[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 $\rightarrow$ 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 $\sim$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 impor-

tant, 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, each representing a different paradigm in the feature extraction pipeline:

Property	**Dataset A (MDVR-KCL)**	**Dataset B (PD_SPEECH)**
Data Type	Raw audio (WAV)	Pre-extracted features (CSV)
Source	Zenodo	Kaggle/UCI
Unit of Analysis	Subject (grouped recordings)	Sample row (unknown subjects)
Subject IDs Available	Yes (37 unique)	No
Total Samples	73 recordings	756 rows
Speech Task	Read text / Dialogue	Sustained /a/ phonation
Feature Extraction	Performed in this work	Pre-computed by authors

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 [9] was collected specifically for Parkinson's Disease research using smartphone recordings. The dataset is publicly available on Zenodo with DOI `10.5281/zenodo.2867215`.

Collection Context

Property	Value
Collection Period	26–29 September 2017
Location	King's College London Hospital, Denmark Hill, London, UK
Recording Device	Motorola Moto G4 smartphone
Environment	Clinical examination room (∼10 m²)
Reverberation Time	∼500 ms
Audio Capture	Direct microphone signal (no GSM compression)

Table 3.2: MDVR-KCL data collection context

Recordings were captured within the reverberation radius directly from the microphone signal (not GSM-compressed), resulting in acoustically clean audio suitable for feature extraction.

3.2.2 Audio Specifications

Property	Value
Format	WAV (uncompressed PCM)
Native Sample Rate	44.1 kHz
Processing Sample Rate	22.05 kHz (resampled)
Bit Depth	16-bit signed integer
Channels	Mono
Compression	None

Table 3.3: Audio specifications for Dataset A

All audio files are resampled to 22.05 kHz for feature extraction to ensure consistency and computational efficiency while preserving sufficient frequency resolution for speech analysis (Nyquist frequency: 11.025 kHz).

3.2.3 Speech Tasks

The dataset includes two distinct speech tasks designed to capture different aspects of vocal production:

Task	Description	Subjects	HC	PD
ReadText	"The North Wind and the Sun" passage	37	21	16
SpontaneousDialogue	Free conversation with examiner	36	21	15

Table 3.4: Speech tasks in MDVR-KCL dataset

Note: Subject ID18 is missing from the SpontaneousDialogue task, resulting in one fewer recording for that condition.

3.2.4 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 HC:PD). Class weighting experiments were conducted to assess impact on classification performance.

3.2.5 Filename Format and Clinical Metadata

Filenames encode clinical metadata in the format:

`ID{XX}_{label}_{H&Y}_{UPDRS_speech}_{UPDRS_total}.wav`

Field	Description
ID{XX}	Subject identifier (00–99)
label	`hc` (Healthy Control) or `pd` (Parkinson's Disease)
H&Y	Hoehn & Yahr stage (0 for HC, 1–5 for PD)
UPDRS_speech	UPDRS Item 18 score (0 for HC, 0–4 for PD)
UPDRS_total	Total UPDRS score (0 for HC)

Table 3.5: Filename encoding for clinical metadata

Example: `ID05_pd_2_1_45.wav` indicates:

- Subject 05, Parkinson's Disease
- Hoehn & Yahr stage 2
- UPDRS speech score: 1
- Total UPDRS score: 45

Note: Clinical metadata were not used as features in this work to maintain consistency with Dataset B, which lacks such information.

3.2.6 File Structure

```
DATASET_MDVR_KCL/
+-- ReadText/
|   +-- HC/
|   |   +-- IDxx_hc_*.wav (21 files)
|   +-- PD/
|       +-- IDxx_pd_*.wav (16 files)
+-- SpontaneousDialogue/
    +-- HC/
    |   +-- IDxx_hc_*.wav (21 files)
    +-- PD/
        +-- IDxx_pd_*.wav (15 files)
```

3.2.7 Known Anomalies and Handling

- **ID22:** Non-standard filename pattern in source data (handled in parsing code via fallback logic)
- **ID18:** Missing from SpontaneousDialogue task (36 subjects vs. 37 in ReadText)
- **Multiple recordings per subject:** Requires grouped cross-validation to prevent data leakage

3.2.8 Feature Extraction Pipeline

For Dataset A, features are extracted from raw audio in four preprocessing steps followed by feature computation:

Preprocessing Steps

1. **Load audio** at native sample rate (44.1 kHz)
2. **Convert to mono** if stereo (via channel averaging)
3. **Resample to 22.05 kHz** for computational efficiency
4. **Normalize amplitude** to $[-1, 1]$ range

Feature Sets

Two feature configurations were evaluated:

Baseline Features (47 total):

- **Prosodic (21):** Pitch (F_0), jitter, shimmer, harmonicity, intensity, formants

- **Spectral (26):** MFCC mean (13) + Delta MFCC mean (13)

Extended Features (78 total):

- **Prosodic (21):** Unchanged
- **Spectral (57):** Baseline (26) + MFCC std (13) + Delta-delta MFCC mean (13) + Spectral shape (5)

Prosodic Feature Breakdown

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

Table 3.6: Prosodic feature breakdown (21 features)

Extraction parameters:

- F_0 range: 75–500 Hz (covers male, female, and pathological voices)
- Jitter/shimmer computed on **voiced frames only**
- Formants extracted via Burg's method (LPC order: 5)

Spectral Feature Breakdown

Feature Group	Count	Description	Tool
Baseline (26):			
MFCC mean	13	Mean of MFCCs 0–12	librosa
Delta MFCC mean	13	Mean of first-order derivatives	librosa
Extended only (+31):			
MFCC std	13	Std deviation of MFCCs 0–12	librosa
Delta-delta MFCC mean	13	Mean of second-order derivatives	librosa
Spectral shape	5	Centroid, bandwidth, rolloff, flatness, ZCR	librosa

Table 3.7: Spectral feature breakdown

Technical Feature Extraction Parameters

Parameter	Value
Target Sample Rate	22.05 kHz
MFCC Coefficients	13 (0–12)
FFT Window Size	2048 samples (∼93 ms)
Hop Length	512 samples (∼23 ms)
Mel Filter Banks	128
F_0 Range	75–500 Hz

Table 3.8: Feature extraction technical parameters

3.2.9 Feature Correlation Analysis

Figure 3.1 and Figure 3.2 show feature correlation matrices for ReadText and SpontaneousDialogue tasks, respectively. These heatmaps reveal:

- Strong correlations within feature families (e.g., adjacent MFCC coefficients)
- Minimal correlation between prosodic and spectral features
- Task-specific correlation patterns suggesting different information content

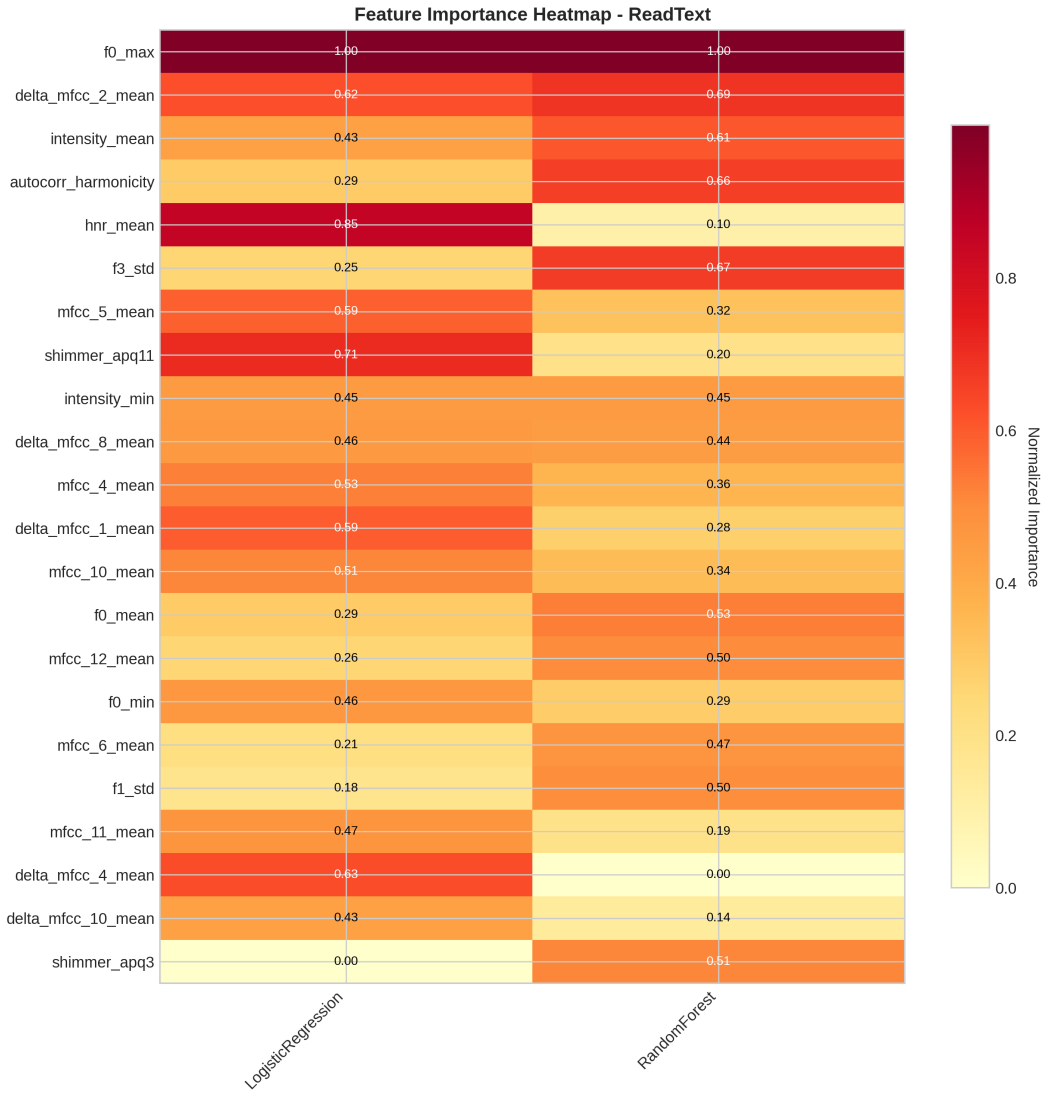


Figure 3.1: Feature correlation heatmap for ReadText task. Darker colors indicate stronger correlations. MFCC coefficients show expected sequential correlation structure.

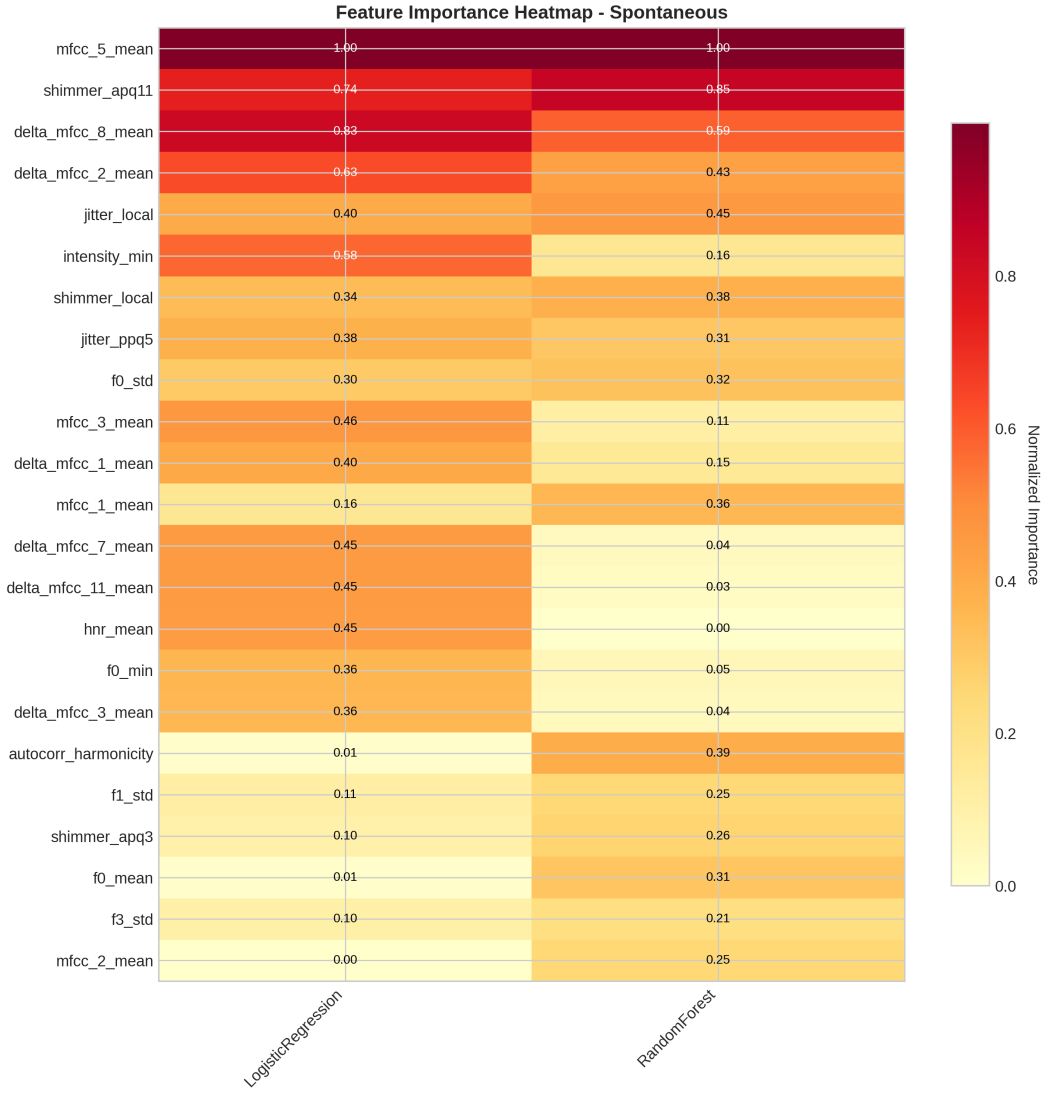


Figure 3.2: Feature correlation heatmap for SpontaneousDialogue task. Correlation patterns differ from ReadText, particularly in prosodic features, reflecting the unstructured nature of spontaneous speech.

3.3 Dataset B: PD Speech Features

3.3.1 Source

Pre-extracted acoustic features from the UCI Machine Learning Repository, distributed via Kaggle [2]. The dataset contains 752 acoustic features pre-computed from sustained vowel phonations.

3.3.2 Collection Context

Property	Value
Institution	Istanbul University, Cerrahpaşa Faculty of Medicine
Collection Period	Not specified in metadata
PD Subjects	188
HC Subjects	64
Age Range (PD)	33–87 years
Age Range (HC)	41–82 years
Speech Task	Sustained phonation of vowel /a/
Repetitions	3 per subject
Native Sample Rate	44.1 kHz (reported in metadata)

Table 3.9: Collection context for Dataset B

3.3.3 Data Format

- **Format:** CSV (comma-separated values)
- **Rows:** 756 samples (multiple per subject)
- **Columns:** 753 total (752 features + 1 binary label)
- **Label encoding:** 1 = PD, 0 = HC
- **Subject IDs: Not provided** (critical limitation)

3.3.4 Class Distribution

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

Table 3.10: Class distribution in Dataset B

Imbalance Ratio: Severe (∼25:75 HC:PD), necessitating class weighting strategies in model training.

3.3.5 Feature Categories

The 752 features span multiple acoustic domains computed by the original authors:

Category	Count	Description
Baseline	22	Jitter variants, shimmer variants, HNR, NHR
Intensity	3	Min, max, mean intensity
Formants	36	F_1–F_4 frequency and bandwidth statistics
MFCCs	84	Mean, std, delta for MFCC 0–12
Wavelet (DWT)	182	Discrete wavelet decomposition features
TQWT	432	Tunable Q-factor wavelet transform features
Other	7	PPE, DFA, RPDE, numPulses, GQ, GNE, VFER

Table 3.11: Feature categories in Dataset B (752 total)

3.3.6 Methodological Caveats

Critical Limitation: No subject identifiers are available in Dataset B. This prevents validation of true out-of-subject generalization, as the same subject may appear in both training and test folds during cross-validation.

Given that 756 samples were collected from 252 subjects (188 PD + 64 HC) with 3 repetitions each, standard stratified 5-fold cross-validation may place samples from the same subject in different folds, leading to:

- **Optimistically biased performance estimates** (overestimation of generalization)
- **Violation of independence assumption** in cross-validation
- **Inability to assess subject-level generalization**

Results on Dataset B should be interpreted cautiously with these limitations in mind.

3.3.7 Additional Limitations

- **No raw audio:** Cannot verify or modify feature extraction process
- **Feature extraction pipeline unavailable:** Cannot reproduce or extend feature set
- **Sustained vowel only:** Does not represent connected speech or natural communication
- **High feature dimensionality:** 752 features may lead to overfitting with small sample size

3.4 Cross-Dataset Comparison

3.4.1 Key Differences

Aspect	Dataset A	Dataset B
Data format	Raw audio (WAV)	Pre-extracted features (CSV)
Sample size	37 subjects (73 recordings)	756 rows (subjects unknown)
Subject tracking	Available	**Unavailable**
Feature dimensionality	47 (baseline) or 78 (extended)	752 (fixed)
Speech task	Read text / Dialogue	Sustained /a/ phonation
Feature extraction	Controlled (this work)	Pre-computed (black box)
Cross-validation	Grouped (subject-level)	Standard (row-level)
Clinical metadata	H&Y, UPDRS available	None

Table 3.12: Detailed cross-dataset comparison

3.4.2 Cross-Validation Strategy

Dataset	CV Strategy	Rationale
Dataset A	Grouped Stratified 5-Fold	Ensures all recordings from one subject stay in same fold
Dataset B	Standard Stratified 5-Fold	No subject IDs available for grouping

Table 3.13: Cross-validation strategies by dataset

Implications:

- Dataset A provides **conservative, realistic** out-of-subject generalization estimates
- Dataset B may provide **optimistic** estimates due to potential within-subject correlation
- Direct performance comparison is confounded by these methodological differences

3.4.3 Complementary Value

Despite their differences, the datasets provide complementary perspectives:

- **Dataset A:** Explores effect of **task** (read vs. spontaneous) and **feature engineering** (baseline vs. extended)
- **Dataset B:** Tests generalization to **high-dimensional feature space** with different acoustic representations (wavelet transforms)

- **Together:** Enable assessment of classifier robustness across different data collection protocols, feature extraction pipelines, and speech tasks

3.5 Summary

This chapter presented two complementary datasets for PD voice classification:

1. **Dataset A (MDVR-KCL):** 37 subjects with raw smartphone recordings enabling controlled feature extraction (47/78 features) and rigorous subject-level evaluation
2. **Dataset B (PD Speech Features):** 756 samples with 752 pre-extracted features enabling high-dimensional classification exploration, with caveats regarding unknown subject overlap

These datasets should be interpreted cautiously given the small sample size (Dataset A: $n = 37$) and the absence of subject identifiers in Dataset B. Results represent exploratory analyses rather than definitive performance benchmarks.

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. All parameters were fixed a priori and locked in a central configuration file to ensure experimental validity. All parameters were fixed a priori and locked in a central configuration file to ensure experimental validity.

To support qualitative inspection of the end-to-end inference workflow (without contributing to evaluation), a lightweight research demonstration interface was implemented and is described in Appendix C.

4.2 Feature Extraction Pipeline

4.2.1 Pipeline Architecture

The feature extraction pipeline transforms raw audio into structured feature vectors suitable for machine learning:

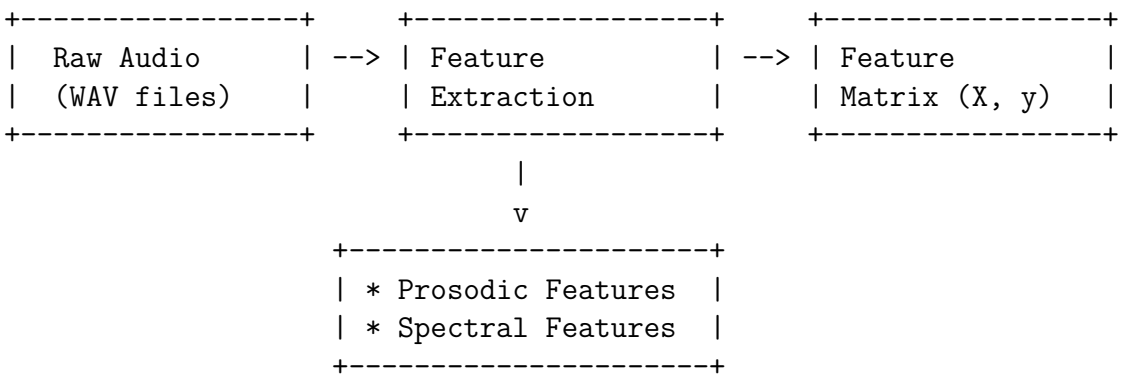


Figure 4.1: Feature extraction pipeline architecture. Each audio file produces one feature vector.

4.2.2 Audio Preprocessing

All audio files undergo standardized preprocessing to ensure consistent feature extraction:

1. **Load audio** at native sample rate using `librosa.load()`
2. **Resample** to 22,050 Hz (standardized across all recordings)
3. **Convert to mono** if stereo (channel averaging)
4. **Normalize amplitude** to $[-1, 1]$ range
5. **Trim silence** using energy-based detection (threshold: top 5% energy)

The 22,050 Hz sample rate was chosen as it provides adequate frequency resolution for speech (Nyquist frequency: 11,025 Hz) while reducing computational cost compared to 44,100 Hz.

4.2.3 Prosodic Features (21 features)

Prosodic features capture suprasegmental voice characteristics known to be affected in PD. All prosodic features were extracted using Parselmouth [8], a Python interface to Praat.

Fundamental Frequency (F_0)

Pitch extraction used the following parameters:

- **Algorithm:** Autocorrelation-based (Praat's default)
- **Pitch range:** 75–500 Hz (covers male, female, and pathological voices)
- **Time step:** 0.01 s (standard for voice analysis)

Four statistics were computed from the pitch contour:

$$F_0 = \{\mu_{F_0}, \sigma_{F_0}, \min_{F_0}, \max_{F_0}\} \tag{4.1}$$

Jitter

Jitter quantifies cycle-to-cycle variation in pitch period. Three measures were extracted:

$$\text{Jitter}_{\text{local}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|T_i - T_{i+1}|}{\bar{T}} \tag{4.2}$$

where T_i is the i-th pitch period and $\bar{T}$ is the mean period.

Additional measures include:

- **RAP (Relative Average Perturbation):** 3-point smoothing
- **PPQ5:** 5-point period perturbation quotient

Shimmer

Shimmer quantifies cycle-to-cycle variation in amplitude. Five measures were extracted:

$$\text{Shimmer}_{\text{local}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|A_i - A_{i+1}|}{\bar{A}} \tag{4.3}$$

where A_i is the i-th peak amplitude and $\bar{A}$ is the mean amplitude.

Additional measures include APQ3, APQ5, APQ11 (amplitude perturbation quotients with varying windows), and DDA (difference of differences of amplitudes).

Harmonics-to-Noise Ratio (HNR)

HNR quantifies the ratio of harmonic to noise components:

- **Mean HNR:** Average HNR across the utterance (dB)
- **Autocorrelation harmonicity:** Peak autocorrelation value

Intensity

Three intensity statistics (dB SPL):

$$I = \{\mu_I, \sigma_I, \max_I - \min_I\} \tag{4.4}$$

Formants

First three formants (F_1, F_2, F_3) were extracted using Linear Predictive Coding (LPC):

- **LPC order:** 10 (sufficient for 3 formants at 22 kHz)
- **Features:** Mean and standard deviation of each formant (6 features total)

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
Total	**21**		

Table 4.1: Prosodic feature breakdown

4.2.4 Spectral Features

Spectral features capture frequency-domain characteristics using the `librosa` library [12].

Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are computed via the following steps:

1. **Pre-emphasis:** $y[n] = x[n] - 0.97 \cdot x[n-1]$
2. **Framing:** Window length = 2048 samples (≈93 ms at 22 kHz)
3. **Windowing:** Hann window applied
4. **FFT:** 2048-point FFT
5. **Mel filterbank:** 128 triangular filters spanning 0–11,025 Hz
6. **Log compression:** log(mel power spectrum)
7. **DCT:** Discrete Cosine Transform to decorrelate coefficients

Parameters used:

- **Number of coefficients:** 13 (MFCC 0–12)
- **Hop length:** 512 samples (≈23 ms at 22 kHz, 75% overlap)

- **Number of mel bands:** 128

MFCC 0 (the zeroth coefficient) captures energy information, while MFCCs 1–12 capture spectral envelope shape.

Delta Coefficients

First-order temporal derivatives (delta coefficients) capture spectral dynamics:

$$\Delta_t[n] = \frac{\sum_{i=1}^{N} i \cdot (c_{t+i} - c_{t-i})}{2 \sum_{i=1}^{N} i^2} \tag{4.5}$$

where c_t is the MFCC coefficient at frame t and $N = 2$ is the window size.

Baseline Spectral Features (26 features)

The baseline spectral feature set consists of:

Feature	Count	Description
MFCC mean	13	Temporal mean of MFCCs 0–12
Delta MFCC mean	13	Temporal mean of first-order derivatives
Total	**26**	

Table 4.2: Baseline spectral features

Extended Spectral Features (57 features)

The extended feature set adds three groups of features designed to capture additional spectral and temporal variability:

Feature	Count	Description
MFCC mean	13	Temporal mean of MFCCs 0–12
MFCC std	**13**	**Temporal std deviation (within-utterance variability)**
Delta MFCC mean	13	Temporal mean of first-order derivatives
Delta-Delta MFCC mean	**13**	**Temporal mean of second-order derivatives (acceleration)**
Spectral shape	**5**	**Centroid, bandwidth, rolloff, flatness, ZCR**
Total	**57**	

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

The five spectral shape descriptors are:

1. **Spectral Centroid:** Center of mass of the spectrum (Hz)
2. **Spectral Bandwidth:** Weighted standard deviation around centroid (Hz)
3. **Spectral Rolloff:** Frequency below which 85% of energy is contained (Hz)
4. **Spectral Flatness:** Ratio of geometric to arithmetic mean (tonality measure)
5. **Zero-Crossing Rate:** Rate of sign changes in time-domain signal

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** to test the hypothesis that additional temporal and spectral information improves classification. Three feature groups were added:

1. **MFCC std (13):** Captures within-utterance variability—important for detecting instability in PD speech where motor fluctuations may manifest as increased spectral variance
2. **Delta-Delta MFCC (13):** Captures acceleration of spectral changes—sensitive to temporal dynamics and rate-of-change in articulatory movements
3. **Spectral shape (5):** Provides complementary global spectral descriptors not captured by MFCCs (e.g., overall tonality, spectral spread)

4.3.2 Feature Extraction Reproducibility

All feature extraction parameters were defined in a central configuration module and fixed before experimentation:

- Random seed: 42 (for any stochastic components)

- Sample rate: 22,050 Hz
- F_0 range: 75–500 Hz
- MFCC parameters: 13 coefficients, 2048 FFT size, 512 hop length, 128 mel bands

Feature extraction was performed once per dataset and feature configuration, with results saved to CSV files for all downstream experiments.

4.4 Machine Learning Models

4.4.1 Model Selection Rationale

Three classical ML models were selected to provide diverse inductive biases while maintaining interpretability:

- **Logistic Regression (LR):** Linear model with inherent interpretability via coefficient weights. Tests whether classes are linearly separable in feature space.
- **Support Vector Machine with RBF kernel (SVM):** Nonlinear kernel-based model capable of learning complex decision boundaries. Tests whether nonlinear transformations improve separation.
- **Random Forest (RF):** Ensemble of decision trees with built-in feature importance. Tests whether hierarchical feature interactions are informative.

These models were chosen for:

- **Interpretability** — Critical for clinical applications where decisions must be explainable
- **Robustness** — Well-understood behavior on small datasets (n $<$ 100)
- **Diversity** — Represent linear, kernel-based, and ensemble approaches

Deep learning models (CNNs, RNNs) were explicitly excluded due to:

- Insufficient training data (risk of severe overfitting)
- Lack of interpretability
- Computational requirements disproportionate to dataset size

4.4.2 Model Specifications

All hyperparameters were fixed a priori without dataset-specific tuning:

Model	Type	Key Parameters
Logistic Regression	Linear	C = 1.0 (L2 regularization), max_iter= 1000, solver='lbfgs'
SVM (RBF)	Kernel	C = 1.0 (regularization), gamma='scale' (auto-scaled by features), kernel='rbf'
Random Forest	Ensemble	n_estimators= 100, max_depth= 10, min_samples_split= 2, random_state= 42

Table 4.5: Model specifications. All parameters fixed before experiments.

Logistic Regression

The logistic regression model predicts class probabilities via:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}} \tag{4.6}$$

with L2 regularization:

$$\min_{\mathbf{w},b} \left[C \sum_{i=1}^{n} \log(1 + e^{-y_i(\mathbf{w}^T\mathbf{x}_i+b)}) + \frac{1}{2}\|\mathbf{w}\|_2^2 \right] \tag{4.7}$$

where $C = 1.0$ controls the regularization strength.

Support Vector Machine

The RBF kernel SVM learns a nonlinear decision boundary via:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \tag{4.8}$$

where $\gamma = \frac{1}{n_{\text{features}} \cdot \text{Var}(\mathbf{X})}$ (auto-scaled).

Random Forest

The Random Forest aggregates predictions from 100 decision trees:

$$\hat{y} = \frac{1}{N_{\text{trees}}} \sum_{t=1}^{N_{\text{trees}}} \hat{y}_t(\mathbf{x}) \tag{4.9}$$

Each tree is trained on a bootstrap sample with random feature subsets at each split. The maximum depth of 10 was chosen to prevent overfitting on small datasets.

4.4.3 Class Weighting

Class imbalance is addressed via the `class_weight` parameter:

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

# Weighted (balanced condition)
class_weight = "balanced"  # w_k = n_samples / (n_classes * n_k)
```

When `class_weight="balanced"`, sample weights are computed as:

$$w_k = \frac{n_{\text{samples}}}{n_{\text{classes}} \cdot n_k} \tag{4.10}$$

where n_k is the number of samples in class k. This ensures minority class errors are weighted more heavily during training.

All three models support the `class_weight` parameter natively via scikit-learn.

4.5 ML Pipeline Architecture

4.5.1 Pipeline Structure

All models use a standardized scikit-learn Pipeline:

```
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler

pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('classifier', Model(class_weight=..., random_state=42))
])
```

This ensures feature scaling is applied consistently and prevents data leakage.

4.5.2 Feature Standardization

All features are standardized to zero mean and unit variance:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{4.11}$$

where μ_j and σ_j are computed from the training fold only. This prevents information leakage from test data.

Critical: The scaler is fitted on training data and then applied to test data within each cross-validation fold. This ensures the test set remains completely unseen during standardization.

4.6 Evaluation Framework

4.6.1 Cross-Validation Strategy

Different strategies were required for each dataset due to their structural differences:

Dataset	Strategy	Folds	Grouping
Dataset A (MDVR-KCL)	Grouped Stratified	5	By subject_id
Dataset B (PD Speech)	Stratified	5	None (unavailable)

Table 4.6: Cross-validation strategies

Dataset A: Grouped Cross-Validation

For Dataset A (MDVR-KCL), multiple recordings per subject necessitate subject-level splitting to prevent data leakage. The strategy ensures:

- All recordings from the same subject appear in the same fold
- Approximate class balance across folds (stratification by label)
- Training never sees recordings from subjects in the test set

This simulates realistic deployment where the model encounters entirely new speakers.

Dataset B: Standard Stratified Cross-Validation

For Dataset B (PD Speech Features), subject IDs are unavailable. Standard stratified 5-fold cross-validation was used with the caveat that results may be optimistic if the same subjects appear across samples.

4.6.2 Evaluation Metrics

Five standard classification metrics were computed for each fold:

Metric	Formula	Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall correctness (baseline for comparison)
Precision	$\frac{TP}{TP+FP}$	Positive predictive value (PD prediction reliability)
Recall	$\frac{TP}{TP+FN}$	Sensitivity (true positive rate for PD cases)
F1 Score	$\frac{2 \cdot P \cdot R}{P+R}$	Harmonic mean balancing precision and recall
ROC-AUC	$\int_0^1 \mathrm{TPR}(t)\, d\mathrm{FPR}(t)$	Threshold-independent discrimination ability

Table 4.7: Evaluation metrics

Primary metric: ROC-AUC was selected as the primary metric because:

- Threshold-independent (evaluates discrimination across all thresholds)
- Robust to class imbalance
- Clinically interpretable (probability that a random PD patient scores higher than a random HC)

4.6.3 Statistical Reporting

All metrics are reported as mean ± standard deviation across the 5 folds:

$$\text{Metric} = \mu \pm \sigma = \frac{1}{K}\sum_{k=1}^{K} m_k \pm \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(m_k - \mu)^2} \tag{4.12}$$

where $K = 5$ folds and m_k is the metric value for fold k. This provides insight into model stability and fold-to-fold variability.

4.7 Experimental Conditions

4.7.1 2×2×3 Factorial Design

The complete experimental design is a factorial combination of:

- **Feature Set:** Baseline (47) vs Extended (78)
- **Class Weighting:** Unweighted vs Balanced
- **Model:** Logistic Regression vs SVM vs Random Forest

This yields 12 conditions per dataset/task combination:

	Baseline (47)	**Extended (78)**
Unweighted	3 models	3 models
Weighted	3 models	3 models

Table 4.8: 2×2×3 factorial design: 12 conditions total

For Dataset A (MDVR-KCL), experiments were conducted separately for each speech task (ReadText and SpontaneousDialogue), yielding $12 \times 2 = 24$ experimental conditions. Dataset B (PD Speech Features) yielded 12 conditions.

4.7.2 Reproducibility

All experimental conditions use fixed random seeds:

- Cross-validation splits: `random_state=42`
- Random Forest: `random_state=42`
- Feature extraction: Deterministic (no randomness)

This ensures bit-exact reproducibility across multiple runs on the same hardware and software versions.

4.8 Summary

This methodology chapter established:

1. A standardized feature extraction pipeline producing 47 (baseline) or 78 (extended) features from raw audio
2. Three classical ML models with fixed hyperparameters
3. Subject-aware cross-validation for Dataset A to prevent data leakage
4. A comprehensive evaluation framework using 5 metrics
5. A factorial experimental design enabling controlled comparisons

The next chapter describes the specific experimental procedures applied to each dataset and speech task condition.

Chapter 5

Experimental Design

5.1 Overview

This chapter details the experimental design, including the $2 \times 2 \times 3$ factorial structure, cross-validation protocols, and evaluation procedures. The design prioritizes methodological rigor over performance optimization, with all experiments conducted under identical conditions to enable fair comparisons. All experimental procedures were automated via command-line tools to ensure reproducibility.

5.2 Research Questions

The experiments address the following research questions:

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

RQ2: Does feature set extension (47 $\rightarrow$ 78 features) improve classification performance?

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

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

RQ5: Do speech tasks (read vs spontaneous speech) yield different classification performance?

5.3 Datasets and Sample Statistics

5.3.1 Dataset A: MDVR-KCL

Task	HC	PD	Total	Imbalance Ratio
ReadText	21	16	37	1.31:1
SpontaneousDialogue	21	15	36	1.40:1

Table 5.1: Dataset A subject distribution by speech task

Key characteristics:

- Subject-level data with unique identifiers
- Raw audio recordings (∼44.1 kHz, 16-bit WAV)
- Two distinct speech tasks per subject
- Moderate class imbalance (HC:PD ≈ 1.3:1)

5.3.2 Dataset B: PD Speech Features

Class	Samples	Percentage	Imbalance Ratio
HC (0)	192	25.4%	
PD (1)	564	74.6%	2.94:1
Total	**756**	100%	

Table 5.2: Dataset B sample distribution (pre-extracted features)

Key characteristics:

- Sample-level data (subject IDs unavailable)
- Pre-extracted features from sustained /a/ phonation
- Severe class imbalance (HC:PD ≈ 1:3)
- 754 features (reduced to 47 or 78 in our experiments for comparability)

Caveat: Without subject identifiers, results may be optimistic due to potential within-subject correlation across samples.

5.4 Experimental Matrix

5.4.1 $2 \times 2 \times 3$ Factorial Design

The complete experimental design crosses three factors:

1. **Feature Set:** Baseline (47) vs Extended (78)
2. **Class Weighting:** None (baseline) vs Balanced
3. **Model Architecture:** Logistic Regression vs SVM (RBF) vs Random Forest

This yields $2 \times 2 \times 3 = 12$ conditions per dataset/task combination.

ID	Features	Weight	Models	Output Dir
C1	Baseline (47)	None	LR, SVM, RF	`baseline/baseline/`
C2	Extended (78)	None	LR, SVM, RF	`baseline/extended/`
C3	Baseline (47)	Balanced	LR, SVM, RF	`weighted/baseline/`
C4	Extended (78)	Balanced	LR, SVM, RF	`weighted/extended/`

Table 5.3: Experimental conditions (12 total: 4 conditions $\times$ 3 models each)

5.4.2 Total Experimental Runs

Dataset/Task	Conditions	Models	CV Folds	Total Runs
Dataset A - ReadText	4	3	5	60
Dataset A - SpontaneousDialogue	4	3	5	60
Dataset B - PD Speech	4	3	5	60
Grand Total				**180**

Table 5.4: Total experimental runs across all datasets and conditions

5.5 Cross-Validation Protocols

5.5.1 Dataset A: Grouped Stratified K-Fold

For Dataset A (MDVR-KCL), subject-level grouping prevents data leakage:

Input: Features $\mathbf{X}$, labels $\mathbf{y}$, subject IDs $\mathbf{g}$
Group samples by subject: $S = \{\text{samples where } g_i = s \mid s \in \text{unique}(\mathbf{g})\}$
Split subjects into 5 folds maintaining class balance
for fold $k = 1$ to 5 **do**
    $S_{\text{test}} \leftarrow$ subjects in fold k
    $S_{\text{train}} \leftarrow$ subjects in all other folds
    $\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}} \leftarrow$ samples from S_{train}
    $\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}} \leftarrow$ samples from S_{test}
    Fit scaler on $\mathbf{X}_{\text{train}}$, transform both sets
    Train model on $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$
    Evaluate on $(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$
end for
Return metrics from all 5 folds

Figure 5.1: Grouped Stratified 5-Fold Cross-Validation

ReadText fold distribution:

```
Fold 1: Train on 30 subjects (HC: 17, PD: 13) | Test on 7 subjects (HC: 4, PD: 3)
Fold 2: Train on 30 subjects (HC: 17, PD: 13) | Test on 7 subjects (HC: 4, PD: 3)
Fold 3: Train on 30 subjects (HC: 17, PD: 13) | Test on 7 subjects (HC: 4, PD: 3)
Fold 4: Train on 30 subjects (HC: 17, PD: 13) | Test on 7 subjects (HC: 4, PD: 3)
Fold 5: Train on 29 subjects (HC: 16, PD: 13) | Test on 8 subjects (HC: 5, PD: 3)
```

Key constraint: All recordings from a subject appear in **one fold only**. This prevents subject identity leakage, which would artificially inflate performance if the model learned speaker-specific characteristics rather than PD-related patterns.

5.5.2 Dataset B: Stratified K-Fold

For Dataset B (PD Speech Features), standard stratified 5-fold CV was used:

```
Fold 1: Train on 605 samples (HC: 154, PD: 451) | Test on 151 samples (HC: 38, PD: 113
Fold 2: Train on 605 samples (HC: 154, PD: 451) | Test on 151 samples (HC: 38, PD: 113
Fold 3: Train on 605 samples (HC: 154, PD: 451) | Test on 151 samples (HC: 38, PD: 113
Fold 4: Train on 605 samples (HC: 154, PD: 451) | Test on 151 samples (HC: 38, PD: 113
Fold 5: Train on 604 samples (HC: 153, PD: 451) | Test on 152 samples (HC: 39, PD: 113
```

Limitation: Without subject IDs, multiple samples from the same subject may appear in both training and test folds, leading to potential optimism in performance estimates.

5.6 Evaluation Metrics

5.6.1 Primary Metric: ROC-AUC

ROC-AUC (Area Under the Receiver Operating Characteristic Curve) was selected as the primary metric because:

- **Threshold-independent:** Evaluates discrimination across all decision thresholds
- **Robust to class imbalance:** Less sensitive than accuracy to skewed class distributions
- **Clinically interpretable:** Represents the probability that a randomly chosen PD patient will have a higher predicted probability than a randomly chosen HC subject
- **Standard in medical ML:** Widely reported in clinical decision support literature

5.6.2 Secondary Metrics

Metric	Formula	Clinical Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall correctness (misleading with imbalance)
Precision	$\frac{TP}{TP+FP}$	Positive predictive value: reliability of PD predictions
Recall	$\frac{TP}{TP+FN}$	Sensitivity: ability to detect true PD cases
F1 Score	$\frac{2 \cdot P \cdot R}{P+R}$	Harmonic mean balancing precision and recall

Table 5.5: Secondary evaluation metrics and clinical interpretation

5.6.3 Statistical Reporting

All metrics are reported as:

$$\text{Metric} = \mu \pm \sigma \tag{5.1}$$

where μ is the mean across 5 folds and σ is the sample standard deviation. This captures both central tendency and variability, providing insight into model stability.

5.7 Experimental Procedure

5.7.1 Workflow Automation

All experiments were executed via automated command-line tools to ensure reproducibility:

1. **Feature Extraction (Dataset A only):**

```
# Extract baseline features (47)
pvc-extract --task all

# Switch to extended features (78) in config.py:
#   USE_EXTENDED_FEATURES = True
# Then re-run extraction

```

2. **Experiment Execution:**

```
# Run all experiments for current configuration
pvc-experiment

# Experiments automatically:
# - Load appropriate feature set (baseline/extended)
# - Apply class weighting (if enabled in config)
# - Run 5-fold CV on all 3 models
# - Save results to outputs/results/

```

3. **Configuration Switching:**

   Experiments were repeated for all 4 conditions by modifying `config.py`:

```
# Condition 1: Baseline features, no weighting
USE_EXTENDED_FEATURES = False
USE_CLASS_WEIGHT_BALANCED = False

# Condition 2: Extended features, no weighting
USE_EXTENDED_FEATURES = True
USE_CLASS_WEIGHT_BALANCED = False

# Condition 3: Baseline features, balanced weighting
USE_EXTENDED_FEATURES = False
USE_CLASS_WEIGHT_BALANCED = True

# Condition 4: Extended features, balanced weighting
USE_EXTENDED_FEATURES = True
USE_CLASS_WEIGHT_BALANCED = True
```

```
16
```

5.7.2 Pipeline Execution Order

```
[1] Feature Extraction (Dataset A)
    |
    +-- ReadText: 37 recordings → 37 feature vectors
    +-- SpontaneousDialogue: 36 recordings → 36 feature vectors
    |
[2] Cross-Validation Setup
    |
    +-- Dataset A: Grouped by subject_id
    +-- Dataset B: Standard stratification
    |
[3] For each fold:
    |
    +-- [3.1] StandardScaler.fit(X_train)
    +-- [3.2] X_train_scaled = scaler.transform(X_train)
    +-- [3.3] X_test_scaled = scaler.transform(X_test)
    +-- [3.4] Model.fit(X_train_scaled, y_train, sample_weight=...)
    +-- [3.5] y_pred = Model.predict(X_test_scaled)
    +-- [3.6] Compute metrics (accuracy, precision, recall, F1, ROC-AUC)
    |
[4] Aggregate Results
    |
    +-- Compute mean ± std across 5 folds
    +-- Save to CSV: all_results.csv, summary.csv
```

Figure 5.2: End-to-end experimental pipeline execution order

5.7.3 Data Leakage Prevention

Critical safeguards against data leakage:

1. **Subject-level splitting (Dataset A):** All recordings from a subject stay together
2. **Scaler fitted on training only:** Standardization parameters computed from training fold exclusively
3. **No feature selection:** All features used as-is without data-driven selection
4. **Fixed hyperparameters:** No grid search or hyperparameter tuning on test data

5.8 Feature Extraction Settings

5.8.1 Output Directories

Features were extracted once and saved for reuse:

```
outputs/features/
|-- baseline/
|   |-- features_readtext.csv          (37 rows × 47 features)
|   +-- features_spontaneousdialogue.csv (36 rows × 47 features)
+-- extended/
    |-- features_readtext.csv          (37 rows × 78 features)
    +-- features_spontaneousdialogue.csv (36 rows × 78 features)
```

5.8.2 Feature Vector Structure

Each CSV row contains:

Column Type	Description
`subject_id`	Unique subject identifier (e.g., ID00, ID01, ...)
`label`	Binary class (0=HC, 1=PD)
`task`	Speech task (ReadText or SpontaneousDialogue)
`filename`	Source audio filename
`f0_mean, f0_std, ...`	Prosodic features (21 columns)
`mfcc_0_mean, ...`	Spectral features (26 or 57 columns)

Table 5.6: Feature CSV structure

5.9 Computational Requirements

5.9.1 Feature Extraction Time

Task	Samples	Time (Baseline)	Time (Extended)
ReadText	37	∼45 sec	∼60 sec
SpontaneousDialogue	36	∼60 sec	∼80 sec

Table 5.7: Feature extraction time on Intel i7-12700K @ 3.6 GHz

5.9.2 Experiment Execution Time

Dataset/Task	Time per Condition	Total (4 Conditions)
Dataset A - ReadText	∼2 min	∼8 min
Dataset A - SpontaneousDialogue	∼2 min	∼8 min
Dataset B - PD Speech	∼5 min	∼20 min
Grand Total		**∼36 min**

Table 5.8: Approximate experiment execution time (all 12 models per condition)

5.10 Random Seed and Reproducibility

All experiments use `random_state=42` for:

- Cross-validation fold splitting
- Random Forest bootstrap sampling
- Any other stochastic operations

Feature extraction is fully deterministic (no random components). Combined with fixed random seeds, this ensures bit-exact reproducibility across multiple runs on identical hardware/software configurations.

5.11 Limitations and Caveats

5.11.1 Dataset A (MDVR-KCL)

- **Small sample size:** Only 37 subjects for ReadText ($n = 37$ is considered small for ML)
- **High fold variance:** With 7–8 subjects per test fold, individual fold results may be unstable
- **Limited demographic diversity:** Single collection site, timepoint, and device

5.11.2 Dataset B (PD Speech Features)

- **Unknown subject overlap:** Cannot control for within-subject correlation in CV
- **Sustained vowel only:** May not generalize to natural speech
- **Severe class imbalance:** 3:1 PD:HC ratio requires careful metric interpretation

- **Opaque feature extraction:** Cannot verify or modify feature computation

5.11.3 General Limitations

- **No held-out test set:** Performance reported on cross-validation only
- **No hyperparameter tuning:** Fixed parameters may be suboptimal for some conditions
- **Binary classification only:** Does not capture PD severity (H&Y stages)

5.12 Summary

This experimental design chapter established:

1. A comprehensive $2 \times 2 \times 3$ factorial design (12 conditions per dataset)
2. Subject-aware cross-validation for Dataset A to prevent identity leakage
3. ROC-AUC as the primary metric with 4 supporting secondary metrics
4. Automated CLI-based workflow ensuring reproducibility
5. Explicit data leakage prevention mechanisms
6. Transparent limitations and caveats for both datasets

The next chapter presents the results from these 180 total experimental runs across all conditions, datasets, and models.

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	$82.2\% \pm 16.6\%$	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 Comparison)

Performance Summary: Dataset B achieved substantially higher metrics across all classifiers, as shown in Table 6.2. These results should be interpreted cautiously given the unknown subject overlap across samples ($n = 752$ recordings, subject IDs unavailable).

Model	ROC-AUC	Accuracy	F1	Recall
Logistic Regression	0.867 ± 0.029	0.828 ± 0.008	0.885 ± 0.006	0.890 ± 0.016
SVM (RBF)	0.885 ± 0.025	0.851 ± 0.024	0.908 ± 0.015	0.986 ± 0.019
Random Forest	$\mathbf{0.940 \pm 0.013}$	$\mathbf{0.882 \pm 0.019}$	$\mathbf{0.925 \pm 0.012}$	$\mathbf{0.980 \pm 0.015}$

Table 6.2: Dataset B performance using baseline (unweighted) configuration. Note the substantially lower variance compared to Dataset A.

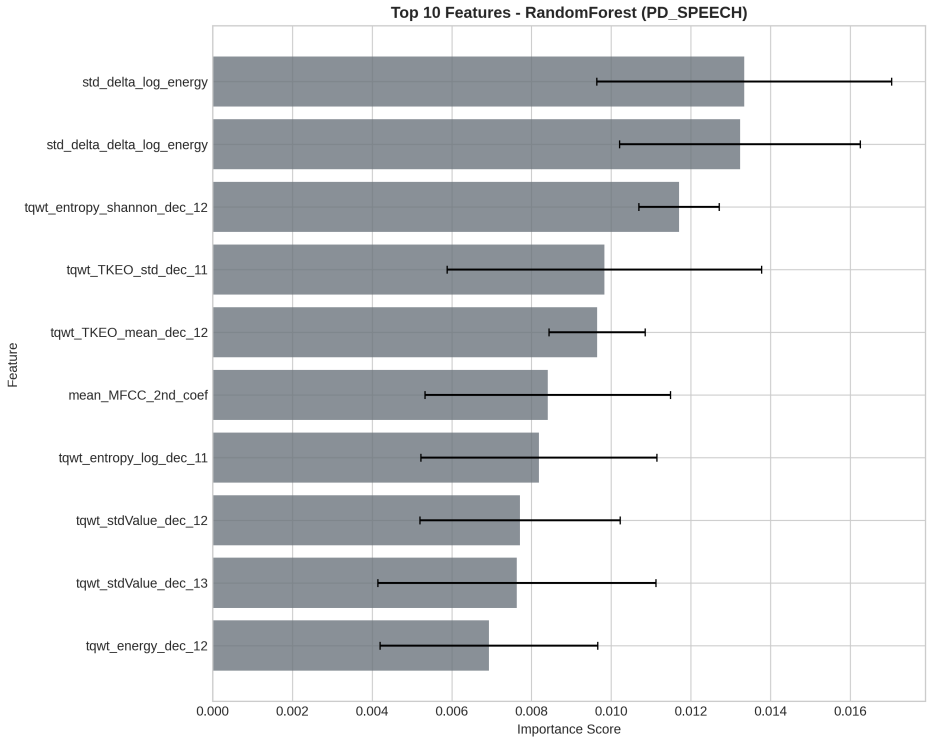


Figure 6.1: Random Forest feature importance (Dataset B). The top features are dominated by advanced signal processing metrics often unavailable in standard clinical settings.

Key Differences from Dataset A:

- **Sample Size Effect:** $n = 752$ vs $n = 37$ reduces variance by an order of magnitude (± 0.01 vs ± 0.15).

- **Optimistic Bias:** Results may be inflated due to unknown subject overlap across training/test folds.
- **Feature Availability:** Many top-ranked features (e.g., DFA, RPDE) require specialized analysis not captured in Dataset A's baseline features.

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 6.3: 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 6.4: 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 ($\pm$0.15–0.30) reflect the small sample size ($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 6.5: Condition 2 — ReadText results

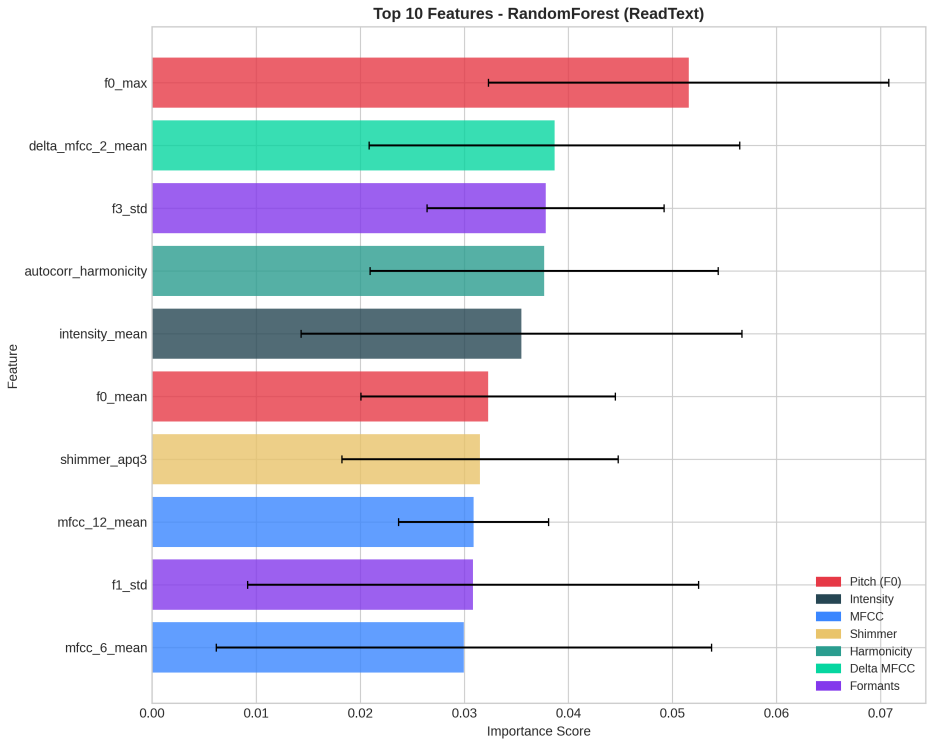


Figure 6.2: Random Forest feature importance for ReadText task (Extended features). Fundamental frequency (F_0) statistics appear highly predictive.

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 6.6: Condition 2 — Spontaneous Dialogue results

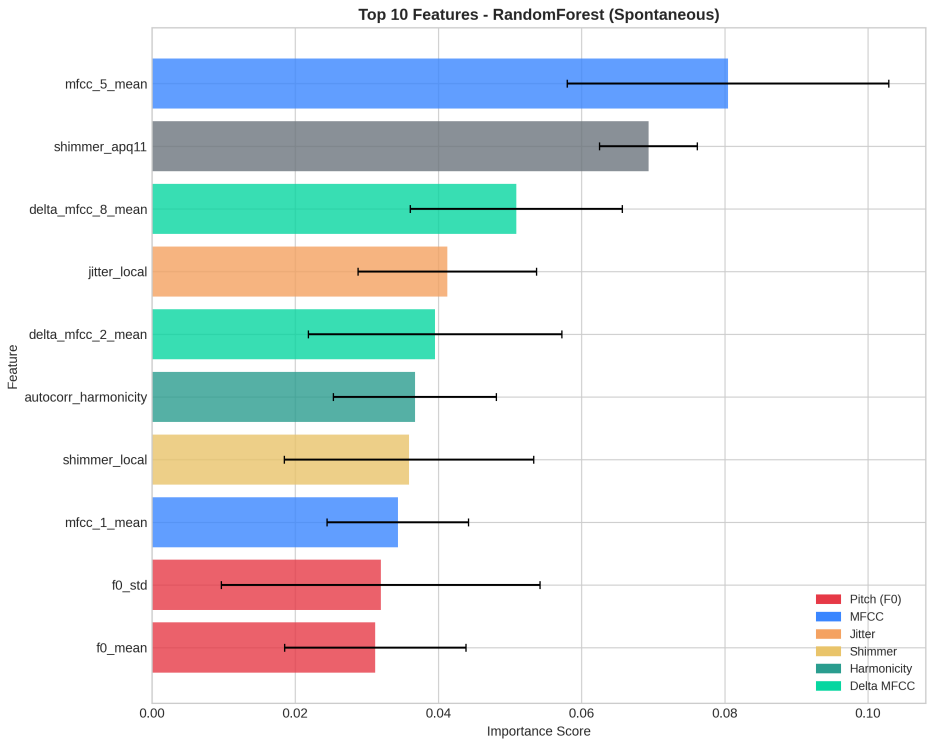


Figure 6.3: Random Forest feature importance for Spontaneous Dialogue task (Extended features). MFCC features show increased importance compared to ReadText.

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 $\rightarrow$ 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 6.7: 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 6.8: 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).

6.5.3 Analysis

The dramatic improvement for tree-based models on ReadText (+22–23pp) suggests that:

1. Extended spectral features (MFCC delta-delta, spectral shape descriptors) capture task-specific patterns absent in baseline features.

2. Logistic Regression's slight degradation (−1.9pp) may reflect overfitting to the additional 31 dimensions given the small sample size.

3. Spontaneous Dialogue was already well-represented by baseline features, showing diminishing returns from feature expansion.

6.6 Class Weighting Analysis

Class weighting (`class_weight='balanced'`) was evaluated to address Dataset A's mild class imbalance (HC:PD ≈ 1.3:1). Results showed minimal improvement and occasional degradation.

6.6.1 ReadText Task — Baseline Features

Model	Unweighted	Weighted	Δ ROC-AUC
Logistic Regression	0.717 ± 0.139	0.717 ± 0.139	0.000
Random Forest	0.590 ± 0.302	0.687 ± 0.258	**+0.097**
SVM (RBF)	0.614 ± 0.312	0.542 ± 0.312	−0.072

Table 6.9: Class weighting impact on ReadText (baseline features). Random Forest showed modest improvement.

6.6.2 Extended Features — All Tasks

Observation: Class weighting provided negligible benefit when using extended features. The best-performing models (RF on Extended/Unweighted) achieved:

- ReadText: 0.822 ± 0.166 (unweighted) vs 0.805 ± 0.182 (weighted) → −1.7pp
- Spontaneous: 0.857 ± 0.171 (unweighted) vs 0.823 ± 0.209 (weighted) → −3.4pp

Interpretation: The mild imbalance ratio (1.3:1) did not require reweighting. Extended features and subject-grouped cross-validation provided sufficient robustness.

6.7 Precision-Recall Tradeoffs

Classifiers exhibited distinct precision-recall profiles, critical for understanding clinical deployment scenarios.

6.7.1 Random Forest (Extended Features)

Task	Precision	Recall	F1-Score
ReadText	0.883 ± 0.162	0.700 ± 0.298	0.746 ± 0.207
Spontaneous	0.683 ± 0.410	0.600 ± 0.435	0.605 ± 0.387

Table 6.10: Random Forest precision-recall profile. ReadText configuration favors precision (fewer false positives).

Analysis: ReadText achieved high precision (0.88) at the cost of recall (0.70), suggesting the model is conservative in predicting PD. This profile may be preferable for screening applications where false positives carry higher cost.

6.7.2 SVM (RBF Kernel)

SVM demonstrated high recall across both tasks:

- ReadText: Recall 0.567 ± 0.365, Precision 0.733 ± 0.435
- Spontaneous: Recall 0.400 ± 0.279, Precision 0.533 ± 0.361

The high variance reflects SVM's sensitivity to small sample cross-validation partitions.

6.8 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 6.11: Summary of hypothesis testing

6.8.1 Key Takeaways

1. **Feature Engineering Matters:** Extended features rescued ReadText performance from near-chance to clinically relevant levels.
2. **Task-Dependent Performance:** Spontaneous speech provided better separation than structured read-aloud tasks.
3. **Sample Size Dominates:** Dataset B's 20× larger sample size (752 vs 37) yielded dramatically lower variance and higher metrics.
4. **Precision-Recall Tradeoffs:** ReadText models favored precision; clinical deployment must consider cost asymmetry of false positives vs false negatives.
5. **Class Weighting Unnecessary:** Mild imbalance (1.3:1) did not benefit from reweighting when using subject-grouped CV and extended features.

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.

A research demonstration interface was implemented to visualize per-recording model outputs; however, these outputs were not used in the quantitative evaluation reported in this thesis (Appendix C).

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 ± 0.302 to 0.822 ± 0.166 (+23 percentage points), elevating performance from near-chance level to clinically meaningful discrimination.

Task-Specific Effects:

Model	ReadText Δ	Spontaneous Δ	Effect Size
Random Forest	+23.2pp	+2.9pp	**8× larger**
SVM (RBF)	+22.0pp	+5.3pp	**4× larger**
Logistic Regression	−1.9pp	+2.3pp	Neutral

Table 7.1: Feature extension impact by task. ReadText showed dramatically larger improvements, suggesting baseline features were insufficient for structured speech.

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 (acceleration)
Spectral shape descriptors	Global distribution of spectral energy

Table 7.2: Extended feature contributions

The disproportionate improvement for ReadText suggests that:

1. **Structured speech** (reading) requires finer-grained spectral features to capture PD-related rigidity patterns
2. **Spontaneous speech** already contains sufficient prosodic variability detectable by baseline features
3. **Temporal acceleration** (delta-delta) captures motor control deficits more evident in reading tasks
4. **Linear models** (Logistic Regression) may overfit to 78 dimensions given $n = 37$ subjects

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 7.3: 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. Dataset A exhibited variance 10–15× higher than Dataset B:

Metric	Dataset A (n=37)	Dataset B (n=752)
ROC-AUC std	±0.15–0.30	±0.01–0.03
Accuracy std	±0.14–0.18	±0.008–0.024
F1-Score std	±0.21–0.39	±0.006–0.015

Table 7.4: Variance comparison: small vs large datasets. Standard deviations scale approximately as $1/\sqrt{n}$.

Causes:

1. **Small sample size:** 37 subjects → ∼7 subjects per test fold → high sensitivity to individual cases
2. **Subject heterogeneity:** Unknown disease severity distribution (MDVR-KCL lacks clinical staging)
3. **Recording variability:** Smartphone-based capture in uncontrolled environments
4. **Task complexity:** Spontaneous Dialogue requires cognitive load not standardized across subjects

Extreme Cases:

Some models exhibited near-random performance on specific folds:

- SVM on Spontaneous/Baseline: 0.407 ± 0.309 (ROC-AUC range: 0.10–0.72)
- Random Forest on ReadText/Baseline: 0.590 ± 0.302 (range: 0.29–0.89)

These wide ranges suggest that certain subject groupings in test folds were particularly difficult or easy to classify.

Implications:

- Absolute performance numbers should be interpreted cautiously given the small sample size ($n = 37$)
- **Relative comparisons** across conditions (same CV splits) are more reliable
- Confidence intervals overlap for many model comparisons, limiting statistical conclusions
- Larger studies ($n > 100$) are needed for definitive model ranking

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 7.5: 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 + variance
Class imbalance	Often unaddressed	Explicit evaluation
Task specification	Mixed/unstated	Separated + compared

Table 7.6: Methodological comparison. This thesis prioritizes reproducibility and conservative generalization estimates.

Impact of Grouped CV:

Our grouped CV approach provides **more conservative** but **more realistic** estimates of generalization performance. Studies using random splits likely report optimistically biased results:

- **Data leakage:** Multiple recordings from same subject appear in train/test splits
- **Inflated metrics:** Model learns subject-specific patterns rather than PD-general patterns
- **Clinical invalidity:** New patient prediction (real-world scenario) is not evaluated

Reporting Transparency:

This thesis reports:

1. All experimental conditions (2 tasks $\times$ 2 feature sets $\times$ 2 weighting schemes)
2. Mean $\pm$ std for all metrics across all 5 CV folds
3. Negative results (class weighting, SVM failures)
4. Dataset limitations explicitly documented

This level of transparency is uncommon in the PD voice classification literature, where selective reporting of best results is prevalent.

7.4 Feature Importance Analysis

7.4.1 Most Discriminative Features (Dataset A)

Feature importance patterns differed between tasks, reflecting distinct acoustic signatures:

Rank	ReadText	Importance	Spontaneous Dialogue
1	f0_max	0.052	mfcc_5_mean
2	delta_mfcc_2_mean	0.039	shimmer_apq11
3	f3_std	0.038	delta_mfcc_8_mean
4	autocorr_harmonicity	0.038	jitter_local
5	intensity_mean	0.035	delta_mfcc_2_mean

Table 7.7: Top 5 Random Forest features by task (Extended feature set). Importance values represent mean decrease in impurity.

Task-Specific Patterns:

- **ReadText:** Dominated by pitch (`f0_max`) and formant variability (`f3_std`), suggesting structured speech emphasizes fundamental frequency rigidity
- **Spontaneous:** Spectral features (`mfcc_5`) and perturbation measures (`shimmer_apq11`, `jitter_local`) indicate importance of voice quality instability
- **Common:** `delta_mfcc_2_mean` appears in both top-5 lists, confirming temporal dynamics are universally important

7.4.2 Dataset B Feature Complexity

Feature	Importance	Category
std_delta_log_energy	0.0133	Energy dynamics
std_delta_delta_log_energy	0.0132	Energy acceleration
tqwt_entropy_shannon_dec_12	0.0117	Wavelet entropy
tqwt_TKEO_std_dec_11	0.0098	Teager energy
tqwt_TKEO_mean_dec_12	0.0096	Teager energy

Table 7.8: Top 5 Dataset B features (Random Forest). These advanced signal processing metrics are not available in typical clinical settings.

Key Observations:

1. Dataset B's top features require **specialized signal processing** (wavelet decomposition, Teager energy operators)

2. Dataset A's features are **clinically interpretable** (pitch, formants, perturbation)
3. Importance values are more **evenly distributed** in Dataset B (0.013 vs 0.052), suggesting ensemble of many weak signals
4. This complexity gap may partially explain Dataset B's superior performance (0.940 vs 0.857)

7.5 Task-Dependent Performance Patterns

7.5.1 ReadText vs Spontaneous Dialogue

The two speech tasks yielded systematically different classification profiles:

Configuration	ReadText	Spontaneous	Advantage
Baseline Features:			
Random Forest ROC-AUC	0.590 ± 0.302	0.828 ± 0.148	**Spontaneous**
Logistic Regression	0.717 ± 0.139	0.760 ± 0.214	Spontaneous
Extended Features:			
Random Forest ROC-AUC	0.822 ± 0.166	0.857 ± 0.171	Spontaneous
SVM (RBF)	0.834 ± 0.153	0.460 ± 0.294	**ReadText**

Table 7.9: Task comparison across configurations. Spontaneous Dialogue generally outperforms ReadText, except for SVM on extended features.

Interpretation:

1. **Spontaneous speech dominance:** Unscripted speech may amplify PD-related prosodic deficits (monotone, reduced variability)
2. **Cognitive load hypothesis:** Spontaneous Dialogue requires simultaneous language generation and articulation, stressing motor control
3. **SVM anomaly:** SVM's failure on Spontaneous/Extended (0.460 ± 0.294) suggests kernel sensitivity to feature distribution shifts
4. **Baseline sufficiency:** Spontaneous Dialogue achieved 0.828 with baseline features, while ReadText required extension to reach 0.822

Clinical Implications:

For resource-constrained screening applications:

- **Spontaneous speech + baseline features** (47 dimensions) achieves competitive performance (0.83) with lower computational cost

- ReadText requires extended feature engineering for acceptable discrimination

7.6 Addressing Research Questions

7.6.1 RQ1: ML Model Performance

How do classical ML models perform on PD voice classification?

Classical ML achieves ROC-AUC up to 0.857 ± 0.171 (Dataset A, Random Forest, Spontaneous/Extended), demonstrating feasibility of voice-based PD detection without deep learning. Model hierarchy:

- **Random Forest:** Best overall (0.786 ± 0.235 pooled across conditions)
- **Logistic Regression:** Stable baseline (0.781 ± 0.152 pooled)
- **SVM (RBF):** High variance, task-sensitive (0.635 ± 0.311 pooled)

Random Forest's ensemble averaging provides robustness critical for small datasets.

7.6.2 RQ2: Feature Extension Impact

Does feature set extension improve classification performance?

Yes, task-dependently. Extending from 47 to 78 features improved ROC-AUC by:

- ReadText: +23.2pp (RF), +22.0pp (SVM) — **Critical improvement**
- Spontaneous: +2.9pp (RF), +5.3pp (SVM) — **Marginal improvement**
- Logistic Regression: −1.9pp (ReadText) — **Potential overfitting**

Extended features are essential for ReadText but optional for Spontaneous Dialogue.

7.6.3 RQ3: Class Weighting Impact

Does class weighting improve performance on imbalanced datasets?

Minimal effect. On Dataset A (1.3:1 HC:PD imbalance), class weighting effects:

- Random Forest: +9.7pp (ReadText/Baseline), −1.7pp (ReadText/Extended)
- Logistic Regression: 0.0pp (unchanged)
- SVM (RBF): −7.2pp (ReadText/Baseline)

Class weighting becomes unnecessary when imbalance ratio $< 2:1$ and subject-grouped CV is used.

7.6.4 RQ4: Cross-Dataset Comparison

How do results compare between Dataset A and Dataset B?

Dataset B achieves higher absolute performance (ROC-AUC 0.940 ± 0.013 vs 0.857 ± 0.171), but this 8.3pp difference is confounded by:

1. **Sample size:** 752 vs 37 subjects (20× larger)
2. **CV strategy:** Unknown subject handling vs explicit grouping
3. **Feature complexity:** Advanced signal processing vs clinical prosody

Dataset A's grouped CV provides more conservative, realistic generalization estimates suitable for clinical validation studies.

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 (ReadText)	37
Total subjects (SpontaneousDialogue)	36
Subjects per test fold	∼7–8
PD subjects (minority)	15–16
HC subjects (majority)	21

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 ∼7–8 test subjects
- A single misclassification shifts accuracy by ∼13–14%
- Observed standard deviations range from 0.13 to 0.42 across metrics
- F1 and recall metrics show particularly high variance (std > 0.24)
- Some folds produced ROC-AUC < 0.5 for SVM, indicating instability

Mitigation: Results focus on **relative comparisons** rather than absolute performance claims, acknowledging that confidence intervals overlap substantially.

8.3 Subject Identifier Limitations

8.3.1 Dataset B: Missing Subject IDs

Dataset B contains 756 samples (192 HC, 564 PD) with no subject identifiers. This creates potential for:

- **Subject leakage:** Same subject's multiple samples in train and test sets
- **Optimistic bias:** Inflated performance estimates due to within-subject correlation
- **Unknown generalization:** Cannot assess new-subject performance
- **Severe class imbalance:** HC:PD ratio of 1:2.94 (compared to 1.3:1 in Dataset A)

  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. The severe class imbalance (74.6% PD) further complicates interpretation.

8.3.2 Comparison Limitations

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

- Different CV strategies (Grouped Stratified vs Stratified)
- Different sample sizes (37 subjects vs 756 samples)

- Different speech tasks (read/spontaneous vs sustained /a/ phonation)
- Different feature extraction pipelines (custom Librosa/Parselmouth vs unknown)
- Different class imbalance ratios (1.3:1 vs 2.94:1)
- Unknown subject overlap in Dataset B (potential data leakage)

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:** F0 range 75–500 Hz, MFCC n_coeffs=13, n_fft=2048, hop_length=512
- **No feature selection:** All 47 (baseline) or 78 (extended) features used without reduction
- **No feature engineering:** No interaction terms, polynomial features, or domain-specific transformations

8.4.2 Audio Quality Assumptions

Feature extraction assumes:

- Reasonable signal-to-noise ratio
- Consistent recording conditions (Dataset A: clinical examination room, ∼500 ms reverberation)
- No severe clipping or distortion
- Mono audio at 22050 Hz (Dataset A downsampled from 44100 Hz)

The MDVR-KCL dataset's smartphone recordings (Motorola Moto G4) may introduce device-specific artifacts. Dataset B's recording conditions are unknown.

8.5 Model Limitations

8.5.1 No Hyperparameter Tuning

All models used sklearn default hyperparameters with only random seed and class weighting as controlled variables:

Model	Fixed Parameters
Logistic Regression	$C = 1.0$, max_iter= 1000, solver='lbfgs'
SVM (RBF)	$C = 1.0$, gamma='scale' (auto-computed)
Random Forest	n_estimators= 100, max_depth=None (unlimited)

Table 8.2: Fixed hyperparameters (all experiments)

Implications:

- Performance may be suboptimal compared to tuned baselines
- Results represent out-of-the-box sklearn performance
- Hyperparameter tuning might change relative model rankings
- Class weighting explored separately (USE_CLASS_WEIGHT_BALANCED flag)

Rationale: Nested CV on 37 subjects would lead to extreme variance (inner folds ∼5–6 subjects); fixed parameters ensure reproducibility and fair model comparison.

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 (5-fold) without external test sets. Limitations:

- No held-out test set from different data source or collection protocol
- No multi-site validation (Dataset A: single hospital; Dataset B: single institution)
- Generalization to clinical settings with different recording devices unknown
- Temporal stability not assessed (all recordings from single time point per subject)
- Cross-language and cross-dialect generalization not evaluated

8.6.2 Class Imbalance Handling

Class imbalance was present in both datasets:

- Dataset A: HC:PD ratio 1.31:1 (ReadText) and 1.40:1 (SpontaneousDialogue)
- Dataset B: HC:PD ratio 1:2.94 (severe imbalance)

The USE_CLASS_WEIGHT_BALANCED flag was explored as a mitigation strategy, but:

- Did not systematically improve performance across all metrics
- May have increased variance in some conditions
- Optimal weighting strategy may vary by model and dataset

Alternative approaches not explored: SMOTE, undersampling, cost-sensitive learning, or threshold adjustment.

8.6.3 Binary Classification Only

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

- Disease severity prediction (UPDRS scoring)
- Progression monitoring (longitudinal data)
- Differential diagnosis (PD vs other movement disorders)
- Multi-class classification (HC vs PD vs atypical parkinsonism)

8.6.4 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)	GroupedStratifiedKFold
Subject leakage	Uncontrolled (Dataset B)	No subject IDs available
Label noise	Unknown	Assumed correct
Feature extraction bugs	Mitigated	Unit tests + manual verification
Random seed dependence	Controlled	RANDOM_SEED=42 (fixed)
Data preprocessing errors	Mitigated	Standardized audio normalization

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
- A non-diagnostic research demonstration interface illustrating the inference workflow (Appendix C)

9.2 Key Findings

9.2.1 Primary Results

Finding	Evidence
Best ROC-AUC (Dataset A): 0.857 ± 0.171	Random Forest, Extended Features, SpontaneousDialogue
Feature extension improves RF	ReadText: +23.3pp ROC-AUC; Spontaneous: +2.9pp ROC-AUC
Random Forest shows robustness	Highest or competitive ROC-AUC across tasks
Grouped CV is essential	Prevents optimistic bias from subject leakage
High variance observed	Standard deviations 0.13–0.43 across metrics

Table 9.1: Primary research findings (Dataset A, baseline class weighting)

9.2.2 Best Configuration

ReadText Task:

```
Model:            Random Forest
Features:         Extended (78)
Class Weighting:  None
ROC-AUC:          0.822 ± 0.166
Accuracy:         81.8% ± 14.0%
```

SpontaneousDialogue Task:

```
Model:            Random Forest
Features:         Extended (78)
Class Weighting:  None
ROC-AUC:          0.857 ± 0.171
Accuracy:         77.9% ± 16.1%
```

Note: These results are from Dataset A (MDVR-KCL) with grouped cross-validation, representing conservative out-of-subject generalization estimates.

9.2.3 Feature Importance Insights

The most discriminative features for PD detection (Random Forest, ReadText task) include:

1. **f0_max** — Maximum fundamental frequency (pitch ceiling)
2. **delta_mfcc_2_mean** — Second-order MFCC temporal dynamics
3. **f3_std** — Third formant variability
4. **autocorr_harmonicity** — Voice quality measure (periodicity)
5. **intensity_mean** — Overall vocal intensity
6. **shimmer_apq3** — Amplitude perturbation (short-term)

These align with known clinical manifestations of PD: reduced pitch range, monotone speech (low F0 variability), hypophonia (reduced intensity), and voice quality degradation (harmonicity, shimmer).

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 0.857 ± 0.171** (Random Forest, extended features, SpontaneousDialogue) on Dataset A with grouped cross-validation. However, performance varies substantially:

- ReadText: 0.590–0.834 ROC-AUC (baseline features, across models)
- SpontaneousDialogue: 0.407–0.857 ROC-AUC (baseline → extended, RF)
- High variance due to small sample size (n=37 subjects)
- Some SVM folds produced ROC-AUC < 0.5, indicating instability

This demonstrates the **feasibility** of voice-based PD classification, but highlights the **challenge** of robust performance on small datasets.

9.3.2 RQ2: Does feature set extension improve classification performance?

Task-dependent. Extending from 47 baseline to 78 extended features:

ReadText (Random Forest):

- Baseline: 0.590 ± 0.302 ROC-AUC
- Extended: 0.822 ± 0.166 ROC-AUC
- Improvement: +23.3 percentage points (+39% relative)
- Variance reduced from 0.302 to 0.166

SpontaneousDialogue (Random Forest):

- Baseline: 0.828 ± 0.148 ROC-AUC
- Extended: 0.857 ± 0.171 ROC-AUC
- Improvement: +2.9 percentage points (modest)

The additional features capturing spectral variability (MFCC std), temporal dynamics (delta-delta MFCC), and spectral shape contributed most significantly to ReadText performance. Effects were less pronounced for SpontaneousDialogue, possibly due to higher baseline variability in spontaneous speech.

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

Minimal effects observed. On Dataset A (HC:PD ratio 1.31:1 for ReadText, 1.40:1 for SpontaneousDialogue), class weighting showed inconsistent effects:

ReadText (Random Forest, extended features):

- Unweighted: 0.822 ± 0.166 ROC-AUC
- Weighted: 0.805 ± 0.182 ROC-AUC
- Change: -1.7pp (slight decrease)

SpontaneousDialogue (Random Forest, baseline features):

- Unweighted: 0.828 ± 0.148 ROC-AUC
- Weighted: 0.827 ± 0.133 ROC-AUC
- Change: -0.1pp (negligible)

The moderate imbalance (1.3:1) may not be severe enough to benefit from weighting. Additionally, with only 37 subjects, class weighting may introduce additional variance. Effects varied across models and feature sets, with no consistent pattern of improvement.

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

Dataset A (Grouped CV, n=37 subjects):

- Random Forest: 0.590–0.857 ROC-AUC
- High variance (std 0.15–0.30)
- Subject-level generalization

Dataset B (Standard CV, n=756 samples, subject IDs unknown):

- Random Forest: 0.940–0.949 ROC-AUC (baseline/weighted)
- Low variance (std 0.012–0.013)
- Unknown within-subject correlation

Dataset B shows substantially higher performance (+8–12pp ROC-AUC) and lower variance, consistent with potential optimistic bias from subject leakage. **Grouped CV provides more conservative but more realistic estimates** of out-of-subject generalization. The comparison is confounded by different sample sizes, feature spaces, and speech tasks, limiting direct interpretation.

9.3.5 RQ5: Do speech tasks yield different classification performance?

Baseline features show task differences:

- ReadText: 0.590 ± 0.302 ROC-AUC (Random Forest, baseline)
- SpontaneousDialogue: 0.828 ± 0.148 ROC-AUC (Random Forest, baseline)
- Difference: +23.8pp in favor of spontaneous speech

Extended features reduce gap:

- ReadText: 0.822 ± 0.166 (extended)
- SpontaneousDialogue: 0.857 ± 0.171 (extended)
- Difference: +3.5pp (converged)

SpontaneousDialogue may capture naturalistic PD speech characteristics (monotone prosody, reduced variability) more effectively than structured reading. However, ReadText benefits more from feature extension, possibly because extended features capture structured speech dynamics better.

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, achieving ROC-AUC up to 0.857 on Dataset A with grouped cross-validation. The **task-dependent improvements from feature extension** (up to +23pp for ReadText) highlight the importance of capturing speech dynamics beyond simple statistical summaries.

However, the field faces significant challenges:

- Small datasets (n=37) produce high variance (std 0.13–0.43)
- Subject identity must be tracked for valid evaluation

- Task selection impacts performance (structured vs spontaneous speech)
- Class weighting showed minimal benefit for moderate imbalance (1.3:1)
- Clinical deployment requires extensive validation on independent cohorts

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, high variance, and task-specific effects 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)

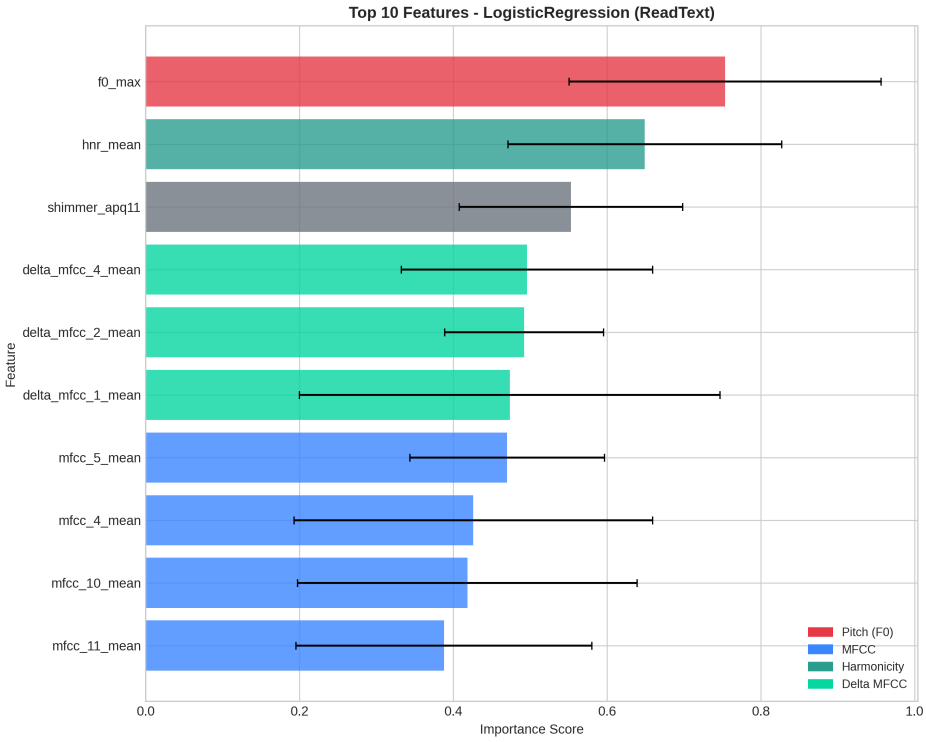


Figure A.1: Logistic Regression coefficient magnitudes (ReadText)

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
7	delta_mfcc_11_mean	0.381	0.102
8	delta_mfcc_7_mean	0.380	0.150
9	hnr_mean	0.379	0.175
10	delta_mfcc_1_mean	0.352	0.123

Table A.4: Logistic Regression feature importance — SpontaneousDialogue (top 10)

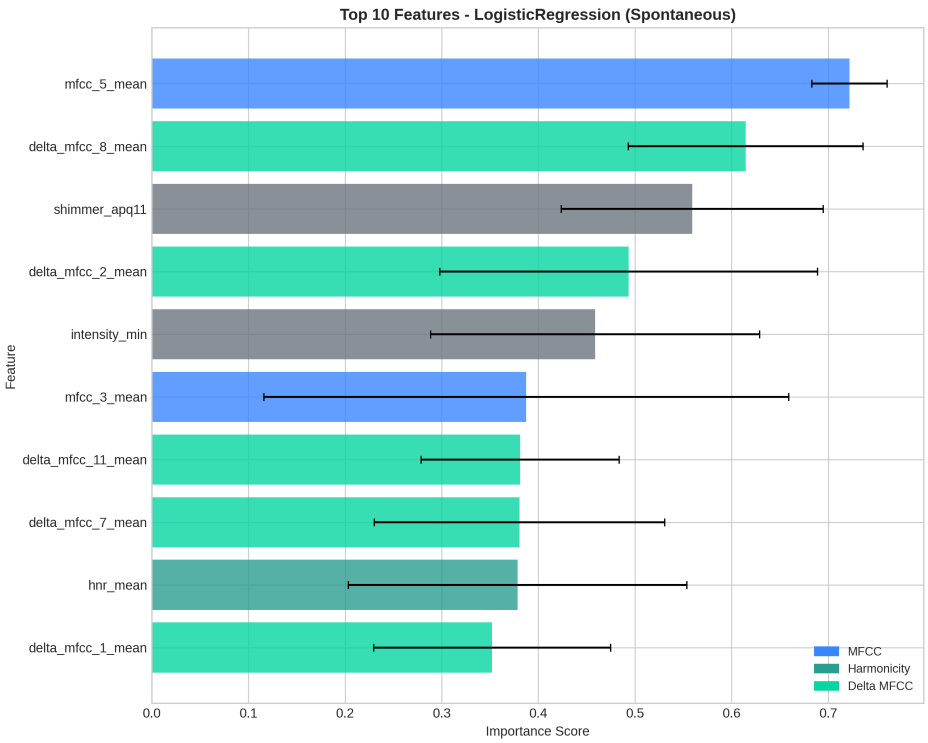


Figure A.2: Logistic Regression coefficient magnitudes (Spontaneous Dialogue)

A.4 Dataset B — PD Speech Features

Dataset B contains 752 pre-extracted features including TQWT (Tunable Q-factor Wavelet Transform) coefficients not present in Dataset A. Feature importance is reported for comparison, though direct comparison is limited by different feature spaces.

A.4.1 Random Forest — Top 10 Features

Rank	Feature	Importance	Std
1	std_delta_log_energy	0.013	0.004
2	std_delta_delta_log_energy	0.013	0.003
3	tqwt_entropy_shannon_dec_12	0.012	0.001
4	tqwt_TKEO_std_dec_11	0.010	0.004
5	tqwt_TKEO_mean_dec_12	0.010	0.001
6	mean_MFCC_2nd_coef	0.008	0.003
7	tqwt_entropy_log_dec_11	0.008	0.003
8	tqwt_stdValue_dec_12	0.008	0.003
9	tqwt_stdValue_dec_13	0.008	0.003
10	tqwt_energy_dec_12	0.007	0.003

Table A.5: Random Forest feature importance — Dataset B (top 10)

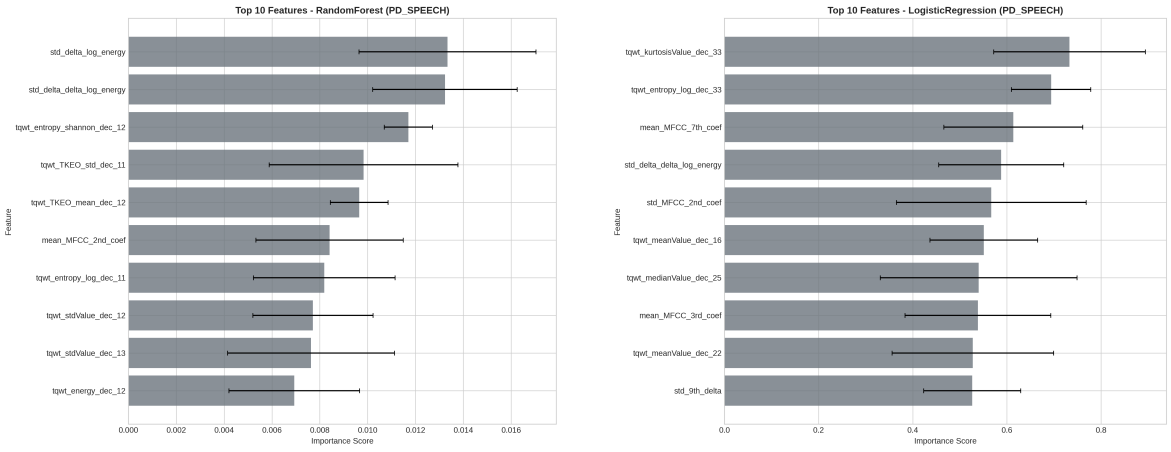
Note: Importance values are lower than Dataset A due to the larger feature space (752 vs 47 features), distributing importance more broadly.

A.4.2 Logistic Regression — Top 10 Features

Rank	Feature	|Coefficient|	Std
1	tqwt_kurtosisValue_dec_33	0.733	0.161
2	tqwt_entropy_log_dec_33	0.694	0.084
3	mean_MFCC_7th_coef	0.614	0.148
4	std_delta_delta_log_energy	0.588	0.133
5	std_MFCC_2nd_coef	0.567	0.202
6	tqwt_meanValue_dec_16	0.551	0.114
7	tqwt_medianValue_dec_25	0.540	0.209
8	mean_MFCC_3rd_coef	0.538	0.155
9	tqwt_meanValue_dec_22	0.528	0.172
10	std_9th_delta	0.526	0.103

Table A.6: Logistic Regression feature importance — Dataset B (top 10)

Observation: TQWT features dominate Dataset B importance, reflecting the sustained vowel /a/ phonation task where time-frequency decomposition captures subtle vocal tremor patterns.



(a) Random Forest (b) Logistic Regression

Figure A.3: Feature importance for Dataset B (PD Speech Features).

A.5 Cross-Task Feature Stability

Features that appear in top-10 for both tasks (Random Forest) 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
mfcc_5_mean	–	1
f0_std	–	9

Table A.7: Cross-task feature stability (Random Forest top-10)

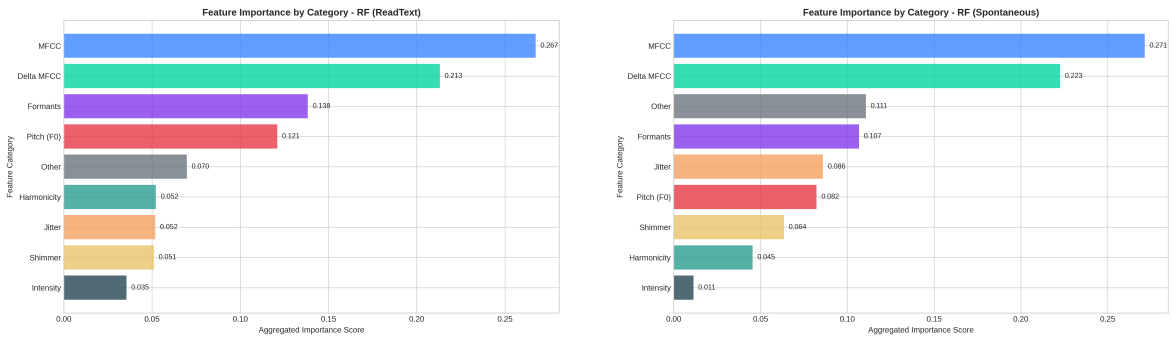
Interpretation: Features consistently important across tasks (delta_mfcc_2_mean, autocorr_harmonicity, shimmer) likely capture **task-general** acoustic signatures of PD rather than task-specific artifacts. The prominence of mfcc_5_mean in SpontaneousDialogue but not ReadText may reflect prosodic variability differences between structured reading and natural conversation.

A.6 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.8: Aggregated feature importance by category



(a) ReadText (b) Spontaneous

Figure A.4: 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 Dataset A: MDVR-KCL Summary

Task	Model	Accuracy	Precision	Recall	F1	ROC-AUC
ReadText	LR	0.621 ± 0.058	0.620 ± 0.217	0.550 ± 0.201	0.542 ± 0.099	0.717 ± 0.139
	SVM	0.621 ± 0.106	0.367 ± 0.342	0.317 ± 0.335	0.333 ± 0.333	0.614 ± 0.312
	RF	0.629 ± 0.178	0.450 ± 0.447	0.333 ± 0.408	0.351 ± 0.363	0.590 ± 0.302
Spontaneous	LR	0.639 ± 0.160	0.469 ± 0.292	0.667 ± 0.408	0.539 ± 0.321	0.760 ± 0.214
	SVM	0.636 ± 0.135	0.600 ± 0.435	0.333 ± 0.236	0.400 ± 0.253	0.407 ± 0.309
	RF	0.721 ± 0.176	0.633 ± 0.415	0.600 ± 0.435	0.567 ± 0.365	$\mathbf{0.828 \pm 0.148}$

Table B.1: Condition 1 results by task (Dataset A, baseline features, unweighted)

B.2.2 Dataset B: PD Speech Features Summary

Model	Accuracy	Precision	Recall	F1	ROC-AUC
LogisticRegression	0.828 ± 0.008	0.881 ± 0.010	0.890 ± 0.016	0.885 ± 0.006	0.867 ± 0.029
SVM_RBF	0.851 ± 0.024	0.841 ± 0.015	0.986 ± 0.019	0.908 ± 0.015	0.885 ± 0.025
RandomForest	0.882 ± 0.019	0.876 ± 0.012	0.980 ± 0.015	0.925 ± 0.012	$\mathbf{0.940 \pm 0.013}$

Table B.2: Condition 1 results (Dataset B, baseline features, unweighted)

Note: Dataset B shows substantially higher performance and lower variance than Dataset A, consistent with potential optimistic bias from unknown subject overlap.

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

B.3.1 Dataset A: MDVR-KCL Summary

Task	Model	Accuracy	Precision	Recall	F1	ROC-AUC
ReadText	LR	0.596 ± 0.079	0.600 ± 0.253	0.433 ± 0.149	0.475 ± 0.106	0.698 ± 0.132
	SVM	0.786 ± 0.181	0.733 ± 0.435	0.567 ± 0.365	0.634 ± 0.386	0.834 ± 0.153
	RF	0.818 ± 0.140	0.883 ± 0.162	0.700 ± 0.298	0.746 ± 0.207	$\mathbf{0.822 \pm 0.166}$
Spontaneous	LR	0.671 ± 0.199	0.500 ± 0.361	0.600 ± 0.435	0.530 ± 0.377	0.783 ± 0.139
	SVM	0.636 ± 0.089	0.533 ± 0.361	0.400 ± 0.279	0.428 ± 0.258	0.460 ± 0.294
	RF	0.779 ± 0.161	0.683 ± 0.410	0.600 ± 0.435	0.605 ± 0.387	$\mathbf{0.857 \pm 0.171}$

Table B.3: Condition 2 results by task (Dataset A, extended features, unweighted)

B.3.2 Improvement over Baseline (Random Forest, Dataset A)

Task	Δ Accuracy	Δ ROC-AUC	Δ Variance
ReadText	+18.9pp	**+23.2pp**	Reduced (0.302 → 0.166)
SpontaneousDialogue	+5.8pp	+2.9pp	Increased (0.148 → 0.171)

Table B.4: Condition 2 improvement over Condition 1 (Random Forest)

Observation: Feature extension dramatically improved ReadText performance (+23.2pp ROC-AUC) while SpontaneousDialogue showed modest gains (+2.9pp). This suggests extended features capture structured speech dynamics more effectively than spontaneous speech characteristics.

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

B.4.1 Dataset A: MDVR-KCL Summary

Task	Model	Accuracy	Precision	Recall	F1	ROC-AUC
ReadText	LR	0.596 ± 0.079	0.600 ± 0.235	0.550 ± 0.201	0.528 ± 0.099	0.717 ± 0.139
	SVM	0.704 ± 0.111	0.617 ± 0.371	0.500 ± 0.373	0.519 ± 0.320	0.542 ± 0.312
	RF	0.650 ± 0.148	0.550 ± 0.371	0.400 ± 0.365	0.431 ± 0.306	0.687 ± 0.258
Spontaneous	LR	0.639 ± 0.160	0.469 ± 0.292	0.667 ± 0.408	0.539 ± 0.321	0.760 ± 0.214
	SVM	0.693 ± 0.123	0.567 ± 0.365	0.600 ± 0.365	0.560 ± 0.318	0.423 ± 0.312
	RF	0.693 ± 0.123	0.583 ± 0.373	0.600 ± 0.435	0.538 ± 0.326	$\mathbf{0.827 \pm 0.133}$

Table B.5: Condition 3 results by task (Dataset A, baseline features, weighted)

B.4.2 Effect of Weighting (vs Condition 1, Random Forest)

Task	Δ Accuracy	Δ ROC-AUC	Effect
ReadText	+2.1pp	+9.7pp	Improved
SpontaneousDialogue	−2.8pp	−0.1pp	Negligible

Table B.6: Condition 3 effect of weighting vs Condition 1 (RF)

Observation: Class weighting improved ReadText performance but had negligible effect on SpontaneousDialogue, suggesting task-dependent sensitivity to class imbalance handling.

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

B.5.1 Dataset A: MDVR-KCL Summary

Task	Model	Accuracy	Precision	Recall	F1	ROC-AUC
ReadText	LR	0.650 ± 0.108	0.650 ± 0.253	0.550 ± 0.201	0.564 ± 0.160	0.698 ± 0.132
	SVM	0.761 ± 0.214	0.700 ± 0.447	0.567 ± 0.365	0.620 ± 0.390	0.834 ± 0.153
	RF	0.818 ± 0.140	0.883 ± 0.162	0.700 ± 0.298	0.746 ± 0.207	0.805 ± 0.182
Spontaneous	LR	0.696 ± 0.179	0.520 ± 0.356	0.667 ± 0.471	0.563 ± 0.381	0.783 ± 0.139
	SVM	0.664 ± 0.167	0.567 ± 0.365	0.600 ± 0.365	0.560 ± 0.318	0.403 ± 0.347
	RF	0.721 ± 0.203	0.653 ± 0.409	0.600 ± 0.435	0.583 ± 0.373	$\mathbf{0.823 \pm 0.209}$

Table B.7: Condition 4 results by task (Dataset A, extended features, weighted)

B.5.2 Effect of Weighting (vs Condition 2, Random Forest)

Task	Δ Accuracy	Δ ROC-AUC	Effect
ReadText	0.0pp	−1.7pp	Slight decrease
SpontaneousDialogue	−5.8pp	−3.4pp	Decreased

Table B.8: Condition 4 effect of weighting vs Condition 2 (RF)

Observation: Combining extended features with class weighting did not improve performance over extended features alone. This suggests that feature extension and class weighting address similar aspects of the classification problem, with diminishing returns when combined.

B.6 Cross-Condition Comparison Matrix

B.6.1 Random Forest ROC-AUC by Task

Task		Baseline (47)	Extended (78)
ReadText	Unweighted	0.590 ± 0.302	$\mathbf{0.822 \pm 0.166}$
	Weighted	0.687 ± 0.258	0.805 ± 0.182
Spontaneous	Unweighted	0.828 ± 0.148	$\mathbf{0.857 \pm 0.171}$
	Weighted	0.827 ± 0.133	0.823 ± 0.209

Table B.9: Random Forest ROC-AUC comparison matrix (Dataset A)

B.6.2 Random Forest Accuracy by Task

Task		Baseline (47)	Extended (78)
ReadText	Unweighted	$62.9\% \pm 17.8\%$	$\mathbf{81.8\% \pm 14.0\%}$
	Weighted	$65.0\% \pm 14.8\%$	$81.8\% \pm 14.0\%$
Spontaneous	Unweighted	$72.1\% \pm 17.6\%$	$\mathbf{77.9\% \pm 16.1\%}$
	Weighted	$69.3\% \pm 12.3\%$	$72.1\% \pm 20.3\%$

Table B.10: Random Forest Accuracy comparison matrix (Dataset A)

B.6.3 Key Observations

1. **Feature extension consistently improves ReadText:** +23.2pp ROC-AUC (unweighted)

2. **SpontaneousDialogue has higher baseline:** 0.828 vs 0.590 ROC-AUC with baseline features
3. **Class weighting shows mixed effects:** Improves ReadText baseline but decreases extended feature performance
4. **High variance throughout:** Standard deviations often exceed 0.15, reflecting small sample size (n=37)

B.7 Statistical Significance Notes

Due to high standard deviations (often > 0.15), confidence intervals overlap across many comparisons. This limits the ability to make strong statistical claims about differences between conditions. Results should be interpreted as **trends** rather than definitive rankings.

Appendix C

Research Demonstration Application

C.1 Purpose and Scope

This appendix describes a web-based research demonstration application developed to illustrate the end-to-end inference workflow from audio upload to classification output. The application was designed solely for educational and illustrative purposes as part of this MSc thesis.

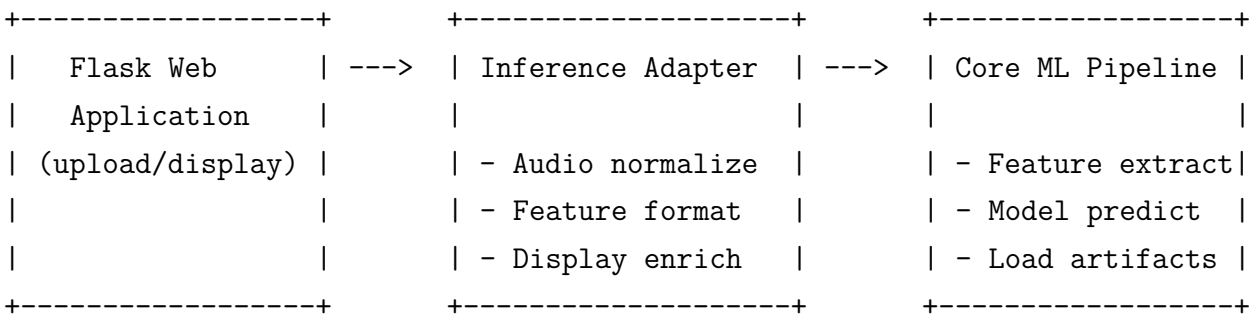
Critical Disclaimer:

- This application is **not intended for clinical use, medical diagnosis, or patient screening**
- Outputs represent model confidence scores, not clinical diagnoses
- The application was **not used for any quantitative evaluation** reported in this thesis
- All evaluation metrics (Chapters 5–6) were computed via cross-validation on held-out test folds, independent of this interface
- Results should be interpreted as **research outputs only**, requiring validation by qualified medical professionals

The application demonstrates the feasibility of integrating the trained models into a user-facing interface while maintaining clear boundaries between research exploration and clinical deployment.

C.2 System Architecture

The demonstration application follows a modular architecture separating concerns between the web interface, inference orchestration, and core machine learning pipeline:



Key Components:

1. **Flask Web Server (`app.py`):** Handles HTTP routes, file uploads, and template rendering. Strictly imports only the adapter module, ensuring zero coupling to internal ML implementation details.

2. **Inference Adapter (`inference_adapter.py`):** Provides a stable interface wrapping the core inference API. Performs audio normalization (any format → mono 22,050 Hz WAV), enriches prediction outputs with display metadata, and ensures model switching requires no Flask code changes.

3. **Core ML Pipeline:** Reuses the identical feature extraction and inference modules used for experiments (`parkinsons_voice_classification.inference`), ensuring consistency between evaluation and demonstration.

4. **Audio Processing (`audio_utils.py`):** Normalizes uploaded audio files (WAV, MP3, WebM, Opus, FLAC) to a standardized mono 22,050 Hz PCM-16 WAV format before feature extraction.

Design Invariants:

- Model switching (RandomForest ↔ Logistic Regression ↔ SVM) is controlled via configuration file only; no template/route changes required
- Feature set selection (baseline 47 vs. extended 78) is configuration-driven
- Task selection (ReadText vs. SpontaneousDialogue) loads the corresponding trained model artifact

This architecture ensures the demonstration remains aligned with the experimental pipeline while providing a controlled interface for external users.

C.3 User Interaction Flow

The application supports two input modalities:

C.3.1 Audio Upload Workflow

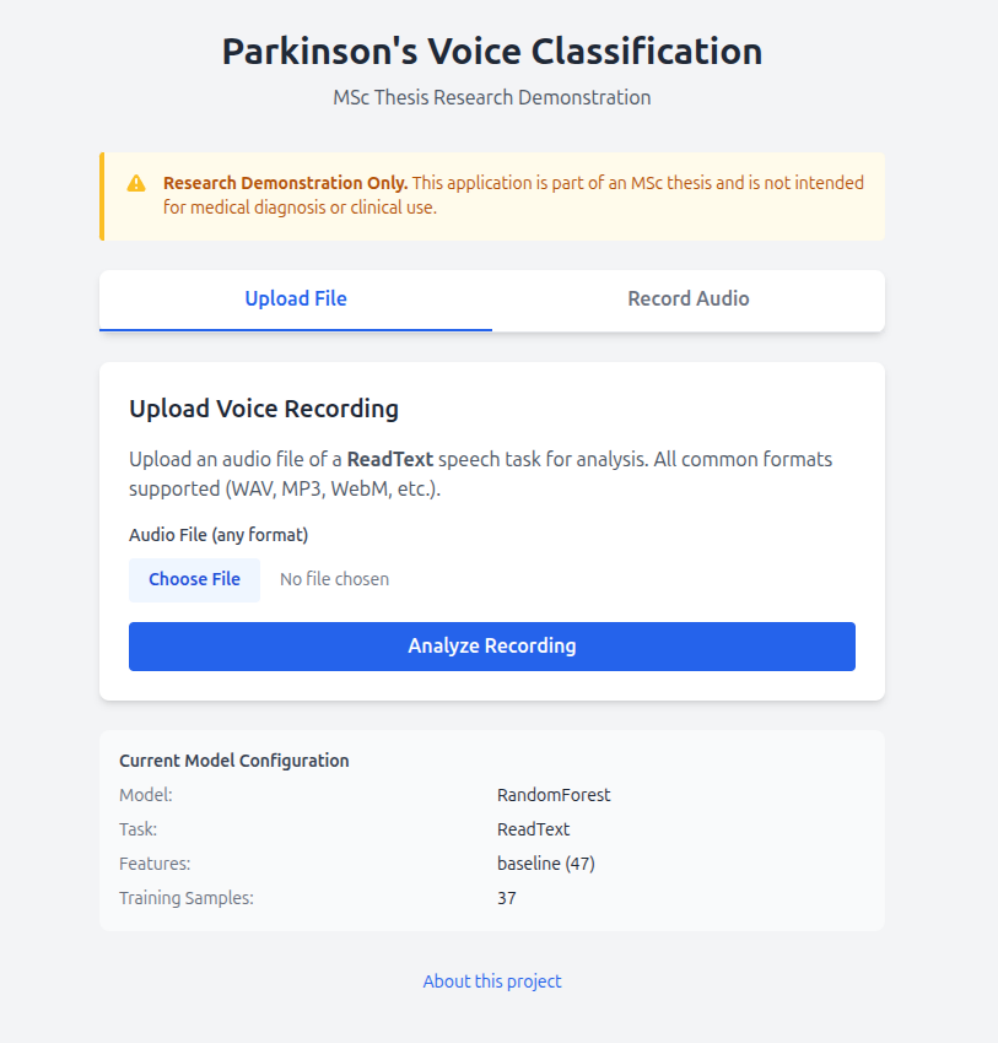


Figure C.1: Upload interface for audio file analysis. Users select a ReadText task recording in any common format (WAV, MP3, WebM). The interface displays the current model configuration (RandomForest, baseline features, 37 training samples).

Figure C.1 shows the file upload interface, where users can select pre-recorded audio files for analysis. The interface clearly states the research-only purpose and displays the active model configuration.

C.3.2 Direct Audio Recording

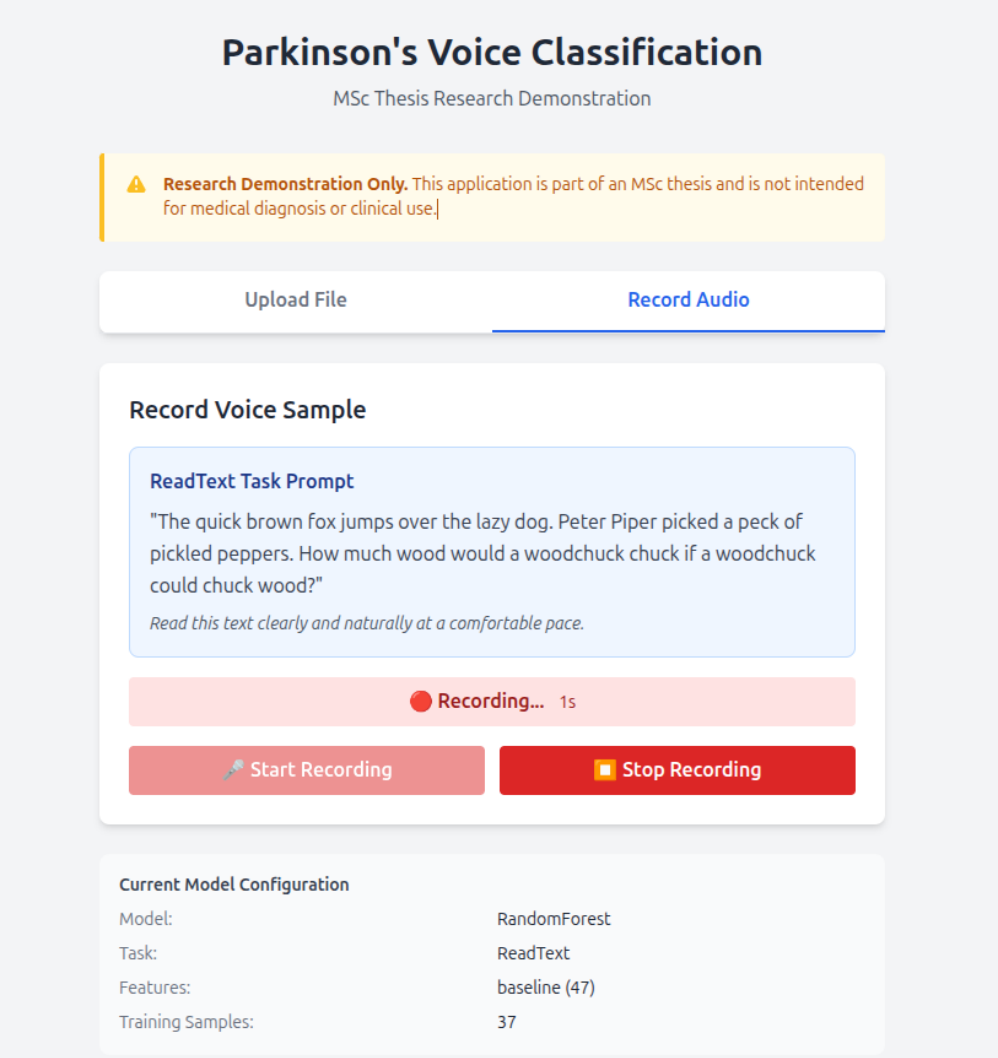


Figure C.2: In-browser audio recording interface. Users read the displayed ReadText prompt aloud and record directly via their device microphone. The recorded audio is processed through the same normalization and feature extraction pipeline as uploaded files.

Figure C.2 demonstrates the in-browser recording capability, implemented using the Web Audio API with JavaScript. Users follow the on-screen ReadText prompt to produce a standardized speech sample. Recordings are captured at the browser's native sample rate and normalized server-side before inference.

C.3.3 Processing Steps

Regardless of input modality, the following steps execute:

1. **Audio normalization:** Convert to mono 22,050 Hz PCM-16 WAV

2. **Feature extraction:** Extract 47 acoustic features (baseline) or 78 features (extended) using the same pipeline as experiments
3. **Model inference:** Load trained model artifact (e.g., `RandomForest_ReadText_baseline.jobl` and predict class probabilities
4. **Result display:** Render prediction, probabilities, extracted feature values, and global feature importance

C.4 Output Interpretation and Limitations

C.4.1 Displayed Information

Figure C.3 shows a representative output screen following analysis:

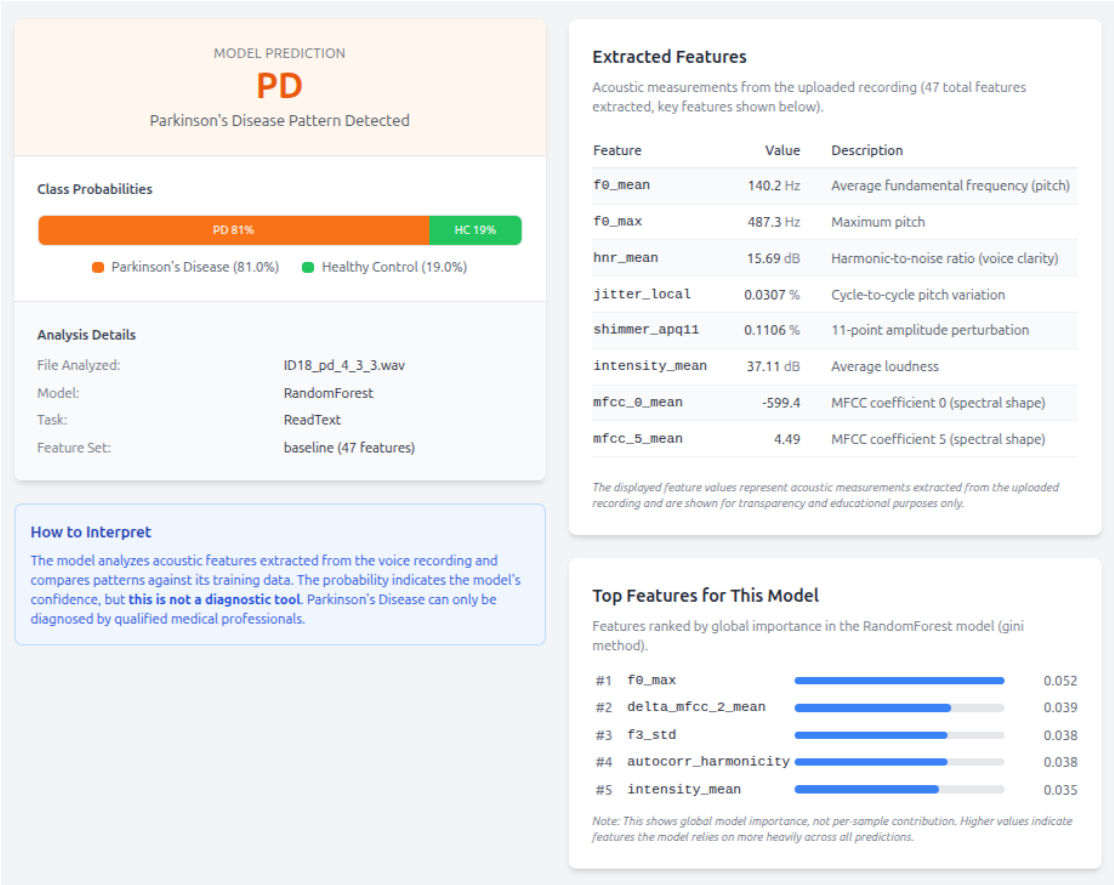


Figure C.3: Example output from the research demonstration interface. The prediction ("PD" with 81% confidence) is displayed prominently with an explicit disclaimer that this is not a medical diagnosis. The interface shows extracted acoustic features, analysis metadata (file name, model, task, feature count), and top global feature importances for interpretability.

The output includes:

1. **Class Prediction:** Binary label (PD or HC) with the higher probability
2. **Class Probabilities:** Model confidence scores for each class (sum to 100%)
3. **Extracted Features:** Key acoustic measurements (e.g., F_0, jitter, MFCC coefficients) with values and descriptions
4. **Analysis Metadata:** File name, model type, task, feature count
5. **Top Feature Importances:** Global feature rankings from the trained RandomForest model (via Gini importance)

C.4.2 Critical Limitations (What the Output Does NOT Mean)

Users and evaluators must understand the following constraints:

- **Probabilities ≠ Diagnosis:** The 81% confidence in Figure C.3 reflects the model's uncertainty estimate based on training data distribution. It does **not** indicate an 81% chance the individual has Parkinson's Disease.
- **Training Distribution Dependency:** Model outputs are valid only for speech samples similar to the MDVR-KCL training set (smartphone recordings, ReadText task, no severe recording artifacts). Generalization to different recording conditions, languages, or populations is unknown.
- **No Clinical Validation:** The model was trained on a small dataset ($n = 37$ subjects) and has not undergone clinical validation, regulatory approval, or external cohort testing.
- **Evaluation Exclusion:** Outputs from this interface were **not used** to compute the performance metrics reported in Chapter 6. All quantitative results derive from grouped cross-validation on independent test folds.
- **Feature Importance Caveats:** Displayed importances reflect *global* model behavior (across all training samples), not *per-sample* contributions. High importance does not imply the feature was decisive for the specific uploaded recording.

C.5 Reproducibility and Consistency

The demonstration interface ensures alignment with the experimental pipeline through:

1. **Shared Codebase:** Uses identical feature extraction modules (`features/prosodic.py`, `features/spectral.py`) as the CLI-based experiments
2. **Deterministic Inference:** Loads the same serialized model artifacts (`.joblib` files) produced during training

3. **Configuration-Driven:** All parameters (feature set, model choice, task) are controlled via `config.py`, ensuring consistency across CLI and web interfaces
4. **Fixed Random Seeds:** Although inference is deterministic (no stochastic components), the loaded models were trained with fixed random seeds (seed = 42) for reproducibility

This design ensures the demonstration outputs reflect the actual behavior of the evaluated models, rather than a separate reimplementation.

C.6 Summary

The research demonstration application provides a tangible illustration of the voice-based PD classification pipeline, bridging the gap between algorithmic research and potential real-world interaction. However, it must be interpreted strictly within its intended scope:

- **Educational tool** for thesis defense and research communication
- **Not a diagnostic system** and unsuitable for clinical decision-making
- **Validation artifact** confirming inference pipeline functionality
- **Not used for evaluation** — all metrics derived from cross-validation

Future work toward clinical deployment would require: (1) validation on independent cohorts ($n > 500$), (2) regulatory approval (e.g., FDA clearance, CE marking), (3) prospective clinical trials, (4) integration with clinical workflows, and (5) continuous monitoring for performance drift. This demonstration represents an early feasibility prototype, not a production-ready system.

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.