	**ExT**
Linear SVC [82]	**LSVC**
SVC [82]	**SVC**
Gaussian NB [65]	**GNB**
AdaBoost Classifier [67]	**ABC**
Bagging Classifier [69, 70]	**BgC**
Random Forest Classifier [72, 73]	**RFC**
Extra Trees Classifier [75, 76]	**ExTsC**
Gaussian Process Classifier [78]	**GPC**
Gradient Boosting Classifier [73, 80]	**GBC**
Linear Discriminant Analysis [65]	**LDA**
Quadratic Discriminant Analysis [65]	**QDA**
XGB Classifier [73],	**XGBC**
MLP Classifier [83]	**MLPC**

4.2 Performance evaluation for 20 ML algorithms

To evaluate the performance of the classifiers and to gain valuable insights into the impact of the number of training and testing samples in small and unbalanced datasets on performance, we used two CV techniques which are Stratified K-Fold CV and LOOCV. The latter validates our ML algorithms more times and giving us very precise estimation due to the strict validation; every sample in the dataset acts as a test set independently, whilst the remaining *n-1* samples use for training in each iteration. On the other hand, Stratified K-Fold CV splits the dataset into k-folds while keeping a symmetrical number of different classes (healthy and PD) in each fold. And in each iteration, one-fold is used for validation, and the other *k-1* Folds collectively are used as a training dataset. The performance of each classifier is the average of these iterations. These strategies have been configured to shuffle the dataset and set k equal to 5 before using them with the classifiers. The performance indicators of these classifiers using Parkinson's dataset I are represented as an average of the accuracy in Fig. 6a, sensitivity (or Recall) in Fig. 6b, the precision in Fig. 6c and the F1-score in Fig. 6d; whilst the map key of the classifiers has been illustrated in Fig. 6e. The x-axis of the figures represents the indicators provided by Stratified K-Fold CV, whilst the y-axis represents the performance that we have got using LOOCV.

Area Under Curve (AUC) is one of the most widely used metrics for evaluation. But it cannot be used directly with LOOCV or Stratified K-Fold CV. Thus, we have shuffled and split the dataset into 4 stratified folds, every fold has proportional samples of people with PD and healthy. Then, we developed a simple algorithm that enables each classifier

Table 5 The optimal hyperparameters of the classifiers that outcome by grid search using dataset I and dataset II

Algorithm	Hyperparameters
Dataset I:	
LR	MaxIter:1000, randomState:1, Solver:'lbfgs'
RidC	Default
SGDC	MaxIter:20000
PAC	C:16, MaxIter:20000
KNN	Neighbors:2, weights: 'distance'
DTC	MinSmplsLeaf&Split:3&9, randState:28, splitter: 'random'
ExT	MaxDepth:5, splitter: 'best'
LSVC	MaxIter: 10000
SVC	C:20, gamma: 'auto', RandomState:1
GNB	varSmoothing: 2.5
ABC	Algorithm: 'SAMME', nEstimators: 74, Lr: 0.7
BgC	nEstimators: 26
RFC	Criterion: 'entropy', maxDepth: 6, nEstimators: 76
ExTsC	Criterion: 'entropy', nEstimators: 103
GPC	Default
GBC	Lr: 0.7, nEstimators: 38
LDA	Default
QDA	Default
XGBC	Lr: 0.7, nEstimators: 9, maxDepth: 3
MLPC	HiddenLayers: (23,17,7), maxIter: 10000, Lr: 0.01
Dataset II:	
LR	C: 0.1 MaxIter:20000, randomState:100
RidC	Alpha: 1.5, fitIntercept: False, randomState: 23
SGDC	Alpha: 1.5, MaxIter: 70000
PAC	C:12, MaxIter:100000
KNN	Neighbors:2, weights: 'distance'
DTC	MinSmplsLeaf:3, randState:10
ExT	criterion: 'entropy', randomState: 31, splitter: 'best'
LSVC	C: 0.1, MaxIter: 100k, randomState: 100
SVC	C:0.04, gamma: 'auto', RandomState:1, krnel='linear'
GNB	varSmoothing: 5.3
ABC	Algorithm: 'SAMME', nEstimators: 197, Lr: 1.0
BgC	nEstimators: 93
RFC	Criterion: 'entropy', maxDepth: 9, nEstimators: 36
ExTsC	Criterion: 'gini', nEstimators: 81
GPC	Default
GBC	Lr: 0.5, maxDepth: 4, nEstimators: 80, Loss:'exp'
LDA	Default
QDA	regParameter: 0.9
XGBC	maxDepth: 4
MLPC	HiddenLayers: (23,17,7), maxIter: 35000, Lr: 0.0001

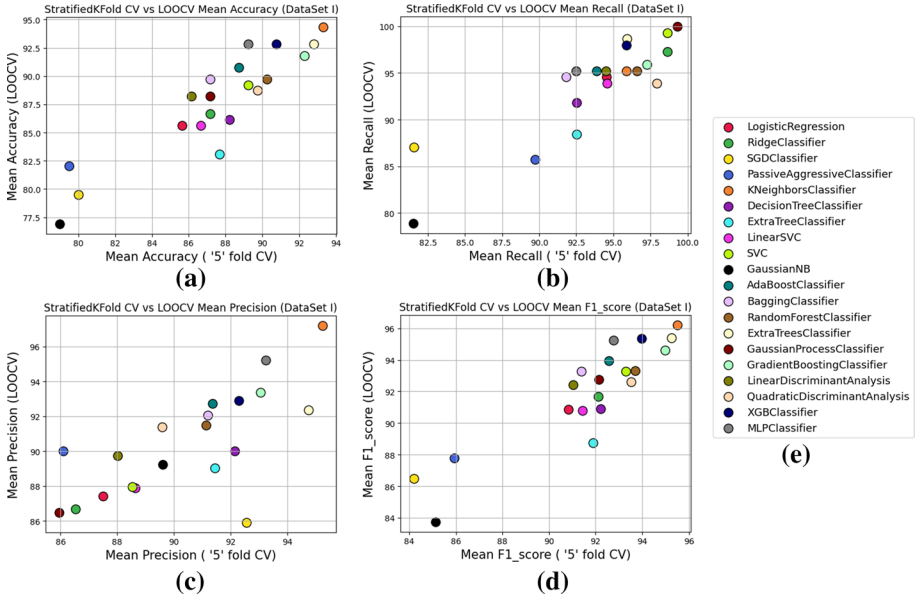

Fig. 6 Representation and comparison of performance evaluation for 20 MLAs including the average of their (**a**): Accuracy, (**b**): Recall, (**c**): Precision and (**d**): F1-score that provided from LOOCV and Stratified k-fold CV with Parkinson's disease dataset I and (**e**): it is the map key

to select randomly one-fold used as validation set from these folds, whilst the remaining folds collectively used for training.

AUC and ROC analysis provide mechanisms or tools to select the most appropriate classifiers and to discard suboptimal ones independently from the cost context or the class distribution. It is related in a direct way to the cost/benefit analysis of diagnostic decision-making. We used it to plot the true positive rate (Recall) against the false positive rate ($1 - specificity$) which is known as the probability of a false alarm as shown in Fig. 7 a, b, c and d.

For evaluating the performance of the classifiers using the second acoustic dataset "II" of Parkinson's disease, we also used the same CV techniques which have been used with the first "dataset I". The two techniques LOOCV and Stratified K-Fold CV have been configured to shuffle the second dataset and split it into 5 splits. Then, we used these splits to evaluate the performance of the twenty classifiers. The performance indicators of these classifiers using Parkinson's dataset II are represented as an average of the accuracy in Fig. 8a, sensitivity or Recall in Fig. 8b, the precision in Fig. 8c and the F1-score in Fig. 8d; whilst the map key of the classifiers has been illustrated in Fig. 8e. The x-axis represents the indicators provided by Stratified K-Fold CV, whilst the y-axis represents the performance that we have got using LOOCV.

We developed a simple algorithm that enables each classifier to choose one split from 5 splits randomly and use it to validate the classifier, whilst the remaining splits are collectively used for training. Taking into consideration shuffling the data, stratified splitting, and ensuring proportional representation of each class (PD, healthy) in each split. Fig. 9a, b, c and d represent the values of AUC and ROC of the classifiers using dataset II. The x-axis represents the false positive rate, whilst the y-axis represents the true positive rate.

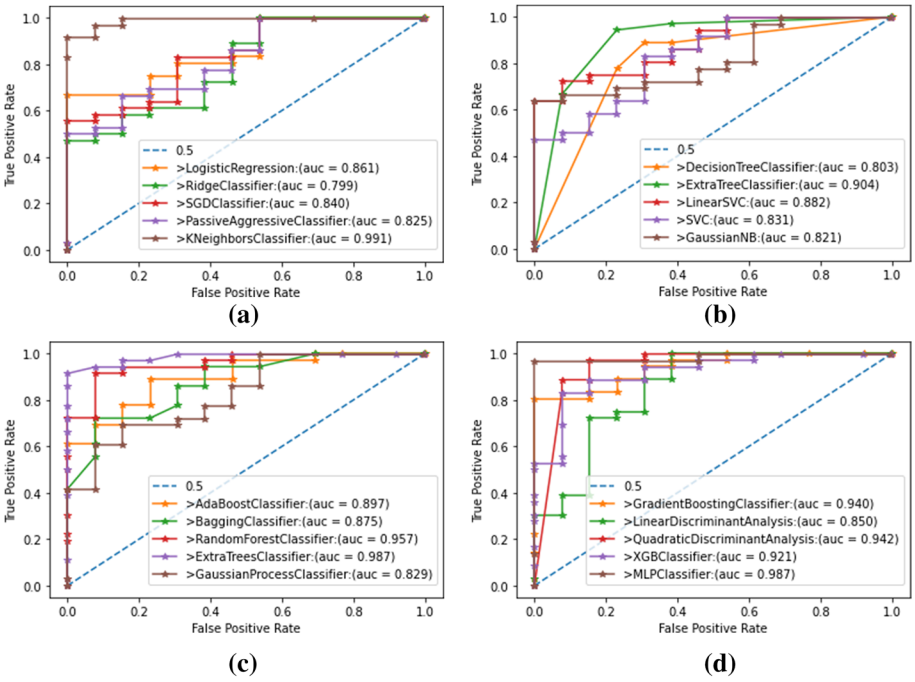

Fig. 7 Representation and comparison of AUC and ROC for 20 ML Algorithms using Parkinson's disease dataset I, each group represent five classifiers (**a**): for LR, RidC, SGDC, PAC and KNN (**b**): for DTC, ExT, LSVC, SVC and GNB (**c**): for ABC, BgC, RFC, ExTsC and GPC (**d**): for GBC, LDA, QDA, XGBC and MLPC

In fact, the train-test split approach has certain limitations specially when the dataset is small because the method is prone to high variance due to the random partition. Moreover, the outcomes can be different for a different test, because some partitions include easy samples to classify, whilst in others partitions or splits include difficult ones. for these reasons, we have used Cross Validation approach to evaluate the performance of our ML algorithms.

As we mentioned, we have used two CV techniques which are LOOCV and Stratified K-fold CV. And to estimate the right performance, we took into consideration a shuffling of the dataset, the stratified, and the number of samples with Parkinson's disease and healthy people in each partition. Thus, Table 6 represents a comparison between the performance of 20 classifiers using the two CVs approaches and the acoustic dataset I.

Also, Table 7 represents a comparison between the performance of 20 classifiers using the two CVs (LOOCV and Stratified K-fold CV) approaches and the second acoustic dataset II.

4.3 Performance evaluation of ensemble voting classifiers

Voting is an ensemble approach that combines the outcomes of various classifiers to decide. Using many classifiers to generate a forecast or decide mitigates the risk of an inaccurate prediction made by a single classifier and makes the estimator more robust. This

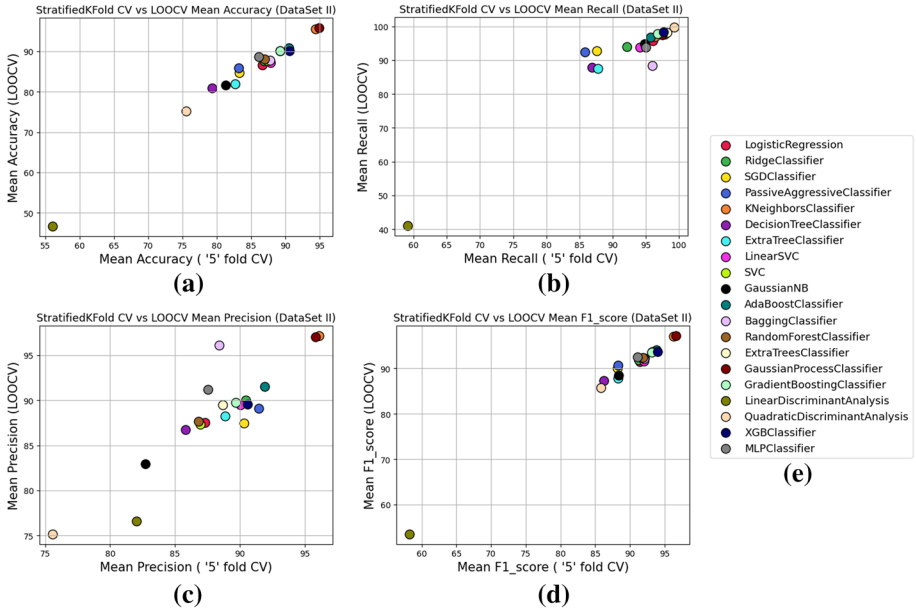

Fig. 8 Representation and comparison of performance evaluation for 20 MLAs including the average of their (**a**): Accuracy, (**b**): Recall, (**c**): precision and (**d**): F1-score that provided from LOOCV and Stratified k-fold CV with Parkinson's disease dataset II and (**e**): it is the map key (legend)

approach can increase the performance of decision-making due to the compensation of the outcomes from classifiers contributed. There are two types of voting which are soft voting and hard voting. Soft voting calculates the total probabilities of each prediction in each model and calculates the average and then chooses the prediction based on the highest probability, whereas Hard voting makes the decision or prediction according to the highest number of votes.

Based on dataset I, we have built an ensemble voting classifier that combines two classifiers which are k-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP) for predicting the people with Parkinson's Disease. The proposed classifier has been tuned on soft voting and the weights of voting are (2.5 and 1) for KNN and MLP respectively. The performance of this classifier can be calculated from the confusion matrix in Fig. 10.

Also, another voting classifier has been built in order to predict the people affected by Parkinson's disease using the acoustic dataset "II". This classifier combines two sub-classifiers which are KNN and XGBC (eXtreme Gradient Boosting), and it has been tuned on soft voting and its weights of voting are (2 and 1) respectively for each sub-classifier. The performance of this classifier can be calculated from the confusion matrix in Fig. 11.

Using Grid Search approach, we repeated the process of hyperparameters' tuning with all classifiers to search for their appropriate (optimal) weights and hyperparameters that control the bias of each sub-classifier and provide good overall performance for detecting Parkinson's disease. Table 8 illustrates the best weighs and hyperparameters that have been chosen for the sub-classifiers that are used for voting and for the ensemble voting classifier.

Moreover, the prediction of Parkinson's disease using the acoustic datasets and ML algorithm was appropriate compared with previous literature due to the strict adjusting and researching about the Optimal hyperparameter, CV, ensemble voting, and some technique

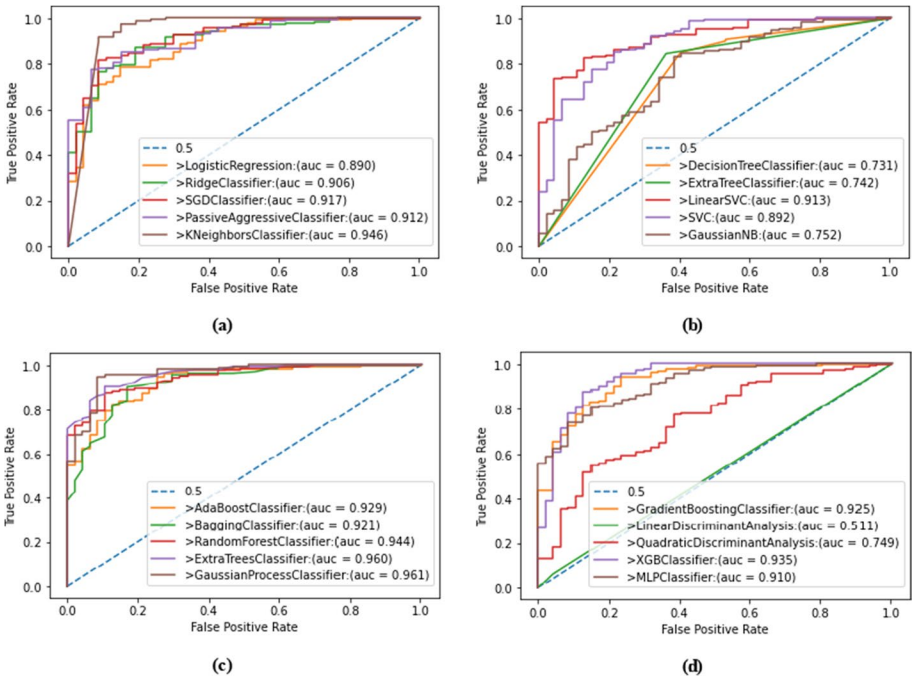

Fig. 9 Representation and comparison of AUC and ROC for 20 ML Algorithms using Parkinson's disease dataset II, each group represent five classifiers (**a**): for LR, RidC, SGDC, PAC and KNN (**b**): for DTC, ExT, LSVC, SVC and GNB (**c**): for ABC, BgC, RFC, ExTsC and GPC (**d**): for GBC, LDA, QDA, XGBC and MLPC

mentioned in this work. Table 9 represents a comparison between the first ensemble voting classifier and the second ensemble voting classifier (1 and 2) used with the dataset I and II, respectively. This comparison includes key indicators of performance which are Accuracy, Sensitivity, Specificity, Precision, F-1 Score and the Matthews correlation coefficient (MCC). For the training and evaluation process, we have used the LOOCV technique to evaluate the performances of the proposed classifiers via cloud-based IDE (Standard Google Collaboratory) with computing resources consisting of an Intel Xeon(R) CPU @ 2.20GHz, equipped with 2 vCPUs (2 virtual CPUs), and 13 GB RAM.

Figure 12 shows an analysis of the performance provided by the voting classifiers after a rigorous search for the best architecture and optimal hyperparameters for each of them. The first voting classifier Consists of two sub-classifiers which are KNN and MLP, it has been used with the acoustic dataset I of Parkinson's disease, whilst the second voting classifier consists of two sub-classifiers which are KNN and XGBC, and it has been trained and evaluated using the second acoustic dataset "II" of Parkinson's disease. In both voting classifiers, the accuracy, sensitivity (recall), and precision are more than 95%.

Some classifiers were good, and some others did not perform well. Moreover, the performance of one classifier was different from one dataset to another. For that we used 20 classifiers, and we compared them using LOOCV, Stratified K-fold CV, and AUC to provide a clear vision of their performances using two different acoustic datasets of Parkinson's disease.

Table 6 Comparison the performance of ML algorithms using dataset I

Model	LOOCV				Stratified k-fold			
	Accuracy	Recall	Precision	F1-score	Accuracy	Recall	precision	F1-score
LR	85.64	94.56	87.42	90.85	85.64	94.55	87.49	90.84
RidC	86.67	97.28	86.67	91.67	87.18	98.62	86.52	92.11
SGDC	79.49	87.07	85.91	86.49	80.00	81.56	92.57	84.19
PAC	82.05	85.71	90.00	87.80	79.49	89.70	86.10	85.94
KNN	94.36	95.24	97.22	96.22	93.33	95.86	95.26	95.52
DTC	86.15	91.84	90.00	90.91	88.21	92.53	92.14	92.22
ExTC	83.08	88.44	89.04	88.74	87.69	92.51	91.44	91.88
LSVC	85.64	93.88	87.90	90.79	86.67	94.57	88.64	91.44
SVC	89.23	99.32	87.95	93.29	89.23	98.64	88.53	93.29
GNB	76.92	78.91	89.23	83.75	78.97	81.52	89.61	85.12
ABC	90.77	95.24	92.72	93.96	88.72	93.86	91.36	92.56
BgC	89.74	94.56	92.05	93.29	87.18	91.82	91.20	91.38
RFC	89.74	95.24	91.50	93.33	90.26	96.57	91.14	93.71
ExTsC	92.82	98.64	92.36	95.39	92.82	95.91	94.75	95.26
GPC	88.21	100.0	86.47	92.74	87.18	99.31	85.94	92.13
GBC	91.79	95.92	93.38	94.63	92.31	97.24	93.06	94.99
LDA	88.21	95.24	89.74	92.41	86.15	94.51	88.02	91.03
QDA	88.72	93.88	91.39	92.62	89.74	97.93	89.59	93.52
XGBC	92.82	97.96	92.90	95.36	90.77	95.89	92.30	93.98
MLPC	92.82	95.24	95.24	95.24	89.23	92.48	93.25	92.76

Table 6 presents a comprehensive performance comparison of the 20 classifiers utilized in this study with dataset I. Additionally, Table 7 provides a detailed analysis of the performance of the same 20 classifiers applied to dataset II. Furthermore, the performance of the Ensemble Voting Classifiers proposed in this work, along with the results of some previous literature, is illustrated in Table 10, offering valuable insights into the effectiveness and reliability of the ML algorithms in the diagnosis of Parkinson's Disease via speech disorders.

Based on the literature, techniques used and the results, we noticed that: **1)** In small and unbalanced datasets, Leave-One-Out Cross-Validation may yield a precise estimation due to each sample in the dataset playing the role of a complete test set. And the omission of a single sample from the training set does not detrimental impact the performance compared to k-fold and train-test split methods. **2)** Stratified sampling and a careful shuffle mechanism may prevent (or mitigate) the bias of the models to correctly forecast one class over the other, particularly in datasets with imbalanced class distribution. **3)** In small datasets, splitting 10% to 20% as a test set may not provide a clear estimation for performance. Using more than 20% as a test set negatively impacts the performance because the training set becomes excessively small and cannot provide the desired diversity that returns a robust model. **4)** While using the train-test split approach, we must take into account all the parameters that return equitable and required splits. **5)** The ML algorithms demonstrate acceptable reliability in diagnosing Parkinson's disease; however, obtaining a large dataset is crucial to enhance the prediction efficiency and strengthen our confidence level.

Table 7 Comparison the performance of ML algorithms using dataset II

Model	LOOCV				Stratified k-fold			
	Accuracy	Recall	Precision	F1-score	Accuracy	Recall	precision	F1-score
LR	86.64	95.74	87.52	91.45	86.64	96.10	87.31	91.47
RidC	87.70	93.97	89.98	91.93	86.90	92.19	90.44	91.27
SGDC	84.66	92.73	87.46	90.02	83.33	87.61	90.31	88.18
PAC	85.85	92.38	89.06	90.69	83.20	85.81	91.46	88.26
KNN	95.63	96.99	97.16	97.07	94.44	96.46	96.12	96.27
DTC	80.95	87.94	86.71	87.32	79.37	86.87	85.80	86.28
ExTC	82.01	87.59	88.21	87.90	82.67	87.77	88.86	88.29
LSVC	87.17	93.79	89.51	91.60	87.83	94.14	90.03	92.01
SVC	87.70	97.70	87.32	92.22	87.44	97.87	86.94	92.08
GNB	81.61	94.86	82.95	88.50	81.35	94.85	82.73	88.36
ABC	90.90	96.80	91.50	94.10	90.50	95.70	91.90	93.80
BgC	87.77	88.42	96.10	92.10	87.76	96.00	88.41	92.04
RFC	88.10	97.87	87.62	92.46	87.04	97.52	86.79	91.82
ExTsC	90.08	98.23	89.50	93.66	89.29	98.23	88.68	93.19
GPC	95.90	97.52	97.00	97.26	94.97	97.52	95.83	96.65
GBC	90.08	97.87	89.76	93.64	89.29	96.81	89.69	93.10
LDA	46.69	41.13	76.57	53.52	56.10	59.11	82.05	58.29
QDA	75.26	99.82	75.17	85.76	75.53	99.29	75.58	85.82
XGBC	90.21	98.40	89.52	93.75	90.61	97.70	90.60	93.97
MLPC	88.62	93.79	91.21	92.48	86.11	95.04	87.54	91.05

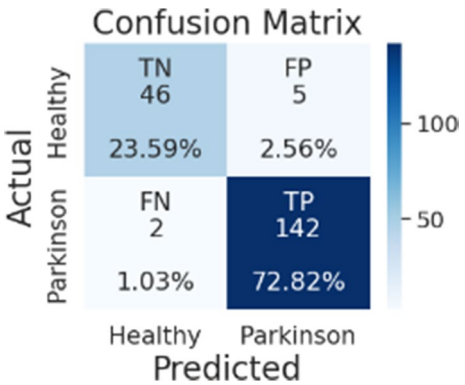

Fig. 10 Confusion matrix of the ensemble voting classifier 1 using Parkinson's disease acoustic dataset I

4.4 Limitations and future directions

Although this study demonstrates a wide range regarding the effectiveness of diagnosing PD via speech disorders using ML algorithms, it is necessary to acknowledge and address some limitations. Limited and unbalanced data of PD is currently an issue and they can lead to biased model training and difficulty in correctly classifying the minority class as shown in the results. On the contrary, large, diverse, high-quality datasets can enhance the

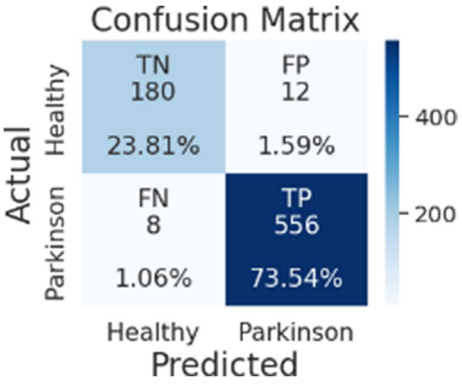

Fig. 11 Confusion matrix of the ensemble voting classifier 2 using Parkinson's disease acoustic dataset II

Table 8 Optimal hyperparameters of the Ensemble Voting Classifiers

Dataset N°	Classifier	Hyperparameters
Dataset I	KNN	K: 4, weights: 'distance', leafSize: 31, p: 1
	MLP	HiddenLayerSizes: (100, 10), maxIter: 700, Lr: 0.01
	VotingClassifier1	Voting: 'soft', weights: (2.5,1)
Dataset II	KNN	K: 2, weights: 'distance', leafSize: 22, p: 1
	XGBC	max_depth: 3
	VotingClassifier2	Voting: 'soft', weights: (2, 1)

Table 9 Performance of the voting classifiers using two datasets of Parkinson's Disease

Dataset N°	Classifier	Accuracy	Sensitivity	Specificity	Precision	F1-score	MCC
I	**VotingClassifier1**	96.41	98.61	90.20	96.60	97.59	90.60
II	**VotingClassifier2**	97.35	98.58	93.75	97.88	98.23	92.98

generalization and performance of ML models to predict the disease, especially rare subtypes and early-stage cases. According to the literature [58], selecting relevant and informative features from a vast range of potential biomarkers is crucial and directly impacts the proposed systems' accuracy and reliability. Choosing the right features that capture the essence of PD while minimizing noise and irrelevant information is a challenge in our future works to build more reliable classifiers based on relevant and informative features.

In fact, PD is a complex disease and cannot be diagnosed via a single symptom such as speech disorders. Thus, we aim to merge together several approaches such as speech disorders [50], handwriting [36], hand tremor [54], spiral-wave sketching [51, 52] and analysing signals from the wearable accelerometer and gyroscope attached to the patient's limbs to build an accurate system for predicting the disease.

Due to the increasing prevalence of PD globally, the development of remote monitoring and early diagnostic systems is crucial for continuous observation based on the Artificial Intelligence of Things (AIoT) and the Internet of Medical Things (IoMT) techniques.

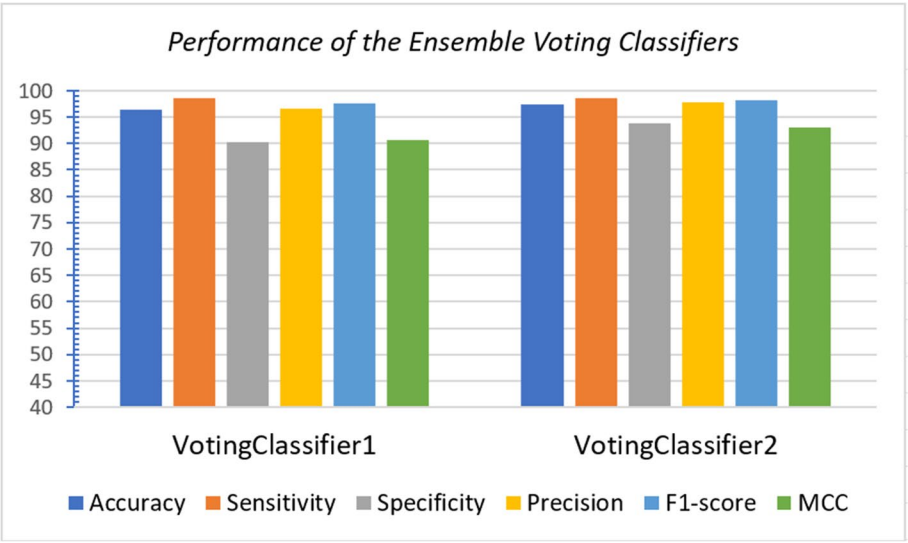

Fig. 12 Comparison between the Performance's key Indicators of the Ensemble Voting Classifiers using two datasets of Parkinson's disease

On the other side, ensuring data security and privacy compliance is equally vital to build trust and encourage broader adoption among healthcare providers and patients, leading to improved outcomes and medical services.

Addressing these limitations requires collaborative efforts between researchers, clinicians and AI experts, along with careful studies design, robust validation, and ongoing refinement of predicting models to improve the accuracy and reliability of diagnostics for many diseases such as Breast Cancer [9], Alzheimer's Disease, Arrhythmia [19], Cervical Cancer [84], [13], Diabetes, Cardiovascular Diseases, Pulmonary Diseases [85], etc.

5 Conclusion

In this study, we focused on predicting people with Parkinson's disease through speech disorders. We used two acoustical datasets provided by UCI Repositor (195 samples with 23 features) and (756 records with 753 features), and we employed twenty classifiers for this mission. Hyperparameters of classifiers play a very important and critical role in obtaining an accurate prediction result because they affect the training. Wherefore, we used the Grid Search to obtain the optimal hyperparameters that enable the classifiers to learn effectively and provide the maximum accurate classification for Parkinson's disease. Stratified K-fold, LOO Cross-Validation, Confusion Matrix, and area under the curve (AUC) have been used to evaluate the performance of these classifiers. To enhance the quality of prediction, we utilized the ensemble voting approach to construct compact classifiers to predict people with Parkinson's disease via speech disorder. The techniques mentioned in this work were very exhaustive, resource-consuming, and expensive during the execution, especially Grid Search and LOOCV. Finally, we achieved an accuracy of 96.41% with the first Ensemble Voting Classifier and 97.35% for the latter.

Table 10 Comparison of the performance with previous literature

Study	Datasets	Algorithm	Accuracy	Sensitivity	Specificity	F1-score
[45]	UCI Repository I 195 records x 23 features	ABO	97.50	97.62	N/A	100.0
		SCFW+KELM	93.82	93.83		93.81
		FCM	91.33	91.34		91.34
		PSO	87.02	87.19		87.13
		ACO	83.17	83.10		83.13
[48]	UCI Repository II 756 records x 753	ANN	79.47	64.51	83.33	56.33
		ANN+PCA	97.35	95.45	97.93	94.38
		RF	89.53	70.17	96.47	77.66
		RF+PCA	76.65	55.55	80.62	43.01
[35]	UCI Repository III 240 records x 45 features	SVM	91.67	83.33	100	90.91
		DA	88.89	83.33	94.44	88.24
		KNN	97.22	100	94.44	97.30
		DT	97.22	94.44	100	97.14
		FR	91.67	83.33	100	90.91
		BT	94.44	88.89	100	94.12
		NB	91.67	83.33	100	90.91
		AB	94.44	88.89	100	94.12
[34]	UCI Repository I 195 records x 23	ANN	82.76	N/A	N/A	N/A
		CNN	93.10			
	UCI Repository III 240 records x 45	ANN	72.22			
		CNN	88.89			
This paper	**UCI Repository I 195 records x 23**	**VotingClassifier1**	**96.41**	**98.61**	**90.20**	**97.59**
	UCI Repository II 756 records x 753	**VotingClassifier2**	**97.35**	**98.58**	**93.75**	**98.23**

Data availability The data used in this paper is available on request to corresponding author.

Declarations

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

References

1. Aouraghe I, Khaissidi G, Mrabti M (2023) A literature review of online handwriting analysis to detect Parkinson's disease at an early stage. Multimed Tools Appl 82:11923–11948. https://doi.org/10.1007/s11042-022-13759-2
2. Kour N, Gupta S, Arora S (2023) Sensor technology with gait as a diagnostic tool for assessment of Parkinson's disease: a survey. Multimed Tools Appl 82:10211–10247. https://doi.org/10.1007/s11042-022-13398-7
3. Montemurro N, Aliaga N, Graff P et al (2022) New targets and new Technologies in the Treatment of Parkinson's disease: a narrative review. IJERPH 19:8799. https://doi.org/10.3390/ijerph19148799
4. Alexander GE (2004) Biology of Parkinson's disease: pathogenesis and pathophysiology of a multisystem neurodegenerative disorder. Dialogues Clin NeuroSci 6:259–280. https://doi.org/10.31887/DCNS.2004.6.3/galexander

5. Poewe W (2006) The natural history of Parkinson's disease. J Neurol 253:vii2–vii6. https://doi.org/10.1007/s00415-006-7002-7
6. Saleh S, Cherradi B, Laghmati S et al (2023) Healthcare embedded system for predicting Parkinson's disease based on AI of things. In: 2023 3rd international conference on innovative research in applied science, engineering and technology (IRASET). IEEE, Mohammedia, Morocco, pp 1–7
7. Saleh S, Cherradi B, El Gannour O, et al (2023) Healthcare monitoring system for automatic database management using mobile application in IoT environment. Bullet EEI 12:1055–1068. 10.11591/eei.v12i2.4282
8. Eskidere Ö, Ertaş F, Hanilçi C (2012) A comparison of regression methods for remote tracking of Parkinson's disease progression. Expert Syst Appl 39:5523–5528. https://doi.org/10.1016/j.eswa.2011.11.067
9. Laghmati S, Cherradi B, Tmiri A et al (2020) Classification of patients with breast cancer using Neighbourhood component analysis and supervised Machine Learning techniques. In: 2020 3rd international conference on advanced communication technologies and networking (CommNet). IEEE, Marrakech, Morocco, pp 1–6
10. Cherradi B, Terrada O, Ouhmida A et al (2021) Computer-aided diagnosis system for early prediction of atherosclerosis using Machine Learning and K-fold cross-validation. In: 2021 international congress of advanced technology and engineering (ICOTEN). IEEE, Taiz, Yemen, pp 1–9
11. Terrada O, Cherradi B, Raihani A, Bouattane O (2020) A novel medical diagnosis support system for predicting patients with atherosclerosis diseases. Inform Med Unlocked 21:100483. https://doi.org/10.1016/j.imu.2020.100483
12. Terrada O, Cherradi B, Raihani A, Bouattane O (2020) Atherosclerosis disease prediction using supervised Machine Learning techniques. In: 2020 1st international conference on innovative research in applied science, engineering and technology (IRASET). IEEE, Meknes, Morocco, pp 1–5
13. Devarajan D, Alex DS, Mahesh TR et al (2022) Cervical cancer diagnosis using intelligent living behavior of artificial jellyfish optimized with artificial neural network. IEEE Access 10:126957–126968. https://doi.org/10.1109/ACCESS.2022.3221451
14. Daanouni O, Cherradi B, Tmiri A (2020) Diabetes diseases prediction using supervised Machine Learning and Neighbourhood components analysis. In: Proceedings of the 3rd international conference on networking. Information Systems & Security. ACM, Marrakech Morocco, pp 1–5
15. Daanouni O, Cherradi B, Tmiri A (2019) Predicting diabetes diseases using mixed data and supervised machine learning algorithms. In: Proceedings of the 4th international conference on Smart City applications. ACM, Casablanca Morocco, pp 1–6
16. Terrada O, Raihani A, Bouattane O, Cherradi B (2018) Fuzzy cardiovascular diagnosis system using clinical data. In: 2018 4th international conference on optimization and applications (ICOA). IEEE, Mohammedia, pp 1–4
17. Terrada O, Cherradi B, Hamida S et al (2020) Prediction of patients with heart disease using artificial neural network and adaptive boosting techniques. In: 2020 3rd international conference on advanced communication technologies and networking (CommNet). IEEE, Marrakech, Morocco, pp 1–6
18. Terrada O, Hamida S, Cherradi B et al (2020) Supervised Machine Learning based medical diagnosis support system for prediction of patients with heart disease. Adv Sci Technol Eng Syst J 5:269–277. https://doi.org/10.25046/aj050533
19. Kumar A, Kumar SA, Dutt V et al (2023) IoT based arrhythmia classification using the enhanced hunt optimization-based deep learning. Expert Syst 40:e13298. https://doi.org/10.1111/exsy.13298
20. Lamrani D, Cherradi B, Gannour OE et al (2022) Brain tumor detection using MRI images and convolutional neural network. IJACSA 13. https://doi.org/10.14569/IJACSA.2022.0130755
21. Moujahid H, Cherradi B, Bahatti L (2020) Convolutional neural networks for multimodal brain MRI images segmentation: a comparative study. In: Hamlich M, Bellatreche L, Mondal A, Ordonez C (eds) Smart applications and data analysis. Springer International Publishing, Cham, pp 329–338
22. Moujahid H, Cherradi B, Gannour OE et al (2020) Convolutional neural network based classification of patients with pneumonia using X-ray lung images. Adv Sci Technol Eng Syst J 5:167–175. https://doi.org/10.25046/aj050522
23. Hamida S, El Gannour O, Cherradi B et al (2021) A novel COVID-19 diagnosis support system using the stacking approach and transfer Learning technique on chest X-ray images. J Healthcare Eng 2021:1–17. https://doi.org/10.1155/2021/9437538
24. Moujahid H, Cherradi B, Al-Sarem M et al (2022) Combining CNN and grad-cam for COVID-19 disease prediction and visual explanation. Intell Auto Soft Comput 32:723–745. https://doi.org/10.32604/iasc.2022.022179

25. El Gannour O, Hamida S, Cherradi B et al (2021) Concatenation of pre-trained convolutional neural networks for enhanced COVID-19 screening using transfer Learning technique. Electronics 11:103. https://doi.org/10.3390/electronics11010103
26. Gannour OE, Hamida S, Saleh S et al (2022) COVID-19 detection on X-ray images using a combining mechanism of pre-trained CNNs. IJACSA 13. https://doi.org/10.14569/IJACSA.2022.0130668
27. Hamida S, Gannour OE, Cherradi B et al (2020) Optimization of Machine Learning algorithms hyper-parameters for improving the prediction of patients infected with COVID-19. In: 2020 IEEE 2nd international conference on electronics, control, optimization and computer science (ICECOCS). IEEE, Kenitra, Morocco, pp 1–6
28. El Gannour O, Hamida S, Cherradi B et al (2020) Performance evaluation of transfer Learning technique for automatic detection of patients with COVID-19 on X-ray images. In: 2020 IEEE 2nd international conference on electronics, control, optimization and computer science (ICECOCS). IEEE, Kenitra, Morocco, pp 1–6
29. El Gannour O, Cherradi B, Hamida S et al (2022) Screening medical face mask for coronavirus prevention using deep Learning and AutoML. In: 2022 2nd international conference on innovative research in applied science, engineering and technology (IRASET). IEEE, Meknes, Morocco, pp 1–7
30. Hasanin T, Kshirsagar PR, Manoharan H et al (2022) Exploration of despair eccentricities based on scale metrics with feature sampling using a deep Learning algorithm. Diagnostics 12:2844. https://doi.org/10.3390/diagnostics12112844
31. Yang Y, Yuan Y, Zhang G et al (2022) Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. Nat Med 28:2207–2215. https://doi.org/10.1038/s41591-022-01932-x
32. Nilashi M, Abumalloh RA, Minaei-Bidgoli B et al (2022) Predicting Parkinson's disease progression: evaluation of ensemble methods in Machine Learning. J Healthcare Eng 2022:1–17. https://doi.org/10.1155/2022/2793361
33. Asmae O, Abdelhadi R, Bouchaib C et al (2020) Parkinson's disease identification using KNN and ANN algorithms based on voice disorder. In: 2020 1st international conference on innovative research in applied science, engineering and technology (IRASET). IEEE, Meknes, Morocco, pp 1–6
34. Ouhmida A, Terrada O, Raihani A et al (2021) Voice-based deep Learning medical diagnosis system for Parkinson's disease prediction. In: 2021 international congress of advanced technology and engineering (ICOTEN). IEEE, Taiz, Yemen, pp 1–5
35. Ouhmida A, Raihani A, Cherradi B, Lamalem Y (2022) Parkinson's disease classification using machine learning algorithms: performance analysis and comparison. In: 2022 2nd international conference on innovative research in applied science, engineering and technology (IRASET). IEEE, Meknes, Morocco, pp 1–6
36. Taleb C, Likforman-Sulem L, Mokbel C, Khachab M (2020) Detection of Parkinson's disease from handwriting using deep learning: a comparative study. Evol Intel. https://doi.org/10.1007/s12065-020-00470-0
37. Jiji GW, Rajesh A, Lakshmi MM (2022) Diagnosis of Parkinson's disease using EEG and fMRI. Multimed Tools Appl. https://doi.org/10.1007/s11042-022-14042-0
38. Arora P, Mishra A, Malhi A (2022) Machine learning Ensemble for the Parkinson's disease using protein sequences. Multimed Tools Appl 81:32215–32242. https://doi.org/10.1007/s11042-022-12960-7
39. Sivaranjini S, Sujatha CM (2020) Deep learning based diagnosis of Parkinson's disease using convolutional neural network. Multimed Tools Appl 79:15467–15479. https://doi.org/10.1007/s11042-019-7469-8
40. Nematzadeh S, Kiani F, Torkamanian-Afshar M, Aydin N (2022) Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: a bioinformatics study on biomedical and biological cases. Comput Biol Chem 97:107619. https://doi.org/10.1016/j.compbiolchem.2021.107619
41. Shahhosseini M, Hu G, Pham H (2022) Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. Mach Learn Appl 7:100251. https://doi.org/10.1016/j.mlwa.2022.100251
42. Senthilkumar B, Zodinpuii D, Pachuau L et al (2022) Ensemble modelling for early breast cancer prediction from diet and lifestyle. IFAC-PapersOnLine 55:429–435. https://doi.org/10.1016/j.ifacol.2022.04.071
43. Tavana P, Akraminia M, Koochari A, Bagherifard A (2023) An efficient ensemble method for detecting spinal curvature type using deep transfer learning and soft voting classifier. Expert Syst Appl 213:119290. https://doi.org/10.1016/j.eswa.2022.119290

44. Kumari S, Kumar D, Mittal M (2021) An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cogn Comput Eng 2:40–46. https://doi.org/10.1016/j.ijcce.2021.01.001
45. Suganya P, Sumathi CP (2015) A novel metaheuristic data mining algorithm for the detection and classification of Parkinson disease. Indian. J Sci Technol 8. https://doi.org/10.17485/ijst/2015/v8i14/72685
46. Chen H-L, Huang C-C, Yu X-G et al (2013) An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert Syst Appl 40:263–271. https://doi.org/10.1016/j.eswa.2012.07.014
47. Shikha S (2020) A comparative analysis for prediction of Parkinson's diseases using classification algorithm. IJRASET 8:349–354. https://doi.org/10.22214/ijraset.2020.7057
48. Gupta I, Sharma V, Kaur S, Singh AK (2022) PCA-RF: an efficient Parkinson's disease prediction model based on random Forest classification. https://doi.org/10.48550/ARXIV.2203.11287
49. García-Ordás MT, Benítez-Andrades JA, Aveleira-Mata J et al (2023) Determining the severity of Parkinson's disease in patients using a multi task neural network. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-14932-x
50. Sonawane B, Sharma P (2021) Speech-based solution to Parkinson's disease management. Multimed Tools Appl 80:29437–29451. https://doi.org/10.1007/s11042-021-11061-1
51. Chakraborty S, Aich S, Jong-Seong-Sim, et al (2020) Parkinson's disease detection from spiral and wave drawings using convolutional neural networks: a multistage classifier approach. In: 2020 22nd international conference on advanced communication technology (ICACT). IEEE, Phoenix Park, PyeongChang,, Korea (South), pp 298–303
52. Zham P, Kumar DK, Dabnichki P et al (2017) Distinguishing different stages of Parkinson's disease using composite index of speed and pen-pressure of sketching a spiral. Front Neurol 8:435. https://doi.org/10.3389/fneur.2017.00435
53. Zhang R, Jia J, Zhang R (2022) EEG analysis of Parkinson's disease using time–frequency analysis and deep learning. Biomed Signal Process Control 78, 103883. https://doi.org/10.1016/j.bspc.2022.103883
54. Fraiwan L, Khnouf R, Mashagbeh AR (2016) Parkinson's disease hand tremor detection system for mobile application. J Med Eng Technol 40:127–134. https://doi.org/10.3109/03091902.2016.1148792
55. Little MA, McSharry PE, Roberts SJ et al (2007) Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed Eng Online 6:23. https://doi.org/10.1186/1475-925X-6-23
56. Little MA, McSharry PE, Hunter EJ et al (2009) Suitability of dysphonia measurements for Telemonitoring of Parkinson's disease. IEEE Trans Biomed Eng 56:1015–1022. https://doi.org/10.1109/TBME.2008.2005954
57. UCI Machine Learning Repository: Parkinsons Data Set. https://archive.ics.uci.edu/ml/datasets/parkinsons.
58. Sakar CO, Serbes G, Gunduz A et al (2019) A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. Appl Soft Comput 74:255–263. https://doi.org/10.1016/j.asoc.2018.10.022
59. Parkinson's Disease (PD) classification. https://www.kaggle.com/datasets/dipayanbiswas/parkinsons-disease-speech-signal-features. Accessed 25 Dec 2022
60. Parkinson's Disease dataset. https://archive.ics.uci.edu/ml/datasets/Parkinson's+Disease+Classification. Accessed 25 Dec 2022
61. Leung PH, Chui KT, Lo K, de Pablos PO (2021) A support vector machine–based voice disorders detection using human voice signal. Artificial Intelligence and Big Data Analytics for Smart Healthcare. Elsevier, In, pp 197–208
62. Berrar D (2019) Cross-validation. Encyclopedia of Bioinformatics and Computational Biology. Elsevier, In, pp 542–545
63. Refaeilzadeh P, Tang L, Liu H (2009) Cross-validation. In: Liu L, Özsu MT (eds) Encyclopedia of database systems. Springer, US, Boston, MA, pp 532–538
64. Rafało M (2022) Cross validation methods: analysis based on diagnostics of thyroid cancer metastasis. ICT Express 8:183–188. https://doi.org/10.1016/j.icte.2021.05.001
65. 1. Supervised learning. In: scikit-learn. https://scikit-learn/stable/supervised_learning.html. Accessed 18 Jul 2023
66. Singh R, Singh R (2023) Applications of sentiment analysis and machine learning techniques in disease outbreak prediction – a review. Mat Today: Proceed 81:1006–1011. https://doi.org/10.1016/j.matpr.2021.04.356
67. Hastie T, Rosset S, Zhu J, Zou H (2009) Multi-class AdaBoost. Statistics and its. Interface 2:349–360. https://doi.org/10.4310/SII.2009.v2.n3.a8

68. van Wieringen WN (2015) Lecture notes on ridge regression. https://doi.org/10.48550/ARXIV.1509.09169.
69. Bühlmann P, Yu B (2002) Analyzing bagging. Ann Stat 30. https://doi.org/10.1214/aos/1031689014
70. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140. https://doi.org/10.1007/BF00058655
71. Du J (2019) The frontier of SGD and its variants in Machine Learning. J Phys Conf Ser 1229:012046. https://doi.org/10.1088/1742-6596/1229/1/012046
72. Parmar A, Katariya R, Patel V (2019) A review on random Forest: an ensemble classifier. In: Hemanth J, Fernando X, Lafata P, Baig Z (eds) International conference on intelligent data communication technologies and internet of things (ICICI) 2018. Springer International Publishing, Cham, pp 758–763
73. Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54:1937–1967. https://doi.org/10.1007/s10462-020-09896-5
74. Crammer K, Dekel O, Keshet J et al (2006) Online passive-aggressive algorithms. J Mach Learn Res 7:551–585
75. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42. https://doi.org/10.1007/s10994-006-6226-1
76. Saeed U, Jan SU, Lee Y-D, Koo I (2021) Fault diagnosis based on extremely randomized trees in wireless sensor networks. Reliability Eng Syst Safety 205:107284. https://doi.org/10.1016/j.ress.2020.107284
77. Müller P (2023) Flexible K nearest neighbors classifier: derivation and application for ion-mobility spectrometry-based indoor localization. https://doi.org/10.48550/ARXIV.2304.10151
78. Morales-Alvarez P, Perez-Suay A, Molina R, Camps-Valls G (2018) Remote sensing image classification with large-scale Gaussian processes. IEEE Trans Geosci Remote Sensing 56:1103–1114. https://doi.org/10.1109/TGRS.2017.2758922
79. Priyanka NA, Kumar D (2020) Decision tree classifier: a detailed survey. IJIDS 12:246. https://doi.org/10.1504/IJIDS.2020.108141
80. Priya SK, Karthikeyan N, Palanikkumar D (2023) Pre screening of cervical cancer through gradient boosting ensemble learning method. Intell autom soft comput 35:2673–2685. https://doi.org/10.32604/iasc.2023.028599
81. sklearn.tree.ExtraTreeClassifier. In: scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.tree.ExtraTreeClassifier.html. Accessed 19 Jul 2023
82. Cunningham P, Cord M, Delany SJ (2008) Supervised Learning. In: Cord M, Cunningham P (eds) Machine Learning techniques for multimedia. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp 21–49
83. Taud H, Mas JF (2018) Multilayer perceptron (MLP). In: Camacho Olmedo MT, Paegelow M, Mas J-F, Escobar F (eds) Geomatic approaches for modeling land change scenarios. Springer International Publishing, Cham, pp 451–455
84. Fekri-Ershad S, Alsaffar MF (2023) Developing a tuned three-layer perceptron fed with trained deep convolutional neural networks for cervical cancer diagnosis. Diagnostics 13:686. https://doi.org/10.3390/diagnostics13040686
85. Khadidos AO, Alshareef AM, Manoharan H et al (2023) Application of improved support vector Machine for pulmonary syndrome exposure with computer vision measures. CBIO 18. https://doi.org/10.2174/1574893618666230206121127