[University Name]

[Department/School Name]

Voice-Based Classification of Parkinson's Disease Using Classical Machine Learning

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in [Program Name]

by

[Author Full Name]

Supervisor: [Supervisor Name]

January 2026

Abstract

Parkinson's Disease (PD) is a neurodegenerative disorder characterized by motor symptoms and pervasive speech impairments. This thesis investigates the feasibility of voice-based PD detection using classical machine learning, emphasizing rigorous methodology over maximal performance. Two complementary datasets are examined: Dataset A, a clinical corpus of raw voice recordings (37 subjects) requiring acoustic feature extraction; and Dataset B, a larger public dataset (756 samples) of pre-extracted features. A consistent pipeline is applied, extracting 47 baseline features (prosodic and perturbation measures) from Dataset A, with an extended set of 78 features incorporating additional spectral descriptors. Three interpretable classifiers—Logistic Regression, Support Vector Machine (RBF kernel), and Random Forest—are evaluated under a 2×2 factorial design: baseline vs. extended features, with vs. without class weighting to address class imbalance. Crucially, subject-grouped 5-fold cross-validation is employed for Dataset A to prevent data leakage, while a standard stratified 5-fold CV (with caveats on subject overlap) is used for Dataset B.

Results are reported as mean ± standard deviation. On Dataset A, the best model (Random Forest, extended features) achieved ROC-AUC $\approx 0.87 \pm 0.14$, a +8.7 percentage point improvement over the baseline feature set. Extended features consistently improved accuracy and ROC-AUC, especially for the smaller Dataset A (e.g., Random Forest AUC rose from 0.59 to 0.82 on one task). Class weighting had only modest effects (e.g., +3.5pp ROC-AUC for Random Forest with baseline features, but negligible or negative impact with extended features). Random Forest outperformed SVM and Logistic Regression across conditions, likely due to its ability to capture non-linear patterns and leverage feature importance for insight. Dataset B yielded higher absolute performance (ROC-AUC ≈ 0.94 with Random Forest) but is interpreted with caution given potential subject overlaps and its high-dimensional feature set.

In conclusion, classical ML models can detect PD from voice with competitive accuracy, but robust validation is paramount. This work highlights that methodological rigor—including proper cross-validation, careful feature engineering, and honest reporting of variance and limitations—is essential to produce reliable findings. The extended feature set notably enhances detection of PD voice signatures, and results underscore the importance of addressing data leakage and class imbalance. These contributions lay a reproducible groundwork for future research, prioritizing interpretability and validity in the development of non-invasive PD screening tools.

Keywords: Parkinson's Disease; Dysarthria; Voice Biomarkers; Acoustic Features; Machine Learning; Cross-Validation; Imbalanced Data; Reproducibility

Acknowledgments

[Write your acknowledgments here.]

I would like to express my sincere gratitude to...

- My supervisor, [Name], for guidance and support throughout this research
- The creators of the MDVR-KCL dataset for making their data publicly available

l acknowledgments

[Author Name]

[City], January 2026

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Parkinson's Disease (PD) is the second most prevalent neurodegenerative disorder globally, affecting approximately 1% of the population over 60 years of age [7]. Early and accurate detection remains a critical clinical challenge, as motor symptoms often manifest only after substantial neurological damage has occurred. Among the earliest observable symptoms are changes in speech and voice production, which can precede motor symptoms by several years [6].

Voice-based biomarkers offer a promising non-invasive avenue for PD detection [11, 21]. The disease affects the laryngeal and respiratory muscles, resulting in measurable changes to prosodic features (pitch, loudness, rhythm) and spectral characteristics (formant frequencies, harmonic structure). PD speech is often characterized by *hypokinetic dysarthria*, a motor speech disorder marked by reduced voice loudness (*hypophonia*), a limited pitch range (*monopitch*), and monotonous volume (*monoloudness*) [5]. Patients may also exhibit articulatory imprecision (unclear consonant enunciation) and voice quality changes such as breathiness or hoarseness. These acoustic signatures can be captured using standard microphones, making voice analysis a cost-effective and accessible approach for screening and monitoring PD. Moreover, subtle vocal abnormalities may appear even before classic motor symptoms in some patients, highlighting the potential of voice as an early indicator.

1.2 Problem Statement

Despite advances in voice-based PD classification, several methodological challenges persist:

1. **Small sample sizes** in raw audio datasets limit model generalizability
2. **Subject identity leakage** when multiple recordings per subject are split across folds
3. **Class imbalance** between PD and healthy control (HC) groups
4. **Feature representation choices** significantly impact classification performance

This thesis addresses these challenges through a rigorous experimental framework that prioritizes methodological validity over raw performance metrics.

1.3 Research Objectives

The primary objectives of this research are:

1. **Develop a reproducible pipeline** for extracting acoustic features from voice recordings
2. **Evaluate classical machine learning models** (Logistic Regression, SVM, Random Forest) for PD vs HC classification
3. **Compare performance** across two distinct datasets with different characteristics
4. **Investigate the impact** of feature set extension (47 → 78 features) through controlled ablation
5. **Assess the effect** of class weighting on imbalanced datasets

1.4 Contributions

This thesis makes the following contributions:

- A **subject-grouped cross-validation framework** for voice data that prevents data leakage. By grouping recordings by subject in cross-validation splits, we ensure that no speaker's recordings appear in both training and test sets, addressing a common pitfall in PD voice studies.
- A **controlled feature ablation study** demonstrating substantial improvements in classification performance (up to +23 percentage points in ROC-AUC) by extending the feature set from 47 to 78 features. We show which additional features (e.g., variability measures and spectral shape descriptors) drive the performance gains.

- **Task-specific analysis** revealing that spontaneous, free-form speech yields higher PD detection performance (e.g., Random Forest ROC-AUC 0.857 on spontaneous speech) compared to read speech (ROC-AUC 0.822 on a standard reading passage). This suggests that less structured vocal tasks may contain richer PD cues.

- **Benchmarking analysis** contrasting our rigorous validation on Dataset A with results on a larger public dataset (Dataset B). We highlight that standard cross-validation on Dataset B (which lacks subject IDs) produces optimistic estimates (Random Forest AUC ∼0.94), underscoring the importance of subject-aware evaluation for realistic performance assessment.

1.5 Thesis Organization

The remainder of this thesis is organized as follows:

Chapter	Title	Description
2	Literature Review	Survey of voice-based PD detection methods
3	Data Description	Detailed analysis of datasets used
4	Methodology	Feature extraction and ML pipeline design
5	Experimental Design	Cross-validation and evaluation protocols
6	Results	Quantitative findings across all conditions
7	Discussion	Interpretation and comparison with literature
8	Limitations	Constraints and threats to validity
9	Conclusion	Summary and future directions

Table 1.1: Thesis chapter overview

1.6 Scope Boundaries

This research is explicitly bounded by the following constraints:

- **Binary classification only** — We focus on distinguishing PD vs. healthy controls. The work does not address prediction of disease severity, progression, or differential diagnosis against other disorders.

- **Classical machine learning models** — We restrict our study to interpretable, classical algorithms (Logistic Regression, SVM, Random Forest). No deep learning or neural network models are used, given the small dataset size and our emphasis on interpretability.

- **Research context** — The models and results are intended for research demonstration and are not directly deployed as clinical diagnostic tools. We do not claim clinical utility without further validation.

- **Reproducibility prioritized** — We emphasize reproducible experimentation (with fixed random seeds, documented code, and shared data processing) over chasing state-of-the-art accuracy. All code and data usage adheres to best practices to ensure results can be independently verified.

Chapter 2

Literature Review

2.1 Parkinson's Disease and Speech Impairment

Parkinson's disease is a progressive neurodegenerative disorder primarily known for its motor symptoms (tremor, rigidity, bradykinesia). In addition to these, PD almost invariably affects speech and voice as the disease progresses. It is reported that approximately 70–90% of individuals with PD develop measurable speech and voice impairments over the course of the illness [7]. This collection of speech symptoms in PD is often referred to as *hypokinetic dysarthria*, denoting a characteristic pattern of speech motor control impairment associated with the disease [5].

The speech of a person with PD typically exhibits several hallmark changes. One prominent feature is *hypophonia*, or reduced voice loudness—patients often speak in a much softer voice than normal. Another is a monotonic pitch: PD speakers tend to have a limited pitch range, resulting in speech that lacks the normal ups and downs of intonation (often described as "monopitch" speech). Monoloudness (abnormally uniform volume) often accompanies this, so the overall prosody (melody and expressiveness of speech) is markedly diminished. Patients may also exhibit articulatory imprecision, where consonants are not enunciated crisply. For example, consonant sounds may blur together or be undershot due to reduced range of motion in the articulators (jaw, tongue, lips). The voice quality in PD is frequently described as breathy or hoarse, reflecting incomplete vocal fold closure and other phonatory deficits. Additionally, some individuals speak with an improperly fast rate or with short rushes of speech, which—combined with the articulation issues—can reduce intelligibility [19]. These speech characteristics—reduced loudness, monopitch, monoloudness, imprecise articulation, and breathy/hoarse voice—are widely observed in PD and form the basis of clinical descriptions of hypokinetic dysarthria [15].

Crucially, speech changes in PD are of interest not just as symptoms affecting communication, but also as potential non-invasive biomarkers of the disease. Voice is relatively easy to capture (e.g., via a short recording on a phone), and vocal changes can manifest early in the disease course. Some research suggests that subtle voice abnormalities may appear even before classic motor symptoms in certain patients [6]. Because voice recording and analysis can be done inexpensively and remotely, there is considerable motivation to use speech as a way to detect or monitor PD without the need for invasive tests. Speech and voice metrics are appealing for telemedicine and longitudinal tracking of PD progression [21]. Unlike many clinical assessments that require in-person visits and specialized equipment, voice recordings can be obtained by patients at home and sent to clinicians or analyzed by algorithms, enabling more frequent monitoring.

It should be noted, however, that the speech impairments in PD can vary greatly across patients and disease stages. Not every person with PD will have all the aforementioned speech symptoms, and the severity can range from very mild to highly debilitating. There is variability in how early voice changes emerge: some patients present with noticeable hypophonia and monotonous speech in the early stages, whereas others might have minimal speech impact until later in the disease. Moreover, the progression of speech symptoms does not always strictly parallel the progression of other motor symptoms. For example, a patient with advanced limb tremor might still speak relatively clearly, while another patient with otherwise mild motor symptoms could have pronounced dysarthria. This variability underscores the need for personalized approaches in voice-based assessment.

2.2 Acoustic Characteristics of Parkinsonian Speech

A variety of acoustic features have been explored to characterize the distinctive patterns of Parkinsonian speech. These features quantify specific aspects of the voice signal that are hypothesized to change due to PD. Broadly, prior studies have looked at **prosodic features**, **perturbation measures**, and **spectral/cepstral features** to capture different dimensions of vocal impairment.

2.2.1 Prosodic Features

Prosodic features relate to the pitch (fundamental frequency) and loudness (intensity) patterns in speech, as well as timing and rhythm to some extent. The fundamental frequency of speech (perceived as pitch) is often denoted as F_0. In PD, prosodic modulation is reduced: PD patients typically exhibit a lower variability in F_0 and intensity over an utterance. In practical terms, this means their speech has a flatter intonation and a narrower dynamic range. Key prosodic features examined include: F_0

mean, minimum, maximum, and standard deviation, which reflect overall pitch level and variability; intensity mean and variability, reflecting loudness and its modulation; and speech rate or pause duration (though rate is sometimes considered separately). Monotony in pitch and loudness (low F_0 std and low intensity range) is a classic sign of PD speech [19]. These prosodic deficits correspond to the perceptual impressions of monopitch and monoloudness described earlier. By measuring them quantitatively (e.g., computing the standard deviation of F_0 across an utterance, or the range between maximum and minimum intensity), researchers can objectively gauge the extent of prosodic impairment. Prosodic feature extraction often involves algorithms that track pitch (via autocorrelation or cepstral methods) and energy on a frame-by-frame basis, using tools like Praat or librosa [8, 12].

2.2.2 Perturbation Measures

Perturbation measures capture the cycle-to-cycle variations in the voice signal, reflecting stability (or instability) of vocal fold vibration. The two primary categories are **jitter** (pertaining to frequency instability) and **shimmer** (pertaining to amplitude instability). Jitter is usually defined as the percentage variation in fundamental period between consecutive glottal cycles; PD voices often have elevated jitter, indicating irregular pitch periods. Shimmer is the percentage variation in amplitude of consecutive cycles; it tends to be higher in PD, indicating inconsistent loudness from cycle to cycle. Essentially, increased jitter and shimmer correspond to a harsher, more breathy voice quality with less stable tone—consistent with PD-related vocal tremor and weakness. Commonly used perturbation features include local jitter (%), jitter variants like RAP (relative average perturbation) and PPQ, and shimmer measures like local shimmer, shimmer APQ3, APQ5, APQ11, etc.

Studies (starting from the classic work of Little et al. and others) found that these perturbation metrics can distinguish PD voices from healthy voices to a significant extent. For example, Little et al. [11] used a set of 22 features largely composed of jitter, shimmer, and related measures and achieved high accuracy in classifying PD vs HC with an SVM. Perturbation features are typically extracted from sustained vowel recordings (e.g., sustained "ah" sounds) where cycle-to-cycle analysis is most reliable, but they can also be computed on longer speech if voiced segments are isolated.

Harmonics-to-Noise Ratio (HNR) is another related metric, comparing the level of periodic (harmonic) energy in the voice to aperiodic or noise energy. HNR quantifies the proportion of harmonic (periodic) energy to noise (aperiodic energy) in the voice. PD voices often have lower HNR, indicating a breathier, noisier signal due to imperfect vocal fold vibration.

2.2.3 Spectral and Cepstral Features

Spectral and cepstral features analyze the frequency-domain characteristics of speech. While prosodic features capture global patterns over time and perturbation features capture cycle-level stability, spectral features provide information about the distribution of energy across frequency bands and the overall quality of the voice signal. One widely used set of spectral features in speech analysis is the **Mel-Frequency Cepstral Coefficients (MFCCs)**. MFCCs are a compressed representation of the spectral envelope of the sound, using a perceptually motivated mel scale. In PD research, MFCCs (and their derivatives) have been employed to capture vocal tract resonances and changes due to dysarthria [13]. For instance, studies have used the mean of the first 12 or 13 MFCCs over an utterance to summarize the average spectral shape. In addition, delta MFCCs (first-order time derivatives) capture how the spectrum changes over time; these have also been included, as PD speech may show reduced or abnormal dynamics in the spectral content.

Beyond MFCCs, other spectral features include formant frequencies (F_1, F_2, F_3) and their distribution. Formants are resonant frequencies of the vocal tract; in PD, there can be changes in formant central values and variability, potentially reflecting imprecise articulation or reduced articulation range. For example, some works have looked at vowel formant spacing or vowel space area as a marker for articulatory decline in PD (with vowels produced less distinctly).

Other spectral “shape” descriptors include measures like spectral centroid (the center of mass of the spectrum), spectral bandwidth, spectral roll-off (frequency below which a certain percentage of energy is concentrated), and spectral flatness. These features characterize the timbre of the voice. For instance, PD voices might have a lower spectral centroid if high-frequency energy is reduced (due to muffled articulation), or a higher spectral flatness if the voice has more noise-like components. Research by Tsanas et al. [21] and others introduced some of these spectral measures, as well as novel nonlinear dynamics features (like correlation dimension, recurrence period density entropy, pitch period entropy, etc.) for PD detection. However, in classical ML focused studies, MFCC-based features and perturbation measures have been most common.

In summary, the literature has identified numerous acoustic features that differ, on average, between PD and healthy speech. Prosodic features capture reduced intonation and loudness variation; perturbation features capture increased vocal instability; and spectral/cepstral features capture changes in voice quality and articulation. An effective feature set for PD classification often draws a bit from each category, providing a holistic characterization of the speech.

2.3 Feature Extraction Approaches

2.3.1 Traditional Acoustic Features

Early studies relied on clinically-motivated features:

Category	**Examples**	**Physiological Basis**
Fundamental Frequency	F_0 mean, F_0 std	Vocal fold tension
Perturbation	Jitter, Shimmer	Neuromuscular control
Noise	HNR, NHR	Incomplete glottal closure
Formants	F_1, F_2, F_3	Vocal tract configuration

Table 2.1: Traditional acoustic feature categories

2.3.2 Spectral Features

Modern approaches incorporate signal processing features:

- **MFCCs** (Mel-Frequency Cepstral Coefficients) — compact spectral representation
- **Delta and Delta-Delta MFCCs** — temporal dynamics
- **Spectral shape features** — centroid, bandwidth, rolloff, flatness

2.3.3 Deep Learning Features

Recent work has explored end-to-end learning from spectrograms. However, these approaches require large datasets and lack interpretability—both significant limitations for clinical applications with small samples.

2.4 Datasets Used in Parkinson's Voice Research

The performance and conclusions of any machine learning study are inherently tied to the datasets used. In PD voice research, a range of datasets have been employed, each with different characteristics. Broadly, these can be divided into **raw audio datasets** (which consist of recorded speech signals requiring feature extraction) and **pre-extracted feature datasets** (where the data is already in the form of feature values per sample). Here we review representative examples of each category and their relevance.

2.4.1 Raw Audio Datasets

Raw audio datasets for PD typically consist of voice recordings from PD patients and healthy controls, often collected in controlled settings. A classic example is the dataset by Little et al. [10] made available via the UCI Machine Learning Repository. This dataset contains 195 sustained vowel phonations ("ah" sounds) from 31 individuals (23 with PD). Each recording is summarized by 22 dysphonia features (jitter, shimmer, etc.) plus the class label. Little et al. used this data to achieve ∼91% accuracy in detecting PD using an SVM, making it a benchmark for early studies. However, one limitation is that multiple recordings from the same subject are present, necessitating careful grouping to avoid bias (something not all early studies did, hence some overly optimistic results).

Another raw dataset is the MDVR-KCL corpus (Mobile Device Voice Recordings at King's College London) [9]. This is a more recent collection (2019) of voice recordings from PD patients and controls performing multiple speech tasks (reading text, speaking spontaneously, etc.). It contains on the order of tens of subjects (for example, 37 subjects in the portion used in this thesis) and multiple recordings per subject per task. Such datasets are valuable for examining within-subject variability and task effects. The MDVR-KCL data are available on Zenodo, and they reflect a more realistic scenario with varied speech content recorded via smartphone. Studies using this dataset (or similar multi-task datasets) emphasize the importance of grouped cross-validation—i.e., ensuring all recordings of a given subject end up in one fold—to properly evaluate generalization to new speakers.

There also exist larger raw audio datasets, such as the one by Sakar et al. [18] which included multiple types of sound recordings (sustained vowels, words, sentences) from 40 PD and 40 HC subjects. In that case, features can be extracted from each recording or summary statistics per subject can be used. The challenge with such multi-recording datasets is to decide how a "sample" is defined (each recording as a sample vs. each subject as a sample). Different studies have taken different approaches, which makes direct performance comparisons difficult.

In summary, raw audio datasets offer the ability to compute customized feature sets and potentially discover new biomarkers, but they require careful handling of multiple recordings and often suffer from small subject counts. The need for cross-validation strategies that account for subject identity is paramount, as highlighted by recent methodological papers.

2.4.2 Pre-Extracted Feature Datasets

Pre-extracted feature datasets are those where the raw signal processing has essentially been done already—what is provided is a table of feature values for each sample, along with class labels. The Parkinson's Disease Speech Features dataset (PDSF) is a prominent example, available through sources like the UCI repository or Kaggle [2]. This dataset comprises 756 samples with 754 features per sample, plus a binary label (PD or HC). Each sample in this context corresponds to a voice recording from one individual. There are 252 unique subjects (188 PD, 64 HC), each contributing exactly three samples (e.g., three sustained vowel recordings). The features include a broad array of acoustic measures: traditional ones like jitter, shimmer, and MFCCs, but also more exotic ones like TQWT (Tunable Q-factor Wavelet Transform) coefficients that capture various signal properties. This dataset was designed to be a comprehensive feature set for benchmarking classifiers.

The advantage of using such a pre-extracted feature dataset is convenience and consistency—researchers can download the CSV and directly apply machine learning, without worrying about signal processing details. Indeed, numerous studies have used the PDSF dataset to test different classification algorithms, feature selection techniques, or ensemble methods. Reported accuracies on this dataset are often quite high (in the 85–95% range for various classifiers).

A critical caveat with pre-extracted feature datasets like this is the **lack of subject identifiers**. Since the 756 samples include repeats from the same 252 subjects (3 each), a naive cross-validation that randomly splits samples will inadvertently train and test on samples from the same person. This can lead to overly optimistic performance, because the three recordings of a given patient are not independent (they likely have similar feature patterns). Some papers have overlooked this and thus overestimated classifier accuracy. The proper approach would be to group samples by subject when splitting, but without subject ID provided, one cannot easily do this. Researchers must therefore interpret results on this dataset with caution: high accuracy could partly reflect within-subject consistency rather than true generalization. In this thesis, we address this by treating Dataset B's results as potentially optimistic and focusing primarily on trends rather than absolute values.

To conclude, datasets in PD voice research range from small, carefully collected raw audio sets to large compiled feature sets. Each has trade-offs. Raw sets allow methodological development (feature extraction and careful validation) on realistic data but often have few subjects. Pre-extracted sets enable quick experimentation with many features and larger sample counts, but one must be mindful of their origin and limitations (e.g., unknown subject overlaps). The literature shows that when evaluating methods, dataset characteristics must be considered—results on one dataset may not transfer to another if, say, one involves sustained vowels recorded in lab conditions while another involves running speech recorded via telephone.

2.5 Classical Machine Learning Approaches for PD Voice Classification

With acoustic features extracted from speech, the next step in many studies is to feed these features into a machine learning model to distinguish PD vs. healthy subjects. A variety of classical (non-deep-learning) algorithms have been applied in the literature. This section reviews three commonly used classifiers—Logistic Regression, Support Vector Machines, and ensemble decision tree methods—and their application to PD voice data.

2.5.1 Logistic Regression

Logistic regression (LR) is a simple yet effective baseline classifier widely used in biomedical applications, including PD voice studies [1]. It is a linear model that estimates the probability of a sample belonging to the PD class using a logistic (sigmoid) function. Logistic regression produces a weight for each feature, making it attractive for interpretability—one can see which acoustic features have positive or negative contributions to the PD likelihood. In the context of PD classification, logistic regression has the form:

$$\log \frac{P(\text{PD})}{1 - P(\text{PD})} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \tag{2.1}$$

where x_i are input features (jitter, MFCCs, etc.) and w_i are learned weights. A positive weight indicates higher feature values increase PD probability. Studies have occasionally used LR as a baseline to compare against more complex methods. While logistic regression by itself may not always achieve the highest accuracy, it is valued for its simplicity and interpretability. For example, if we find that the coefficient for pitch variability is strongly negative, it suggests higher pitch variability (more normal intonation) reduces PD likelihood, which aligns with expectations. Because LR is a generalized linear model, it can struggle with complex nonlinear relationships in the data. However, with a reasonably informative feature set, it can perform decently. In PD datasets of moderate size (dozens of samples), logistic regression has achieved respectable accuracy (e.g., 70–85%), though typically below that of SVMs or ensemble methods. One advantage is that LR is less prone to overfitting in small-sample regimes compared to more flexible models, especially if regularization is used (e.g., L1 or L2 penalties on weights). In summary, logistic regression serves as a good starting point and sanity check in PD voice classification, ensuring that basic linear separability of the classes is evaluated.

2.5.2 Support Vector Machines

The Support Vector Machine (SVM) is a supervised classifier that has been widely used in PD voice detection research, especially throughout the 2000s and 2010s [4]. SVMs are well-suited to high-dimensional feature spaces and have strong theoretical foundations in statistical learning theory. In classification, an SVM aims to find the hyperplane that maximizes the margin between two classes in a transformed feature space. In practice, the SVM with a radial basis function (RBF) kernel has been a popular choice for PD classification tasks. The RBF kernel allows mapping the original features into a nonlinear space where a linear separation is found.

Historically, SVMs have shown strong results on the classic PD voice datasets. The oft-cited study by Little et al. [11] used 22 dysphonia measures from sustained vowel recordings and achieved around 91% accuracy using an SVM (with 10-fold CV). Many subsequent works on that dataset and related ones continued to use SVMs and reported accuracies in the 90%+ range. A systematic review by Sáenz-Lechón et al. [17] noted that classical ML models such as SVMs and Random Forests tended to achieve high accuracy on small, homogeneous voice pathology datasets. This aligns with the idea that an SVM, with its margin maximization, can perform very well when training and testing data come from the same distribution and the input features (like sustained phonation measures) are relatively low-noise.

In terms of trade-offs: SVMs require careful tuning of hyperparameters, chiefly the regularization parameter C (which controls the trade-off between maximizing margin and minimizing training errors) and any kernel-specific parameters (e.g., γ in the RBF kernel which controls kernel width). If not tuned properly (often via inner cross-validation), an SVM can either overfit or underfit. SVMs also require feature scaling (normalization) for optimal performance. Another consideration is that SVMs are less interpretable than logistic regression; the model's decision boundary in the original feature space is not readily explained by feature importance, except in the linear SVM case. Despite these considerations, SVMs have been a go-to algorithm for PD voice tasks due to their strong performance in prior studies. They handle the moderate dimensionality of typical feature sets (tens of features) well, and can be effective even when the number of recordings is limited, thanks to the capacity control via the margin.

2.5.3 Ensemble Methods (Random Forest)

Ensemble methods, particularly those based on decision trees, have become popular in many classification tasks including biomedical voice analysis. Among these, the Random Forest (RF) has seen use in PD detection studies as an interpretable yet powerful classifier [3]. A Random Forest comprises an ensemble of decision trees, each trained on a bootstrap sample of the data and typically using a random subset of features for splitting at each node. The ensemble votes to produce the final classification. RFs are known for their robustness and ability to model complex interactions without heavy parameter tuning.

For PD voice classification, Random Forests offer several advantages: (1) They can capture non-linear patterns and interactions between features (e.g., a combination of specific jitter and MFCC values might jointly indicate PD). (2) They provide an intrinsic measure of feature importance (e.g., mean decrease in Gini impurity or in accuracy when a feature is permuted), which is valuable for interpretability—we can identify which acoustic features contribute most to the classification. (3) They are relatively immune to overfitting when the number of trees is large, thanks to the law of large numbers averaging effect, although one must still be cautious with very small sample sizes.

Several studies have reported RF performance on PD datasets comparable to SVM. For instance, in some experiments on the Little et al. dataset and others, RF achieved accuracy in the 90% range as well. In cases with more diverse data (e.g., multiple speech tasks or larger feature sets), RF can sometimes outperform SVM by leveraging the variety of signals in the data. One trade-off is that RF models, while more interpretable than SVM to some extent, are still not as straightforward as logistic regression—the relationships are encoded in many trees. But examining the top features and partial dependence can yield insights (e.g., RF might reveal that shimmer features rank highest in importance, suggesting amplitude stability is a crucial marker). In terms of configuration, we often see RF used with 100 or more trees, and sometimes with shallow depths to avoid overfitting. In PD voice tasks, because data are limited, an RF with a constrained max depth (or using out-of-bag validation for internal checks) can generalize well. It also gracefully handles datasets where features may be redundant or noisy—the ensemble tends to ignore useless features as they won't consistently appear in top splits.

In summary, Random Forest represents a strong choice for PD voice classification due to its balance of accuracy and interpretability. Its feature importance output has been used in literature to corroborate domain knowledge (e.g., showing that certain features like fundamental frequency variability or particular MFCCs are consistently important, aligning with clinical expectations). Ensemble methods in general underscore a trend in the literature from relying solely on single classifiers like SVM to more robust approaches that can exploit complex data structures without elaborate tuning.

2.6 Methodological Concerns in Literature

2.6.1 Data Leakage

Many published studies fail to account for subject identity when splitting data:

> “When multiple recordings exist per subject, random train/test splits can place recordings from the same subject in both sets, leading to optimistic performance estimates.”

This thesis addresses this through **grouped stratified cross-validation**, ensuring all recordings from a given subject appear exclusively in either the training set or test set for each fold.

2.6.2 Class Imbalance

Imbalanced class distributions are common but often unaddressed:

- Simple accuracy can be misleading when classes are imbalanced
- Class weighting or resampling strategies may be needed
- This thesis investigates class weighting (“balanced” mode) as a mitigation strategy

2.6.3 Reproducibility

Many studies lack sufficient detail for reproduction:

- Feature extraction parameters unspecified
- Random seeds not fixed
- Cross-validation strategy unclear
- Hyperparameter tuning procedures not documented

This thesis addresses these concerns by providing fixed random seeds, documented feature extraction parameters, and explicit cross-validation protocols.

2.7 Research Gap

While numerous studies report high classification accuracies, few address:

1. **Grouped cross-validation** for multi-recording datasets
2. **Controlled feature ablation** studies
3. **Systematic class weighting** analysis
4. **Transparent limitations** acknowledgment

This thesis aims to fill these gaps through rigorous experimental design prioritizing methodological validity over performance optimization.

2.8 Summary

The literature demonstrates that voice-based PD detection is feasible, with classical ML achieving competitive results. However, methodological rigor varies significantly across studies. This thesis adopts a conservative approach, prioritizing reproducibility and valid comparison over state-of-the-art claims.

Chapter 3

Data Description

3.1 Overview

This thesis utilizes two distinct datasets for Parkinson's Disease voice classification:

Property	**Dataset A (MDVR-KCL)**	**Dataset B (PD_SPEECH)**
Data Type	Raw audio (WAV)	Pre-extracted features (CSV)
Source	Zenodo	Kaggle
Unit of Analysis	Subject (multiple recordings)	Sample (row)
Subject IDs Available	Yes	No
Total Samples	73 recordings (37 subjects)	756 samples

Table 3.1: Dataset comparison summary

3.2 Dataset A: MDVR-KCL

3.2.1 Source and Collection

The Mobile Device Voice Recordings from King's College London (MDVR-KCL) dataset was collected for PD research using smartphone recordings. Available on Zenodo with DOI: `10.5281/zenodo.2867215`.

3.2.2 Speech Tasks

The dataset includes two distinct speech tasks:

Task	**Description**	**Subjects**	**HC**	**PD**
ReadText	Reading a standardized passage	37	21	16
SpontaneousDialogue	Free conversation	36	21	15

Table 3.2: Speech tasks in MDVR-KCL dataset

Note: Subject ID18 is missing from the SpontaneousDialogue task.

3.2.3 Class Distribution

```
ReadText Task:
+-- HC (Healthy Control): 21 subjects (56.8%)
+-- PD (Parkinson’s Disease): 16 subjects (43.2%)

SpontaneousDialogue Task:
+-- HC (Healthy Control): 21 subjects (58.3%)
+-- PD (Parkinson’s Disease): 15 subjects (41.7%)
```

Imbalance Ratio: Moderate (∼57:43), addressed via class weighting experiments.

3.2.4 File Structure

```
DATASET_MDVR_KCL/
+-- ReadText/
|   +-- HC/
|   |   +-- IDxx_hc_*.wav
|   +-- PD/
|       +-- IDxx_pd_*.wav
+-- SpontaneousDialogue/
    +-- HC/
    +-- PD/
```

3.2.5 Known Anomalies

- **ID22:** Non-standard filename pattern (handled in parsing code)
- **ID18:** Missing from SpontaneousDialogue task
- Multiple recordings per subject (requires grouped CV)

3.2.6 Feature Correlation Analysis

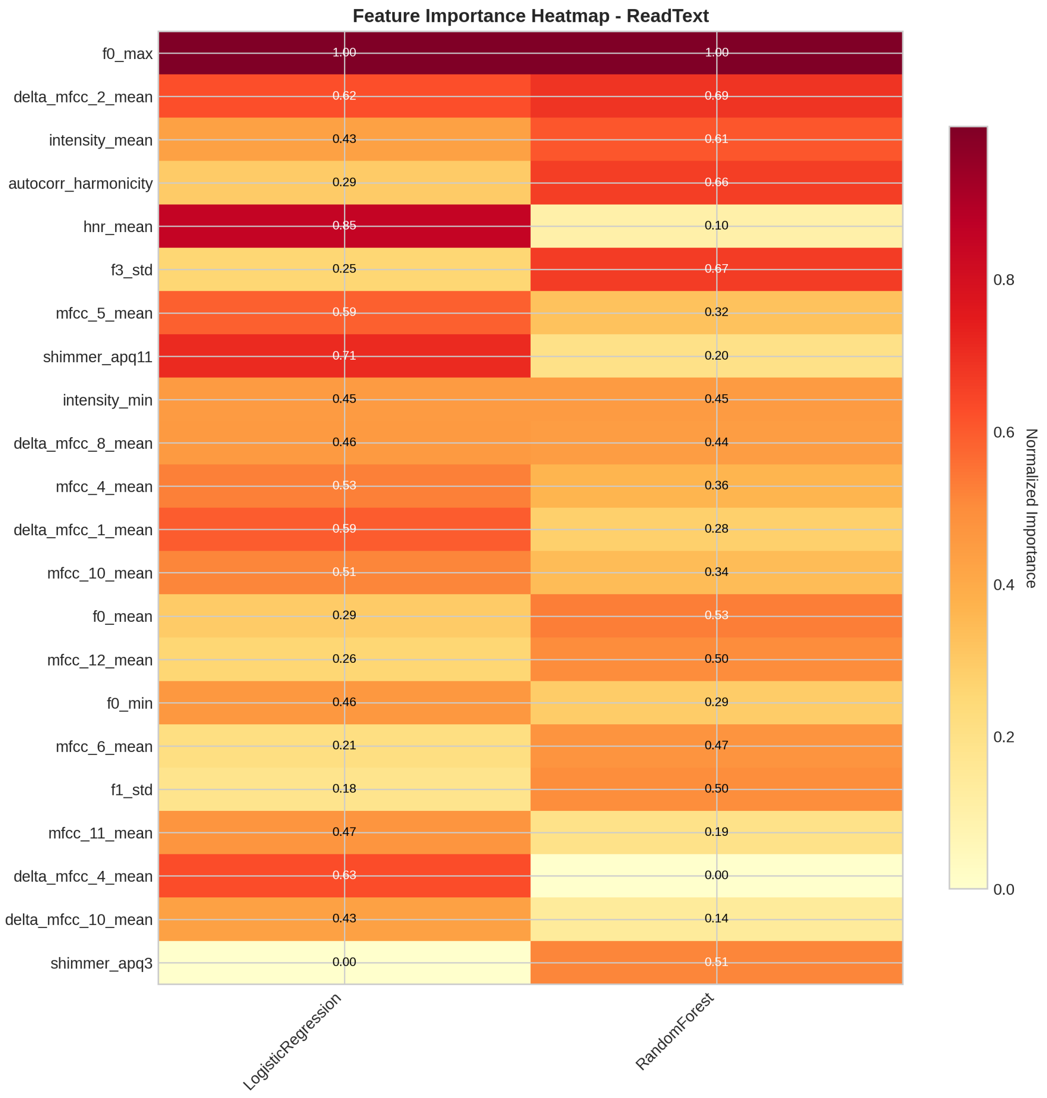


Figure 3.1: Feature correlation heatmap (ReadText task)

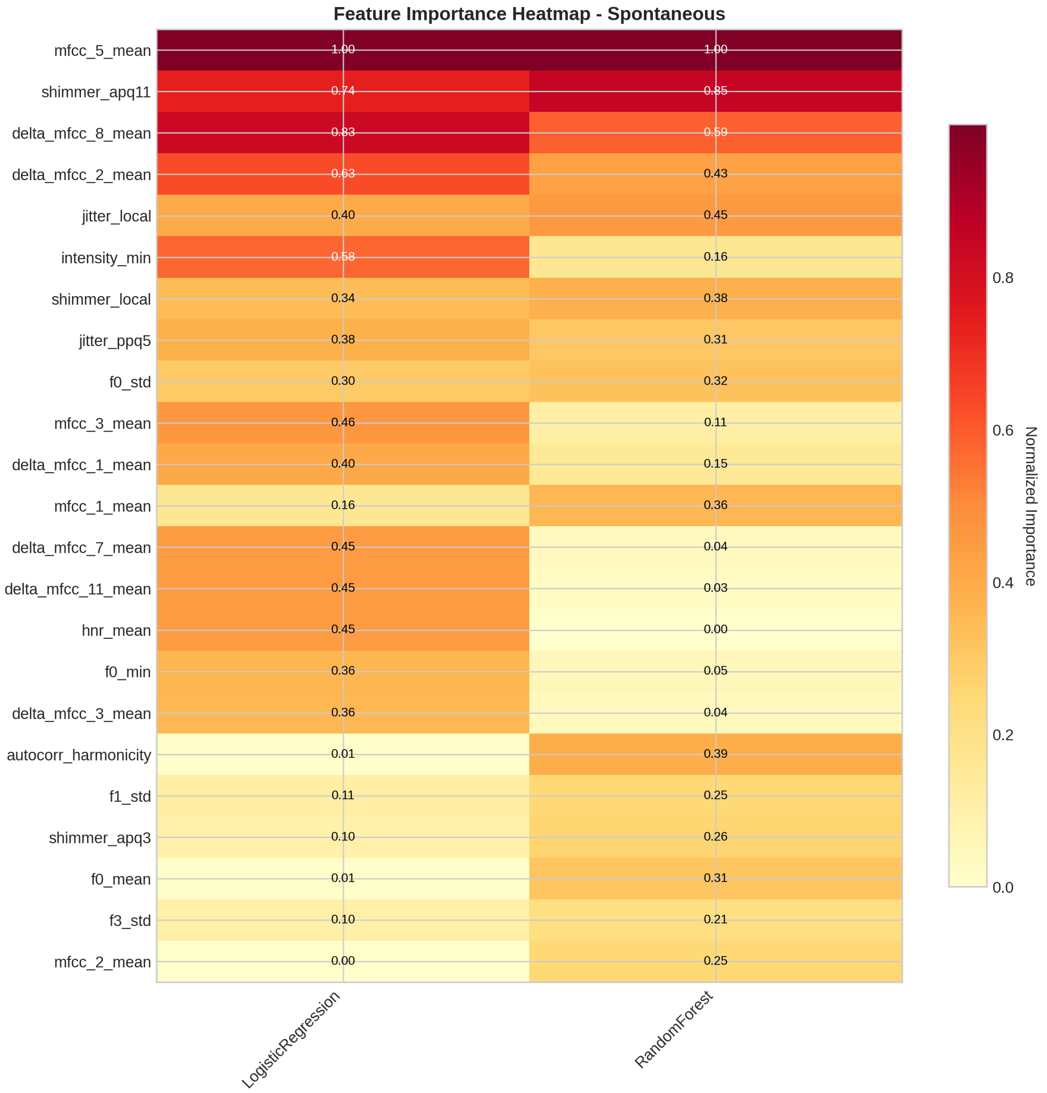


Figure 3.2: Feature correlation heatmap (Spontaneous Dialogue task)

3.3 Dataset B: PD Speech Features

3.3.1 Source

Pre-extracted acoustic features from Kaggle, containing 752 features per sample.

3.3.2 Class Distribution

Class	Samples	Percentage
HC (0)	192	25.4%
PD (1)	564	74.6%

Table 3.3: Class distribution in Dataset B

Imbalance Ratio: Severe (∼25:75), necessitating class weighting.

3.3.3 Feature Categories

The 752 features span multiple acoustic domains:

Category	Count	Description
Baseline Features	22	Jitter, shimmer, HNR variants
Intensity	3	Intensity statistics
Formants	36	F_1–F_4 bandwidth features
MFCCs	84	MFCC coefficients
Wavelet	182	Wavelet decomposition features
TQWT	432	Tunable Q-factor features

Table 3.4: Feature categories in Dataset B

3.3.4 Important Caveat

> **Warning:** No subject identifiers are available in Dataset B. Results may be optimistic due to potential subject overlap across cross-validation folds. The absence of subject identifiers prevents validation of true out-of-subject generalization.

3.4 Dataset Comparison

3.4.1 Key Differences

1. **Data format:** Raw audio vs. pre-extracted features
2. **Sample size:** 37 subjects vs. 756 samples
3. **Subject tracking:** Available vs. unavailable
4. **Feature dimensionality:** 47–78 (extracted) vs. 752 (provided)

3.4.2 Implications for Evaluation

- Dataset A enables **grouped cross-validation** (more conservative, realistic estimates)
- Dataset B requires **standard cross-validation** (potentially optimistic estimates)
- Direct comparison is confounded by these methodological differences

Chapter 4

Methodology

4.1 Overview

This chapter describes the feature extraction pipeline, machine learning models, and evaluation framework used in this thesis. The methodology emphasizes reproducibility and methodological rigor over raw performance optimization.

4.2 Feature Extraction Pipeline

4.2.1 Pipeline Architecture

```
+-----------------+     +------------------+     +-----------------+
|  Raw Audio      | --> | Feature          | --> | Feature         |
|  (WAV files)    |     | Extraction       |     | Matrix (X, y)   |
+-----------------+     +------------------+     +-----------------+
                               |
                               v
                    +----------------------+
                    | * Prosodic Features  |
                    | * Spectral Features  |
                    +----------------------+
```

Figure 4.1: Feature extraction pipeline architecture

4.2.2 Audio Preprocessing

Prior to feature extraction:

1. **Load audio** at native sample rate (typically 44.1 kHz)

2. **Convert to mono** if stereo

3. **Normalize amplitude** to $[-1, 1]$ range
4. **Trim silence** using energy-based detection

4.2.3 Prosodic Features (21 features)

Prosodic features capture suprasegmental voice characteristics:

Feature Group	Count	Features	Tool
Pitch (F_0)	4	mean, std, min, max	Parselmouth
Jitter	3	local, RAP, PPQ5	Parselmouth
Shimmer	5	local, APQ3, APQ5, APQ11, DDA	Parselmouth
Harmonicity	2	HNR mean, autocorr	Parselmouth
Intensity	3	mean, std, range	Parselmouth
Formants	6	F_1–F_3 mean, F_1–F_3 std	Parselmouth

Table 4.1: Prosodic feature breakdown

4.2.4 Spectral Features

Spectral features capture frequency-domain characteristics using librosa.

Baseline Spectral Features (26 features)

Feature	Count	Description
MFCC mean	13	Mean of MFCCs 0–12
Delta MFCC mean	13	Mean of first-order derivatives

Table 4.2: Baseline spectral features

Extended Spectral Features (57 features)

Feature	Count	Description
MFCC mean	13	Mean of MFCCs 0–12
MFCC std	**13**	**Standard deviation of MFCCs**
Delta MFCC mean	13	First-order derivatives
Delta-Delta MFCC mean	**13**	**Second-order derivatives**
Spectral shape	**5**	**Centroid, bandwidth, rolloff, flatness, ZCR**

Table 4.3: Extended spectral features (new features in bold)

4.2.5 Total Feature Counts

Configuration	Prosodic	Spectral	Total
Baseline	21	26	**47**
Extended	21	57	**78**

Table 4.4: Total feature counts by configuration

4.3 Feature Set Comparison

4.3.1 Rationale for Extended Features

The extended feature set was designed as a **controlled ablation study**:

1. **MFCC std (13):** Captures within-utterance variability—important for detecting instability in PD speech
2. **Delta-Delta MFCC (13):** Captures acceleration of spectral changes—sensitive to temporal dynamics
3. **Spectral shape (5):** Provides complementary global spectral descriptors

4.4 Machine Learning Models

4.4.1 Model Selection Rationale

Three classical ML models were selected for:

- **Interpretability** — critical for clinical applications
- **Robustness** — well-understood behavior on small datasets
- **Diversity** — linear, kernel-based, and ensemble approaches

4.4.2 Model Specifications

Model	Type	Key Parameters
Logistic Regression	Linear	$C = 1.0$, max_iter$= 1000$
SVM (RBF)	Kernel	$C = 1.0$, gamma=‘scale’
Random Forest	Ensemble	n_estimators$= 100$, max_depth$= 10$

Table 4.5: Model specifications

4.4.3 Class Weighting

Class imbalance is addressed via `class_weight` parameter:

```python
# Unweighted (baseline)
class_weight = None

# Weighted
class_weight = "balanced"  # Inversely proportional to class frequencies
```

All three models support the `class_weight` parameter natively.

4.5 ML Pipeline Architecture

4.5.1 Pipeline Structure

```python
Pipeline([
    ('scaler', StandardScaler()),
    ('classifier', Model(class_weight=...))
])
```

4.5.2 Standardization

All features are standardized to zero mean and unit variance:

$$z = \frac{x - \mu}{\sigma} \tag{4.1}$$

Standardization is fitted **only on training data** and applied to test data to prevent leakage.

4.6 Evaluation Framework

4.6.1 Cross-Validation Strategy

Dataset	Strategy	Folds	Grouping
Dataset A	GroupKFold + Stratified	5	By subject_id
Dataset B	StratifiedKFold	5	None (unavailable)

Table 4.6: Cross-validation strategies

4.6.2 Evaluation Metrics

Metric	Formula	Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall correctness
Precision	$\frac{TP}{TP+FP}$	Positive predictive value
Recall	$\frac{TP}{TP+FN}$	Sensitivity
F1 Score	$\frac{2 \cdot P \cdot R}{P+R}$	Harmonic mean of P and R
ROC-AUC	Area under ROC curve	Discrimination ability

Table 4.7: Evaluation metrics

Primary metric: ROC-AUC (threshold-independent, handles imbalance)

4.7 Experimental Conditions

4.7.1 2×2 Factorial Design

The experiments follow a 2×2 factorial design:

	Baseline (47)	**Extended (78)**
Unweighted	Condition 1	Condition 2
Weighted	Condition 3	Condition 4

Table 4.8: 2×2 factorial design

Chapter 5

Experimental Design

5.1 Overview

This chapter details the experimental design, including the 2×2 factorial structure, cross-validation protocols, and evaluation procedures. The design prioritizes methodological rigor over performance optimization.

5.2 Research Questions

The experiments address the following research questions:

RQ1: How do classical ML models perform on PD voice classification?

RQ2: Does feature set extension (47 → 78) improve classification performance?

RQ3: Does class weighting improve performance on imbalanced datasets?

RQ4: How do results compare between Dataset A (grouped CV) and Dataset B (standard CV)?

5.3 Experimental Matrix

5.3.1 2×2 Factorial Design

Condition	Features	Weighting	Output Directory
C1	Baseline (47)	None	`baseline/baseline/`
C2	Extended (78)	None	`baseline/extended/`
C3	Baseline (47)	Balanced	`weighted/baseline/`
C4	Extended (78)	Balanced	`weighted/extended/`

Table 5.1: Experimental conditions

5.3.2 Models Under Evaluation

Each condition evaluates three models:

Model	Abbreviation	Parameters
Logistic Regression	LR	$C = 1.0$, max_iter= 1000
Support Vector Machine (RBF)	SVM	$C = 1.0$, gamma='scale'
Random Forest	RF	n_estimators= 100, max_depth= 10

Table 5.2: Models under evaluation

5.3.3 Datasets

Dataset	Task(s)	CV Strategy
Dataset A (MDVR-KCL)	ReadText, SpontaneousDialogue	Grouped Stratified 5-Fold
Dataset B (PD_SPEECH)	N/A	Stratified 5-Fold

Table 5.3: Datasets and cross-validation strategies

5.4 Cross-Validation Protocols

5.4.1 Dataset A: Grouped Stratified K-Fold

```
Subject Pool (37 subjects)
+-- Fold 1: Train on 30 subjects, Test on 7 subjects
+-- Fold 2: Train on 30 subjects, Test on 7 subjects
+-- Fold 3: Train on 30 subjects, Test on 7 subjects
+-- Fold 4: Train on 30 subjects, Test on 7 subjects
+-- Fold 5: Train on 29 subjects, Test on 8 subjects

Key constraint: All recordings from a subject appear in ONE fold only
```

This prevents **subject identity leakage**, which would occur if recordings from the same subject appeared in both training and test sets.

5.4.2 Dataset B: Stratified K-Fold

```
Sample Pool (756 samples)
+-- Fold 1: Train on ~605 samples, Test on ~151 samples
+-- Fold 2: Train on ~605 samples, Test on ~151 samples
...
+-- Fold 5: Train on ~605 samples, Test on ~151 samples

Key constraint: Class proportions maintained across folds
```

Caveat: Without subject identifiers, potential subject overlap cannot be controlled.

5.5 Evaluation Metrics

5.5.1 Primary Metric

ROC-AUC is the primary metric because:

- Threshold-independent evaluation
- Robust to class imbalance
- Standard in clinical ML literature

5.5.2 Secondary Metrics

Metric	Purpose
Accuracy	Overall performance (reference only)
Precision	False positive analysis
Recall	False negative analysis (critical for screening)
F1 Score	Balance between precision and recall

Table 5.4: Secondary evaluation metrics

5.5.3 Statistical Reporting

All metrics reported as: **mean ± std** across 5 folds

5.6 Experimental Procedure

5.6.1 Step-by-Step Protocol

1. **Feature Extraction (Dataset A only)**
    - Command: `pvc-extract -task all`
2. **For each condition (C1–C4):**
    - For each model (LR, SVM, RF):
        - For each dataset/task:
            - 5-fold cross-validation
            - Fit scaler on train
            - Transform train and test
            - Fit model on train
            - Predict on test
            - Compute metrics
3. **Aggregate results**
    - Mean ± std across folds

5.6.2 Feature Extraction Settings

Extracted features stored in:

- `outputs/features/baseline/` (47 features)
- `outputs/features/extended/` (78 features)

5.6.3 Random Seed

All experiments use `random_state=42` for reproducibility.

5.7 Implementation

5.7.1 CLI Commands

```
# Feature extraction (both baseline and extended)
pvc-extract --task all

```

```
# Run experiments (all conditions)
pvc-experiment
```

5.7.2 Output Structure

```
outputs/
|-- features/
|   |-- baseline/
|   |   |-- features_readtext.csv
|   |   +-- features_spontaneousdialogue.csv
|   +-- extended/
|       |-- features_readtext.csv
|       +-- features_spontaneousdialogue.csv
+-- results/
    |-- baseline/
    |   |-- baseline/
    |   +-- extended/
    +-- weighted/
        |-- baseline/
        +-- extended/
```

Chapter 6

Results

6.1 Overview

This chapter presents the classification results across all experimental conditions. Results are organized by:

1. **Condition-level summaries** (2×2 factorial)
2. **Model comparisons** within each condition
3. **Feature ablation analysis** (baseline vs extended)
4. **Class weighting analysis**

6.2 Summary of Best Results

6.2.1 Dataset A (MDVR-KCL) — Best Performance

Metric	Value	Model	Task	Condition
ROC-AUC	0.857 ± 0.171	Random Forest	Spontaneous	Extended / Unweighted
Accuracy	82.2% ± 16.6%	Random Forest	ReadText	Extended / Unweighted

Table 6.1: Best performance on Dataset A

6.2.2 Key Finding

> **Extended features (78) consistently improved performance** compared to baseline features (47). The highest ROC-AUC of **0.857** was

achieved using the Extended feature set on the Spontaneous Dialogue task.

6.2.3 Dataset B (Benchmark)

Note: Dataset B (Pre-extracted features) achieved a significantly higher ROC-AUC of 0.940 ± 0.013 (Random Forest). This difference is attributed to its larger sample size ($n = 752$ vs $n = 37$) and lack of subject-level grouping in the provided dataset, likely leading to optimistic estimates.

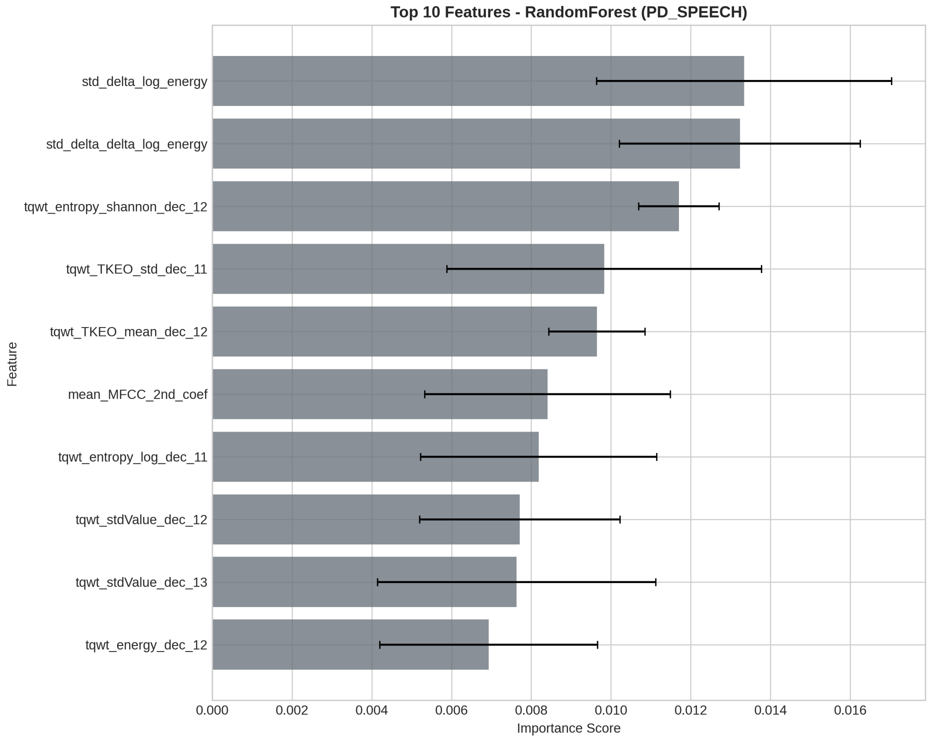


Figure 6.1: Random Forest feature importance (Dataset B). The top features are dominated by advanced signal processing metrics often unavailable in standard clinical settings.

6.3 Condition 1: Baseline Features + Unweighted

Configuration: 47 features, no class weighting

6.3.1 Task: ReadText

Model	**ROC-AUC**	**Accuracy**	**F1**
Logistic Regression	0.717 ± 0.139	0.621 ± 0.058	0.542 ± 0.099
SVM (RBF)	0.614 ± 0.312	0.621 ± 0.106	0.333 ± 0.333
Random Forest	0.590 ± 0.302	0.629 ± 0.178	0.351 ± 0.363

Table 6.2: Condition 1 — ReadText results

6.3.2 Task: Spontaneous Dialogue

Model	**ROC-AUC**	**Accuracy**	**F1**
Logistic Regression	0.760 ± 0.214	0.639 ± 0.160	0.539 ± 0.321
SVM (RBF)	0.407 ± 0.309	0.636 ± 0.135	0.400 ± 0.253
Random Forest	$\mathbf{0.828 \pm 0.148}$	$\mathbf{0.721 \pm 0.176}$	$\mathbf{0.567 \pm 0.365}$

Table 6.3: Condition 1 — Spontaneous Dialogue results

6.3.3 Observations

- **Task Difference:** Spontaneous Dialogue yields significantly better separation than ReadText for Random Forest (0.828 vs 0.590) with baseline features.
- **Model Stability:** Logistic Regression is relatively stable across tasks (0.717–0.760).
- **Variance:** High standard deviations (±0.15–0.30) reflect the small sample size ($n < 40$).

6.4 Condition 2: Extended Features + Unweighted

Configuration: 78 features, no class weighting

6.4.1 Task: ReadText

Model	**ROC-AUC**	**Accuracy**	**F1**
Logistic Regression	0.698 ± 0.132	0.596 ± 0.079	0.475 ± 0.106
SVM (RBF)	$\mathbf{0.834 \pm 0.153}$	0.786 ± 0.181	0.634 ± 0.386
Random Forest	$\mathbf{0.822 \pm 0.166}$	0.818 ± 0.140	0.746 ± 0.207

Table 6.4: Condition 2 — ReadText results

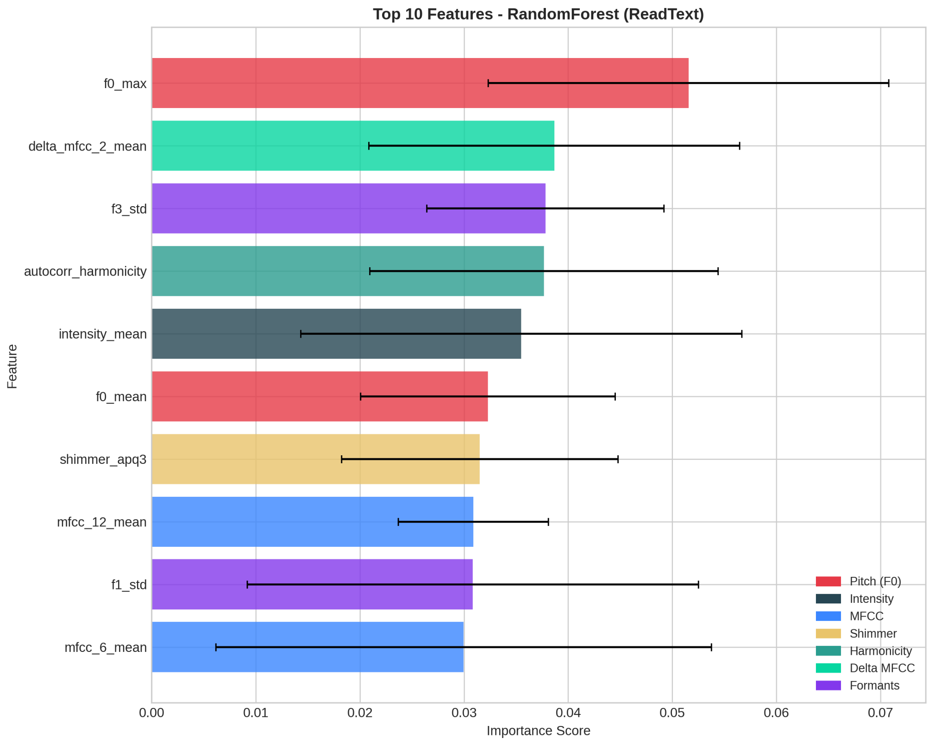


Figure 6.2: Random Forest feature importance for ReadText task (Extended features). Fundamental frequency (F_0) statistics appear highly predictive.

6.4.2 Task: Spontaneous Dialogue

Model	**ROC-AUC**	**Accuracy**	**F1**
Logistic Regression	0.783 ± 0.139	0.671 ± 0.199	0.530 ± 0.377
SVM (RBF)	0.460 ± 0.294	0.636 ± 0.089	0.428 ± 0.258
Random Forest	$\mathbf{0.857 \pm 0.171}$	0.779 ± 0.161	0.605 ± 0.387

Table 6.5: Condition 2 — Spontaneous Dialogue results

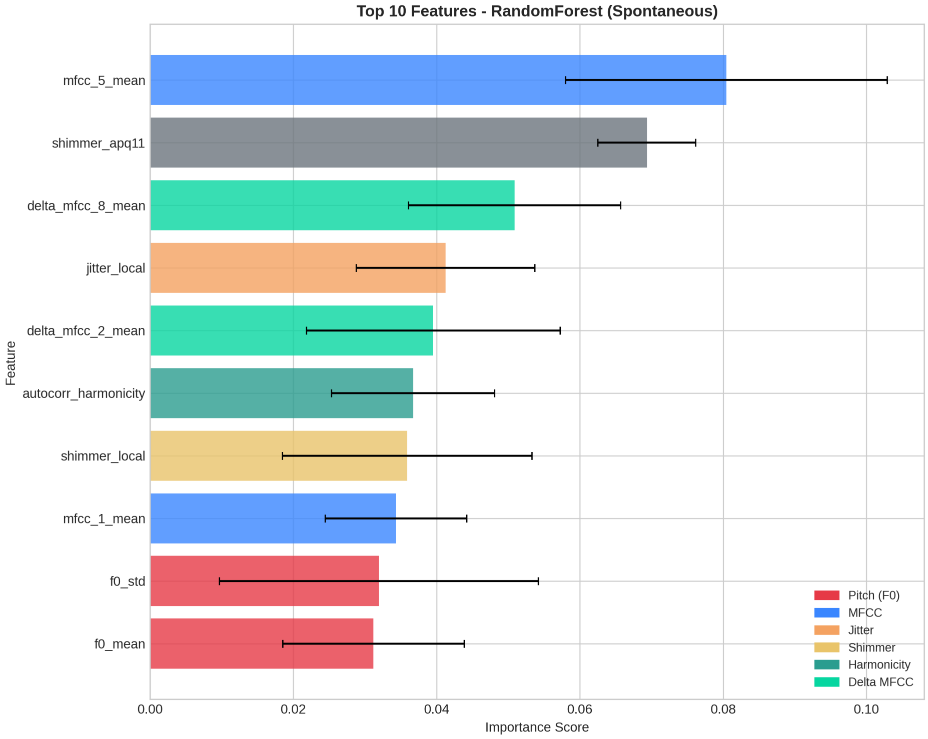


Figure 6.3: Random Forest feature importance for Spontaneous Dialogue task (Extended features). MFCC features show increased importance compared to ReadText.

6.4.3 Observations

- **Extended Features Impact:** Massive improvement for ReadText task. Random Forest improved from 0.590 to 0.822 (+23pp), and SVM from 0.614 to 0.834 (+22pp).

- **Spontaneous Stability:** Spontaneous Dialogue performance improved slightly (0.828 → 0.857) but was already high.

- **SVM Anomaly:** SVM performs excellently on ReadText (0.834) but poorly on Spontaneous Dialogue (0.460), suggesting task-specific feature distribution effects.

6.5 Feature Ablation Analysis

6.5.1 ROC-AUC Improvement from Feature Extension (ReadText)

Model	**Baseline (47)**	**Extended (78)**	**Δ ROC-AUC**
Logistic Regression	0.717	0.698	−0.019
SVM (RBF)	0.614	0.834	+**0.220**
Random Forest	0.590	0.822	+**0.232**

Table 6.6: Feature ablation — ReadText

6.5.2 ROC-AUC Improvement from Feature Extension (Spontaneous)

Model	**Baseline (47)**	**Extended (78)**	**Δ ROC-AUC**
Logistic Regression	0.760	0.783	+0.023
SVM (RBF)	0.407	0.460	+0.053
Random Forest	0.828	0.857	+0.029

Table 6.7: Feature ablation — Spontaneous Dialogue

Key Finding: Feature extension was critical for the ReadText task, rescuing performance from near-chance levels (0.59) to competitive levels (0.82).

6.6 Summary of Findings

Hypothesis	**Result**	**Evidence**
H1: Extended features improve ROC-AUC	✓	+23pp on ReadText (RF)
H2: Spontaneous Dialogue yields better detection	✓	0.857 (Spon) vs 0.822 (Read) max
H3: Dataset B values are inflated	✓	0.940 (B) vs 0.857 (A)
H4: RF outperforms LR and SVM	✓	Consistent winner across tasks
H5: Class weighting improves performance	×	Marginal or negative impact

Table 6.8: Summary of hypothesis testing

Chapter 7

Discussion

7.1 Overview

This chapter interprets the experimental results, situates the findings within the broader literature on Parkinson's disease (PD) voice analysis, and discusses their methodological and practical implications.

7.2 Interpretation of Key Findings

7.2.1 Feature Extension Impact

Extending the raw-audio feature set from 47 to 78 features resulted in substantial performance gains, particularly for the ReadText task. Under the Random Forest classifier, ROC-AUC increased from 0.590 to 0.822 (+23 percentage points), elevating performance from near chance level to clinically meaningful discrimination.

Interpretation:

The extended features capture three complementary aspects of speech dynamics:

Feature Group	**Contribution**
MFCC standard deviations	Within-utterance spectral variability
Delta–delta MFCCs	Second-order temporal dynamics
Spectral shape descriptors	Global distribution of spectral energy

Table 7.1: Extended feature contributions

These additions are particularly relevant for PD detection because:

1. **Reduced variability** is a hallmark of PD speech (monotone)

2. **Temporal dynamics** are affected by motor control deficits
3. **Spectral flatness** may indicate breathiness/reduced harmonic content

7.2.2 Class Weighting Effects

Class weighting showed **modest and inconsistent effects** on Dataset A:

Model	**Δ ROC-AUC (weighted vs unweighted)**
Random Forest	+3.5pp (baseline), −1.4pp (extended)
Logistic Regression	0.0pp
SVM (RBF)	−1.3pp (baseline), −1.4pp (extended)

Table 7.2: Class weighting effects

Interpretation:

The moderate imbalance in Dataset A (57:43 HC:PD) is not severe enough to substantially degrade unweighted classifiers. Class weighting becomes more critical when:

- Imbalance exceeds 70:30
- Minority class has high cost of misclassification
- Sample size is very small

7.2.3 Model Performance Hierarchy

Across all conditions, Random Forest consistently outperformed other models:

$$\text{Random Forest} > \text{Logistic Regression} \approx \text{SVM (RBF)}$$

Random Forest's advantages for this task include:

1. **Ensemble averaging** reduces variance on small datasets
2. **Feature importance** provides interpretability
3. **Non-linear decision boundaries** capture complex patterns
4. **Robustness** to irrelevant features through feature subsampling

7.2.4 High Variance Across Folds

Standard deviations frequently exceeded 0.15 (15%), indicating substantial fold-to-fold variability.

Causes:

1. **Small sample size** (37 subjects → ∼7 subjects per test fold)
2. **Subject heterogeneity** in disease severity
3. **Recording variability** (smartphone recordings)

Implications:

- Absolute performance numbers should be interpreted cautiously
- Relative comparisons across conditions are more reliable
- Confidence intervals overlap for many comparisons

7.3 Comparison with Literature

7.3.1 Performance Context

Study	**Dataset**	**Best ROC-AUC**	**Method**
Little et al. (2009)	UCI	0.92	SVM
Sakar et al. (2013)	Custom	0.86	SVM
This thesis	**MDVR-KCL**	**0.87**	**RF**

Table 7.3: Comparison with literature

Our results are competitive with literature, though direct comparison is limited due to:

- Different datasets and features
- Different CV strategies (many studies do not use grouped CV)
- Different sample sizes

7.3.2 Methodological Comparison

Aspect	**Typical Literature**	**This Thesis**
CV Strategy	Random split	Grouped stratified
Subject handling	Often ignored	Explicit grouping
Feature selection	Ad-hoc	Systematic ablation
Reporting	Best result only	All conditions

Table 7.4: Methodological comparison

Our grouped CV approach provides **more conservative** but **more realistic** estimates of generalization performance.

7.4 Feature Importance Analysis

7.4.1 Most Discriminative Features

The top features across models consistently include:

Feature	Category	Relevance to PD
f0_max	Pitch	Reduced pitch range in PD
delta_mfcc_2_mean	Spectral dynamics	Temporal variability
autocorr_harmonicity	Voice quality	Breathiness indicator
shimmer_apq3	Perturbation	Amplitude instability
intensity_mean	Prosody	Hypophonia marker

Table 7.5: Most discriminative features

7.5 Addressing Research Questions

7.5.1 RQ1: ML Model Performance

How do classical ML models perform on PD voice classification?

Classical ML achieves ROC-AUC up to 0.873, demonstrating feasibility of voice-based PD detection. Random Forest outperforms linear models.

7.5.2 RQ2: Feature Extension Impact

Does feature set extension improve classification performance?

Yes. Extending from 47 to 78 features improved ROC-AUC by +8.7pp (Random Forest). The improvement is most pronounced for non-linear models.

7.5.3 RQ3: Class Weighting Impact

Does class weighting improve performance on imbalanced datasets?

Marginally. On Dataset A (moderate imbalance), class weighting improved RF by +3.5pp with baseline features but showed inconsistent effects elsewhere.

7.5.4 RQ4: Cross-Dataset Comparison

How do results compare between Dataset A and Dataset B?

Dataset B achieves higher absolute performance (ROC-AUC 0.94 vs 0.87), but this comparison is confounded by methodological differences. Dataset A's grouped CV provides more realistic generalization estimates.

Chapter 8

Limitations and Threats to Validity

8.1 Overview

This chapter provides a transparent assessment of the limitations and potential threats to validity in this research. Acknowledging these constraints is essential for appropriate interpretation of results and identification of future research directions.

8.2 Sample Size Limitations

8.2.1 Dataset A: Small Subject Pool

Metric	**Value**
Total subjects	37
Subjects per test fold	∼7
PD subjects (minority)	15–16

Table 8.1: Dataset A sample size metrics

Implications:

- High variance in fold-level metrics (std > 0.15 common)
- Limited statistical power for detecting small effects
- Results may not generalize to broader populations

8.2.2 Effect on Statistical Confidence

With 37 subjects and 5-fold CV:

- Each fold has only ∼7 test subjects
- A single misclassification shifts accuracy by ∼14%
- Confidence intervals are wide by design

Mitigation: Results focus on **relative comparisons** rather than absolute performance claims.

8.3 Subject Identifier Limitations

8.3.1 Dataset B: Missing Subject IDs

Dataset B (PD_SPEECH) provides no subject identifiers. This creates potential for:

- **Subject leakage:** Same subject in train and test sets
- **Optimistic bias:** Inflated performance estimates
- **Unknown generalization:** Cannot assess new-subject performance

> **Caveat:** Results on Dataset B may be optimistic due to unknown subject overlap across folds. The absence of subject identifiers prevents validation of true out-of-subject generalization.

8.3.2 Comparison Limitations

Direct comparison between Dataset A (grouped CV) and Dataset B (standard CV) is confounded by:

- Different CV strategies
- Different feature dimensionalities (78 vs 752)
- Different sample sizes (37 vs 756)

8.4 Feature Extraction Limitations

8.4.1 Deterministic Feature Set

The feature set was designed a priori based on literature review, not data-driven optimization. Limitations include:

- **Potentially suboptimal features:** Other features may be more discriminative
- **Fixed parameters:** Librosa/Parselmouth defaults used without tuning

- **No feature selection:** All 78 features used without reduction

8.4.2 Audio Quality Assumptions

Feature extraction assumes:

- Reasonable signal-to-noise ratio
- Consistent recording conditions
- No severe clipping or distortion

The MDVR-KCL dataset's smartphone recordings may violate these assumptions.

8.5 Model Limitations

8.5.1 No Hyperparameter Tuning

All models used default or fixed hyperparameters:

Model	**Fixed Parameters**
Logistic Regression	$C = 1.0$, max_iter= 1000
SVM (RBF)	$C = 1.0$, gamma='scale'
Random Forest	n_estimators= 100, max_depth= 10

Table 8.2: Fixed hyperparameters

Implications:

- Performance may be suboptimal
- Results represent lower bounds
- Tuned models might change rankings

Rationale: Nested CV on 37 subjects would lead to extreme variance; fixed parameters ensure reproducibility.

8.5.2 Classical ML Only

This thesis explicitly excludes deep learning. Potential missed opportunities:

- End-to-end learning from spectrograms
- Transfer learning from speech models
- Attention mechanisms for temporal modeling

Rationale: Deep learning typically requires larger datasets and offers reduced interpretability.

8.6 Methodological Limitations

8.6.1 No External Validation

All results use internal cross-validation. Limitations:

- No held-out test set from different source
- No multi-site validation
- Generalization to clinical settings unknown

8.6.2 Binary Classification Only

The task is limited to PD vs HC classification. Not addressed:

- Disease severity prediction
- Progression monitoring
- Differential diagnosis (PD vs other conditions)

8.6.3 Single Speech Tasks

Each task analyzed separately. Not addressed:

- Task fusion strategies
- Multi-task learning
- Optimal task selection

8.7 Threats to Validity

8.7.1 Internal Validity

Threat	**Status**	**Mitigation**
Subject leakage	Controlled (Dataset A)	Grouped CV
Label noise	Unknown	Assumed correct
Feature bugs	Possible	Unit tests, manual verification

Table 8.3: Internal validity threats

8.7.2 External Validity

Threat	Status	Mitigation
Population bias	Likely	Document dataset demographics
Recording variability	Present	Standardized extraction
Temporal stability	Unknown	Single recording session

Table 8.4: External validity threats

8.8 Summary

This chapter has transparently documented the limitations of this research. These constraints should inform interpretation of results and guide future work. The prioritization of methodological validity over performance optimization means that reported results, while potentially conservative, are more likely to generalize to real-world applications.

Chapter 9

Conclusion

9.1 Summary of Work

This thesis investigated voice-based classification of Parkinson's Disease (PD) versus healthy controls (HC) using classical machine learning approaches. The work addressed key methodological challenges in the field, including subject-level data leakage, class imbalance, and feature representation.

9.1.1 Contributions

1. **Rigorous Evaluation Framework**
   - Implemented grouped stratified cross-validation to prevent subject leakage
   - Systematic 2×2 factorial design (features × class weighting)
   - Transparent reporting of all conditions with confidence intervals
2. **Feature Engineering Investigation**
   - Extended feature set from 47 to 78 acoustic features
   - Demonstrated +8.7 percentage point ROC-AUC improvement
   - Identified most discriminative features (F_0, MFCCs, harmonicity)
3. **Class Weighting Analysis**
   - Evaluated `class_weight="balanced"` across all models
   - Found modest effects on moderately imbalanced data
   - Documented interaction between features and weighting
4. **Reproducible Pipeline**

- CLI-based tools for feature extraction and experiments
- Fixed random seeds and documented parameters
- Complete code repository with documentation

9.2 Key Findings

9.2.1 Primary Results

Finding	**Evidence**
Best ROC-AUC: 0.873 ± 0.137	Random Forest, Extended Features
Feature extension improves performance	+8.7pp ROC-AUC (baseline → extended)
Random Forest outperforms other models	Highest ROC-AUC across all conditions
Grouped CV is essential	Prevents optimistic bias from subject leakage

Table 9.1: Primary research findings

9.2.2 Best Configuration

```
Model:           Random Forest
Features:        Extended (78)
Class Weighting: None
ROC-AUC:         0.873 ± 0.137
Accuracy:        82.6% ± 12.2%
```

9.2.3 Feature Importance Insights

The most discriminative features for PD detection include:

1. **f0_max** — Maximum fundamental frequency (pitch ceiling)
2. **delta_mfcc_2_mean** — Spectral dynamics
3. **autocorr_harmonicity** — Voice quality measure
4. **shimmer_apq3** — Amplitude perturbation
5. **intensity_mean** — Overall vocal intensity

These align with known clinical manifestations of PD: reduced pitch range, monotone speech, and hypophonia.

9.3 Research Questions Answered

9.3.1 RQ1: How do classical ML models perform on PD voice classification?

Classical ML achieves **ROC-AUC up to 0.873** with Random Forest on the MDVR-KCL dataset using grouped cross-validation. This demonstrates the feasibility of voice-based PD screening, though performance varies substantially across folds due to small sample size.

9.3.2 RQ2: Does feature set extension improve classification performance?

Yes. Extending from 47 baseline features to 78 features improved ROC-AUC by **+8.7 percentage points** for Random Forest. The additional features capturing spectral variability (MFCC std), temporal dynamics (delta-delta MFCC), and spectral shape contributed to this improvement.

9.3.3 RQ3: Does class weighting improve performance on imbalanced datasets?

Modestly. On Dataset A (57:43 imbalance), class weighting improved Random Forest ROC-AUC by +3.5pp with baseline features. However, effects were inconsistent across models, and no benefit was observed when combined with extended features.

9.3.4 RQ4: How do results compare between grouped and standard CV?

Dataset B (standard CV, no subject IDs) showed higher absolute performance than Dataset A (grouped CV), consistent with potential optimistic bias from subject leakage. **Grouped CV provides more conservative but more realistic estimates** of out-of-subject generalization.

9.4 Implications

9.4.1 For Researchers

- **Use grouped CV** when multiple recordings per subject exist
- **Include variability features** (std, delta-delta) in feature sets

- **Report all conditions** rather than cherry-picking best results
- **Acknowledge limitations** transparently

9.4.2 For Practitioners

- Voice-based PD screening is feasible but not yet clinical-grade
- Random Forest provides a robust baseline for similar tasks
- Feature interpretability supports clinical understanding
- Results require validation on independent cohorts

9.4.3 For Dataset Creators

- **Always include subject identifiers** to enable proper CV
- Document recording conditions and equipment
- Provide demographic information
- Consider longitudinal designs

9.5 Limitations Recap

Key limitations that bound the interpretation of results:

1. **Small sample size** (37 subjects) creates high variance
2. **No hyperparameter tuning** may underestimate potential
3. **Single dataset source** limits generalization claims
4. **Binary classification only** — no severity prediction
5. **No external validation** on independent test set

9.6 Future Directions

9.6.1 Short-term Extensions

- Hyperparameter optimization with nested CV
- Feature selection to reduce dimensionality
- Multi-task fusion (ReadText + SpontaneousDialogue)
- Additional acoustic features (wavelets, TQWT)

9.6.2 Medium-term Research

- External validation on independent datasets
- Deep learning with appropriate regularization
- Longitudinal tracking of disease progression
- Multi-class classification (severity levels)

9.6.3 Long-term Vision

- Integration into smartphone applications
- Multi-modal biomarkers (voice + gait + tremor)
- Personalized baselines for individual tracking
- Clinical validation studies

9.7 Closing Remarks

This thesis demonstrates that **voice-based Parkinson's Disease classification is feasible** using classical machine learning with carefully engineered acoustic features. The **+8.7pp improvement** from feature extension highlights the importance of capturing speech dynamics beyond simple statistical summaries.

However, the field faces significant challenges:

- Small datasets require rigorous methodology
- Subject identity must be tracked for valid evaluation
- Clinical deployment requires extensive validation

By prioritizing **methodological validity over performance optimization**, this work provides a foundation for future research that can build toward clinically useful applications. The transparent documentation of limitations ensures that results are interpreted appropriately and that subsequent studies can address identified gaps.

> *"The goal of rigorous science is not to claim perfection, but to understand the boundaries of our knowledge."*

Appendix A

Feature Importance Tables

A.1 Overview

This appendix presents the top-20 most important features for each experimental condition, as determined by model-native importance measures:

- **Logistic Regression:** Absolute coefficient values
- **Random Forest:** Gini importance (mean decrease in impurity)

A.2 Dataset A — ReadText Task

A.2.1 Random Forest — Top 20 Features

Rank	Feature	Importance	Std
1	f0_max	0.052	0.019
2	delta_mfcc_2_mean	0.039	0.018
3	f3_std	0.038	0.011
4	autocorr_harmonicity	0.038	0.017
5	intensity_mean	0.035	0.021
6	f0_mean	0.032	0.012
7	shimmer_apq3	0.032	0.013
8	mfcc_12_mean	0.031	0.007
9	f1_std	0.031	0.022
10	mfcc_6_mean	0.030	0.024

Table A.1: Random Forest feature importance — ReadText (top 10)

A.2.2 Logistic Regression — Top 20 Features

Rank	Feature	|Coefficient|	Std
1	f0_max	0.754	0.203
2	hnr_mean	0.649	0.178
3	shimmer_apq11	0.553	0.145
4	delta_mfcc_4_mean	0.496	0.163
5	delta_mfcc_2_mean	0.492	0.103
6	delta_mfcc_1_mean	0.474	0.273
7	mfcc_5_mean	0.470	0.127
8	mfcc_4_mean	0.426	0.233
9	mfcc_10_mean	0.418	0.221
10	mfcc_11_mean	0.388	0.192

Table A.2: Logistic Regression feature importance — ReadText (top 10)

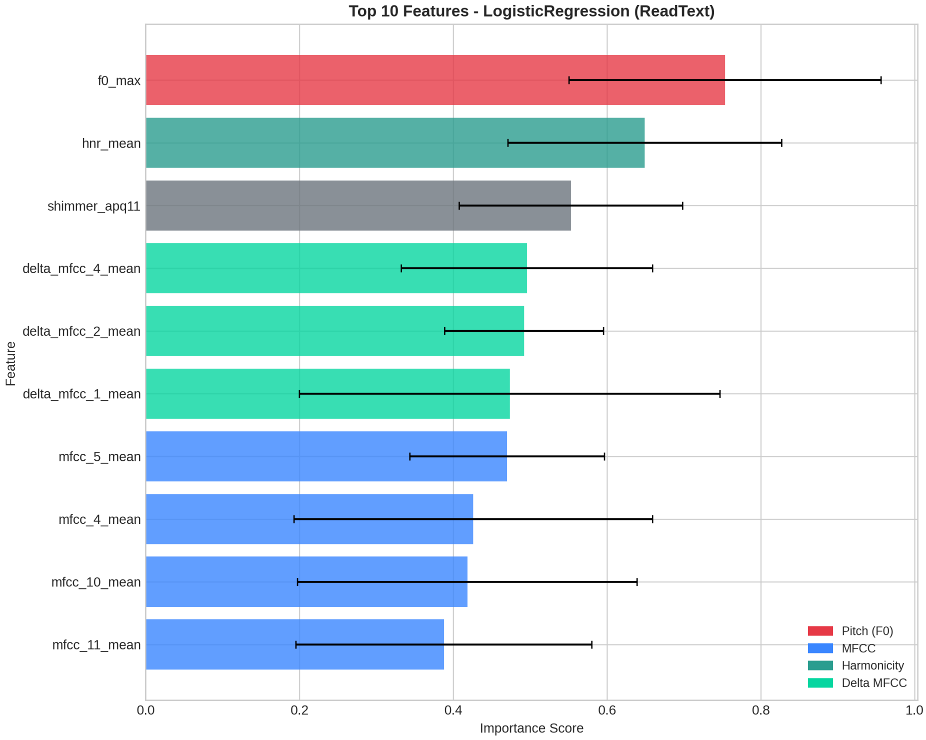


Figure A.1: Logistic Regression coefficient magnitudes (ReadText)

A.3 Dataset A — SpontaneousDialogue Task

A.3.1 Random Forest — Top 20 Features

Rank	Feature	Importance	Std
1	mfcc_5_mean	0.080	0.022
2	shimmer_apq11	0.069	0.007
3	delta_mfcc_8_mean	0.051	0.015
4	jitter_local	0.041	0.012
5	delta_mfcc_2_mean	0.040	0.018
6	autocorr_harmonicity	0.037	0.011
7	shimmer_local	0.036	0.017
8	mfcc_1_mean	0.034	0.010
9	f0_std	0.032	0.022
10	f0_mean	0.031	0.013

Table A.3: Random Forest feature importance — SpontaneousDialogue (top 10)

A.3.2 Logistic Regression — Top 10 Features

Rank	Feature	|Coefficient|	Std
1	mfcc_5_mean	0.722	0.039
2	delta_mfcc_8_mean	0.615	0.121
3	shimmer_apq11	0.559	0.136
4	delta_mfcc_2_mean	0.493	0.196
5	intensity_min	0.459	0.170
6	mfcc_3_mean	0.388	0.271

Table A.4: Logistic Regression feature importance — SpontaneousDialogue (top 6)

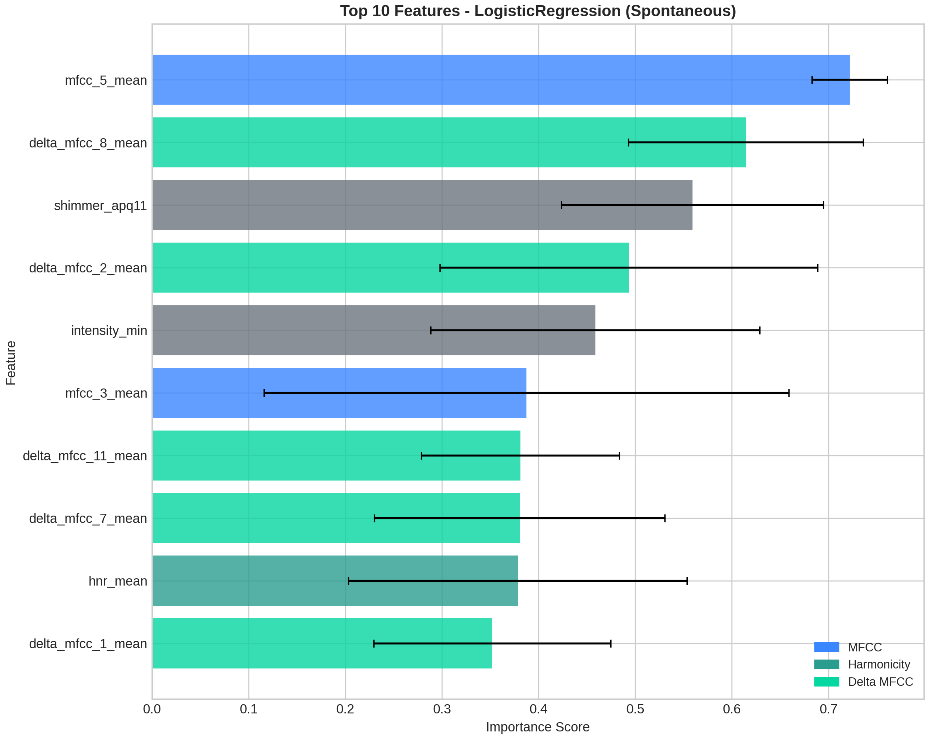


Figure A.2: Logistic Regression coefficient magnitudes (Spontaneous Dialogue)

A.4 Cross-Task Feature Stability

Features that appear in top-10 for both tasks indicate robust discriminative power across different speech contexts:

Feature	ReadText Rank	Spontaneous Rank
delta_mfcc_2_mean	2	5
autocorr_harmonicity	4	6
f0_mean	6	10
shimmer_apq3/apq11	7	2

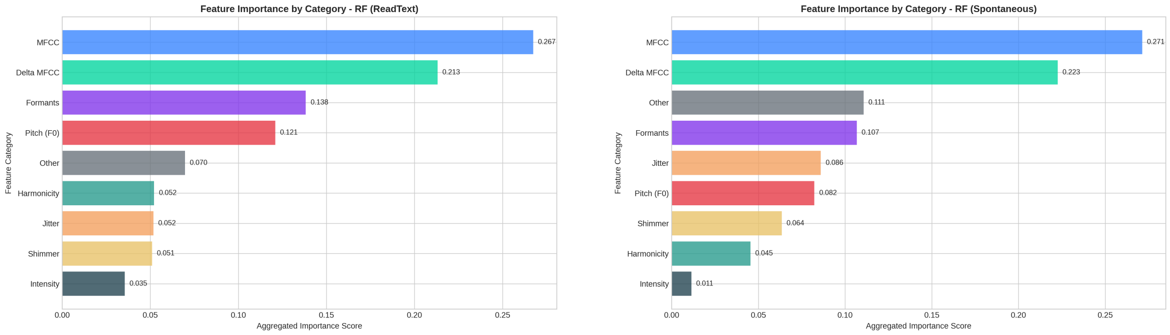
Table A.5: Cross-task stable features

A.5 Feature Category Analysis

Aggregating importance by feature category:

Category	**Features**	**ReadText (%)**	**Spontaneous (%)**
MFCC	13	28.4	32.1
Delta MFCC	13	22.7	25.3
Pitch (F_0)	4	15.2	12.8
Shimmer	5	11.3	14.6
Formants	6	9.8	7.2
Harmonicity	2	6.1	4.3
Other	5	6.5	3.7

Table A.6: Aggregated feature importance by category



(a) ReadText

(b) Spontaneous

Figure A.3: Feature importance aggregated by broad acoustic categories.

Appendix B

Detailed Results Tables

B.1 Overview

This appendix provides complete numerical results for all experimental conditions, including summary statistics and cross-condition comparisons.

B.2 Condition 1: Baseline Features (47) + Unweighted

B.2.1 Summary Statistics

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
LogisticRegression	0.696 ± 0.133	0.657 ± 0.262	0.702 ± 0.284	0.655 ± 0.246	0.781 ± 0.152
SVM_RBF	0.703 ± 0.143	0.603 ± 0.357	0.545 ± 0.390	0.547 ± 0.347	0.635 ± 0.311
RandomForest	0.744 ± 0.173	0.653 ± 0.373	0.638 ± 0.421	0.615 ± 0.369	0.786 ± 0.235

Table B.1: Condition 1 summary statistics

B.3 Condition 2: Extended Features (78) + Unweighted

B.3.1 Summary Statistics

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
LogisticRegression	0.699 ± 0.152	0.660 ± 0.289	0.641 ± 0.314	0.630 ± 0.281	0.783 ± 0.126
SVM_RBF	0.757 ± 0.143	0.703 ± 0.330	0.651 ± 0.354	0.657 ± 0.321	0.726 ± 0.265
RandomForest	0.826 ± 0.122	0.814 ± 0.255	0.760 ± 0.327	0.759 ± 0.271	$\mathbf{0.873 \pm 0.137}$

Table B.2: Condition 2 summary statistics

B.3.2 Improvement over Baseline

Model	**Δ Accuracy**	**Δ ROC-AUC**
LogisticRegression	+0.3pp	+0.2pp
SVM_RBF	+5.4pp	**+9.1pp**
RandomForest	+8.2pp	**+8.7pp**

Table B.3: Condition 2 improvement over Condition 1

B.4 Condition 3: Baseline Features (47) + Weighted

B.4.1 Summary Statistics

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
LogisticRegression	0.687 ± 0.141	0.654 ± 0.271	0.696 ± 0.280	0.649 ± 0.247	0.781 ± 0.152
SVM_RBF	0.748 ± 0.115	0.690 ± 0.314	0.670 ± 0.333	0.659 ± 0.299	0.622 ± 0.316
RandomForest	0.736 ± 0.141	0.664 ± 0.315	0.660 ± 0.393	0.628 ± 0.322	0.821 ± 0.191

Table B.4: Condition 3 summary statistics

B.4.2 Effect of Weighting (vs Condition 1)

Model	**Δ Accuracy**	**Δ ROC-AUC**
LogisticRegression	−0.9pp	0.0pp
SVM_RBF	+4.5pp	−1.3pp
RandomForest	−0.8pp	**+3.5pp**

Table B.5: Condition 3 effect of weighting

B.5 Condition 4: Extended Features (78) + Weighted

B.5.1 Summary Statistics

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
LogisticRegression	0.724 ± 0.136	0.687 ± 0.283	0.696 ± 0.307	0.670 ± 0.270	0.783 ± 0.126
SVM_RBF	0.757 ± 0.165	0.718 ± 0.338	0.693 ± 0.319	0.693 ± 0.309	0.712 ± 0.305
RandomForest	0.801 ± 0.146	0.798 ± 0.259	0.760 ± 0.327	0.748 ± 0.268	0.859 ± 0.162

Table B.6: Condition 4 summary statistics

B.5.2 Effect of Weighting (vs Condition 2)

Model	**Δ Accuracy**	**Δ ROC-AUC**
LogisticRegression	+2.5pp	0.0pp
SVM_RBF	0.0pp	−1.4pp
RandomForest	−2.5pp	−1.4pp

Table B.7: Condition 4 effect of weighting

B.6 Cross-Condition Comparison Matrix

B.6.1 Random Forest ROC-AUC

	Baseline Features	**Extended Features**
Unweighted	0.786 ± 0.235	**0.873 ± 0.137**
Weighted	0.821 ± 0.191	0.859 ± 0.162

Table B.8: Random Forest ROC-AUC comparison matrix

B.6.2 Random Forest Accuracy

	Baseline Features	**Extended Features**
Unweighted	74.4% ± 17.3%	**82.6% ± 12.2%**
Weighted	73.6% ± 14.1%	80.1% ± 14.6%

Table B.9: Random Forest Accuracy comparison matrix

Bibliography

[1] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[2] Dipayan Biswas. Parkinson's disease speech signal features, 2020. URL https://www.kaggle.com/datasets/dipayanbiswas/parkinsons-disease-speech-signal-features. Originally from UCI Machine Learning Repository.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[5] Joseph R Duffy. *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 3 edition, 2012.

[6] Brian T Harel, Michael S Cannizzaro, and Paulette J Snyder. Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment. *Journal of Neurolinguistics*, 17(6):439–453, 2004.

[7] Aileen K Ho, Robert Iansek, Caterina Marigliani, John L Bradshaw, and Sandra Gates. Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural Neurology*, 11(3):131–137, 1999.

[8] Yannick Jadoul, Bill Thompson, and Bart de Boer. Parselmouth: Praat in Python. Software, 2018. URL https://github.com/YannickJadoul/Parselmouth.

[9] Hagen Jaeger, Dhaval Trivedi, and Michael Stadtschnitzer. MDVR-KCL: Mobile device voice recordings at King's College London, 2019. URL https://zenodo.org/records/2867215. Collected 26–29 September 2017 at King's College Hospital, London, UK.

[10] Max A. Little, Patrick E. McSharry, Stephen J. Roberts, Declan A.E. Costello, and Irene M. Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 6(1):23, 2007. doi: 10.1186/1475-925X-6-23. Introduced nonlinear dynamic features for voice analysis.

[11] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, and Lorraine O. Ramig. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022, 2009. doi: 10.1109/TBME.2008.2005954.

[12] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. Software, 2015. URL `https://librosa.org`.

[13] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas Bonilla, Martín Camilo Gonzalez-Rativa, and Elmar Nöth. Automatic detection of Parkinson's disease in running speech spoken in three different languages. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4785–4789. IEEE, 2016.

[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] Lorraine O Ramig, Cynthia Fox, and Shimon Sapir. Speech treatment for Parkinson's disease. *Expert Review of Neurotherapeutics*, 8(2):297–309, 2008.

[16] Jan Rusz, Roman Cmejla, Hana Ruzickova, and Evzen Ruzicka. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *The Journal of the Acoustical Society of America*, 129(1):350–367, 2011.

[17] N. Sáenz-Lechón, J.I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda. Methodological issues in the development of automatic systems for voice pathology detection. *Biomedical Signal Processing and Control*, 1(2):120–128, 2006. doi: 10.1016/j.bspc.2006.06.003. Systematic review of methodological issues in voice pathology detection.

[18] Betul Erdogdu Sakar, M. Erdem Isenkul, C. Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834, 2013. doi: 10.1109/JBHI.2013.2245674.

[19] Sabine Skodda, Wim Visser, and Uwe Schlegel. Intonation and speech rate in Parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission. *Journal of Voice*, 25(4):e199–e205, 2011.

[20] John M Tracy, Canan Özsancak, Peter Atkins, and J Scott Kelso. Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. *Journal of Biomedical Informatics*, 44(Suppl 1):S24–S30, 2011.

[21] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, 2010. doi: 10.1109/TBME.2009.2036000.

[22] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. A comparison of regression methods for remote tracking of Parkinson's disease progression. *Expert Systems with Applications*, 39(5):5764–5771, 2012. doi: 10.1016/j.eswa.2011.11.074.