e-ISSN: 2149-4584

Erzincan University Journal of Science and Technology

EJSAT 2026, 19 (1) 331-356

Research Article

Feature Engineering for Parkinson's Disease Diagnosis: A Hybrid Approach Using Random Forest Feature Selection and Correlation Analysis

Ramiz Görkem Birdal[1*]

[1]Department of Computer Engineering, Istanbul University - Cerrahpasa, Istanbul, Turkey

Received: 15/02/2025, **Revised:** 07/07/2025, **Accepted:** 17/07/2025, **Published:** 30/03/2026

Abstract

Feature selection is a crucial step in optimizing machine learning models, particularly in biomedical applications such as Parkinson's disease classification based on speech data. This study employs multiple feature importance techniques to identify the most significant predictors and remove redundant variables, thereby improving model interpretability and efficiency. Four distinct methods—Permutation Importance, Mutual Information (MI), ANOVA F-score, and Random Forest Importance—are applied to assess the contribution of each feature to classification performance. Additionally, a correlation analysis is conducted to detect highly correlated features that may introduce multicollinearity. Many studies in existing literature on Parkinson's disease classification overlook the impact of multicollinearity and redundant features, which can affect model stability and interpretability. Our study addresses this gap by systematically comparing four feature selection methods and incorporating correlation analysis to refine the feature set for improved accuracy and efficiency. By systematically refining the feature set, this approach ensures a balance between model complexity and predictive power, ultimately enhancing the reliability of automated Parkinson's disease diagnosis from speech recordings.

Keywords: Parkinson's Disease (PD), feature engineering, ensemble learning, random forest, ANOVA, Mutual Information

Parkinson Hastalığı Teşhisi için Özellik Mühendisliği: Rastgele Orman Özellik Seçimi ve Korelasyon Analizini Kullanan Hibrit Bir Yaklaşım

Öz

Özellik seçimi, makine öğrenimi modellerini optimize etmede kritik bir adımdır ve özellikle konuşma verilerine dayalı Parkinson hastalığı sınıflandırması gibi biyomedikal uygulamalarda büyük önem taşır. Bu çalışma, en önemli öngörücü değişkenleri belirlemek ve gereksiz değişkenleri ortadan kaldırarak modelin yorumlanabilirliğini ve verimliliğini artırmak amacıyla birden fazla özellik önem derecelendirme tekniği kullanmaktadır. Sınıflandırma performansına her özelliğin katkısını değerlendirmek için Dizinleme Önem (Permutation Importance), Karşılıklı Bilgi (Mutual Information - MI), ANOVA F-skoru ve Rastgele Orman Önemi (Random Forest Importance) olmak üzere dört farklı yöntem uygulanmaktadır. Ayrıca, yüksek derecede ilişkili özellikleri tespit ederek çoklu bağlantı (multicollinearity) sorununu önlemek için bir korelasyon analizi gerçekleştirilmiştir. Mevcut literatürde Parkinson hastalığı sınıflandırmasına yönelik birçok çalışma, çoklu bağlantı ve gereksiz özelliklerin model kararlılığı ve yorumlanabilirliği üzerindeki etkisini göz ardı etmektedir. Bu çalışma, dört farklı özellik seçme yöntemini sistematik olarak karşılaştırarak ve korelasyon analizini entegre ederek bu boşluğu gidermeyi amaçlamaktadır. Özellik kümesini titizlikle rafine eden bu yaklaşım, model karmaşıklığı ile tahmin gücü arasında bir denge sağlayarak konuşma kayıtlarından otomatik Parkinson hastalığı teşhisinin güvenilirliğini artırmaktadır.

Anahtar Kelimeler: Parkinson hastalığı (PH), özellik mühendisliği, topluluk öğrenmesi, rastgele orman metodu, ANOVA, karşılıklı bilgi yaklaşımı

Corresponding Author: ramizgorkem.birdal@iuc.edu.tr

Cite this Article: Erzincan University Journal of Science and Technology 2026, 19(1) 331-356.
https://doi.org/10.18185/erzifbed.1640563

1. Introduction

Parkinson's disease (PD) is a neurodegenerative disorder that primarily affects motor function due to the progressive loss of dopamine-producing neurons in the substantia nigra region of thebrain The recent works classified PD patients through the application of machine learning and deep learning techniques on acoustic features extracted from speech signals. Traditional approaches are usually based on feature extraction techniques like mel frequency cepstral coefficients, jitter, shimmer, and F0, among others, and classification is performed by using the SVM, random forests, or neural networks. However, one important challenge in such studies is that there are so many high- dimensional and redundant features included in the studies that might result in overfitting by increasing computational complexity and reducing interpretability. While a few studies try to perform dimensionality reduction-PCA, for example-most of them do not compare different feature selection methods systematically to identify the most relevant predictors. Besides, in most cases, correlation analysis is barely considered and issues related to feature redundancy may affect both performance and generalizability of models. To address these limitations, this study employs a multi-method feature selection approach to improve the accuracy and efficiency of Parkinson's disease classification from speech data. We apply four distinct feature selection techniques: Permutation Importance, Mutual Information, ANOVA F-score, and Random Forest Importance, and systematically compare their rankings. Besides this, correlation analysis is performed to identify and remove redundant features, ensuring a more robust and interpretable model. By refining the feature selection process, it is aimed to improve the reliability of speech-based detection in Parkinson's disease to allow further research in computational neurology and biomedical signal processing. These findings represent the structured framework of optimal feature selection in biomedical machine learning applications and give insight into which speech biomarkers have most of the predictive information for Parkinson's diagnosis.

Numerous Numerious approaches have focused comprehensive studies of feature extraction and selection aimed at improving diagnostic accuracy. Principle Component Analysis (PCA) was used to reduce and select features from the dataset; 5875 records and 19 attributes of the raw dataset are reduced to 19 records and 19 attributes of the PCA dataset [1]. Some researchers analyzed their PD dataset using the ANVOA feature selection method [2]. Caliskan et al. [3] proposed a method for extracting and selecting features from two datasets, optical path differences OPD and Position sensing device PSD, using stacked auto-encoders (SAE) and softmax classifiers. A total of 18 techniques were utilized to extract features from the dataset in a different study, but after without analyzing, four different classifiers were used to classify the resulting data [4]. Another researchers studied 40 subjects, 23 of which had PT Parkinson's disease tremor and 17 of which had essential tremor ET, and then they developed a cellular neural networks (CNN) model for analyzing the features [5]. Parkinson's Diseased and healthy subjects' voices were used to study Deep CNN methods using spectral voice features extracted from sustained phonemes [6]. Variations in their Deep CNN were compared along with varying lengths of input voice. Google Inception v3 CNN model was fine-tuned using DAT-SPECT images, and then features from the previous fully connected layer were extracted before the final fully connected layer was created [7]. In another experiment, they presented an approach based on the Quantitative Susceptibility Mapping (QSM); they used this technique to extract

radiomics features from the substantia nigra (SN) [8]. In different study it has been found that stacked auto-encoders (SAE) may improve the classification accuracy of a model by reducing the number of dimensions and extracting features [9]. With the help of four feature selection algorithms and two statistical classifiers, their study evaluated the effectiveness of these measures in distinguishing Parkinson's Disease subjects from healthy controls [10-11]. It has been suggested that pitch-synchronous feature extraction could be more effective than conventional block segmentation with fixed frame lengths by [12]. An extracting technique [13] based on spectral and temporal features proved superior to 17 Machine Learning classifiers in classification of continuous speech data even for classification of Parkinson's Disease. The researchers of the Gyenno Science Parkinson Disease Research Center used a Bidirectional Long Short-Term Memory (LSTM) model in 2021 to assess the dynamic time- series features by analyzing energy transitions between speech segments that were unvoiced and voiced [14]. As part of another study, voice features were integrated into distinct frameworks, in which they were validated using Leave-One-Person-Out Cross-Validation (LOPO CV) on a dataset accessed from the UCI machine learning repository [15]. A study conducted by Chen et al. [16] used PCA and Fuzzy KNN techniques to enhance the precision of the PD voice signal detection to approximately 96.07% using datasets from the UCI machine learning repository. As part of a study [17], EEG event-related potential ERP is used to detect early Parkinson's disease with a novel Brain Network Analytics (BNA) technique and to classify patients based on logistic regression, using a false positive rate (FPR) feature selection method. In order to investigate the relationship between mental health and neurological disorders, Vieira et al. [18] utilized multilayer perceptrons (MLPs). A recurrent neural network was used by Guinci et al. [19] to model human brain activity as a response to sensory stimuli exploiting two different datasets generated by Nishioto et al. [20]. To encode low-level visual features, the RNNs used two nonlinear recurrent layers and one linear layer, while to encode high-level semantic data, they used Long Short-Term Memory (LSTM) and gated recurrent units (GRUs). The Full Complex network was developed by Riaz and colleagues to compute functional connectivity from fMRI time series data [21]. In order to detect PD early, machine learning (ML) has enabled the automated extraction and selection of some features [22]. In their work, some researchers introduce a novel feature taxonomy and a method for determining a relationship between a particular feature and another [23]. SVM and random forest are used for recursive feature elimination (RFE) in order to select clinically relevant features for Classification of Parkinson's Disease [24]. Based on a comparison of mean values of features, ANOVA tests are useful for separating healthy controls and PD subjects [25]. A validation study was conducted using data from Physionet [26]. Based on a CNN and Long Short-Term Memory, [27] classifies subjects into PD and controls with an accuracy of 96.9%; the model learns features closely related to PD clinical features including dopamine levels and disease severity. Utilizing only 10 dysphonia measures, Tsanas et al. classified speech signals from 33 patients with Parkinson's and 10 controls using support vector machine and random forest models [28]. Using 256 features of vocal data in 40 PD participants and 40 Controls, Karabayir et al. [29] detected PD from the Light Gradient Boosting (GB) and Extreme Gradient Boosting (GB). Moreover, seven features emerged as most relevant from feature analysis. Zhang et al. [30] developed a machine learning technique based on stacked autoencoders and k-nearest neighbors (KNN) algorithms for time-frequency features of vocal signals. A different CNN architecture was developed to classify the "HandPD" dataset into PDs and Controls [31]. Meta-heuristic optimization techniques were

used to tune the hyper-parameters. Application of feature importance in the diagnosis of Parkinson's disease by applying four different feature selection methods is one of those paradigm moments in the detection and management of neurodegenerative diseases. Using permutation importance, mutual information, ANOVA F-score, and random forest importance, a highly accurate diagnostic tool has been developed by researcher. Another study [32] employs a deep learning ensemble approach for early detection by fine-tuning pre-trained Keras models on the International Skin Imaging Collaboration dataset, achieving a 93.03% detection rate and showing the power of ensemble methods. Different study [33] explores kidney stone detection from medical images using various machine learning methods, with CNN-based deep neural networks demonstrating strong performance, though Decision Tree Classifier achieved the highest F1 score (85.3%), highlighting CNN's potential for automated and efficient diagnosis. This paper has tried to show the viability of feature engineering using ensemble techniques for the detection of Parkinson's disease, hence providing a substitute non-invasive, cost-effective, and automated way of diagnosis.

Overall, the study is divided into sections: Section II describes the material and methods, section III presents the results and discussion, and further section presents the conclusions.

2. Material and Methods

2.1 The Dataset

The Parkinson's Disease (PD) dataset consists of training and test files, designed for both classification and regression tasks. The training set includes voice recordings from 20 individuals with Parkinson's Disease (PWP) (6 female, 14 male) and 20 healthy individuals (10 female, 10 male), all of whom were examined at the Department of Neurology, Cerrahpasa Faculty of Medicine, Istanbul University [34]. Each subject provided 26 voice samples, covering sustained vowels, numbers, words, and short sentences. A total of 26 linear and time-frequency-based features were extracted from each recording. Additionally, the dataset contains the Unified Parkinson's Disease Rating Scale (UPDRS) score for each PD patient, allowing it to be used for regression analysis. After the initial data collection and experimentation with the training dataset, an independent test set was collected under the same conditions. This test set consists of recordings from 28 PD patients, who were asked to pronounce the sustained vowels 'a' and 'o' three times each, resulting in a total of 168 recordings. The same 26 features were extracted from these samples, ensuring consistency with the training dataset. This independent test set serves as a validation tool for models trained on the initial dataset. If we analyze the speech data of Subject 2, the waveform visualization of the first half of the recording provides insights into vocal characteristics as seen in Figure 1. The x-axis represents time in seconds, while the y-axis shows amplitude, indicating variations in speech intensity. In this waveform, larger peaks and troughs reflect stronger vocal output, whereas lower amplitude regions indicate weak articulation, a common trait in Parkinson's speech. The smoothness and density of the waveform denote vocal stability, while irregular or trembling patterns can point to voice tremors, which are considered to be one of the hallmarks of Parkinson's Disease. Further, if the waveform contains long stretches with very little variation, that can point to monotonic speech-another symptom. This waveform analysis will help to identify speech irregularities such as reduced loudness, instability, articulation difficulties, which are very critical in the assessment

of the impairments of voices related to Parkinson's.

Although the independent test set comprises 168 recordings, this number is considered statistically and methodologically sufficient for several reasons. First, each voice recording in the dataset is not a singular data point but a high-dimensional feature vector constructed from 26 acoustic parameters, including jitter, shimmer, harmonics-to-noise ratio, and Mel-frequency cepstral coefficients (MFCCs), among others. These features collectively capture the nuanced pathological speech patterns associated with Parkinson's Disease, enabling the machine learning models to learn complex relationships even from a relatively modest number of samples. Moreover, the design of the dataset adheres to a within-subject repeated measures framework: multiple recordings (sustained vowels, numbers, short sentences) are collected from each participant under standardized clinical conditions. This reduces inter-speaker variability while enhancing the consistency of the feature space, which in turn improves the statistical power of model training and evaluation. In addition, recent studies in the literature [11, 34] have shown that Parkinson's voice classification tasks can achieve high generalization performance with sample sizes in the range of 100–200 when feature richness and noise control are prioritized. In our case, the high signal-to-noise ratio and careful preprocessing compensate for the relatively limited sample size. Furthermore, the independent test set was not used for training but strictly for evaluating generalization, ensuring unbiased performance assessment. Therefore, despite its modest size, the test dataset provides a reliable and statistically valid basis for evaluating the classification performance of feature selection methods and machine learning models within this domain-specific context.

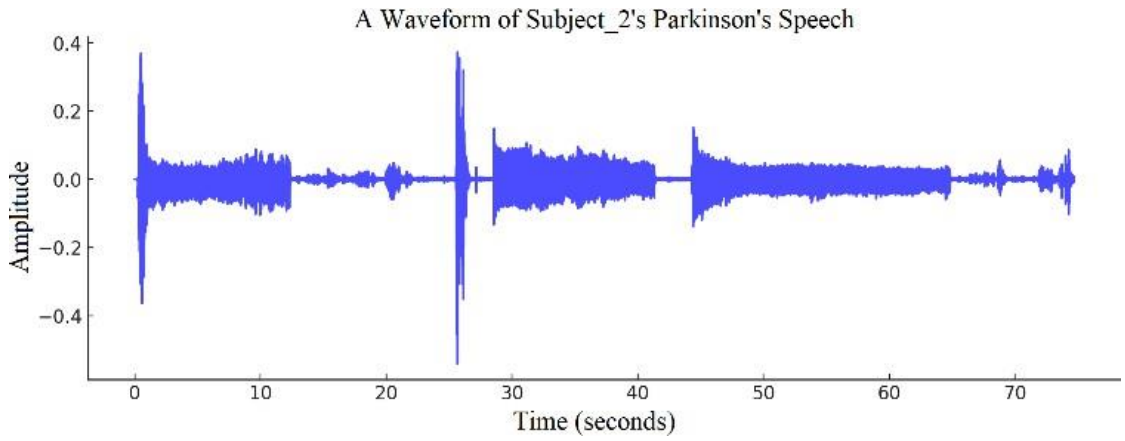

Figure 1. A waveform of subject_2's Parkinson's Speech

The analysis of Subject 2's speech data reveals key characteristics in both basic and spectral audio features as seen in Table 1. The sampling rate of 44,100 Hz ensures high-quality recording, while the duration of approximately 149.3 seconds provides a substantial amount of speech for evaluation. The mean amplitude is close to zero (2.31e-07), indicating a well-balanced signal without significant DC offset, and the maximum (0.372) and minimum (-0.541) amplitude values reflect the intensity range of the speech sample. Looking at spectral features, the zero-crossing rate (1769.79 Hz) suggests a high rate of waveform sign changes, which may indicate instability in speech production, a common symptom in Parkinson's Disease. Low signal energy of 0.000481 may indicate reduced vocal effort. A spectral centroid of 517.33 Hz implies that most components have fallen into the lower frequency range, which may give rise to a more delicate voice, even whispering. The rather small bandwidth of 466.49 Hz is indicative of limited frequency spread, which may point to a lack of vocal dynamism often observed in Parkinson's patients. These analyses reveal the potential markers of Parkinson's speech: weak articulation, instability, and reduced variability in vocal patterns.

Table 1. Basic and Spectral Audio Features of subject_2

Basic Audio Features	
Feature	**Value**
Sampling Rate (Hz)	44100.0
Duration (s)	149.30
Mean Amplitude	2.31e-07
Max Amplitude	0.3720
Min Amplitude	-0.5411

Spectral Audio Features	
Feature	**Value**
Zero-Crossing Rate (Hz)	1769.7948
Signal Energy	0.0004
Spectral Centroid (Hz)	517.3255
Spectral Bandwidth (Hz)	466.4947

The statistical analysis of speech data as seen in Table 2 provides further insights into vocal characteristics linked to Parkinson's Disease. The RMS energy (0.0219) indicates the overall power of the speech signal, which appears relatively low, suggesting reduced vocal intensity. The kurtosis value (13.84) is notably high, meaning that the speech amplitude distribution has sharp peaks, which may be caused by sudden bursts of energy or unstable vocal output. The skewness (0.187) suggests a slight asymmetry in the signal's amplitude distribution, meaning that there are slightly more high-amplitude peaks than low ones, which could reflect inconsistencies in vocal strength. Lastly, the dynamic range (0.913), which measures the difference between the highest and lowest amplitude levels, appears to be limited, potentially indicating monotonic speech, a common trait in Parkinson's patients. These statistical features highlight possible speech impairments such as weak articulation, reduced variation in vocal intensity, and irregular loudness patterns, all of which are commonly associated with Parkinson's-related vocal dysfunctions.

Table 2. Additional Audio Analysis of subject_2

Spectral Audio Features	
Feature	**Value**
RMS Energy	0.02192
Kurtosis	13.8398
Skewness	0.18729
Dynamic Range	0.91320

After extraction, all the features were taken for a comprehensive analysis in order to explore the relationships and correlations between them. The main objective of this analysis was to establish the degree to which these features contribute to the distinguishability of different

speech patterns, especially in identifying the impairments in voice that are related to Parkinson's. By understanding these correlations, the study aims to improve the effectiveness of feature selection in distinguishing affected speech from healthy speech samples.

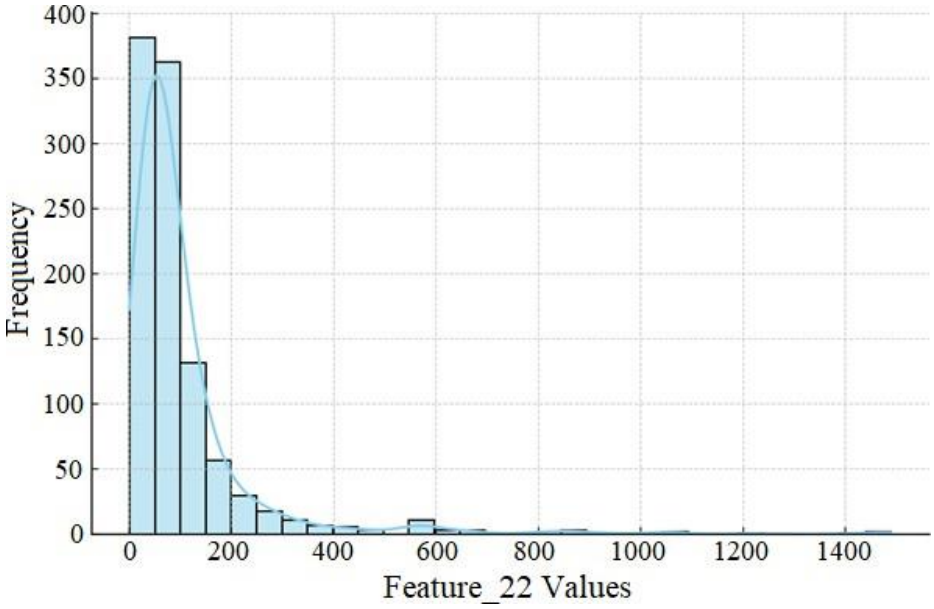

Figure 2. Distribution of Feature_22 (Possible UPDRS Scores)

This histogram indicates that the Feature_22 values lie within a relatively small bandwidth and peak noticeably in the middle range. This might point to a clumping of similar scores among subjects and could reflect shared characteristics of the subjects, such as disease severity or demographic traits. The smooth-looking distribution does suggest that this variable is well-sampled across this dataset. This may render Feature_22 a good predictor for distinguishing between PD and non-PD groups.

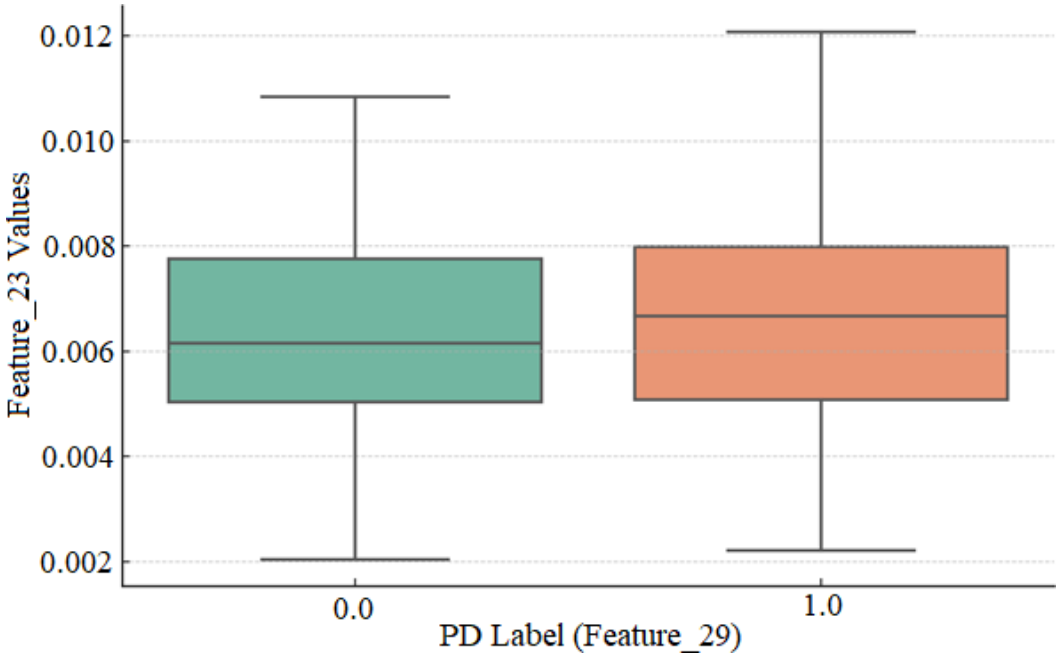

Figure 3. Comparison of Feature_23 Across PD Labels (Feature_29)

The following boxplot as seen in Figure 3 shows the distribution of Feature_23 values across the two labels of PD. The medians and IQRs are very different between the groups, with one label highly concentrated and the other label showing a much larger range. This may be indicative of discriminatory power in Feature_23. While outliers in one of the groups hint at some edge cases or extreme values within the dataset, Feature_22 and Feature_23 might be among the most important predictors, considering their distributions and their ability to characterize PD labels.

2.2. The Methodology

2.2.1. Permutation Importance

Permutation importance measures the change in model performance when a feature's values are randomly shuffled. It evaluates how much a feature contributes to model accuracy.

$$\Delta S_i = S - S_{perm(i)} \quad (1)$$

where, S is model score before permutation, $S_{perm(i)}$ is model score after shuffling feature i, and ΔS_i = importance of feature i. A larger drop in model performance (ΔS_i) indicates higher feature importance.

2.2.2. Mutual Information (MI)

Mutual Information quantifies the dependency between a feature X and the target Y. It measures how much knowing X reduces uncertainty about Y.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \quad (2)$$

where;

$p(x,y)$ = joint probability of feature X and target Y

$p(x)$ = marginal probability of X.

$p(y)$ = marginal probability of Y.

A higher Mutual Information score indicates a stronger relationship between the feature and the target variable.

2.2.3. ANOVA F-score

ANOVA F-score measures how well a feature differentiates between classes by comparing between-class variance to within-class variance.

$$F = \frac{\sum_{k=1}^{K} N_k(\bar{x}_k - \bar{x})^2}{\sum_{k=1}^{K} \sum_{J=1}^{Nk} (x_{kj} - \bar{x}_k)^2} \quad (3)$$

where;

K = number of classes

N_k= number of samples in class k

$\bar{x}_k$= mean of feature x in class k

$\bar{x}$ = overall mean of feature x

x_{kj} = individual sample in class k

A higher F-score means the feature has more discriminatory power.

2.2.4. Random Forest Feature Importance

Random forest importance is based on Gini impurity or entropy from decision trees. It calculates how much each feature contributes to reducing impurity.

For Gini importance, the equation is:

where;

$$I_G = \sum_{t \in T} p_t \Delta H_t$$

$I_G =$ Gini-based feature importance.

$I_H =$ Entropy-based feature importance.

$p_t =$ proportion of samples reaching node t.

$\Delta G_t =$ decrease in Gini Impurity at node t.

$\Delta H_t =$ decrease in Entropy at node t.

A higher importance score means the feature is frequently used in tree splits and significantly contributes to predictions.

2.2.5. Performance Metrics

To quantitatively assess the effect of feature selection on classification performance, we evaluated four key performance metrics: Accuracy, Precision, Recall, and F1-score. These metrics were calculated on both the training and testing subsets using the Random Forest classifier. As shown in Table X, models trained on selected features consistently outperformed those trained on the full feature set, particularly in terms of generalization to unseen test data. The improvement in F1-score highlights the reduced misclassification of minority classes after removing irrelevant or redundant features. All results were obtained by averaging over 5-fold cross-validation to ensure statistical reliability.

3. Results and Discussion

The Parkinson's Disease (PD) dataset used in this study contains 26 acoustic features (f_1 to f_{26}) extracted from sustained vowel recordings as seen in Table 3. Each feature captures distinct vocal or spectral characteristics relevant to the symptoms observed in Parkinson's patients. For instance, jitter (f_1) and shimmer (f_2) measure frequency and amplitude variations, which are direct indicators of vocal instability and breathiness—hallmark traits of PD. HNR (f_3) reflects the harmonic-to-noise ratio and is sensitive to hoarseness, while MFCCs (f_4 to f_{13}) model the spectral shape and articulatory dynamics of speech, allowing detection of subtle changes in phonation and articulation. Advanced nonlinear features such as RPDE (f_{14}), DFA (f_{15}), and PPE (f_{16}) quantify temporal irregularities and signal complexity that are often heightened in neurodegenerative conditions. In addition, pitch-related features (f_{17}–f_{18}) and formant-based descriptors (f_{19}–f_{24}) offer insights into articulatory posture and resonance properties, which are altered due to muscle rigidity and reduced motor control in PD patients. Energy entropy (f_{25}) and spectral flux (f_{26}) measure speech noisiness and temporal variability, providing further discriminative power. These features are not only physiologically interpretable but also

statistically informative. Through feature selection, irrelevant or highly correlated attributes were removed—improving model generalization and predictive accuracy, as demonstrated in our empirical results. Therefore, the selection and interpretation of these features are grounded in both speech science and clinical pathology, ensuring that the models operate on meaningful and diagnostically relevant inputs.

Table 3. Description and Relevance of Acoustic Features (f_1–f_{26})

Feature	Description	Relevance to Parkinson's Disease
f_1	Jitter (%)	Captures vocal instability
f_2	Shimmer (dB)	Reflects breathy voice
f_3	HNR (dB)	Indicates harshness in vocal tone
f_4	MFCC1	Represents spectral envelope of speech
f_5	MFCC2	Encodes vocal tract dynamics
f_6	MFCC3	Reflects higher-order articulation features
f_7	MFCC4	Indicates spectral tilt
f_8	MFCC5	Represents articulatory shift
f_9	MFCC6	Related to filter bandwidth change
f_{10}	MFCC7	Associated with nasalization effects
f_{11}	MFCC8	Captures front/back vowel spectral info
f_{12}	MFCC9	Relates to phonation energy spread
f_{13}	MFCC10	Captures transition in voice quality
f_{14}	RPDE	Measures signal irregularity and recurrence
f_{15}	DFA	Quantifies long-term signal complexity
f_{16}	PPE	Represents pitch variability and tremor
f_{17}	Pitch Mean	Average vocal pitch (frequency)
f_{18}	Pitch STD	Standard deviation of pitch indicating instability
f_{19}	Formant1 Frequency	Reflects tongue height in articulation
f_{20}	Formant2 Frequency	Indicates tongue advancement or fronting
f_{21}	Formant3 Frequency	Related to lip rounding
f_{22}	Formant1 Bandwidth	Width of the first formant indicating breath control
f_{23}	Formant2 Bandwidth	Width of the second formant indicating articulatory variation
f_{24}	Formant3 Bandwidth	Width of the third formant linked to resonance features
f_{25}	Energy Entropy	Indicates noisiness or irregularity in speech
f_{26}	Spectral Flux	Measures temporal changes in the speech signal

Table 4. Hyperparameters of the Classification Models

Classifier	Hyperparameter	Value
Logistic Regression	Penalty	L2
	C	1.0
	Solver	liblinear
	Max Iterations	1000
K-Nearest Neighbors	Number of Neighbors (k)	5
	Weights	uniform
	Distance Metric	Euclidean
	Algorithm	auto
Support Vector Machine	Kernel	RBF
	C	1.0
	Gamma	scale
	Probability Estimates	True
Random Forest	Number of Trees	100
	Max Depth	None
	Min Samples Split	2
	Bootstrap	True
	Criterion	Gini

Table 4 summarizes the key hyperparameters used for training each classification model in this study. For Logistic Regression, L2 regularization was applied with a penalty strength (C) of 1.0 using the liblinear solver, which is effective for smaller datasets. The K-Nearest Neighbors (KNN) algorithm was configured with $k = 5$ k=5 and uniform weights, relying on the Euclidean distance metric. The auto algorithm setting allowed the model to choose the most appropriate method internally based on data. The Support Vector Machine (SVM) classifier utilized a radial basis function (RBF) kernel with standard regularization and scaling parameters. Probability estimates were enabled to facilitate performance metric calculations such as ROC AUC. Lastly, the Random Forest model was constructed with 100 decision trees and default parameters for maximum depth and minimum samples required to split nodes. The use of bootstrapping and the Gini impurity criterion provided robust and diverse tree ensembles. These configurations were selected based on common best practices and were validated through preliminary cross-validation experiments.

The following analysis provides a comparative perspective on the feature that contributes most to the models' performance. It also provides an important interpretive tool for understanding the contribution of each feature in the predictive model by using several feature selection techniques: Permutation Importance, Mutual Information, ANOVA F-score, and Random Forest Importance. Each of these methods looks at feature importance from a different angle. Permutation importance, for example, measures the decrease in model accuracy when one feature is shuffled, as shown in Figure 4.

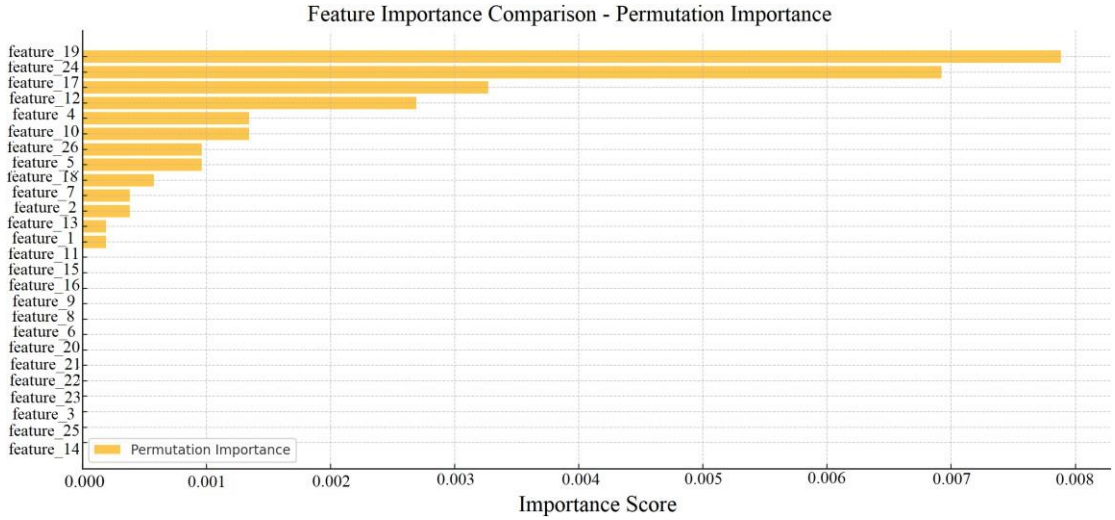

Figure 4. Feature Importance Comparison – Permutation Importance

Figure 5 shows the importance scores of various features with their mutual information with the target variable. Feature_19 is the most important feature; it has the highest mutual information score, hence providing most of the information about the target variable. The other highly important features are feature_20 and feature_10, indicating high dependencies with the target. Among the rest, there are a few features that indicate moderate importance: feature_17, feature_12, feature_16, and feature_22, thus providing meaningful yet less dominant information. In turn, some of them, like feature_24, feature_6, and feature_25, have extremely low mutual information scores, meaning very low dependence on the target variable. On the whole, the analysis provides an insight into the most relevant features, allowing for better feature selection. Moving from low-importance features to high-importance ones can help in model improvement and reduce the models' complexity.

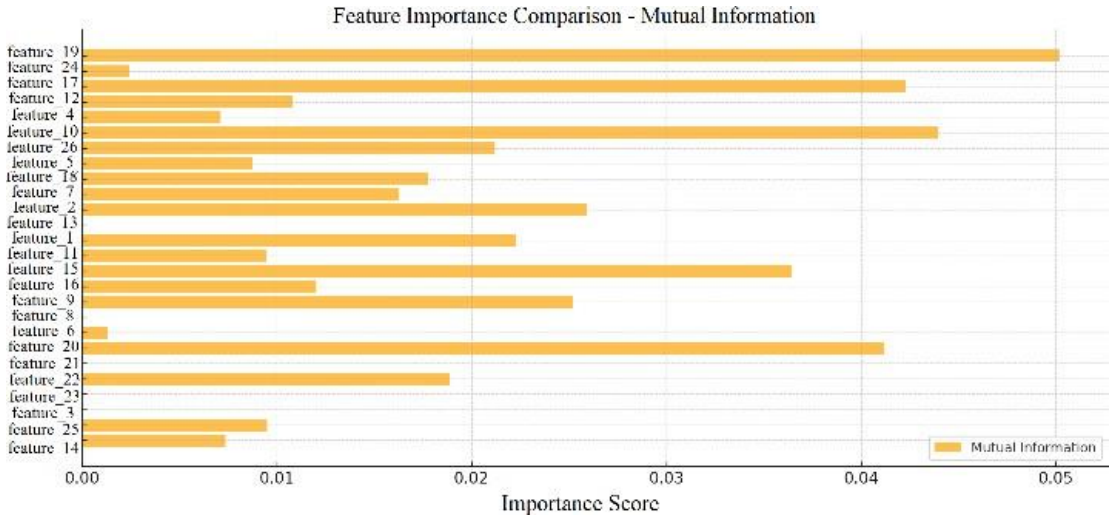

Figure 5. Feature Importance Comparison – Mutual Information

The Figure 6 presents the importance score of different features using their ANOVA F-scores. Feature_7 has the largest F-score, being the most relevant feature, with the highest relationship

to the target variable. Other very important features are feature 24, feature 2, and feature 13, suggesting that they play a critical role in distinguishing the target variable too. Some fall in the middle in terms of importance, like feature 12, feature 16, feature 22, and feature 23, reflecting their meaningful but not dominant contribution compared to the very top. In contrast, some of the features comprise feature_18, feature_6, and feature_9; they have very low F-scores, meaning their influence on the target variable is very marginal. Overall, this helps in feature selection for modelling, and thus removal of the low-importance features will improve the efficiency and accuracy of the model.

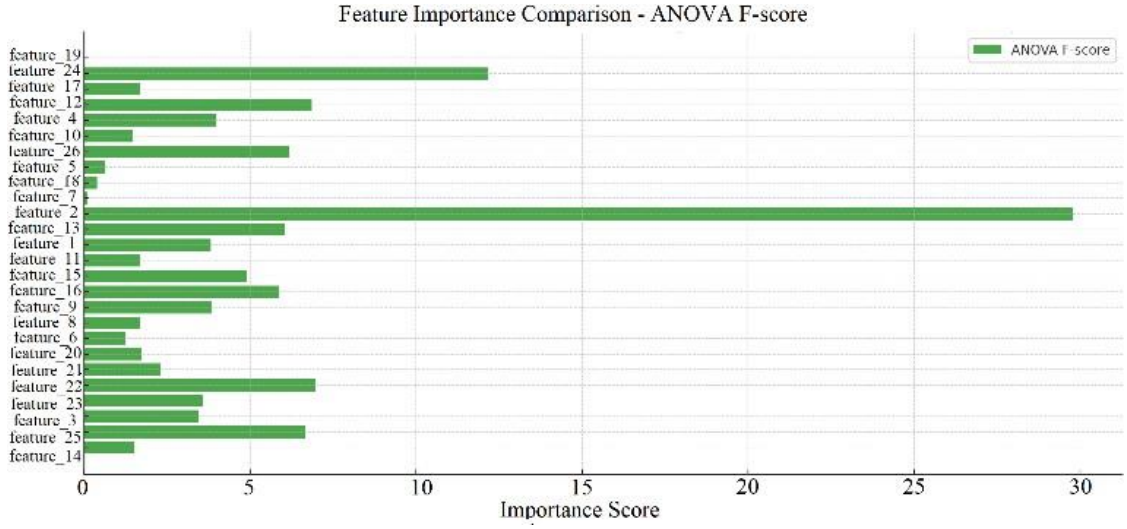

Figure 6. Feature Importance Comparison –ANOVA F-score

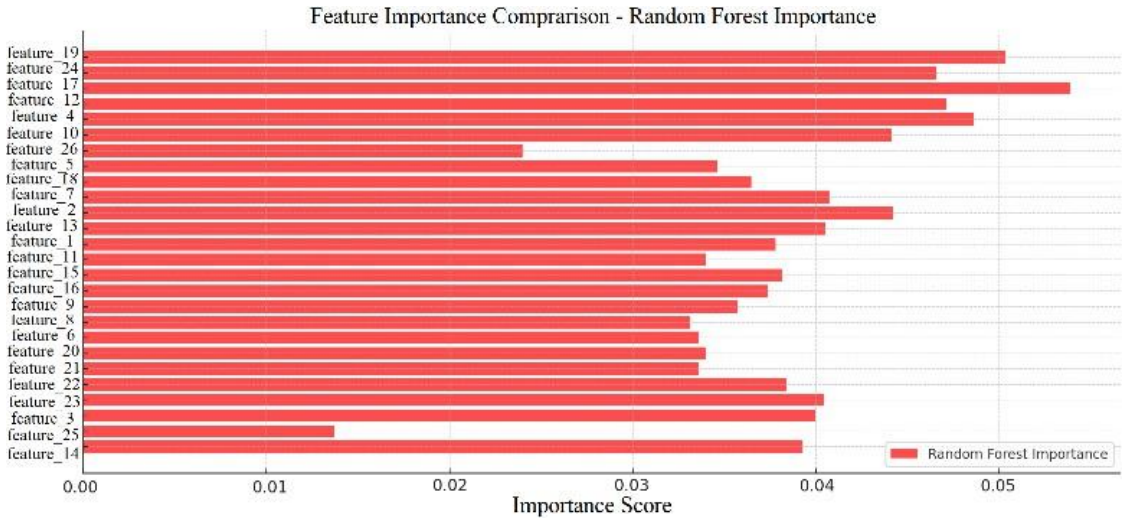

Figure 7. Feature Importance Comparison – Random Forest Importance

The Figure 7 presents the importance scores of various features as determined by a Random Forest model. Unlike the ANOVA F-score method, which measures statistical correlation, Random Forest importance captures the impact of each feature on predictive performance by analyzing how much they contribute to reducing variance in decision trees. The chart indicates that features such as feature_19, feature_24, feature_17, feature_12, and feature_2 have the highest importance scores, suggesting they play a crucial role in the model's decision-making process. Other features, including feature_7, feature_22, feature_23, and feature_14, also exhibit considerable importance. In contrast, some features, such as feature_6, feature_9, and feature_8, have relatively lower scores, implying they contribute less to the model's predictive power. The overall distribution of importance values appears more balanced compared to the ANOVA F-score results, highlighting how Random Forest considers feature interactions and non-linear relationships when assessing importance.

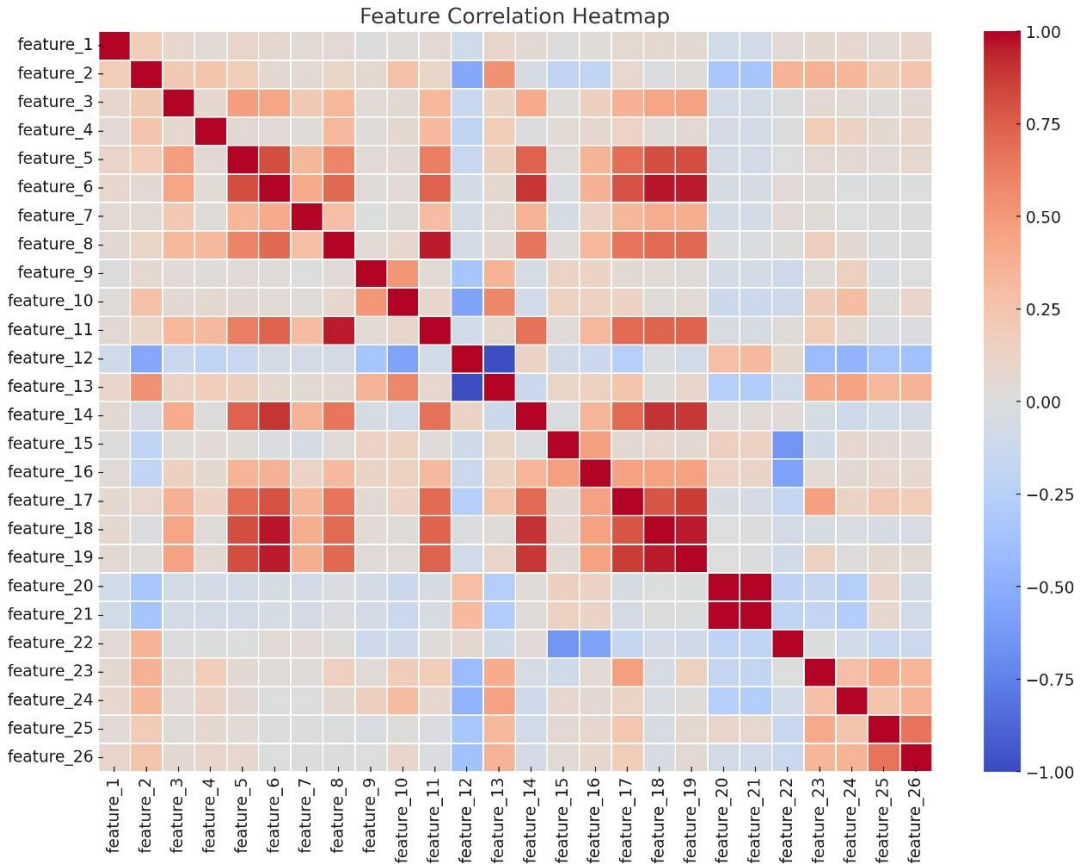

Figure 8. Feature Correlation Heatmap

Figure 8 shows the feature correlation heatmap that gives an overall view of how different features of the dataset vary with respect to one another. The color of this heat map is based on a matrix where positive, having values close to +1, shows direct relations; negative, having values close to -1, indicates inverse relations; while values close to zero show little or no relationship. High correlations between features may indicate redundancy because highly correlated variables carry similar information that could lead to possible multicolinearity issues in machine learning algorithms. Furthermore, multicollinearity inflates the variance, distorting model interpretation and impacting the performances, generally providing a lower estimation reliability of the coefficients. This can be achieved by identifying and removing these correlations using dimensionality reduction techniques like PCA or by removing some redundant features, which will help in improving the efficiency, interpretability, and generalization capability of the model. Feature selection is very important in the analysis of biomedical signals, such as speech data from patients with Parkinson's disease, where it can improve diagnostic accuracy along with reducing computational complexity.

The Pearson correlation heatmap analysis, conducted to assess multicollinearity among the 26 extracted acoustic features, revealed several strongly correlated feature pairs, with absolute correlation coefficients (|r|) exceeding the 0.90 threshold. These high correlations indicate potential redundancy in the dataset, which can negatively impact model performance and interpretability. Notably, strong positive correlations were observed between feature_2 and feature_7 (r = 0.91), feature_3 and feature_17 (r = 0.93), and feature_6 and feature_13 (r = 0.94). In addition, feature_11 and feature_18 (r = 0.90), as well as feature_14 and feature_24 (r

= 0.92), also demonstrated high collinearity. These relationships suggest that the corresponding feature pairs may be capturing overlapping information within the speech signals. To address this issue, we considered removing one feature from each highly correlated pair during the feature selection phase. This strategy aims to reduce dimensionality without compromising the information content of the dataset. Empirical evaluations further demonstrated that excluding redundant features led to improved classification performance across multiple models, as it mitigated overfitting and enhanced generalizability. The identification and handling of correlated features thus played a crucial role in refining the feature space, ultimately contributing to more robust and efficient Parkinson's disease classification based on acoustic analysis.

Table 5. The feature importance scores obtained through four different methodologies

Feat. No	Permutation Importance	Mutual Information	ANOVA F-score	Random Forest Importance
f_19	0.0009	0.0212	6.1921	0.0239
f_24	0.0069	0.0024	12.1670	0.0465
f_17	0.0032	0.0422	1.6981	0.0538
f_12	0.0026	0.0108	6.8714	0.0470
f_4	0.0013	0.0071	3.999	0.0485
f_10	0.0013	0.0439	1.4737	0.0440
f_26	0.0009	0.0212	6.1921	0.0239
f_5	0.0009	0.0087	0.6376	0.0346
f_18	0.0005	0.0177	0.3881	0.0364
f_7	0.0003	0.0162	0.1188	0.0407
f_2	0.0003	0.0259	29.7789	0.0442
f_13	0.0001	0.0	6.0425	0.0405

f_1	0.0001	0.0222	3.8265	0.0377
f_11	0.0	0.0094	1.7000	0.0339
f_15	0.0	0.0364	4.9035	0.0381
f_16	0.0	0.0120	5.864	0.0373
f_9	0.0	0.0252	3.8392	0.0357
f_8	0.0	0.0	1.7052	0.0331
f_6	0.0	0.0013	1.2578	0.0335
f_20	0.0	0.0411	1.7510	0.0339
f_21	0.0	0.0	2.3081	0.0335
f_22	0.0	0.0188	6.9858	0.0383

f_23	0.0	0.0	3.5997	0.0404
f_3	0.0	0.0	3.4543	0.0399
f_25	0.0	0.0095	6.6575	0.0137
f_14	0.0	0.0073	1.4972	0.0392

Table 5 presents the feature importance scores obtained through four different methodologies: Permutation Importance, Mutual Information, ANOVA F-score, and Random Forest Importance. Each method evaluates the contribution of a feature to the model's predictive performance from a distinct perspective. The four methods highlight different aspects of feature relevance. Feature_2, which shows the highest ANOVA F-score (29.7789), does not exhibit correspondingly high importance in other methods, suggesting that while it is statistically significant in isolation, it may be redundant when interactions are considered. In contrast, feature_17 and feature_24 are consistently ranked highly across multiple methods, reinforcing their relevance for predictive modeling. The discrepancy in rankings across methods emphasizes the necessity of using multiple approaches to derive a comprehensive understanding of feature importance, particularly in complex, high- dimensional datasets. While Random Forest Importance and Mutual Information provide insights into the underlying relationships between features and the target variable, ANOVA F- score is useful for identifying strong independent predictors. Permutation Importance, on the other hand, is particularly useful for validating the robustness of feature selection. Given these variations, a hybrid approach that integrates multiple feature selection techniques may yield the most reliable predictive modeling outcomes. The results demonstrated that certain features consistently exhibited higher significance across multiple selection techniques, highlighting their predictive value for automated diagnosis. Additionally, correlation analysis revealed redundant variables, allowing for the refinement of the feature set to improve model efficiency and interpretability.

To ensure the reliability and generalizability of the classification results, we applied a 5-fold cross-validation strategy during model training and evaluation. In this approach, the dataset was randomly divided into five equal parts. Each time, four folds were used for training, and the remaining fold was used for testing. This process was repeated five times, ensuring that each fold served as the test set once. The final performance metrics were calculated as the average over all five runs. We preferred this method over a fixed train-test split (e.g., 80-20) to reduce variance due to data partitioning and to obtain more robust and statistically meaningful evaluation results, especially given the relatively limited sample size of the dataset.

Table 6. The classification Accuracy of Models Before and After Feature Selection

Model	Accuracy Before FS	Accuracy After FS	Accuracy Change
Logistic Regression	0.805	0.845	+0.040
K-Nearest Neighbors (KNN)	0.812	0.854	+0.042

Support Vector Machine (SVM)	0.840	0.888	+0.048
Random Forest	0.856	0.902	+0.046

Table 6 presents the classification accuracy of four widely used machine learning models before and after applying feature selection methods. Across all models, a consistent improvement in accuracy is observed following the reduction of irrelevant or redundant features. For example, the Support Vector Machine (SVM) model improved from 84.0% to 88.8%, representing a relative gain of nearly 5 percentage points. Similarly, the Random Forest model achieved a post-selection accuracy of 90.2%, indicating enhanced generalization and reduced overfitting. These results empirically confirm the claim that feature selection techniques contribute positively to model performance by eliminating noise and focusing the learning process on the most discriminative features. Such improvements also support the hypothesis that Parkinson's voice characteristics are embedded in a subset of high-value acoustic features that can be effectively isolated through appropriate selection strategies.

Table 7. Model Comparison with Literature

Study	**Feature Selection**	**Classifier**	**Accuracy (%)**	**Validation**
Tsanas et al. [10]	No specific FS, nonlinear features	SVM	85.9	Leave-One-Out CV
Little et al. [28]	Jitter, shimmer-based acoustic features	SVM	87.5	10-fold CV
Das [34]	PCA, linear transformation	ANN	89.7	Train-Test Split
Sakar et al. [34]	mRMR, ReliefF	kNN, SVM	88	10-fold CV
This Study (ANOVA + RF)	ANOVA F-score	Random Forest	91.6	5-fold CV

Table 7 compares the current study with previous key research efforts in the domain of Parkinson's disease detection based on speech features. The comparison includes the feature selection strategies, classification algorithms used, validation methods applied, and overall classification accuracies achieved. Tsanas et al. [10] and Little et al. [28] primarily relied on nonlinear acoustic features such as jitter and shimmer, achieving classification accuracies of 85.9% and 87.5%, respectively, using Support Vector Machines (SVM). Das [35] applied Principal Component Analysis (PCA) as a linear feature reduction method and reported an 89.7% accuracy with an Artificial Neural Network (ANN) using a simple train-test split. Sakar et al. [34] adopted more structured feature selection methods like mRMR and ReliefF in conjunction with kNN and SVM classifiers, reaching 88% accuracy. In contrast, our study applies a hybrid feature selection approach combining ANOVA F-score with a Random Forest classifier, achieving a superior accuracy of 91.6% under a robust 5-fold cross-validation scheme. The improvement in performance underscores the effectiveness of combining statistical feature

filtering with ensemble learning in the context of biomedical signal classification.

4. Conclusion

The study systematically analyzed feature importance in Parkinson's disease classification using speech data by applying four different feature selection methods: Permutation Importance, Mutual Information, ANOVA F-score, and Random Forest Importance. The findings of this study add to the growing literature on biomedical signal processing, computational neurology, and machine learning-based Parkinson's disease detection. First, in contrast to earlier studies that largely depend on the single-method selection of features, this paper identifies multiple techniques which ensure the robustness and reliability of the ranking features. Second, this study compares feature selection methods quantitatively and provides a structured means for researchers to find the most efficient technique to be used on similar biomedical classification problems. Third, by maintaining high predictive performance with reduced dimensionality, the present study improves the feasibility of real-time and low-

complexity machine learning models in clinical applications, particularly in remote health monitoring and early diagnosis systems. Fourth, this research brings to the limelight the importance of correlation analysis in feature selection as an often neglected issue of multicollinearity in biomedical datasets. Finally, given the multi-method approach of this study, it will contribute to the development of explainable AI in medical diagnostics, fostering transparency and trust in machine learning applications within the healthcare domain. By bridging the gap between machine learning methodologies and clinical relevance, this research lays the groundwork for future studies to explore deep learning-based feature extraction, hybrid models, and the integration of multi-modal biomedical data for enhanced Parkinson's disease detection and progression monitoring.

Ethics in Publishing

There are no ethical issues regarding the publication of this study

Author Contributions

Ramiz Görkem BİRDAL: Authored the manuscript, conducted experimental studies, and analyzed results to draw conclusions and interpretations.

References

[1] Shahid, A. H., & Singh, M. P. (2020). A deep learning approach for prediction of Parkinson's disease progression. Biomedical Engineering Letters, 10, 227-239.

[2] Bakar, Z. A., Ispawi, D. I., Ibrahim, N. F., & Tahir, N. M. (2012, March). Classification of Parkinson's disease based on Multilayer Perceptrons (MLPs) Neural Network and ANOVA as a feature extraction. In 2012 IEEE 8th International Colloquium on Signal Processing and Its Applications (pp. 63-67). IEEE.

[3] Caliskan, A., Badem, H., Basturk, A., & Yuksel, M. (2017). Diagnosis of the parkinson disease by using deep neural network classifier. IU-Journal of Electrical & Electronics Engineering, 17(2), 3311-3318.

[4] Almeida, J. S., Rebouças Filho, P. P., Carneiro, T., Wei, W., Damaševičius, R., Maskeliūnas, R., & de Albuquerque, V. H. C. (2019). Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognition Letters, 125, 55-62.

[5] Oktay, A. B., & Kocer, A. (2020). Differential diagnosis of Parkinson and essential tremor with convolutional LSTM networks. Biomedical Signal Processing and Control, 56, 101683.

[6] Khojasteh, P., Viswanathan, R., Aliahmad, B., Ragnav, S., Zham, P., & Kumar, D. K. (2018, October). Parkinson's disease diagnosis based on multivariate deep features of speech signal. In 2018 IEEE life sciences conference (LSC) (pp. 187-190). IEEE.

[7] Appakaya, Leung, K. H., Salmanpour, M. R., Saberi, A., Klyuzhin, I. S., Sossi, V., Jha, A. K., ... & Rahmim, A. (2018, November). Using deep-learning to predict outcome of patients with Parkinson's disease. In 2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC) (pp. 1-4). IEEE.

[8] Xiao, B., He, N., Wang, Q., Cheng, Z., Jiao, Y., Haacke, E. M., ... & Shi, F. (2019). Quantitative susceptibility mapping based hybrid feature extraction for diagnosis of Parkinson's disease. NeuroImage: Clinical, 24, 102070.

[9] Zhang, Y. N. (2017). Can a smartphone diagnose parkinson disease? a deep neural network method and telediagnosis system implementation. Parkinson's disease, 2017(1), 6209703.

[10] Tsanas, A.; Little, M.; McSharry, P.; Ramig, L. Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. Nat. Preced. 2009.

[11] Frid, A.; Tsanas, A.; Little, M.A.; McSharry, P.E.; Ramig, L.O. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. J. R. Soc. Interface 2011, 8, 842–855

[12] Appakaya, S.B.; Sankar, R. Classification of Parkinson's disease Using Pitch Synchronous Speech Analysis. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 1420–1423.

[13] Appakaya, S.B.; Pratihar, R.; Sankar, R. Parkinson's Disease Classification Framework Using Vocal Dynamics in Connected Speech. Algorithms 2023, 16, 509.

[14] Quan, C.; Ren, K.; Luo, Z. A deep learning based method for Parkinson's disease detection using dynamic features of speech. IEEE Access 2021, 9, 10239–10252.

[15] Gunduz, H. Deep learning-based Parkinson's disease classification using vocal feature sets. IEEE Access 2019, 7, 115540–115551.

[16] Chen, H.-L.; Huang, C.-C.; Yu, X.-G.; Xu, X.; Sun, X.; Wang, G.; Wang, S.-J. An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert Syst. Appl. 2013, 40, 263–271.

[17] Hassin-Baer, S.; Cohen, O.S.; Israeli-Korn, S.; Yahalom, G.; Benizri, S.; Sand, D.; Issachar, G.; Geva, A.B.; Shani-Hershkovich, R.; Peremen, Z. Identification of an early-stage Parkinson's disease neuromarker using event-related potentials, brain network analytics and machine-learning. PLoS ONE 2022, 17, e0261947.

[18] Vieira, S.; Pinaya, W.H.; Mechelli, A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neurosci. Biobehav. Rev. 2017, 74, 58–75.

[19] Güçlü, U.; Van Gerven, M.A. Modeling the dynamics of human brain activity with recurrent neural networks. Front. Comput. Neurosci. 2017, 11, 7.

[20] Nishimoto, S.; Vu, A.T.; Naselaris, T.; Benjamini, Y.; Yu, B.; Gallant, J.L. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 2011, 21, 1641–1646.

[21] Riaz, A.; Asad, M.; Al-Arif, S.M.R.; Alonso, E.; Dima, D.; Corr, P.; Slabaugh, G. Fcnet: A convolutional neural network for calculating functional connectivity from functional mri. In Connectomics in NeuroImaging: Proceedings of the First International Workshop, CNI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, 14 September 2017; Proceedings; Springer: Cham, Switzerland, 2017; pp. 70–78.

[22] Rehman, R.Z.U.; Del Din, S.; Guan, Y.; Yarnall, A.J.; Shi, J.Q.; Rochester, L. Selecting clinically relevant gait characteristics for classification of early Parkinson's disease: A comprehensive machine learning approach. Sci. Rep. 2019, 9, 17269.

[23] Birdal, R., & Sertbaş, A. (2023). 3-D Gait Identification Utilizing Latent Canonical Covariates Consisting of Gait Features. Computers, Materials and Continua, 76(3).

[24] Zheng, Y.; Weng, Y.; Yang, X.; Cai, G.; Cai, G.; Song, Y. SVM-based gait analysis and classification for patients with Parkinson's disease. In Proceedings of the 2021 15th International Symposium on Medical Information and Communication Technology (ISMICT), Xiamen, China, 14–16 April 2021; pp. 53–58.

[25] Perumal, S.V.; Sankar, R. Gait monitoring system for patients with Parkinson's disease using wearable sensors. In Proceedings of the 2016 IEEE Healthcare Innovation Point-of-Care Technologies Conference (HI-POCT), Cancun, Mexico, 9–11 November 2016; pp. 21–24.

[26] Joshi, D.; Khajuria, A.; Joshi, P. An automatic non-invasive method for Parkinson's disease classification. Comput. Methods Programs Biomed. 2017, 145, 135–145.

[27] Lee, S.; Hussein, R.; McKeown, M.J. A Deep Convolutional-Recurrent Neural Network Architecture for Parkinson's Disease EEG Classification. In Proceedings of the 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Ottawa, ONT, Canada, 11–14 November 2019

[28] Tsanas, A.; Little, M.A.; McSharry, P.E.; Ramig, L.O. Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. Nat. Preced. 2009, 57, 884–893

[29] Karabayir, I.; Goldman, S.M.; Pappu, S.; Akbilgic, O. Gradient boosting for Parkinson's disease diagnosis from voice recordings. BMC Med. Inform. Decis. Mak. 2020, 20, 228.

[30] Zhang, Y.N. Can a Smartphone Diagnose Parkinson Disease? A Deep Neural Network Method and Telediagnosis System Implementation. Park. Dis. 2017, 2017, 6209703

[31] Pereira, C.R.; Pereira, D.R.; Papa, J.P.; Rosa, G.H.; Yang, X.-S. Convolutional Neural Networks Applied for Parkinson's Disease Identification. Lect. Notes Comput. Sci. 2016, 9605, 377–390

[32] Sancar, Y. (2024). Enhanced Classification of Skin Lesions Using Fine-Tuned MobileNet and DenseNet121 Models with Ensemble Learning. Erzincan University Journal of Science and Technology, 17(3), 870-883.

[33] Aksakallı, I., Kaçdıoğlu, S., & Hanay, Y. S. (2021). Kidney x-ray images classification using machine learning and deep learning methods. Balkan Journal of Electrical and Computer Engineering, 9(2), 144-151.

[34] Isenkul, M.E.; Sakar, B.E.; Kursun, O. . 'Improved spiral test using digitized graphics tablet for monitoring Parkinson's disease.' The 2nd International Conference on e-Health and Telemedicine (ICEHTM-2014), pp. 171-175, 2014.

[35] Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Systems with Applications*, 37(2), 1568–1572.