

Early Diagnosis of Parkinson's Disease Through Voice Analysis and Machine Learning

Amrita Naik, Mohd Sahil Khan, Akhilesh Manoj Saraf, Kashyap Chodankar and Umang Dholu

Department of Computer Engineering, Don Bosco College of Engineering, Fatorda, Goa, India

Article history

Received: 19-07-2025

Revised: 16-04-2026

Accepted: 22-09-2026

Corresponding Author:

Amrita Naik

Department of Computer

Engineering, Don Bosco

College of Engineering,

Fatorda, Goa, India

Email:

dramritanaikdbce@gmail.com

Abstract: Parkinson's Disease (PD) is a progressive neurological disorder, and its prevalence has increased considerably in recent years. Early and accurate diagnosis plays a crucial role in effective treatment, but identifying the disease in its early stages remains challenging, especially for non-specialist clinicians. This study presents a predictive analytics framework for detecting Parkinson's disease using voice-based features. To improve the diversity and reliability of the data, datasets from UCI and Kaggle were combined. The data were preprocessed, and important vocal features such as jitter, shimmer, and pitch were extracted using Praat. Several machine learning models, including KNN, SVM, Random Forest, Logistic Regression, and ANN, were applied and evaluated independently. Among these, Logistic Regression showed the best performance, achieving an accuracy of 98.3% and a sensitivity of 97%. The results indicate that combining datasets can improve model performance and reliability. Overall, this study highlights the potential of using multiple datasets and machine learning models for more reliable Parkinson's disease prediction.

Keywords: Machine Learning, Parkinson, Voice Analysis, Logistic Regression

Introduction

Parkinson's Disease (PD) is a degenerative disease that, at present, cannot be reversed and affects millions of people across the world. It most notably affects the central nervous system by damaging the nerve cells that help with movement by lowering dopamine production. This results in motor problems that can consist of tremors, rigidity, slow movement, and unstable posture. PD is the second most common neurodegenerative disorder after dementia and is more prevalent in those above the age of 60. Also, in the last few years, there has been a sharper rise in the number of PD cases. This is paired with a higher incidence rate of morbidity and mortality. Though it is important to manage PD early on, it is a challenge to detect it in its early stages, which makes it even more important to find ways to detect PD early in order to improve the management of Parkinson's disease.

Due to advancements in healthcare and technology, it is now possible to provide services, such as remote diagnostics via telediagnosis and remote monitoring (telemonitoring), which are now effective and less costly for managing PD by avoiding many of the in-person visits that are typically required. Also, of the many non-invasive

methods of diagnostics available, voice analysis is the most common and least expensive. This is mainly because more than 90% of patients with Parkinson's disease have, in one form or another, some voice disorder, and more than 90% of patients have voice problems early in the disease process. A number of studies have been conducted on the use of artificial intelligence processes in the voices of patients with PD. Although it has been possible to achieve some level of satisfaction with some of these methods, many of the studies have been conducted using a single dataset and have used very few models, which has significantly limited the overall robustness and generalizability of the results.

To overcome these limitations, this study presents a framework for predictive analytics incorporating different datasets and assessing different machine learning algorithms for the detection of Parkinson's disease. This framework encompasses data collection, data preprocessing, feature extraction, merging of datasets, training of models, and evaluation of models.

The main contributions of this study comprise:

- (i) Fusion of disparate datasets for enhanced diversity and robustness
- (ii) Extraction and study of distinctive voice characteristics for detection of Parkinson's disease

- (iii) Empirical study of different machine learning algorithms
- (iv) Selection of the most accurate model for prediction

Related Work

Recent developments in Artificial Intelligence (AI) have greatly improved disease diagnosis by allowing different types of data to be combined and different types of machine learning and deep learning models to be used. Dysphonia measures, which are speech signal processing algorithms, were created in a study by Tsanas et al. (2012) to predict the severity of Parkinson's Disease (PD) symptoms from speech signals. Researchers discovered that these algorithms distinguished between PD patients and healthy controls. The study showed the possibility for enhanced dysphonia-based noninvasive diagnostic decision support for PD. In his research, Braga et al. (2019) created a strategy for using free-speech recordings in natural settings to detect Parkinson's Disease (PD) early on using machine learning algorithms and signal/speech processing techniques. The results demonstrated that when appropriately adjusted, the Random Forest (RF) and Support Vector Machine (SVM) approaches were highly effective at detecting Parkinson's disease. RF had an accuracy of 94.77%, and SVM had an accuracy of 90.93%. The use of authentic speech, which more closely resembles real-world situations, is the study's strength. However, the method's reliance on dataset-specific tuning may diminish robustness. The study focuses on the potential of voice analysis as a digital biomarker to help in Parkinson's disease diagnosis.

Solana-Lavalle and Rosas-Romero (2021) conducted research to identify Parkinson's Disease (PD) utilizing voice-based analysis. The approach used four classifiers and featured subset selection on speech recordings from the UCI repository (Kursun et al., 2013). Compared to conventional methods, this methodology enhanced classification performance; however, because it required intricate feature engineering, it was less useful for real-time clinical applications. Sakar et al. (2013) investigated issues with the motor system in a study on telemonitoring and telediagnosis techniques for Parkinson's Disease (PD). Research on voice impairment detection for telediagnosis has been spurred by early-stage PD vocal abnormalities. The study used the TQWT (tunable Q-factor wavelet transform) to extract features from the voice samples of PD patients in contrast to more contemporary feature extraction techniques. The voice recordings of subjects were used to generate many feature subsets, and ensemble learning was applied with a variety of different classifiers. According to the results, KNN had an accuracy of 86%, SVM (linear) had 85%, and SVM (RBF) had 86% for Logistic Regression (RF). The findings demonstrated that TQWT performed better

than previous methods, especially when combined with Mel-frequency cepstral and tunable-Q wavelet coefficients and a filter feature selection strategy, which enhanced PD classification accuracy.

Sakar et al. (2019) conducted research on the use of speech pattern analysis in Parkinson's disease using a telemonitoring voice dataset. Two significant problems were resolved:

- (i) Examining how well different voice sample types, such as sustained vowels versus words, can diagnose Parkinson's disease
- (ii) Evaluating the effectiveness of metrics for central tendency and dispersion in describing all sample recordings from a specific subject. According to the findings, SVM has an accuracy of 85%, and KNN is 82.5% accurate

The method relies on iterative feature selection, which increases computational overhead even if it produced better classification results.

Wrobel (2021) extracted patient speech recordings and used an ensemble classification technique to identify Parkinson's disease. The relevant features were selected using a wrapper method called Sequential Backward Selection (SBS), and the outcomes from the feature reduction set include SVM of 87.5%, k-NN of 90%, and Random Forest (RF) of 87.5%. Tai et al. (2021) conducted a study on the use of speech patterns to recognize Parkinson's Disease (PD) in its early stages. For the study, 36 subjects, including PD patients and controls, were preprocessed, and 70 features were extracted from the recordings. Using supervised learning methods like Multilayer Perceptron (MLP), Random Forest (RF), and Support Vector Machines (SVM), the data were split into patients and controls. SVM scored 88%, which was higher than the other models tested in terms of accuracy. It used smartphone-based data gathering, which is useful for monitoring in the real world; however, the variability and amount of the dataset were constrained.

Sajal et al. (2020) developed a technique that uses smartphone sensors to collect data on vowel phonation and rest tremor from 31 individuals with Parkinson's Disease (PD) and healthy individuals. According to the data, SVM has an accuracy of 87% and KNN of 90.5%. By retraining the model with the most recent data and incorporating patient feedback, the model is continuously improved. Sriram et al. (2015), investigated Parkinson's disease in the UCI repository. The Sieve multigram and Survey graph were used in the statistical analysis of speech data. KStar and NNge (Non-Nested Generalisation exemplar) demonstrated good classification accuracy. The study's generalizability is constrained by the limited dataset of 40 subjects. Tsanas et al. (2012)

showed that dysphonia measures are very good at finding Parkinson's disease, the study depended on a single dataset, constraining its generalizability. Most of the studies above have worked on limited datasets. A study by Pati et al. (2025) shows how using a machine learning algorithm on integrated datasets can make diagnoses more accurate. Even though machine learning has made it much easier to find Parkinson's disease, there are still some problems with the current studies. A lot of methods only use one dataset, which makes the models less reliable and generalizable. Also, the fact that there aren't many different types of datasets and that there aren't many different types of classifiers makes predictions less reliable. So, we need a complete framework that combines different datasets and tests different machine learning models to make detection more accurate and reliable. Additional research is required to study the system's clinical viability and usefulness, as well as to test the system on a wider and more varied dataset.

Materials and Methods

This study involved the collection of data from a variety of sources, including the UCI repository and Kaggle. To prepare the data for analysis, data preprocessing and cleaning techniques were utilized, including the removal of duplicate records, filling in missing values, and handling outliers. Data transformation was also performed to create a consistent view of the data. To select the most relevant subset of data, data reduction techniques such as feature selection and extraction were used. For classification, three types of classification algorithms were utilized, including binary classification, multiclass classification, and multilabel classification. The training phase involved the use of various classification algorithms such as Logistic Regression, K-Nearest Neighbor Algorithm, SVM, and Random Forest. The steps used in training include:

- (i) Data Collection
- (ii) Data Preprocessing
- (iii) Data Cleaning
- (iv) Feature Extraction
- (v) Classification

Data Collection

In the process of developing an ML model, Data Collection is crucial because the amount and quality of the data utilized will determine how successful the model is expected to be. Data can be obtained from already-existing databases or created from scratch.

There are three datasets used from the following sources:

1. Dataset 1- UCI Repository Dataset (Kursun et al., 2013)
2. Dataset 2- Kaggle Dataset (Ukani, 2018)
3. Dataset 3- Kaggle Dataset (Singh, 2021)

Data Preprocessing

The collected data underwent preprocessing and cleaning to improve its quality and make it suitable for data mining tasks. This involved identifying and removing missing, inconsistent, or irrelevant data, as well as combining data from multiple sources and selecting a relevant subset of the data. Data preprocessing plays a crucial role in preparing the data for the categorization task. The steps involved in data preprocessing, such as encoding categorical variables, scaling, or normalizing features, and handling missing data, are essential for ensuring the quality and reliability of the data used in the project. The data must be prepared for the categorization task in this step. The work may involve encoding categorical variables, scaling, or normalizing features, and handling missing data.

Data cleaning is a crucial phase in the data analysis and machine learning processes, commonly referred to as data cleansing or data pretreatment. For the raw data to be accurate, dependable, and fit for analysis or modeling, errors, inconsistencies, and inaccuracies must be found and corrected. Data cleaning is essential since the accuracy and efficiency of any downstream analysis or machine learning model directly depend on the quality of the data.

Feature Extraction

The key features are identified in this phase, which are later given to the classification stage. Since the PD patients exhibit speech anomalies as one of the key features, we extract the vocal features such as jitter local, jitter absolute, jitter rap, jitter ppq5, jitter ddp, shimmer local, shimmer dB, shimmer APQ3, shimmer APQ5, shimmer APQ11, shimmer DDA, NTH, HTN, mean pitch, minimum pitch, maximum pitch. Prat Acoustic software (Boersma et al., 1992) was used to extract the mentioned mean habitual frequency features. Statistical information of each of the features is given in Fig. 1. The description of each of the features is explained below:

- i) Jitter Local: Jitter local refers to the variation from one cycle to the next in the basic period of vocal fold vibration. It measures pitch instability by calculating the average absolute difference between consecutive periods, adjusted by the mean period
- ii) Jitter Absolute: Jitter absolute is the average difference in seconds between two consecutive pitch periods. It is not normalized like local jitter, so it gives a direct measure of how the vibration of the vocal folds changes over time

	count	mean	std	min	max
ID	1039.0	20.518768	11.538633	1.000000	40.000000
Jitter (local)	1039.0	2.680669	1.765515	0.100000	14.376000
Jitter (local, absolute)	1039.0	0.000170	0.000106	0.000006	0.000777
Jitter (rap)	1039.0	1.247387	0.979874	0.062000	8.016000
Jitter (ppq5)	1039.0	1.348961	1.139160	0.081000	13.542000
Jitter (ddp)	1039.0	3.742164	2.939680	0.185000	24.048000
Shimmer (local)	1039.0	12.922803	5.452672	1.185000	41.137000
Shimmer (local, dB)	1039.0	1.195296	0.420075	0.103000	2.721000
Shimmer (apq3)	1039.0	5.700738	3.016411	0.496000	25.820000
Shimmer (apq5)	1039.0	7.986800	4.842089	0.708000	72.860000
Shimmer (apq11)	1039.0	12.220556	6.016810	0.517000	44.764000
Shimmer (dda)	1039.0	17.102254	9.049223	1.488000	77.459000
AC	1039.0	0.845956	0.085730	0.539566	0.997904
NTH	1039.0	0.231487	0.151316	0.002106	0.869277
HTN	1039.0	9.998454	4.293227	0.695000	28.418000
Median pitch	1039.0	163.365240	56.048571	81.460000	468.618000
Mean pitch	1039.0	168.731419	55.996733	82.363000	470.456000
Standard deviation	1039.0	27.564113	35.686430	0.533000	293.877000
Minimum pitch	1039.0	134.530699	47.080115	67.957000	452.083000
Maximum pitch	1039.0	234.921514	121.590903	85.541000	597.974000
Number of pulses	1039.0	109.595891	150.091841	0.000000	1490.000000
Number of periods	1039.0	105.918191	149.479959	0.000000	1489.000000
Mean period	1039.0	0.006547	0.001876	0.002039	0.012070
Standard deviation of period	1039.0	0.000843	0.000723	0.000055	0.006371
Fraction of locally unvoiced frames	1039.0	27.709500	20.967781	0.000000	88.158000
Number of voice breaks	1039.0	1.135707	1.615157	0.000000	12.000000
Degree of voice breaks	1039.0	12.381948	15.164353	0.000000	69.117000
UPDRS	1039.0	12.990375	15.899368	1.000000	55.000000

Fig. 1: Statistical information of feature set

- iii) Jitter RAP (Relative Average Perturbation): Jitter RAP tells you how far apart a period is from the average of its neighboring periods (usually one before and one after), divided by the mean period. It gives a smoothed estimate of how much the pitch changes and is less affected by random changes
- iv) Jitter PPQ5 (Pitch Perturbation Quotient): Jitter PPQ5 figures out the average absolute difference between a pitch period and the average of five nearby periods, then normalizes it by the mean period. It records pitch changes over a longer period of time and gives a more stable measure than RAP
- v) Jitter DDP (Difference of Differences of Periods): Jitter DDP comes from RAP and shows the average absolute difference between the differences in periods. It focuses on quick changes in the vibration of the vocal folds and is often used to find small problems with phonation
- vi) Shimmer local: A measurement of the fluctuation in the amplitude or intensity of the speech signal at the local level is known as shimmer local. It evaluates any inconsistencies or variations in the vibrations of the vocal folds while speech is being produced. Shimmer is a common acoustic indicator for vocal irregularities or disorders and is used to assess voice quality

- vii) Shimmer dB(decibel): Shimmer is an acoustic metric used to examine the fluctuation in the amplitude or intensity of the spoken signal as well as voice quality. Shimmer is the relative change in the voice signal's peak-to-peak amplitude, expressed in decibels (dB). Greater variance in vocal fold vibrations is indicated by higher shimmer dB. values, which can also signify a rougher or more unstable voice
- viii) Shimmer APQ3 (Amplitude Perturbation Quotient 3): Shimmer APQ3 is a particular technique used to assess voice shimmer. By computing the average absolute difference between successive amplitude measurements taken over a given time (usually within a 3-millisecond frame), it quantifies the variation in the amplitude or intensity of the spoken signal. The size of the amplitude disturbances in the speech signal is expressed numerically by Shimmer APQ3 in terms of the voice signal.
- ix) Shimmer APQ5: Shimmer APQ5 (Amplitude Perturbation Quotient) is a particular technique used to assess voice shimmer. By computing the average absolute difference between successive amplitude measurements taken over a given time (usually within a 5-millisecond frame), it measures the variation in the amplitude or intensity of the spoken signal. The amount of amplitude disturbances in the voice signal is quantified by Shimmer APQ5 using a numerical value
- x) Shimmer APQ11 (Amplitude Perturbation Quotient 11): The Shimmer APQ11 method is a specialized way to gauge voice shimmer. By computing the average absolute difference between successive amplitude measurements taken over a given time (usually within an 11-millisecond frame), it quantifies the fluctuation in the amplitude or intensity of the spoken signal. The speech signal amplitude perturbation magnitude is quantified by Shimmer APQ11 as a numerical value
- xi) Shimmer DDA (Difference of difference of Amplitude): Shimmer DDA is a metric used to assess the change in the sound signal's amplitude or intensity. It shows the variation in maximum and minimum amplitudes over a given time frame. Shimmer DDA is a measure of the voice's dynamic range, with higher values suggesting more amplitude change and maybe more instability or roughness in the voice
- xii) NHR (Noise to Harmonics Ratio): NHR measures the amount of noise present in the voice signal relative to its harmonic components. A higher NHR value generally indicates greater vocal noise and irregularity, which may be associated with reduced voice quality or impaired phonation.

- xiii) HNR (Harmonics to Noise Ratio): HNR measures the proportion of periodic harmonic energy in the voice signal compared with noise energy. A higher HNR value generally reflects a clearer and more periodic voice signal, whereas a lower HNR may indicate increased hoarseness, breathiness, or vocal instability
- xiv) Mean pitch: The average pitch value of a speech signal is calculated using the measure of mean pitch in voice analysis. It is calculated by adding up all the pitch values over a certain time frame and dividing the result by the overall number of pitch values. To analyze voice characteristics and fluctuations, mean pitch provides a general approximation of the pitch level in a speaker's voice
- xv) Minimum pitch: The lowest pitch value noticed in a spoken signal is determined using the voice analysis metric minimum pitch. It stands for the lowest frequency at which speech is produced by the vocal folds. To research voice features like vocal range or pitch modulation, the minimum pitch can be used to provide information about the lower limit or range of pitch in a speaker's voice
- xvi) Maximum pitch: The highest pitch value that can be heard in a spoken signal is identified using the maximum pitch measurement in voice analysis. It stands for the highest frequency at which speech is produced by the vocal folds

Data Preprocessing

To guarantee consistency and quality, the gathered data were preprocessed. All features were standardized to a similar scale, and noise and missing values were handled properly. Praat was used to extract acoustic features like pitch, shimmer, and jitter from the MDVR-KCL Dataset. To ensure consistency, 16 common sound features were chosen for each dataset. In order to guarantee a balanced class distribution during training and validation, stratified sampling was used once the datasets were combined based on these shared characteristics.

Classification

A classification algorithm is a machine learning algorithm that is used to categorize or assign predefined labels to input data based on its characteristics or features. The goal of a classification algorithm is to build a model that can learn patterns from labeled training data and then apply that knowledge to predict the class or category of new, unseen instances. The steps followed to evaluate the model are:

- (i) Splitting training/testing data
- (ii) Model training
- (iii) Model evaluation and prediction

Each of these steps is explained below.

Splitting Training/ Testing Data

A training set and a test set are typically created from the labeled training data. The classification model is trained on the training set, and its performance is assessed on the test set after hyperparameter tuning.

Model Training

To discover the underlying patterns or correlations between the input features and the class labels, the classification algorithm is applied to the training data. There are various classification algorithms available, each with its own strengths and weaknesses. Based on the training data, the model modifies its internal parameters to minimize a particular goal function, such as maximizing accuracy or minimizing error. Some of the classification algorithms used in this study are KNN, SVM, random forest, logistic regression, and artificial neural network. To make sure the model was strong, k-fold cross-validation was used. This meant that the dataset was split into k subsets and the model was trained and tested over and over again. The final performance metrics were calculated by averaging the results from all folds. Each of the classification algorithms is explained in detail below.

ANN (Artificial Neural Network)

The form and operation of biological neural networks in the human brain serve as the basis for Artificial Neural Networks (ANNs), also known as neural networks. ANNs are strong machine learning models that can extract complicated correlations and patterns from data. An artificial neural network is made up of nodes or artificial neurons, which are interconnected processing units. An input layer, one or more hidden layers, and an output layer are only a few of the layers that these nodes are arranged into. Each node in the network receives input signals, processes them, and then outputs signals to other nodes in the network. In numerous fields, such as speech and image identification, natural language processing, recommendation systems, and many more jobs requiring pattern detection or prediction, artificial neural networks are frequently utilized. Based on the problem and dataset, neural network design might change, including the number of layers, nodes, and activation functions.

KNN (K-Nearest Neighbor)

Machine learning tasks that require both classification and regression can use the K-Nearest Neighbors (KNN) non-parametric classification technique. KNN is a straightforward and understandable algorithm that bases its predictions on the idea of similarity. The main principle of KNN is to categorize

new data points according to how close they are to existing data points in the training dataset. The algorithm assumes that similar instances will probably have related labels. The number of closest neighbors utilized to create a forecast is indicated by the "K" in KNN.

KNN equation can be written as:

S_x is defined as $S_x \subseteq D$ such that $|S_x|=k$ and $\forall (x', y') \in D$,
 $\text{dist}(x, x') \geq \max(x'', y'') \in S_x \text{dist}(x, x'')$

KNN is a flexible algorithm that can be used in many different sectors; however, it has significant drawbacks. It is sensitive to the distance metric and K choices. In addition, KNN is less comprehensible than some other algorithms because it doesn't explicitly state the connections between features and labels.

LR (Logistic Regression)

Logistic regression's main goal is to represent the link between a group of independent variables (also known as predictors or features) and a dependent variable (the binary result or the target variable). Common binary values for the dependent variable are 0 or 1, true or false, yes, or no, etc. It is a statistical modeling technique that is used to categorize data into distinct groups or forecast the likelihood of a binary outcome. It is a well-liked approach for resolving classification issues in statistics and machine learning.

The logistic regression equation can be written as:

$$P(y=1|x) = 1 / (1 + e^{(-z)})$$

The logistic function, also referred to as the sigmoid function, is used by the logistic regression model to adapt the linear regression model. Any real-valued number can be represented by the sigmoid function as a value between 0 and 1, which can be thought of as the probability of falling into a specific class. Maximum Likelihood Estimation (MLE) is a popular method for estimating the coefficients in logistic regression. Considering the model, the MLE identifies the set of coefficients that maximizes the likelihood of the observed data. The coefficients are modified iteratively by the optimization process until convergence is reached.

RF (Random Forest)

Random Forest is a flexible ensemble learning technique used in machine learning for both classification and regression tasks. To produce more precise and reliable forecasts, it combines the predictions of various separate decision trees.

The main goal of Random Forest is to assemble a group of decision trees into an ensemble. A random portion of the training data is used to train each decision tree, and a random subset of features is considered at each split. The trees become more diverse because of the

randomness, which also helps prevent overfitting. When compared to individual decision trees, random forest seems to be more resistant to overfitting. The collection of various trees aids in lowering variance and enhancing generalization abilities. Random Forest offers a measure of feature relevance that identifies the features that have the greatest influence on prediction.

SVM (Support Vector Machine)

The supervised machine learning technique Support Vector Machines (SVM) is utilized for both classification and regression tasks. SVM can be expanded to address multi-class classification problems; however, it is particularly good at handling binary classification issues.

Finding an ideal hyperplane that divides the data into distinct classes while maximizing the margin between the classes is the basic aim of SVM. A decision boundary that effectively divides the data points depending on their attributes is the hyperplane. SVM has several benefits, including the capacity to handle high-dimensional data, the efficiency in handling both cases of linear and non-linear separability, and the resistance to overfitting. SVM, however, may be sensitive to the regularization parameter and kernel function selection, requiring careful tweaking.

Model Evaluation and Prediction

The overall process of model development is explained and shown in Figure 2. The process of model development starts with collecting data from UCI (Kursun et al., 2013), Kaggle1 (Ukani, 2018), and Kaggle2 (Singh, 2021). The data that has been collected is then prepared in such a way that the missing data are addressed, and the data are normalized so that the datasets can be used consistently.

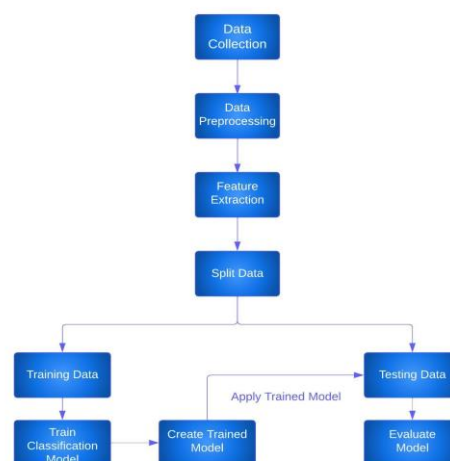


Fig. 2: Predictive analytics framework for Parkinson's disease detection

Using the Praat software, we extract relevant vocal features, including jitter, shimmer, and pitch, specifically for the Kaggle2 dataset (Singh, 2021). These features, when integrated with other datasets, give us the means to enhance diversity and articulation in the dataset's overall robustness. The enhanced dataset is split into training and testing datasets, and various machine learning classifiers will be utilized, including KNN, SVM, Random Forest, Logistic Regression, and ANN. The dataset was split into ten subgroups, and 10-fold cross-validation was used to guarantee robustness and generalization. Nine subsets were used to train the model, and the remaining subset was used to validate it. This procedure was repeated iteratively. The average of all folds was used to get the final performance metrics. We employed grid search to determine each classifier's optimal configuration. To provide fair model evaluation during training and validation, stratified sampling was employed. K-Nearest Neighbors (KNN) identifies local similarities in the dataset. Support Vector Machine (SVM) identifies separation margins between the classes. Random Forest (RF) increases prediction accuracy by aggregating several decision trees. Logistic Regression (LR) offers a classification model based on probability for a binary outcome, and the Artificial Neural Network (ANN) is applied to determine the intricacies within the dataset that are non-linear and difficult to grasp. The proposed framework provides the opportunity for the training and testing of each model for better comparison of the models and for the merged dataset. As a last point of this discussion, we evaluate the models based on their accuracy, sensitivity, specificity, and AUC (Area Under the Curve), as these metrics will provide the best means of ascertaining the most suitable

Results and Discussion

Experimental Setup

We tested our suggested model using the Nvidia GeForce RTX 2070 machine, an i7 processor, and libraries such as Keras 2.2.4 with TensorFlow as the backend. We also used the scikit-learn library for implementing the classification algorithm and the praat library to extract the mean habitual frequency, which is a measure commonly used in phonetics.

Data Description

Experiments were conducted on publicly available datasets, namely UCI (University of California, Irvine), Kaggle1, and Kaggle2. The dataset has a fairly balanced distribution of Parkinson's and healthy samples.

UCI: Twenty Parkinson patients (6 female, 14 male) and twenty healthy persons (10 female, 10 male) who applied to the Department of Neurology at the Cerrahpasa Faculty of Medicine, Istanbul University,

were part of the training data. Twenty-six voice samples of various sounds, including words, numerals, sustained vowels, and brief sentences, were recorded from each subject. Each speech sample is used to extract a set of 26 linear and time-frequency-based characteristics. The UPDRS score, which is established by a board-certified physician, is further included in this dataset.

Kaggle1: The dataset consists of biomedical voice metrics from 31 subjects (23 having a Parkinson's disease diagnosis) from Istanbul University. There are 195 voice recordings (one per row), and the recordings in the dataset are broken down by feature (column). 16 important voice features are extracted from the recordings and included in the dataset. While there are differences in feature naming across datasets, the same voice measures are captured, and this allows for features to be compared and integrated across datasets.

Kaggle2: The MDVR-KCL Dataset captures high quality voice of healthy and Parkinson's patients recorded via smartphone in a simulated call scenario in King's College Hospital, London. The dataset consists of 37 subjects, 16 of which are Parkinson's disease patients and the other 21 are healthy control subjects. The voice recordings also capture 16 critical voice measures (jitter, shimmer, pitch, and noise, etc.), which are captured using the Praat software.

A summary of the datasets used in this study and the resulting merged dataset is presented in Table 1.

Simulation

To minimize overfitting, the dataset was divided into independent training and testing subsets. A total of 6000 samples are used to evaluate the suggested model. 4500 and 1500 samples are used for training and testing, respectively. The dataset is analyzed using 6 classification algorithms, namely KNN, SVM, random forest, logistic regression, and artificial neural network.

These three measures are used to assess the classifier's correctness: Accuracy, sensitivity, specificity, and positive predictive value. An accurate test can be determined by how well it can distinguish between patients and healthy cases. Specificity reveals the tendency for misleading negative results. Sensitivity reveals the possibility of erroneous negatives. It is the proportion of patients with valid diagnoses among all patients with positive test results. Table 2 provides the evaluation metrics for every CNN model. The following are the formulas for each of these measurements for your convenience:

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

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

$$\text{Positive Predictive Value} = \text{TP} / (\text{TP} + \text{FP})$$

Table 1: Summary of datasets used and merged dataset

Dataset	No. of Subjects	PD / Healthy	No. of Recordings	No. of Features	Type of Data
UCI (Kursun et al., 2013)	40	20 PD / 20 Healthy	~1040 (26 per subject)	26	Pre-extracted
Kaggle1 (Ukani, 2018)	31	23 PD / 8 Healthy	195	16(selected)	Pre-extracted
Kaggle2 (Singh, 2021)	37	16 PD / 21 Healthy	74(2 per subject)	16(extracted using Praat)	Raw audio
Merged Dataset	108	59 PD / 49 Healthy	1308	16 (common)	Combined

Table 2: Comparison of machine learning algorithms for Parkinson disease detection

Classification Model	Accuracy	Sensitivity	Specificity	Positive Predictive Value (PPV)
KNN	86.6%	87.1%	85%	86%
SVM	95.2%	90%	100%	100%
RF	87.2%	91.2%	86%	89%
LR	98.3%	97%	100%	100%
ANN	90.1%	88.2%	92.3%	92%

Where TP, FN, FP, and TN are true positive, false negative, false positive, and true negative, respectively, obtained from the confusion matrix table.

All the machine learning algorithms were experimented with on the merged dataset. The evaluation of each of the algorithms is listed in Table 2.

The combined datasets were evaluated using classification algorithms like KNN, SVM, random forest, logistic regression, and artificial neural network. We used 10-fold cross-validation to ensure robustness. We used grid search to find the best-performing configuration for each classifier by optimizing hyperparameters. Every dataset's class distribution was examined. The combined dataset produced a more balanced distribution (59 PD vs. 49 healthy), despite one dataset showing a considerable class imbalance. During training and validation, stratified sampling was used to guarantee equitable model evaluation. Among all the algorithms, logistic regression has shown better accuracy of 98.3% and sensitivity of 97% when compared to all other algorithms, as shown in Figure 3. SVM has also shown better accuracy of 95.2% and sensitivity of 90%. The specificity and positive predictive value of both logistic regression and support vector machine are 100%, indicating that the positive cases are correctly detected.

The entire machine learning algorithm is compared with all three datasets: UCI, Kaggle1, and Kaggle 2 repositories. Logistic Regression and SVM have shown better results with all three datasets, as shown in Figure 3.

Ten subsets of the dataset were created, and the machine learning model was gradually applied to each of these subgroups. A total of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100% of the data were used to train and assess the model. As demonstrated in Figure 4, it was observed that the AUC score of Logistic regression grows with dataset size. AUC score of SVM also grows with dataset size, as seen in Figure 5.

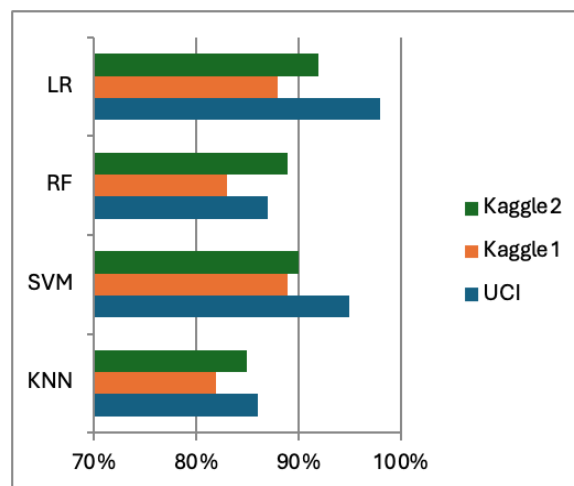


Fig. 3: Accuracy of different classification algorithms with different datasets

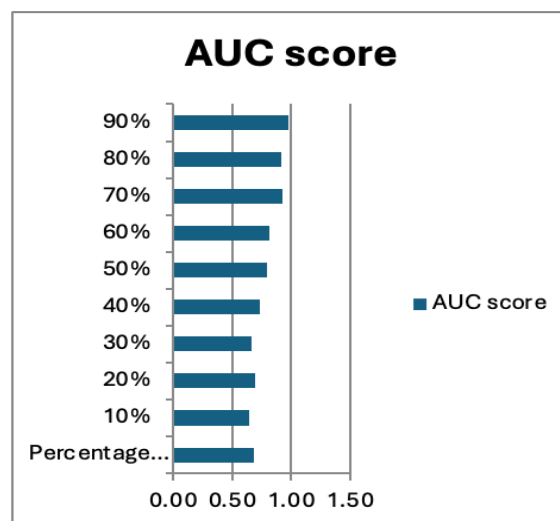


Fig. 4: AUC score of KNN

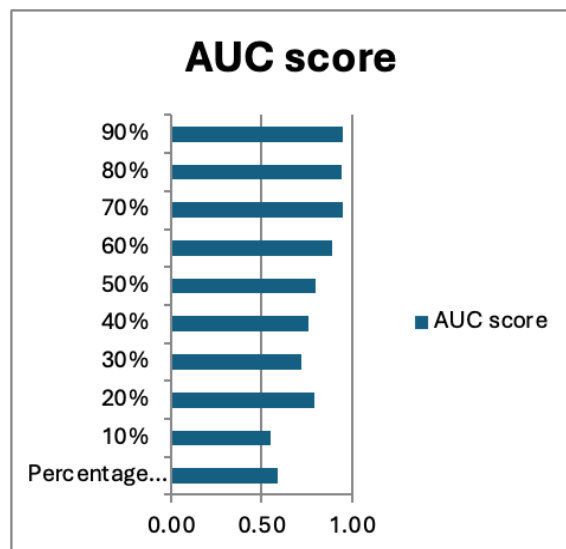


Fig. 5: AUC score of LRC

Table 3: Comparative Analysis of proposed work with existing work

State of the art work	Accuracy	Year
SVM (RBF) (Sakar et al., 2013)	86%	2013
RF (Braga et al., 2019)	94.77%	2019
SVM (Sakar et al., 2019)	85.00%	2019
SVM (Tai et al., 2021)	88%	2021
KNN (Solana-Lavalle and Rosas-Romero, 2021)	95.05%	2021
Recursive feature elimination algorithm with SVM (Aich et al., 2017)	97.37%	2017
Multi-layer Perceptron (Wrobel, 2021)	90.6%	2021
Multiple ANN (Berus et al., 2018)	86.47%	2018
CNN (Murari et al., 2025)	96.85%	2025
XGBoost (Shyamala and Navamani, 2024)	96.61%	2024
Stacked Ensemble (Al-Tam et al., 2024)	96.27%	2024
SVM (Laudis et al., 2024)	93.2%	2024
Proposed work	98%	

Despite achieving 100% specificity and Positive Predictive Value (PPV), the Logistic Regression and Support Vector Machine models may be overfitting, or worse, the models may be trained on unbalanced datasets. The better the data balance, the better the models will be. Although the use of multiple datasets, including a merged dataset, will improve the models' ability to generalize, some influence of data distribution cannot be completely ignored. The AUC scores will always improve as datasets of increasing size are used, and thus the models will be less likely to overfit. In subsequent work, the models will be tested on larger and more diverse datasets to continue to assess their reliability and generalizability.

The Logistic Regression model performed well. This is because the voice features we extracted are well-organized and can be easily separated into groups. Also, by combining datasets, we got a more varied set of data. This helped reduce bias in the model. Made it better at handling new, unseen data. The Logistic Regression model benefited a lot from these voice features and diverse data. Clinical decision-making benefits from the intrinsic interpretability of logistic regression, which permits feature contribution analysis.

The proposed work was also compared with other state-of-the-art work in the literature, as shown in Table 3. It was noticed that the proposed model has shown better results when compared to all the existing work in the literature.

Conclusion

This study introduces a comprehensive machine learning framework for the early detection of Parkinson's disease through voice signal analysis. The suggested method improves model generalization and data diversity by combining data from UCI and Kaggle. Logistic Regression outperformed the other models, achieving an accuracy of 98.3%, a sensitivity of 97%, a specificity of 100%, and a positive predictive value of 100%. The model's high performance is due to the discriminative power of extracted vocal features like jitter, shimmer, and pitch, as well as good preprocessing and dataset fusion methods. A comparison with other methods shows that the proposed method is more accurate and sensitive than many of the best ones currently available. Even though the model achieved high accuracy and specificity, evaluating it on larger and more diverse datasets will help ensure its robustness and reduce the chances of overfitting.

Despite these promising results, the study recognizes specific limitations, including the utilization of restricted and publicly accessible datasets, which may influence real-world applicability. Consequently, subsequent efforts should concentrate on validating the model using larger, diverse datasets and investigating its incorporation into real-time diagnostic systems. Overall, the proposed system shows a lot of promise as a cheap and non-invasive way to find out if someone has Parkinson's disease early on. Future work may encompass the integration of deep learning architectures, including CNN and LSTM, to enhance feature extraction, the integration of multimodal data sources, and the implementation of the system in real-time clinical settings for empirical validation. Future research will concentrate on assessing the suggested model's resilience in real-world scenarios, such as changes in recording environments, background noise, and device variances. Further research will also take into account a variety of demographic parameters,

including age, gender, and linguistic differences, in order to enhance the system's clinical application and generality.

Acknowledgment

The authors have not received any financial support or funding to report.

Funding Information

No funding was received for conducting this study.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Aich, S., Sain, M., Park, J., Choi, K., & Kim, H. (2017). A mixed classification approach for the prediction of Parkinson's disease using nonlinear feature selection technique based on the voice recording. *2017 International Conference on Inventive Computing and Informatics (ICICI)*, 959–962.
- Al-Tam, R. M., Hashim, F. A., Maqsood, S., Abualigah, L., & Alwhaibi, R. M. (2024). Enhancing Parkinson's Disease Diagnosis Through Stacking Ensemble-Based Machine Learning Approach. *IEEE Access*, 12, 79549–79567. <https://doi.org/10.1109/access.2024.3408680>
- Berus, L., Klancnik, S., Brezocnik, M., & Ficko, M. (2018). Classifying Parkinson's Disease Based on Acoustic Measures Using Artificial Neural Networks. *Sensors*, 19(1), 16. <https://doi.org/10.3390/s19010016>
- Boersma, P., Weenink, D., & Shchupak, A. (1992). Praat: doing phonetics by computer. *Institute of Phonetic Sciences*. <https://www.fon.hum.uva.nl/praat/>
- Braga, D., Madureira, A. M., Coelho, L., & Ajith, R. (2019). Automatic detection of Parkinson's disease based on acoustic analysis of speech. *Engineering Applications of Artificial Intelligence*, 77, 148–158. <https://doi.org/10.1016/j.engappai.2018.09.018>
- Kursun, O., Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., & Gurgun, F. (2013). Parkinson's Speech with Multiple Types of Sound Recordings. *UCI Machine Learning Repository*. <https://doi.org/10.24432/C5NC8M>
- Laudis, L. L., Fred, L. A., & Jambek, A. B. (2024). A Nature Inspired Optimization Algorithm for Parkinson's Disease Classification Through Speech Analysis. *Procedia Computer Science*, 235, 840–851. <https://doi.org/10.1016/j.procs.2024.04.080>
- Murari, T. B., Monteiro, R. L. S., Silva, A. do N., Santos, L. B., Fontoura, J. R. de A., & Figueredo, M. B. (2025). An Approach for Parkinson's Disease Detection Based on Artificial Intelligence Techniques. *IEEE Access*, 13, 214761–214771. <https://doi.org/10.1109/access.2025.3626204>
- Pati, A., Addula, S. R., Panigrahi, A., Sahu, B., Nayak, D. S. K., & Dash, M. (2025). Artificial intelligence in improving disease diagnosis: A case study of cardiovascular disease prediction. *Artificial Intelligence in Medicine and Healthcare*, 1–26. <https://doi.org/10.1201/9781003508595-2>
- Sajal, Md. S. R., Ehsan, Md. T., Vaidyanathan, R., Wang, S., Aziz, T., & Mamun, K. A. A. (2020). Telemonitoring Parkinson's disease using machine learning by combining tremor and voice analysis. *Brain Informatics*, 7(1), 12. <https://doi.org/10.1186/s40708-020-00113-1>
- Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgun, F., Delil, S., Apaydin, H., & Kursun, O. (2013). Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4), 828–834. <https://doi.org/10.1109/jbhi.2013.2245674>
- Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., Tutuncu, M., Aydin, T., Isenkul, M. E., & Apaydin, H. (2019). A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Applied Soft Computing*, 74, 255–263. <https://doi.org/10.1016/j.asoc.2018.10.022>
- Shyamala, K., & Navamani, T. M. (2024). Design of an Efficient Prediction Model for Early Parkinson's Disease Diagnosis. *IEEE Access*, 12, 137295–137309. <https://doi.org/10.1109/access.2024.3421302>
- Singh, N. (2021). MDVR-KCL Dataset. *Kaggle*. <https://www.kaggle.com/datasets/nutansingh/mdvr-kcl-dataset/data>
- Solana-Lavalle, G., & Rosas-Romero, R. (2021). Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation. *Biomedical Signal Processing and Control*, 66, 102415. <https://doi.org/10.1016/j.bspc.2021.102415>

- Sriram, T. V. S., Rao, M. V., Narayana, G. V. S., & Kaladhar, D. S. V. G. K. (2015). Diagnosis of Parkinson Disease Using Machine Learning and Data Mining Systems from Voice Dataset. *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*, 151–157. https://doi.org/10.1007/978-3-319-11933-5_17
- Tai, Y. C., Bryan, P. G., Loayza, F., & Peláez, E. (2021). A voice analysis approach for recognizing Parkinson's disease patterns. *IFAC-PapersOnLine*, 54(15), 382–387. <https://doi.org/10.1016/j.ifacol.2021.10.286>
- Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease. *IEEE Transactions on Biomedical Engineering*, 59(5), 1264–1271. <https://doi.org/10.1109/tbme.2012.2183367>
- Ukani, V. (2018). Parkinson's Disease Data Set. *Kaggle*. <https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set/data>
- Wrobel, K. (2021). Diagnosing Parkinson's disease by means of ensemble classification of patients' voice samples. *Procedia Computer Science*, 192, 3905–3914. <https://doi.org/10.1016/j.procs.2021.09.165>