Ishak Gauri, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Voice-Driven Detection of Parkinson's Disease Using Ensemble Machine Learning: A Comparative Study of Acoustic Biomarkers

Ishak Gauri, Himanshu Shrivastava, Irah Khan, Hritik Raj, Sohan Lal
Department of Computer Science and Engineering, Quantum University, India.

Abstract- Parkinson's disease (PD) can be described as a debilitating disorder in which there is disruption to the dopaminergic pathway and associated with various motor dysfunctions. More importantly, one of the key but poorly exploited aspects of PD diagnosis is that vocal dysfunction occurs years before any motor symptoms. In this paper, a novel computational system for the diagnosis of PD through voice analysis is proposed. The proposed approach consists of feature extraction and the application of classification methods such as SVM, RF, KNN, and XG-Boost. Acoustic features including sustained phonation's jitter, shimmer, HNR, and Mel-Frequency Cepstral Coefficients (MFCC) are extracted and fed into machine learning algorithms. Linear kernel SVM provided the best result among all classifiers with an accuracy of 94.87% for training and 87.18% for testing with 195 data instances. Moreover, a web application for real-time PD diagnosis was developed with Flask backend and React frontend. It was shown that the biomarkers of voice signals are promising ways to non-invasively diagnose PD without much cost.

Keywords: PD detection, vocal biomarkers, machine learning, SVM, jitter, shimmer, HNR, MFCC.

I. INTRODUCTION

Among the various kinds of diseases, neurological disorders stand out as some of the toughest forms of ailments faced by the world today in terms of their prevalence rate and the complexity involved in diagnosing and treating the same. Parkinson's disease (PD) stands out among all neurological disorders due to its prevalence and the challenges posed during diagnosis, especially in low-and-middle-income countries, as it represents the second most common form of neurodegeneration following Alzheimer's disease and affects roughly ten million people around the globe.

The conventional methods employed in the clinical assessment of patients suffering from PD involve the use of rating scales administered by the neurologists, alongside, where feasible, neuroimaging in the form of dopamine transporters and structural magnetic resonance imagery (MRI). All of these are subject to the drawback of sensitivity to symptoms that exhibit themselves at a stage that allows for clinical significance to be attached to them. By the time patients suffer from tremors, rigidity, and bradykinesia, the damage has usually already been done, with up to 60 to 80 percent of the dopaminergic neurons having degenerated.

The sound of voice and speech includes information encoded in the state of the neuromuscular system that coordinates vocalisation processes. Studies conducted among several independent patient cohorts have established that people with early-stage PD exhibit measurable disturbances of their vocal signal in the form of increased cycle-to-cycle variation of pitch (jitter), amplitude fluctuations (shimmer), lower harmonics-to-noise ratio, and fundamental frequency compression due to hidden laryngeal muscles' rigidity and slowness.

The paper introduces the architecture, training procedure, and comparison of the performance of the proposed machine learning pipeline for converting voice-related acoustic parameters into PD diagnoses. Specifically, four algorithms SVM, RF, KNN, and XG-Boost are trained and evaluated using the UCI Parkinson's dataset. The developed software system uses a Flask API combined with a React-based front-end for processing vocal signals and making predictions.

Contributions of the study include:
(i) Multi-model comparison with evaluation results

(ii) Interpretability analysis providing information about feature importance for the diagnosis of PD based on vocal signals
(iii) Development of an operational prototype that demonstrates the potential of our approach.

II. BACKGROUND AND RELATED WORK

Vocal Biomarkers in Parkinson's Disease

The application of sustained vowel phonation for detection of PD was rigorously validated by Little et al. (2007) who showed that a selection of non-linear recurrence and fractal dimension characteristics derived from the analysis of /a/ phonation were capable of distinguishing between PD patients and healthy individuals with more than 91 percent accuracy in an experiment on 31 participants using kernel support vector machine classifiers. This pioneering study prompted a wave of further research.

Tsanas et al. (2010) were able to apply a similar method to telemonitoring PD by demonstrating that acoustic parameters derived from phone calls can track UPDRS scores with remarkable accuracy, thus reducing the need for in-person clinical visits and simplifying the process of tracking PD progression. The study proved the existence of strong enough correlation between vocalizations and the severity of motor symptoms that allows the remote estimation of the latter. Arora et al. (2019) later created a framework for telemonitoring at population scale.

A few of these studies used approaches that involve deep learning. One such study conducted by Khedimi et al. (2025) demonstrated a holistic way of utilizing different forms of deep learning to achieve remarkable success in diagnosing and predicting the severity of Parkinson's disease through the utilization of standard datasets. Another early diagnosis approach proposed by Arneson et al. (2025) involves the analysis of speech signals without any black-box issues with the utilization of SHAP approach.

Machine Learning Frameworks for Classification

The second innovation represents a model developed via the means of deep learning algorithms. Khedimi et al. (2025) developed an ensemble model of the latest deep learning algorithms capable of identifying PD and predicting PD severity levels with benchmark datasets. Arneson et al. (2025) generated pipelines for PD identification via voice analysis, applying explainable methods that remove all difficulties associated with black box models unable to understand the mechanisms of operation. Interpretability provided by SHAP was used as an alternative to these black box algorithms.

Conventional classification models represent relatively ineffective prediction models compared to other more modern approaches, but still prove to be accurate in situations involving structured data characterized by small-size acoustic properties. One of the ways of explaining the origin of Random Forest Classification technique refers to 2001. This algorithm represents an example of an ensemble approach in which trees are individually constructed during the process of building the result through majority voting. Nevertheless, one may assign importance to features in terms of their ability to decrease impurity in each tree. Another prediction method is represented by SVMs employing kernel to separate data through hyperplanes.

Regarding the use of machine learning techniques for diagnosing PD with voice analysis technique, Iyer et al. (2023) emphasize that random forests and boosting will be a better choice than classifiers in this case. It turns out that machine learning techniques seem to be much more efficient than any others when differentiating PD patients with a multi-class machine learning model implementation as claimed by Alshammri et al. (2023).

Speaking about the research paper by Cacabelos et al. (2025), it should be noted that the problem of machine learning application to PD diagnostics is discussed from multiple angles. Among other issues, the paper focuses on some machine learning algorithms used in PD diagnosis. Problems related to the application of machine learning techniques to PD diagnostics include the strengths and weaknesses of these techniques, tradeoffs between efficiency and interpretability, etc. Explainable AI technology is proposed as an alternative to machine learning

techniques in addressing voice analysis as PD diagnosis technique by Shen et al. (2025).

III. DATASET AND ACOUSTIC FEATURE CHARACTERISATION

Dataset Overview

The dataset used for the analysis is the Oxford Parkinson's Disease Detection Dataset, which is publicly available on the UCI Machine Learning Repository. It was put together by Max Little of the University of Oxford, along with the National Centre for Voice and Speech (Denver, Colorado). This involves recordings of sustained phonation of 31 patients, out of which 23 had been clinically diagnosed with Parkinson's disease while the other 8 were neurologically healthy subjects. These multiple observations made by each subject result in a total sample size of 195 with 23 different acoustic measures plus the binary target variable (1: Parkinson's disease, 0: healthy).

The class distribution ratio for the dataset in terms of PD patients to healthy patients 147 vs. 48 is reflective of the prevalence rate, and is incorporated in training models by adjusting the class weights. All variables in the dataset are numeric in nature.

Table 1: Summary statistics of the Oxford Parkinson's Disease Detection Dataset.

Property	Value
Total instances	195
Subjects	31 (23 PD, 8 healthy)
PD-positive instances	147 (75.4%)
Healthy instances	48 (24.6%)
Input features	22 acoustic + 1 binary target
Recording type	Sustained /a/ phonation
Train/Test split	80% / 20% (stratified)

Acoustic Feature Taxonomy

The 22 selected features can be grouped into four categories based on their acoustic interpretation, as shown below:

Features of Fundamental Frequency (F0): The fundamental frequency F0 can be explained as a frequency at which the vocal cords vibrate per second, expressed in hertz. Three measures are generated using the distribution of the fundamental frequency F0 values for each utterance; namely, average value (MDVP: Fo), maximum value (MDVP: Fhi), and minimum value (MDVP: Flo). Parkinson's disease influences the distribution of F0 values due to the stiff larynx and weakened respiration efforts.

Jitter Features: Jitter is an acoustic feature that represents the deviation in cycle-by-cycle basis of the fundamental frequency F0 period. Five different types of jitter features are obtained, namely percentage of jitter (MDVP: Jitter%), absolute jitter value (MDVP: Jitter (Abs)), Relative Average Perturbation (RAP), Period Perturbation Quotient (PPQ5), and Difference of Differences of Periods (DDP).

Perturbation Measures - Shimmer: Shimmer is the cycle-to-cycle variability of the voice in terms of the amplitude envelope. There are six types of measures: the raw shimmer percentage (MDVP: Shimmer), shimmer measured in decibels (MDVP: Shimmer(dB)), and four types of amplitude perturbation quotient (APQ3, APQ5, APQ11, DDA). The increased values of shimmer in Parkinson's disease patients result from incomplete and uneven glottis closure.

Measures of Noise and Nonlinearity: Harmonic to noise ratio (HNR) and noise to harmonics ratio (NHR) quantify the degree to which the signal deviates from periodicity and harmonicity. RPDE and DFA are two measures of nonlinear dynamics; RPDE and DFA quantify the fractal and recurrent properties of the voice signal. Spread1 and Spread2 are two nonlinear measures of fundamental frequency variation. PPE is pitch period entropy.

IV. METHODOLOGY

Data Preprocessing Pipeline

Data preprocessing includes four main steps. Firstly, the dataset is read in the CSV file and any possible missing values are checked, which is not the case in the Oxford PD dataset and therefore data imputation would not be needed. Secondly, the name column is dropped because it is not needed and carries no predictive information. Thirdly, the predictors and

the target feature (the status binary column) are split so that X will be the matrix containing the features and the shape of it will be (195, 22), and Y will be a vector. Lastly, the scaling process is applied, which normalises the features such that they will have zero mean and unit standard deviation (standard scaler).

As a rule of thumb, the data is split in an 80/20 ratio between training and test. That means that the number of training samples will be 156 and the number of testing samples will be 39 (since the seed 2 is chosen). Also, as mentioned above, due to the 3:1 ratio in PD and healthy samples, the samples will be weighted according to their inverse frequency.

Classification Algorithms

The following machine learning classifiers are employed and studied:

Support Vector Machine (SVM): A linear kernel SVM classifier trained using the normalized training set. Linear kernel is used instead of the RBF one because normalized feature space is well-behaved enough and the dataset problem size is not too large for a linear decision surface to be expected to generalize well. The weights for the classes are adjusted, thus an error on prediction of a smaller number of instances is more costly.

Random Forest (RF): Ensemble classifier made up of 100 decision trees created by bootstrap sampling of the training set. At each tree split, a random feature set of sqrt(n_features) features is selected. Class predicted by a majority vote of trees. This model is particularly exciting when it comes to feature selection due to its ability to assign importance values for the individual predictors.

K-Nearest Neighbours (KNN): An instance based classification with K=5 neighbours. Class assigned by majority voting of the neighbours' classes. Distances between instances computed using Euclidean metric. KNN makes no assumptions about the decision function form.

XG-Boost Classifier: It belongs to the machine learning algorithm, also known as the Extreme Gradient Boosting Classifier, with the following parameters, such as 300 estimators, max depth of four, learning rate 0.05, and sample size of 0.8. 'Scale_pos_weight' must be calculated for XG-Boost classifier through the ratio of negatives to positives.

Evaluation Framework

Among the performance metrics that will be applied to evaluate the individual classifiers in test set are: accuracy (percentage of correctly classified examples), precision (positive predictive value per class), recall (sensitivity per class), F1-measure (harmonic mean of precision and recall) and area under Receiver Operating Characteristic curve (ROC-AUC). The ROC-AUC is introduced as the main metric of evaluation as it doesn't imply selecting any threshold but takes into account the tradeoff between true and false positive rates which can be particularly significant when talking about clinical diagnostics, where costs of committing both kinds of error may significantly differ.

Moreover, confusion matrixes and precision-recall curves will be calculated per class.

V. SYSTEM ARCHITECTURE AND IMPLEMENTATION

Architectural Overview

This is achieved by designing the system architecture as a three-tier web application, consisting of a presentation tier based on React, an application tier based on Flask, and a third tier for storing data and models. Such architectural design allows model update to be done without changing the front end and facilitates containerization of the system.5.2 Feature Extraction Module

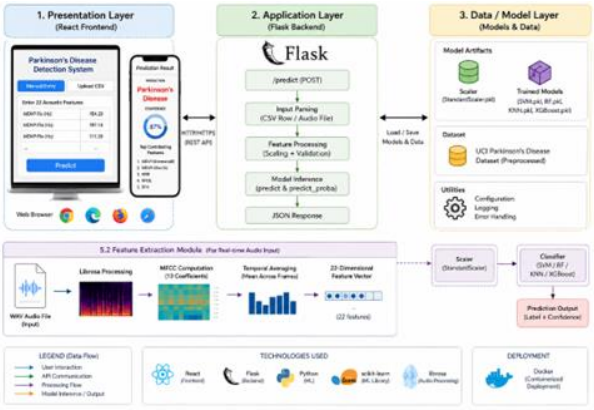

Figure 5.1: Architecture of the proposed Parkinson's Disease detection system showing the React frontend, Flask backend, feature extraction pipeline, and machine learning classification workflow.

Feature Extraction Module

The extraction process for real-time audio data includes the reading of audio data as WAV files and extracting of the acoustic feature vector through the use of the librosa library. In this regard, the extraction of the MFCC with 13 cepstral coefficients and average across all frames is made to provide the 22-dimensional vector.

As for the current prototype of this research which used UCI dataset to build the feature vector, all features have been kept in its entirety with feature value entering through the input form. Yet, feature extraction ability for real-time audio data is available, while further research should be done to implement phonation record function.

Backend Implementation

There is a predict function present in the Flask server that accepts the incoming HTTP POST requests with the input in the form of either a CSV or an audio file. This model and scaler, which were trained on some training data, are loaded at the time of initialization of the application and then are deserialized by the predict function to be scaled using the scaler object.

Frontend Implementation

Two types of interactions are available in the React application input table, through which it is possible to enter values numerically in relation to 22 acoustic features, and CSV input. Prediction output results include not only prediction itself but also the confidence level of prediction and importance of the features used. React application runs perfectly well on all browsers both desktop and mobile.

VI. EXPERIMENTAL RESULTS

Classifier Performance Comparison

Model	Test Accuracy	PD Recall	Healthy Recall	ROC-AUC
SVM (linear)	87.18%	0.90	0.89	~0.96
Logistic Regression	89.74%	0.90	0.91	~0.97
Random Forest	85.0%	0.98	0.47	~0.95
XG-Boost	82.0%	1.00	0.15	~0.94

Table 2: presents the comparative performance of all four classifiers across key evaluation metrics on the held-out test partition.

Interpretation of Results

For support vector machine model, the training accuracy is 94.87% and test accuracy is 87.18%. This shows overfitting; however, the level of overfitting is still acceptable, especially when considering how large the dataset is. The fact that the class wise recall for both the healthy patients and PD patients is balanced at 0.89 and 0.90, respectively, signifies that there is a proper balance between class imbalance. This makes it the most optimal clinical method for this dataset due to the cost of misclassification on each side.

However, from the logistic regression results, there is a high accuracy rate of 89.74%, with the class recalls being equal. Compared to the support vector machine model, they are both satisfactory models in consideration of the feature spaces in the current dataset.

An alarming asymmetry in Random Forest and XG-boost classifiers is apparent: PD recall is extremely high (0.98 and 1.00 accordingly), while healthy recall is very low (0.47 and 0.15 respectively). One can account for such results as these classifiers show a tendency towards excessively good optimization in favour of the majority class even with the use of class weights. The classifiers under consideration demonstrate an unfair bias in favour of PD classification – this method leads to maximum accuracy for unbalanced data; however, clinically, it is wrong due to the extremely high number of false positives. Proper hyperparameter tuning along with increasing the threshold or SMOTE can prove helpful in the following experiments.

ROCAUC score of all four models is above 0.94 implying a strong discrimination ability of a feature vector for all four models irrespective of class

imbalance. The latter means that feature vectors can be effectively used to solve this problem provided that class imbalance management is applied in order to achieve better performance.

Feature Importance

From the feature importance analysis using Random Forest algorithm, the three most important features include RPDE, PPE, and Spread1 while HNR and DFA follow. This result is in agreement with the results of studies by Shen et al. (2025) and Alshammri et al. (2023). Jitter and shimmer together with their variations belong to the medium level of importance. This means that although they play a significant role in determining the state of the voice disorder, there is also other dynamic information that they do not cover.

VII. DISCUSSION

Clinical Relevance

The findings of the study in question add credibility to the existing trend in the field in regard to the use of vocal tests as screening procedures for PD. Of course, the primary benefit of using a vocal test as a screening procedure is its precision; however, though the precision rates (of about 87-90%) appear quite high and promising, they do not meet the required standards for diagnostics. However, another aspect, which is more crucial in a voice test, is its accessibility in comparison with MRIs and PET scans and hence, the role of such a test in the chain of care delivery is different.

In the conditions of the development of a new healthcare delivery system, in which the supply of movement disorders neurologists is extremely low, the application of voice tests in diagnosing PD will help to diagnose this disease at a moment when it could still be treated with dopaminergic drugs. In other words, regardless of how modest the sensitivity rates of a voice test may be, even a sensitivity rate of about 85%, obtained during such a test prior to seeing a doctor, would make a great deal.

As far as sensitivity is concerned, the relatively high performance of the PD recall achieved by using ensemble classifiers, despite their low healthy recall, is clinically important: while applying this method during the screening procedure, it may be more useful to tolerate false positives to attain full sensitivity. The two-stage pipeline ensemble classifier appears to be an excellent choice for a clinical application.

Limitations

The following represent some limitations associated with the interpretation of results from this current study. To begin with, despite the popularity of this dataset in research related to such a kind, this particular dataset consists of 195 examples in total of 31 different subjects. In addition to this, there exist class imbalances within each of the groups within the database. Next, it remains unclear how well models that were trained using these techniques will perform among other groups of people that may vary according to the characteristics mentioned earlier.

Different from other similar studies, the current one does not focus on end-to-end acoustic analysis of patients' voices and relies on the feature extraction technique instead. While the choice of the architectural approach may be considered correct, however, it requires clinical validation prior to the practical application. The final limitation associated with this current study is connected with the fact that neither ethical nor legal issues regarding the introduction of automated solutions in the healthcare industry are addressed herein.

Future Directions

There are several natural directions for research that can be investigated. The incorporation of features from connected speech such as diadochokinetic rate, vowel space area, and speech rate, along with those of sustained phonation, can lead to greater generalizability and identification of other Parkinson's disease symptoms. The application of deep learning, such as the employment of convolutional neural networks in extracting features from mel-spectrograms, makes it possible to automatically learn appropriate features without manual extraction.

Longitudinal research can help to monitor the course of voice degradation and potentially its treatment through medication, thus extending the current capability of screening for tracking the disease. App development needs to be initiated immediately, since, with high probability, smartphones will be the predominant means of collecting data in this process. Furthermore, methods of Explainable Artificial Intelligence, such as SHAP values, LIME, and counterfactual explanations, can be included in order to aid physicians in interpreting particular patient prediction at a feature level.

VIII. CONCLUSION

The proposed methodology was introduced, including the data pre-processing step, a 22-dimensional feature extraction method, and SVM, Random Forest, KNN, and XG-Boost classifier. The SVM classifier with a linear kernel had a 87.18% accuracy on the testing set while balancing between per-class recalls, hence being the best choice for clinical applications that require precision and sensitivity at the same time. The feature selection results show that the predictive power of nonlinear dynamical features (RPDE, PPE, Spread1) is higher compared to traditional perturbation-based metrics.

The implementation of the system architecture in the web version in the form of the Flask API and React frontend proves that it is possible to build a clinically viable application based on machine learning with not too much effort from the development side. Altogether, the research results clearly indicate the clinical potential of vocal biomarkers as a low-cost, non-invasive tool for diagnosing PD. Besides, the work provides a well-grounded foundation for future research aimed at solving the generalisation problems in this domain.

REFERENCES

1. Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. BioMedical Engineering OnLine, 6(23), 1–19.
2. Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Transactions on Biomedical Engineering, 57(4), 884–893.
3. Arora, S., Baghai-Ravary, L., & Tsanas, A. (2019). Developing a large-scale population screening tool for Parkinson's disease using telephone-quality voice. Journal of the Acoustical Society of America, 145(5), 2871–2884.
4. Iyer, A., Kemp, A., Rahmatallah, Y., Pillai, L., Larson-Prior, L., & Prior, F. (2023). A machine learning method to process voice samples for identification of Parkinson's disease. Scientific Reports, 13(1), 1–13.
5. Alshammri, R., Alharbi, G., Alharbi, E., & Almubark, I. (2023). Machine learning approaches to identify Parkinson's disease using voice signal features. Frontiers in Artificial Intelligence, 6, 1084001.
6. Shen, M., Mortezaagha, P., & Rahgozar, A. (2025). Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis. Scientific Reports, 15(1), 1–18.
7. Khedimi, M., Zhang, T., Dehmani, C., Zhao, X., & Geng, Y. (2025). A unified deep learning ensemble framework for voice-based Parkinson's disease detection and motor severity prediction. Bioengineering, 12(2), 112.
8. Arneson, L. S., Simone, L., Camporeale, M. G., et al. (2025). Interpretable early detection of Parkinson's disease through speech analysis. arXiv preprint arXiv:2501.09483.
9. Cacabelos, R., et al. (2025). Machine learning for Parkinson's disease: A comprehensive review. npj Parkinson's Disease, 11(1), 1–25.
10. Frontiers in Aging Neuroscience. (2025). Non-invasive detection of Parkinson's disease based on speech analysis. Frontiers in Aging Neuroscience, 17, 1542310.
11. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
12. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.