Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 09 March 2026
Revised 12 April 2026
Accepted 25 May 2026

Natural Resources for Human Health 2026, Vol. 6 (8s): 708–717

KEYWORDS:
Parkinson's disease, Speech-based detection, pathological speech processing, Acoustic features, Subject-independent validation, Nonlinear speech features, Machine learning.

Subject-Independent Parkinson's Disease Detection Using Acoustic and Nonlinear Speech Features: A Robust Evaluation of Machine-Learning Classifiers

Ashwini D. Bhople[1], Avinash Kapse[2], Pravin A. Kharat[3]

[1]Assistant Professor, Department of Computer Science and Engineering, Padmashri Dr. V. B. Kolte College of Engineering, Malkapur, India (Research Scholar, Department of Computer Science and Engineering, Anuradha College of Engineering and Technology, Chikhli, India), ashwinibhople@gmail.com
[2]Professor, Department of Computer Science and Engineering, Anuradha College of Engineering and Technology, Chikhli, India, askapse@gmail.com
[3]Professor, Department of Computer Science and Engineering, Padmashri Dr. V. B. Kolte College of Engineering, Malkapur, India, pravinakharat82@gmail.com

ABSTRACT: Early detection of Parkinson's disease can facilitate timely intervention. Speech is a noninvasive and cost-effective biomarker for Parkinson's disease detection. In this study, a subject-independent speech-based pipeline was developed using acoustic and nonlinear speech features. To avoid subject-level information leakage, a strict fivefold subject-independent cross-validation procedure was performed. Eight machine-learning classifiers were evaluated using five speech feature representations: acoustic, spectral, nonlinear, hybrid, and standardized minimal acoustic features. Among these, the combination of standardized minimal acoustic features and logistic regression achieved the highest accuracy of 86.49%, with a sensitivity of 75.00% and a specificity of 95.24%. The hybrid feature representation achieved an accuracy of 83.78% and the highest area under the receiver operating characteristic curve (AUROC) of 0.8958. The nonlinear feature sets combined with tree-based and instance-based classifiers achieved accuracies of up to 81.08%. Error analysis showed that probabilistic classifiers produced more false-negative predictions, which is an important consideration for Parkinson's disease screening. The findings indicate that the proposed pipeline provides a noninvasive and interpretable approach for Parkinson's disease screening, with potential applications in remote speech-based assessment.

1. INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative illness that progressively affects the ability to control movements which lead to various symptoms such as tremors, bradykinesia and postural instability. Moreover, with respect to classical motor symptoms, the involvement of speech, referred to collectively as hypokinetic dysarthria, affects approximately 90% of individuals with PD and frequently appears during the early stages of the disease [1]. These voice changes usually present as decreased volume, breathy or hoarsely vocal quality and reduced clarity of articulation.

Conventional clinical diagnosis depends on the Unified Parkinson's Disease Rating Scale (UPDRS) score and expert observation. However, these assessments are subjective and often not available to patients in rural or resource-poor areas [2]. Therefore, there is increasing interest in developing objective, noninvasive and automatic screening methods to detect the presence of PD at early stages via acoustic analysis of speech [3].

One of the core issues to be addressed in automated PD prediction is the reliability and generalizability of machine learning models. Many existing works follow a segment-level evaluation protocol, which is likely to cause data leakage as speech segments of the same subject can be found in both training and testing. Such a practice can result in overoptimistic performance estimation that does not generalize to novel subjects [4], [5]. To overcome this problem, the current study embraces a strict subject-level diagnostic paradigm with a subject-independent evaluation protocol.

Here, we explore the discriminative power of a heterogeneous set of vocal attributes including traditional acoustic feature bases and advanced nonlinear speech characterizers. Specifically, the following five feature sets are considered: Conventional Acoustic and Spectral Speech Features (CASSF), Statistically Pooled acoustic and Dynamic Speech Features (SP-ADSF), Hybrid Correlation and Nonlinear Acoustic Features (HCNAF), Hybrid Acoustic and Dynamical Nonlinear Speech Features (HADNSF) while comparing with extended Geneva Minimalistic acoustic parameter Set (eGeMAPS) [6], [35]. These features are systematically benchmarked under eight machine learning classifiers covering MLP, SVM and ensemble-based voting classifiers to identify feature classifiers, which may be appropriate for robust clinical screening.

The main novelty of this work is the complete comparative study performed under challenging subject-independent evaluation conditions. In contrast to most previous studies validating on a per-segment basis, the approach guarantees that subject identities are cleanly separated for performance evaluation to be trustworthy and clinically useful.

Additionally, this effort combines various acoustic and nonlinear speech feature descriptors and employs a cycle of consolidated statistical validation of binary-level as well as aggregate-level performance metrics. By focusing on clinically significant misclassification patterns, namely false-negative errors, the proposed approach enhances the credibility of speech-based PD screening.

2. RELATED WORK

2.1 Speech-based Parkinson's disease detection

Speech abnormalities are among the common manifestations of Parkinson's disease (PD), making speech analysis a promising noninvasive and cost-effective approach for disease screening and monitoring. Consequently, automatic PD detection using speech has received considerable attention in the research community. Early studies primarily focused on sustained vowel phonation and prosodic characteristics, demonstrating the potential of vocal features to distinguish individuals with PD from healthy controls because speech abnormalities are associated with motor and articulatory impairments [7]. With the increasing availability of public speech datasets, subsequent studies expanded the analysis to continuous speech and task-independent recordings and incorporated more advanced signal-processing and machine-learning techniques. These developments have enabled the extraction of diverse acoustic, spectral, and nonlinear characteristics from speech signals. However, reported performance varies considerably across studies because of differences in datasets, speech tasks, feature-extraction procedures, validation strategies, and classifier configurations [9]. In particular, the use of recordings from the same subjects across training and testing partitions can lead to subject-level information leakage and may result in overly optimistic performance estimates. Therefore, the generalization of speech-based PD detection models to previously unseen subjects remains an important challenge.[8], [9] These limitations highlight the need for systematic evaluation using subject-independent validation and diverse speech feature representations.

2.2 Acoustic and nonlinear feature representations

Various acoustic feature representations have been explored for PD speech analysis. Historically, low-level features such as MFCCs, jitter, shimmer, and the harmonic-to-noise ratio continue to be popular choices not only because of their interpretability but also because they are effective in tracking the spectral, prosodic and phonatory deviations associated with PD [10]. However, such features capture only linear and short-term spectral aspects and may not account for the complex phonatory instabilities that are present in hypokinetic dysarthria [11].

To overcome these shortcomings, enriched and standardized feature sets have been developed. Among the latter, a set of acoustic voice features known as the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) has been widely used for its clinical relevance [12]. Concurrently, nonlinear and dynamic characteristics, that is, entropy inspired measures, complexity-based indices and fractal type descriptors have been utilized to simulate erratic and chaotic vocal cord vibrations occurring in neurodegenerative diseases.

Recent research has shown that both acoustic and nonlinear characteristics are different aspects of a speech signal degradation. However, most studies evaluate particular sets of either acoustic or nonlinear features while ignoring the information contained in the other category by not combining them in a hybrid feature space [13].

2.3 Classification Models for PD Detection

A variety of machine learning models have been employed for speech-based PD detection, such as support vector machines, random forests, k-nearest neighbors, logistic regression, neural networks, ensembles and probability methods. Although sophisticated models, such as deep neural networks and ensemble classifiers, may achieve strong performance, they are often sensitive to feature type, feature selection, data amount, and experimental design [14], [15].

On the other hand, there are various studies reporting competitive or even better performance with simpler probabilistic and linear classifiers for cases where features are easy to separate and have clinical relevance. At the same time, however, a direct comparison is often lacking in terms of more than one classifier per study using the same subject-level evaluation protocol, thereby hindering confidence in the robustness of a specific classifier and its practical relevance for routine clinical practice [16], [17],[18].

2.4 Methodological challenges and motivation

Despite the great achievements, there are still some limitations and challenges in PD speech analysis, which need to be addressed by further research. Firstly, most of the existing methods are trained at the recording or segment level with randomly divided partitions, which contain independent segments. This approach risks producing subject-level data leakage if multiple segments from one subject are included in either training or test sets. As a result, models can capture speaker-specific characteristics unrelated to the disease state and show inflated performance on such data, providing little generalization power to new records[19],[20].

Second, most studies focused on only acoustic or nonlinear features and did not investigate the discriminative power of their combination in a systematic way by means of hybrid feature engineering. Third, there is little or no statistical validation of classifier differences, undermining confidence in performance improvement for PD diagnosis [20, 21].

Inspired by these concerns, the current work strictly follows a subject-level-style evaluation paradigm in which group K-fold cross-validation is performed to ensure the full dissociation of subjects between the training and test folds [21]. Moreover, five diverse types of feature sets—CASSF, SP-ADSF, HCNAF, HADNSF and eGeMAPS—have been systematically tested on various machine learning classifiers. By utilizing hybrid acoustic and nonlinear traits and employing stringent statistical testing, this study was designed to generate a clinically accurate and methodologically rigorous analysis of speech-based PD detection systems. The proposed method is described in the following section.

3. METHODOLOGY

The entire experimental pipeline of the dataset creation, feature extraction, subject-level evaluation and statistical validation will be described in this section. The experimental pipeline in Figure 1 includes speech processing transformation, feature extraction, training the classifiers, aggregation and statistical analysis at the participant level.

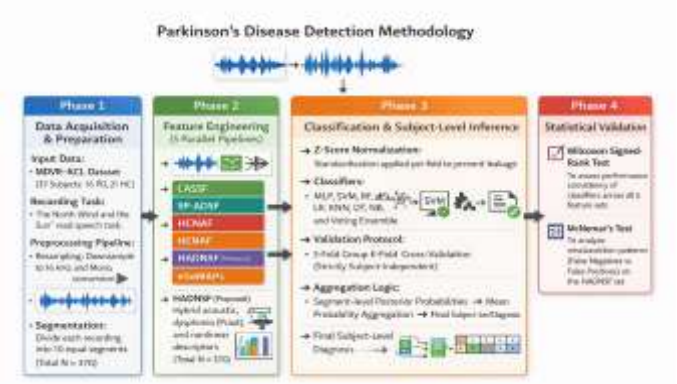

Figure 1: Overview of the Subject-Level PD Detection Framework

3.1 Dataset Description

In this study, we utilized the MDVR–KCL dataset, a public PD speech corpus available on Zenodo. The recordings were made in September 2017 at King's College London (KCL) Hospital and were voice data for 37 participants, 16 with PD and 21 healthy controls (HC) [34]. All the participants read a standardized passage for a read speech task, "The North Wind and the Sun", commonly used in speech pathology studies which encourages a consistent phonetic and linguistic experience among speakers. Clinical experts annotated each recording. Annotation was carried out using standard clinical rating scales, including the Hoehn and Yahr scale (H&Y), the UPDRS II part 5, and the UPDRS-III part 18. These rating scales are characterized by their consistency in measuring disease severity and motor impairment. The voice was recorded using a Motorola Moto G4 smartphone. That audio was captured uncompressed from a device microphone, with a sample rate of the recordings at 44.1 kHz and recorded with a resolution of 16 bits. The audio files were saved in wave format on the storage of the handset. This configuration yielded good quality speech recordings, suitable for acoustic and pathological speech evaluation.

3.2 Speech Preprocessing

To obtain a uniform signal before feature extraction, all the recordings were downmixed to mono and resampled to 16 kHz. Selective silencing removal was subsequently performed to attenuate the effect of uninformative regions and to retain the natural speech structure [22].

Suppose that x(n) represents a discrete-time speech signal. The short-term root mean square (RMS) energy was calculated for every frame k as:

$$E_k = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_k^2(n)} \quad (1)$$

where N is the length of the frame (approximately 30 ms). All frames with RMS energies lower than a threshold value were considered silent. Silent frames were not cut altogether, but only those with continuous silence longer than 500 ms were removed. To prevent speech regions from being fragmented by sudden silence boundaries, the time-frequency mask of silence was smoothed with morphology.

After silence filtering, each cleaned recording was split into 10 equal nonoverlapping segments of the same length. Based on the 37 original recordings, we obtained a total of 370 labels speech segments. The class label (PD or HC) of the corresponding subject was assigned to each segment. Data augmentation was not used. This preprocessing continued with the aim of providing a standardized repetition that would form the basis for further feature extraction and classification.

3.3 Feature Extraction

To capture Parkinsonian speech, five mutually orthogonal representations of features were generated. These parameters represent the acoustic, dynamic, dysphonia-related and nonlinear characteristics of speech signals.

3.3.1 Conventional Acoustic–Spectral Speech Features (CASSF)

The CASSF feature set is taken as an input acoustic vector of speech in all systems. It comprises time-domain energy features like the zero-crossing rate (ZCR) and RMS energy; spectral attributes encompassing the spectral centroid, spectral bandwidth and spectral roll-off, mel-frequency cepstral coefficients (MFCCs 1–13), and harmonic chromatic features (Chroma 1–12) [23] . In total, CASSF consists of 30 features and represents a benchmark for comparing advanced feature representations.

3.3.2 Statistically Pooled Acoustic–Dynamic Speech Features (SP-ADSF)

The SP-ADSF representation highlights the temporal dynamics of speech with statistical summarization. Features Static MFCCs and their first order delta (Δ) and second-order delta-delta (ΔΔ) derivatives were computed and summarized with mean/standard-deviation statistics [24],[25]. For a feature at frame-level f_t, we calculate pooled statistics as:

$$\mu_f = \frac{1}{T}\sum_{t=1}^{T} f_t \quad (2)$$

$$\sigma_f = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(f_t - \mu_f)^2} \quad (3)$$

where T is the number of frames in the segment. The other pooled features are chroma descriptors, the fundamental frequency (F0), the RMS energy and the ZCR. The final form of the SP-ADSF vector contains 108 features, and is representative of both dynamic articulation and prosodic variation.

3.3.3 Hybrid correlation–nonlinear acoustic features (HCNAFs)

The HCNAF feature set was developed to model interfeature dependencies and nonlinear dynamics that characterize neuromotor deficits. The linear-correlation-based features are MFCC–pitch coupling, energy–pitch correlations, and pairwise inter-MFCC correlations [10]. The correlation coefficients were computed as

$$\rho_{xy} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \quad (4)$$

The nonlinear indices used are the Hurst exponent, sample entropy, permutation entropy and detrended fluctuation analysis (DFA) as well as measures from recurrence quantification analysis (RQA). The HCNAF consists of 99 features and is designed to model complex relationships due to impaired neuromotor control.

3.3.4 Hybrid acoustic–dysphonia–nonlinear speech features (HADNSF)

We propose hybrid set of features called HADNSF, composed of acoustic, dysphonia related and nonlinear parameters. They comprise Mel cepstral coefficients and their first and second order temporal derivatives, energy and its temporal derivative, spectral moments and Chroma features. Furthermore, we included dysphonia features extracted using Praat software and a shortened set of nonlinear features [26],[27]. In total, there are 118 features in the proposed set. We aimed to obtain features that are clinically interpretable while being able to capture the nonlinear characteristics of the speech signals of persons with PD.

3.3.5 Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)

The eGeMAPS feature set is a clinically validated baseline for the analysis of pathological speech. It includes pitch-based features, formant and spectral shape descriptors, voice quality acoustical measures and energy statistics in the time domain resulting in 88 features. All eGeMAPS features were extracted via openSMILE and were used as an acoustic representation of reference for comparison [12].

3.4 Feature normalization

Prior to classification, all feature vectors were standardized via z score normalization [28], [29]:

$$z = \frac{x-\mu}{\sigma} \quad (5)$$

where μ and σ denote the mean and standard deviation computed from the training data only. The same normalization parameters were applied to the test data to prevent information leakage.

3.5 Classification models

After training the classifier, the we performed the prediction step independently for each subject to evaluate performance at the individual level. Because the speech input was segmented, the initial predictions took the form of individual segment predictions. The posterior probabilities of each segment were then averaged across all subjects. Overall, averaging reduces the variance between subjects, which can occur because of variations in demographics or other factors during data collection. In doing so, the procedure enhances the accuracy and reliability of the prediction results while maintaining clinical relevance. Lastly, the classification was performed for each individual subject through the application of a fixed decision threshold to the averaged probabilities.

The ensemble probability was computed as:

$$p_i = \frac{1}{M}\sum_{m=1}^{M} p_i^{(m)} \quad (6)$$

3.6 Subject-level evaluation protocol

For independent performance testing, predictions were assessed at the subject level after classifier training. Since speech signals were segmented, initial predictions were obtained for each segment. Posterior probabilities were then aggregated across segments for each subject. Mean probability aggregation was used to minimize subject-level bias. This step also improved the clinical relevance of the findings. Final subject-level decisions are made by applying a fixed decision threshold to the aggregated probabilities.

All the experiments were conducted via strict subject-wise cross-validation. All speech segments from a given speaker were assigned exclusively to either the training or the testing fold in each trial. This evaluation strategy prevents record-linkage leakage. It also provides an accurate estimate of the generalization performance on unseen subjects.

4. RESULTS AND DISCUSSION

In this section, we present the results for all eight of our classifiers for machine learning for Parkinson's detection using various features sets that contain acoustic features. All of our experiments were performed at the subject level to ensure that there is no leakage in the tests and trainings sets. For all the experiments, the classification performance was measured in terms of accuracy, precision, sensitivity, specificity, F1-score, and ROC-AUC. We believe that all the classification performance measures are important for biomedical signal screening since the metrics account for both the magnitude of correct outcomes and the discriminatory performance of the resulting model. All the feature sets used in the experiments were CASSF, SP-ADSF, HCNAF, HADNSF, and eGeMAPS. The used classifiers were MLP, SVM, RF, LR, KNN, DT, NB, and Voting Ensemble Classifier.

4.1 Performance comparison across feature sets

The best performing classifiers for each feature set according to the accuracy, F1 score, and ROC-AUC metrics are summarized in the Table 1 below. From the results, it is evident that Naive Bayes showed its robustness and efficiency for most of the tested feature sets including CASSF and HADNSF. In the contrary, the eGeMAPS feature set allows one to build linear classification models with the highest accuracy, logistic regression being the most accurate algorithm for this feature set.

Table 1: Best Classifier Performance for Each Feature Set

Feature Set	Best Classifier	Accuracy	F1-Score	ROC-AUC
CASSF	Naive Bayes	0.8378	0.8000	0.8750
SP-ADSF	Naive Bayes	0.7838	0.7333	0.8810
HCNAF	Random Forest / KNN	0.8108	0.7407 / 0.7586	0.8646
HADNSF	Naive Bayes	0.8378	0.8125	0.8958
eGeMAPS	Logistic Regression	0.8649	0.8276	0.8661

4.2 Classifier Wise Comparative Analysis

In order to judge the impact of feature selection process on the results independently, in Table 2 the best obtained subject level accuracy for each classifier is presented. Even though the accuracies achieved by the ensemble and nonlinear classifiers are indeed competitive, the probabilistic approaches were not consistently outperformed. Thus, the contribution of the selected features plays a more significant role in comparison to the choice of the classifier itself.

Table 2: Overall Best Performance of Each Classifier

Classifier	Best Accuracy	Best F1-Score	Best ROC-AUC	Feature Set
MLP	0.7838	0.7143	0.8363	eGeMAPS
SVM	0.8108	0.7879	0.8423	HADNSF
Random Forest	0.8378	0.8000	0.8571	eGeMAPS

Logistic Regression	0.8649	0.8276	0.8661	eGeMAPS
KNN	0.8108	0.7586	0.8646	HCNAF
Decision Tree	0.7568	0.7429	0.8631	HCNAF
Naive Bayes	0.8378	0.8125	0.8958	HADNSF
Voting Ensemble	0.8108	0.7586	0.8780	eGeMAPS

4.3 Statistical significance analysis

4.3.1 Wilcoxon signed-rank test

To judge the statistical significance of the observed differences in performance of the methods, the nonparametric statistical tests were applied at the 5% significance level ($\alpha = 0.05$) that is common for biomedical classification studies with small sample sizes [30]. Specifically, to compare classifiers on the same feature set under similar experimental conditions, paired ROC-AUC values between classifiers were evaluated using the Wilcoxon signed-rank test [31]. Despite the higher average ROC-AUC values in different feature representations of the data, the Naive Bayes classifier did not show significant improvement over random in any of the pairwise comparisons ($p > 0.05$ for all tests). This is possibly due to the relatively small number of pairwise comparisons made (five feature sets) and the conservative nature of nonparametric statistical testing, both of which decrease the statistical power. However, the observed trends are ranked the Naive Bayes classifier first in all the feature sets, surpassing Decision Tree, k-Nearest Neighbors and Logistic Regression classifiers. This indicates that despite the lack of statistical significance, Naive Bayes performed consistently against multiple classifiers using various feature representations. Its selection as the best classifier is thus based on global indicators of the model performance, such as the ROC-AUC and F1-score rather than on individual statistical significance. The p-values of the Wilcoxon signed-rank test are shown in Table 3 with naive Bayes as the baseline classifier. All comparisons were done under a single experimental condition. Although no statistically significant improvement was observed, the consistently superior performance of Naive Bayes across all feature sets supports the feasibility and potential usefulness of this particular classifier.

Table 3: Wilcoxon signed-rank test results (ROC-AUC)

Reference Classifier	Compared Classifier	p-value (ROC-AUC)
Naive Bayes	MLP	0.1562
Naive Bayes	SVM	0.1562
Naive Bayes	Random Forest	0.3125
Naive Bayes	Logistic Regression	0.0938
Naive Bayes	KNN	0.0938
Naive Bayes	Decision Tree	0.0938
Naive Bayes	Voting	0.1562

4.3.2 McNemar's test

We checked our prediction results with the HADNSF feature set using the McNemar's test for possible misclassification discrepancies at an individual level [32]. The differences between naive Bayes and k-Nearest Neighbors ($p = 0.0074$) and between Naive Bayes and Multilayer Perceptron ($p = 0.0312$) were found to be statistically significant. These results mean that Naive Bayes leads to a significantly smaller number of misclassifications on the same subjects as compared to the other classifiers. In particular, it prevents a higher number of false-negative errors, which is vital for identifying patients suffering from PD who may need timely clinical attention. Comparisons with Logistic Regression were almost significant ($p = 0.0625$). Random Forest, Decision Tree and voting ensemble classifiers, on the other hand, showed no statistically significant difference. It is important to note that the decision boundaries of these models are not the same, but that their prediction distributions overlap to some extent. Table 4 summarizes the results of the McNemar's test of the HADNSF feature set for subject-level

predictions with Naive Bayes as reference classifier. Values that differ significantly ($p<0.05$) are identified as such and marked in bold. Overall, the findings indicate that Naive Bayes exhibited a distinct misclassification pattern compared with kNN and MLP, with fewer false-negative predictions observed in the evaluated subjects.

Table 4: McNemar's test results for subject-level classifier comparisons using the HADNSF feature set.

Reference Classifier	Compared Classifier	NB correct / Other wrong	NB wrong / Other correct	p-value
Naive Bayes	KNN	13	2	**0.0074**
Naive Bayes	MLP	6	0	**0.0312**
Naive Bayes	Logistic Regression	5	0	0.0625
Naive Bayes	Decision Tree	8	3	0.2266
Naive Bayes	Voting	4	1	0.3750
Naive Bayes	SVM	2	1	1.0000
Naive Bayes	Random Forest	2	1	1.0000

5. CONCLUSION

In this study, we performed a comprehensive evaluation of different machine learning classifiers for detecting PD using various acoustic feature representations on a subject-wise basis. As a result, it was found that both the feature engineering and the choice of the classifier play a vital role in determining the accuracy of diagnosis. It was observed that Naive Bayes was the most consistent and the highest performing algorithm across all the feature sets with an accuracy of 83.78%, an F1-score of 0.8125, and the highest overall ROC-AUC value of 0.8958. However, the application of the Naive Bayes algorithm showed good discriminative power for most of the considered feature sets. Therefore, for actual screening, when the feature set is not fully predetermined, the second model could be more relevant. Considering all the calculations, it is possible to propose several recommendations regarding the choice of the best classifier, configuration with accuracy-oriented considerations, and generalizable approach for the case under consideration. Thus, the best-performing classifier was Naive Bayes, while the most accuracy-oriented configuration was Logistic Regression with eGeMAPS features. Finally, the most generalizable solution was Naive Bayes with HADNSF features. Based on the obtained results and detailed statistical analysis, it is possible to conclude that the careful consideration of acoustic features along with the application of less complex algorithms could provide more reliable and accurate results than advanced configurations of probabilistic models. Moreover, the current research demonstrated that the Naive Bayes approach could be a suitable and computationally efficient method for the initial screening of PD. Finally, the results of the study prove that the subject-oriented approach to the feature selection could be successfully applied in subsequent studies of speech analysis.

6. LIMITATIONS AND CLINICAL RELEVANCE

6.1 Limitations

The results are good so far, at the same time it is important to report the following issues. For one, we used a small set of public data which may not represent the large and very diverse clinical population. Also, we did not pay special attention to language variables and differences in recording settings and microphone quality which may have played a role in the results. Also the study looked at binary classification only and did not look at the degree of disease or the disease's progression over time. Also, we did not look into deep learning methods which may have improved results above what we got with tailored features.

6.2 Clinical relevance

From a clinical perspective, it is of great importance that PD be detected early and accurately, which in turn improves patient care [33]. We put forward an approach which focuses on high sensitivity which at the same time maintains an acceptable level of specific performance thus greatly reduce false negatives which is a great need in a screening setting. Also, the Naive Bayes

classifier's probability based and, easy to interpret features improve clinical transparency and trust which in turn makes it very much a fit for use in clinical decision support systems. We put forth a method which is also very efficient from a computation point of view and at the same time very robust which makes it a perfect fit for telemedicine platforms, mobile health apps and low-income settings. That said the system is not to be used for a clinical diagnosis but to serve as a prescreen which will help clinicians to identify which individuals require more in depth neurologic work up.

REFERENCES:

[1] M. Sapmaz Atalar, "Hypokinetic Dysarthria in Parkinson's Disease: A Narrative Review," Sisli Etfal, vol. 57, no. 2, pp. 163–170, Jan. 2023, doi: 10.14744/semb.2023.29560.

[2] M. Disease, "The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations.," Movement Disorders, vol. 18, no. 7, pp. 738–750, June 2003, doi: 10.1002/mds.10473.

[3] A. S. Gullapalli and V. K. Mittal, "Early Detection of Parkinson's Disease Through Speech Features and Machine Learning: A Review," Springer Singapore, 2021, pp. 203–212. doi: 10.1007/978-981-16-4177-0_22.

[4] I. Ahammad, "Generalization over accuracy: A cross-dataset, explainable, and federated learning framework for Parkinson's disease detection.," Digital health, vol. 12, Feb. 2026, doi: 10.1177/20552076261467837.

[5] T. A. Vu et al., "A Comparison of Machine Learning Algorithms for Parkinson's Disease Detection," Institute Of Electrical Electronics Engineers, Nov. 2023, pp. 316–321. doi: 10.1109/iccais59597.2023.10382406.

[6] M. Labied and A. Belangour, "Automatic Speech Recognition Features Extraction Techniques: A Multi-criteria Comparison," IJACSA, vol. 12, no. 8, Jan. 2021, doi: 10.14569/ijacsa.2021.0120821.

[7] Z.-H. Xu, M.-H. Zhang, and J. Wang, "Diagnostic Value of Speech Acoustic Analysis in Parkinson's Disease," Sichuan da xue xue bao. Yi xue ban = Journal of Sichuan University. Medical science edition, vol. 53, no. 4, pp. 726–731, July 2022, doi: 10.12182/20220760304.

[8] C. Bunterngchit and Y. Bunterngchit, "A Comparative Study of Machine Learning Models for Parkinson's Disease Detection," Institute Of Electrical Electronics Engineers, Mar. 2022, pp. 465–469. doi: 10.1109/dasa54658.2022.9765159.

[9] G. Aqil Ali, L. Sharifi, P. Rashidi-Khazaee, and H. Nahid-Titkanlue, "A Mutual-Information-Guided and ADASYN-Augmented Machine Learning Framework for Early Prediction of Parkinson's Disease," msesj, pp. 1–12, Jan. 2026, doi: 10.61838/msesj.313.

[10] S. Bouagina, M. Naouara, S. Hafsi, and S. Djaziri-Larbi, "MFCC-Based Analysis of Vibratory Anomalies in Parkinson's Disease Detection using Sustained Vowels," Institute Of Electrical Electronics Engineers, Dec. 2023, pp. 1–5. doi: 10.1109/amcai59331.2023.10431494.

[11] H. Zhang, N. Yan, L. Wang, and M. L. Ng, "Energy distribution analysis and nonlinear dynamical analysis of phonation in patients with Parkinson's disease," Institute Of Electrical Electronics Engineers, Dec. 2017, pp. 630–635. doi: 10.1109/apsipa.2017.8282102.

[12] F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," IEEE Trans. Affective Comput., vol. 7, no. 2, pp. 190–202, Apr. 2016, doi: 10.1109/taffc.2015.2457417.

[13] L. J. Bronakowski, "Nonlinear dynamics approach to speech detection in noisy signals," in Proceedings of SPIE, the International Society for Optical Engineering/Proceedings of SPIE, Chalmers University Of Technology, June 2009, p. 75021M. doi: 10.1117/12.837837.

[14] A. Ratnakar, "EVALUATING SPEECH ANALYSIS TECHNIQUES FOR PARKINSONS DISEASE DETECTION: A COMPARISON OF MACHINE LEARNING AND DEEP LEARNING ALGORITHMS," IJAR, vol. 12, no. 05, pp. 1118–1137, May 2024, doi: 10.21474/ijar01/18827.

[15] P. Patil, "Comparative Investigation of Classification Algorithms using Parkinson's Disease Acoustic Analysis Dataset to Choose Best Classifier for Best Result," jisem, vol. 10, no. 14s, pp. 126–131, Feb. 2025, doi: 10.52783/jisem.v10i14s.2168.

[16] A. Singh, Z. Hasan, V. Paranjape, V. Vashishtha, and S. Chhabra, "Application of Machine Learning in Early Detection of Parkinson's disease Using Vocal Features," ijirset, vol. 10, no. 11, pp. 14454–14460, Nov. 2023, doi: 10.15680/ijirset.2021.1011128.

[17] H. Mishra, K. Verma, H. Yadav, and A. Kumar, "Parkinson's Disease Detection System Using Machine Learning Techniques," Institute Of Electrical Electronics Engineers, Aug. 2024, pp. 1–5. doi: 10.1109/acet61898.2024.10730078.

[18] B. A. Tama and S. Lim, "A Comparative Performance Evaluation of Classification Algorithms for Clinical Decision Support Systems," Mathematics, vol. 8, no. 10, p. 1814, Oct. 2020, doi: 10.3390/math8101814.

[19] S. E. Davis, M. E. Matheny, S. Balu, and M. P. Sendak, "A framework for understanding label leakage in machine learning for health care.," Journal of the American Medical Informatics Association : JAMIA, vol. 31, no. 1, pp. 274–280, Sept. 2023, doi: 10.1093/jamia/ocad178.

[20] N. Bussola, G. Jurman, A. Marcolini, V. Maggio, and C. Furlanello, "AI slipping on tiles: data leakage in digital pathology," Sept. 14, 2019, Cornell University. doi: 10.48550/arxiv.1909.06539.

[21] M. - and D. -, "A Comprehensive Review of Cross-Validation Techniques in Machine Learning," IJSAT, vol. 16, no. 1, Jan. 2025, doi: 10.71097/ijsat.v16.i1.1305.

[22] P. J. Pompei, "Preprocessing method for nonlinear acoustic system," J. Acoust. Soc. Am., vol. 120, no. 6, p. 3452, Jan. 2006, doi: 10.1121/1.2409452.

[23] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," IEEE Access, vol. 10, pp. 122136–122158, Jan. 2022, doi: 10.1109/access.2022.3223444.

[24] O. Ichikawa, T. Fukuda, and M. Nishimura, "Dynamic Features in the Linear-Logarithmic Hybrid Domain for Automatic Speech Recognition in a Reverberant Environment," IEEE J. Sel. Top. Signal Process., vol. 4, no. 5, pp. 816–823, Oct. 2010, doi: 10.1109/jstsp.2010.2057191.

[25] L. Gongora, O. Ramos, and J. Mauricio, "Comparative Analysis of the Mel Frequency Cepstral Coefficients for Voiced and Silent Speech," IRECAP, vol. 6, no. 5, p. 255, Oct. 2016, doi: 10.15866/irecap.v6i5.10271.

[26] P. Mahesha and D. S. Vinod, "Combining Cepstral and Prosodic Features for Classification of Disfluencies in Stuttered Speech," Springer India, 2014, pp. 623–633. doi: 10.1007/978-81-322-2012-1_67.

[27] P. Vizza, G. Tradigo, P. H. Guzzi, and P. Veltri, "Dysphonia discovering using a Goertzel algorithm implementation for vocal signals analysis," Biocybernetics and Biomedical Engineering, vol. 45, no. 3, pp. 469–475, July 2025, doi: 10.1016/j.bbe.2025.07.001.

[28] M. Z. Al-Faiz, A. A. Ibrahim, and S. M. Hadi, "The effect of Z-Score standardization (normalization) on binary input due the speed of learning in back-propagation neural network," Iraqi Journal of ICT, vol. 1, no. 3, pp. 42–48, Feb. 2019, doi: 10.31987/ijict.1.3.41.

[29] N. Fei, Y. Gao, Z. Lu, and T. Xiang, "Z-Score Normalization, Hubness, and Few-Shot Learning," Institute Of Electrical Electronics Engineers, Oct. 2021, pp. 142–151. doi: 10.1109/iccv48922.2021.00021.

[30] A. K. Dwivedi, I. Mallawaarachchi, and L. A. Alvarado, "Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method.," Statistics in Medicine, vol. 36, no. 14, pp. 2187–2205, Mar. 2017, doi: 10.1002/sim.7263.

[31] P. K. Singh, R. Sarkar, and M. Nasipuri, "Significance of non-parametric statistical tests for comparison of classifiers over multiple datasets," IJCSM, vol. 7, no. 5, p. 410, Jan. 2016, doi: 10.1504/ijcsm.2016.080073.

[32] J. F. Troendle, K. F. Yu, P. H. Westfall, G. Pennello, and E. F. Schisterman, "Comparing the Expected Misclassification Cost for Two Classifiers Based on Estimates From the Same Sample," Statistics in Biopharmaceutical Research, vol. 4, no. 3, pp. 301–312, July 2012, doi: 10.1080/19466315.2012.695263.

[33] Z. Gao, T. Chen, W. Wang, J. Wang, S. Hong, and Z. Wang, "Advances in the diagnosis and treatment of Parkinson's disease," Chinese Journal of Contemporary Neurology and Neurosurgery, vol. 15, no. 10, pp. 777–781, Oct. 2015, doi: 10.3969/cjcnn.v15i10.1281.

[34] Jaeger, Hagen, Dhaval Trivedi, and Michael Stadtschnitzer. 2019. Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls. *OPAL (Open@LaTrobe) (La Trobe University)*. La Trobe University. https://doi.org/10.5281/zenodo.2867216.

[35] Ashwini D. Bhople, Avinash Kapse, & Pravin A. Kharat. (2026). A Comparative Study of Acoustic and Clinical Speech Feature Vectors for Parkinson's Disease Detection with a Multilayer Perceptron Classifier. *International Journal of Computer Information Systems and Industrial Management Applications*, *18*(1s), 708–715. https://doi.org/10.70917/ijcisim-2026-1965