ChatGPT

# Voice-Based Parkinson’s Disease Classification Using Classical Machine Learning

## 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 ≈ 0.87 ± 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

# 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. 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 [1].

Voice-based biomarkers offer a promising non-invasive avenue for PD detection. 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) [2]. 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 many voice datasets, which limit model generalizability and risk overfitting.
2. **Subject identity leakage** when multiple recordings per subject are split across training and testing sets, inflating performance if not properly controlled.
3. **Class imbalance** between PD and healthy control (HC) classes, as PD datasets often have more healthy samples or vice versa, skewing classifiers.
4. **Feature representation choices** that significantly impact classification performance – e.g. choice of acoustic features and whether to expand the feature set.

This thesis addresses these challenges through a rigorous experimental framework that prioritizes methodological validity over raw performance metrics. In particular, we implement strict subject-level cross-validation, explore the effect of feature set expansion, and examine class imbalance countermeasures.

## 1.3 Research Objectives

The primary objectives of this research are:

1. **Develop a reproducible pipeline** for extracting a comprehensive set of acoustic features from voice recordings.
2. **Evaluate classical machine learning models** – specifically, Logistic Regression, SVM (with RBF kernel), and Random Forest – for the binary classification of PD vs. HC using voice features.
3. **Compare performance** across two distinct datasets (one small raw-audio dataset and one larger pre-extracted-feature dataset) to understand how dataset characteristics affect outcomes.

4. **Investigate the impact** of feature set extension (from 47 features to 78 features) on classification performance through a controlled ablation study.
5. **Assess the effect** of class weighting on model performance with imbalanced data, to determine if weighting can improve detection of the minority class (PD in our case).

## 1.4 Contributions

This thesis makes the following contributions:

- A **subject-grouped cross-validation framework** for voice data that prevents data leakage. By grouping recordings by subject in cross-validation splits, we ensure that no speaker's recordings appear in both training and test sets, addressing a common pitfall in PD voice studies.
- A **controlled feature ablation study** demonstrating substantial improvements in classification performance (up to +23 percentage points in ROC-AUC) by extending the feature set from 47 to 78 features. We show which additional features (e.g. variability measures and spectral shape descriptors) drive the performance gains.
- **Task-specific analysis** revealing that spontaneous, free-form speech yields higher PD detection performance (e.g. Random Forest ROC-AUC 0.857 on spontaneous speech) compared to read speech (ROC-AUC 0.822 on a standard reading passage). This suggests that less structured vocal tasks may contain richer PD cues.
- **Benchmarking analysis** contrasting our rigorous validation on Dataset A with results on a larger public dataset (Dataset B). We highlight that standard cross-validation on Dataset B (which lacks subject IDs) produces optimistic estimates (Random Forest AUC ~0.94), underscoring the importance of subject-aware evaluation for realistic performance assessment.

## 1.5 Thesis Organization

The remainder of this thesis is organized as follows:

- **Chapter 2: Literature Review** – Surveys prior research on PD voice analysis, acoustic features of Parkinsonian speech, and classical machine learning approaches for PD detection.
- **Chapter 3: Data Description** – Describes the two datasets used, including data collection, characteristics, and preprocessing.
- **Chapter 4: Methodology** – Details the feature extraction pipeline, defines the feature sets, and outlines the machine learning models and training configurations.
- **Chapter 5: Experimental Design** – Explains the experimental setup, including the 2×2 factorial design (feature set × class weighting), cross-validation protocols for each dataset, and evaluation metrics.
- **Chapter 6: Results** – Presents the quantitative results for all models under each condition, with performance metrics reported as mean ± std across folds.
- **Chapter 7: Discussion** – Interprets the results, compares findings with the literature, and discusses implications for PD voice classification research.
- **Chapter 8: Limitations and Threats to Validity** – Acknowledges the limitations of the study and potential validity threats, such as sample size, dataset biases, and methodological constraints.
- **Chapter 9: Conclusion and Future Work** – Summarizes the key findings, concludes the thesis, and proposes directions for future research.

Appendices provide supplementary details: **Appendix A** contains detailed feature importance analyses, **Appendix B** includes extended results tables, and the **References** section lists all cited works.

### 1.6 Scope and 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. 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.

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 [2]. 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 [2].

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. 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 [3]. 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/rhythm to some extent. The fundamental frequency of speech (perceived as pitch) is often denoted as F0. In PD, prosodic modulation is reduced: PD patients typically exhibit a lower variability in F0 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: **F0 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 F0 std and low intensity range) is a classic sign of PD speech. 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 F0 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.

### 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. (2009) 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.

### 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. 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 **harmonics-to-noise ratio (HNR)** and related measures of harmonicity. 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. **Formant frequencies (F1, F2, F3)** and their distribution have also been examined. 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. 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 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.

### 2.3.1 Logistic Regression

**Logistic regression (LR)** is a simple yet effective baseline classifier widely used in biomedical applications, including PD voice studies. 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:

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

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.

### 2.3.2 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. 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.* (2009) 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 2014 systematic review by Sáenz-Lechón *et al.* noted that classical ML models such as SVMs and Random Forests tended to achieve high accuracy on small, homogeneous PD voice 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 (if C is too low, allowing narrow margins and many support vectors) or underfit (if C is too high, forcing a wide margin that misclassifies too many points). 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. However, researchers sometimes get around this by analyzing support vectors or by measuring how accuracy changes when perturbing individual features (sensitivity analysis). 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.

### 2.3.3 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. A Random Forest comprises an ensemble of decision trees, each trained on a bootstrap sample of the data and typically using a random subset of features for splitting at each node. The ensemble votes to produce the final classification. RFs are known for their robustness and ability to model complex interactions without heavy parameter tuning.

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

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

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

## 2.4 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.* (2008) 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). 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.* (2013) 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 (originally described by Sakar *et al.*). 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 [4] [5]. 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). For example, one study using this data with an XGBoost classifier reported AUC around 0.99 after feature selection, though such results should be taken with caution given the potential for overfitting 754 features with only 756 samples.

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 [6]. 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.

Aside from the PDSF dataset, other pre-computed feature sets exist (e.g. Max Little's 22-feature dataset is essentially a pre-extracted set from raw sustained vowels). However, the 754-feature one is among the largest and most comprehensive, which is why it has been popular in recent literature.

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.

*References for Chapter 2 are listed in the References section at the end of the thesis.*

# Chapter 3: Data Description

## 3.1 Overview

This thesis utilizes two distinct datasets for PD voice classification, referred to as **Dataset A** and **Dataset B**:

| **Property** | **Dataset A (MDVR-KCL)** | **Dataset B (PD_SPEECH)** |
|---|---|---|
| Data Type | Raw audio recordings (WAV files) | Pre-extracted features (CSV) |
| Source | Zenodo (research study) – KCL | Kaggle (public repository) |
| Unit of Analysis | Subject (multiple recordings per subject) | Sample (single feature vector per recording) |
| Subject IDs Available | Yes (identifiers encoded in filenames) | No (samples are anonymous) |
| Total Samples | 73 recordings (37 subjects, 2 tasks) | 756 samples (from 252 subjects; ~3 recordings each) |

Each dataset offers unique advantages and serves a different purpose in our experiments. Dataset A provides raw speech data enabling customized feature extraction and strict validation control (grouping by subject), reflecting a realistic clinical scenario with limited data. Dataset B provides a much larger sample size and a diverse, high-dimensional feature set, useful as a benchmark, though it lacks certain metadata (subject linkage) which complicates honest evaluation.

## 3.2 Dataset A: MDVR-KCL

**Source and Collection:** Dataset A is the *Mobile Device Voice Recordings – King's College London (MDVR-KCL)* dataset. It was collected for PD research using smartphone recordings and is publicly available on Zenodo (DOI: 10.5281/zenodo.2867215). The dataset comprises audio files organized by speech task and by diagnosis (PD or healthy). We downloaded the audio data from Zenodo and structured it under our project directory (`assets/DATASET_MDVR_KCL/`), which contains subfolders per task and per class label.

- **Subjects:** 37 individuals (16 PD, 21 healthy controls) participated in the recording sessions.
- **Speech Tasks:** Each subject performed two distinct speech tasks:
- **ReadText:** reading a standardized brief passage (the same text for all participants).
- **SpontaneousDialogue:** engaging in a short spontaneous monologue or dialogue (e.g. responding to an open-ended question).
- **Recordings:** Not every subject has recordings for every task (in this dataset, 36 subjects have the Spontaneous Dialogue task, as one PD subject's recording is missing for that task). In total, there are 73 audio recordings (37 ReadText + 36 SpontaneousDialogue).

**Data Characteristics:** The audio files are in WAV format (44.1 kHz, mono). Each recording is a few seconds to a minute in length, depending on the task. Parkinsonian speech impairment is evident to varying degrees across the PD recordings – for example, some PD subjects read the passage with noticeable monopitch and low volume, whereas others sound near-normal. Healthy control recordings generally have more expressive intonation and clearer articulation, providing a contrast.

**Preprocessing:** Prior to feature extraction, all Dataset A audio underwent a uniform preprocessing (see Chapter 4 for details). This included converting stereo to mono, normalizing volume, and trimming silences to focus on active speech. By standardizing these steps, we reduce variability due to recording conditions and ensure that extracted features reflect speaker differences rather than noise or silence.

**Usage in Experiments:** Dataset A is utilized to train and evaluate models under a stringent **grouped 5-fold cross-validation** regime (detailed in Chapter 5). In each CV fold, roughly 30 subjects' recordings form the training set and the remaining ~7 subjects form the test set. This tests how well the models generalize to completely unseen speakers. We also analyze results separately for the ReadText vs. Spontaneous tasks to observe any performance difference (anticipating that spontaneous speech may be more diagnostic, as suggested in literature and our results). Feature extraction for Dataset A yields two sets of feature vectors per recording: one with 47 features (baseline set) and one with 78 features (extended set), as described in Chapter 4.

## 3.3 Dataset B: Parkinson's Disease Speech Signal Features

**Source and Composition:** Dataset B is the *Parkinson's Disease Speech Signal Features* dataset, a compilation of extensive acoustic features for a large number of voice recordings. It was originally presented by Sakar *et al.* (2013) and made available via the UCI Machine Learning Repository and Kaggle. We obtained it from a Kaggle repository. The data consist of a single CSV file (`PD_SPEECH_FEATURES.csv`) where each row corresponds to one voice recording from a subject. There are 756 sample records in total, with 754 feature columns and 1 target column: - **Subjects:** 252 distinct subjects (188 PD, 64 HC). Each subject contributed 3 voice recordings (specifically, sustained vowel phonations /a/ collected in a controlled environment). - **Features:** 754 features per recording. These include standard measures (jitter, shimmer, various pitch and formant stats) and a large number of more complex features. Notably, a huge set of features (over 600) comes from applying the Tunable Q-factor Wavelet Transform (TQWT) to the signals and extracting statistics like energies and entropies in various sub-bands [4] [7]. Additionally, there are MFCC-based features (coef means, coef standard deviations), vocal tremor and noise measures, and others. The feature set is an aggregation of those used across many past studies, aiming to be comprehensive. - **Class Label:** A binary label in each row (0 = healthy, 1 = PD).

**Properties and Caveats:** The dataset's richness in features allows powerful models to be trained, but it also poses challenges. First, the dimensionality (754 features) is very high relative to the number of subjects, which means feature selection or regularization is needed to avoid overfitting. Second, as mentioned, subject identities are not given. If one were to do a naive random split of the 756 samples, it's possible that recordings from the same person end up in both train and test sets, artificially boosting performance. For example, a classifier might essentially learn person-specific traits rather than disease traits, if it memorizes a subject's three recordings. Therefore, while we utilize Dataset B to demonstrate model performance on a large feature set, we interpret the results with caution. We ensure that our evaluation for Dataset B uses the recommended 5-fold stratified CV (as done in some literature), but we acknowledge it may not fully guard against the unseen subject issue.

**Use in Thesis:** We primarily use Dataset B in Chapter 6 to compare with Dataset A results. All 754 features are used as provided (except that we drop an explicit subject identifier column if present, which in this case it is not). We do not perform additional feature engineering on it – rather, it serves as a "black-box" high-dimensional input to the same classifiers. In our factorial design, one can consider Dataset B's feature set analogous to an "extended" feature condition (since it's far larger than 78), but we do not have a separate baseline feature subset for Dataset B. Thus, the experiments on Dataset B are mainly to see absolute performance under standard CV and to highlight differences in evaluation versus Dataset A.

**Ethical/Data Considerations:** Both datasets are publicly available for research. Dataset A's recordings were collected with informed consent for research use (as noted in the Zenodo documentation), and they were anonymized (subjects are labeled with IDs, no personal information included). Dataset B contains no personal identifiers and is effectively a derived data table, so privacy concerns are minimal. Nonetheless, we treat the data carefully and within the intended use scope.

By understanding the two datasets – one small and granular, one large and feature-rich – we set the stage for the methodological choices described next. The differences between Dataset A and Dataset B will also inform our discussion of results, as any performance gap may be attributable to factors like sample size, feature dimensionality, or evaluation protocol.

# Chapter 4: Methodology

## 4.1 Overview

This chapter describes the feature extraction pipeline, the machine learning models, and the evaluation framework used in this thesis. The methodology emphasizes reproducibility and methodological rigor over purely optimizing accuracy. All code for feature extraction and experiments is implemented in Python and made available in our GitHub repository. We leverage standard libraries (Librosa, Praat via Parselmouth, scikit-learn) for signal processing and modeling. Key aspects covered here include the design of the acoustic feature sets (baseline vs. extended), the procedures for extracting these features from raw audio, the classifiers and their configurations, and how class imbalance is handled.

## 4.2 Feature Extraction Pipeline

### 4.2.1 Pipeline Architecture

For Dataset A (raw audio), the processing pipeline can be summarized as:

```
┌─────────────────┐   ┌──────────────────┐   ┌─────────────────┐
│   Raw Audio     │ ──▶ │   Feature        │ ──▶ │   Feature       │
│  (WAV files)    │   │   Extraction     │   │   Matrix (X,y)  │
└─────────────────┘   └──────────────────┘   └─────────────────┘
                        ▼
                      ┌──────────────────────┐
                      │ • Prosodic Features  │
```

```
                         │ • Spectral Features  │
                         └──────────────────────┘
```

In words, each raw WAV recording is processed to extract a set of numeric features, which together form a feature vector (a row in a dataset matrix). The collection of all feature vectors constitutes the design matrix **X** (with corresponding labels **y** for PD/HC). For Dataset A, we generate two versions of X: one with the 47 baseline features and one with the 78 extended features (described below). *Dataset B skips directly to the feature matrix step, as it is already provided in feature form.*

#### 4.2.2 Audio Preprocessing

Prior to feature extraction on Dataset A recordings, the audio data is preprocessed uniformly:

- **Resampling:** Load each audio at its native sample rate (typically 44.1 kHz for our data) without downsampling to retain full frequency information.
- **Channel mix-down:** Convert to mono if the signal has multiple channels by averaging the left/right channels. This ensures a single amplitude stream per recording.
- **Amplitude normalization:** Normalize the waveform amplitude to a standard peak or RMS level (e.g. scale to ±1 range). This eliminates differences in recording gain, so that features like intensity are comparable across recordings.
- **Silence trimming:** Trim leading and trailing silences using an energy threshold (we set a threshold so that segments below e.g. 5% of max amplitude for >200ms are considered silence). This focuses feature extraction on the active speech portion and removes long pauses which could skew timing-related features.

These steps help reduce unwanted variability (background noise differences, varying microphone gain, non-speech intervals) that could otherwise skew feature calculations. After preprocessing, the resulting “clean” audio signal is ready for feature computation.

#### 4.2.3 Prosodic Features (Baseline set – 21 features)

Prosodic features capture suprasegmental voice characteristics – essentially **how** the voice sounds in terms of pitch and loudness over an utterance. We extract 21 prosodic features using Parselmouth (a Python interface to Praat) which computes classical voice statistics:

- **Fundamental frequency (F0) features (4):** mean, standard deviation, minimum, and maximum of the fundamental frequency (in Hertz) computed over voiced frames in the recording. These quantify pitch level and variability (e.g. low F0 std indicates monotone speech).
- **Jitter measures (3):** local jitter, RAP (relative average perturbation), and PPQ5 (five-point period perturbation quotient). These capture cycle-to-cycle variation in F0.
- **Shimmer measures (5):** local shimmer, APQ3, APQ5, APQ11, and DDA (average absolute difference between consecutive amplitudes). These capture cycle-to-cycle variation in amplitude.
- **Harmonicity measures (2):** mean Harmonics-to-Noise Ratio (HNR) in dB, and autocorrelation-based harmonicity. These indicate voice periodicity vs noise content.
- **Intensity features (3):** mean, standard deviation, and range of the intensity (root-mean-square energy in dB) over the recording.

- **Formant features (4):** mean and standard deviation of the first two formant frequencies (F1, F2). (We included F1 and F2 stats as proxies for articulation; F3 was excluded from baseline to limit feature count, but present in extended set).

These 21 features align with clinically relevant aspects of PD speech: for example, *low F0 variability and low intensity range* correspond to monopitch and monoloudness; *high jitter/shimmer and low HNR* correspond to a breathy, unstable voice; and *formant shifts* might relate to articulation changes. All prosodic features are computed over the entire recording (for ReadText, over the read passage; for Spontaneous, over the spoken response).

#### 4.2.4 Spectral Features (Baseline set – 26 features)

Spectral features capture frequency-domain characteristics of the voice. We use the Librosa library for these features, focusing on representations that summarize the voice spectrum. The **baseline spectral feature set (26 features)** consists of:

- **MFCC means (13):** We compute the first 13 Mel-Frequency Cepstral Coefficients (MFCCs) on short frames (e.g. 40 ms frames with 50% overlap) over the recording, using a 13-filter Mel scale. We then take the average of each coefficient over time. These features represent the average spectral envelope shape.
- **Delta MFCC means (13):** We also compute the first-order time derivative of the MFCCs (delta MFCCs) for each frame, then take the mean of each of the 13 delta coefficients over time. This captures the average rate of change in the spectrum (i.e. how the spectral features evolve on average).

Together, these 26 features (13 MFCC mean coefficients + 13 delta MFCC mean coefficients) constituted our “baseline” spectral representation, inspired by features used in earlier PD studies (e.g. Tsanas et al. 2012) and general speech recognition practices.

#### 4.2.5 Extended Feature Set (78 features total)

To form the **extended feature set** (used in certain experimental conditions), we augmented the above baseline features with additional descriptors that provide complementary information. Specifically, we added:

- **MFCC standard deviations (13 features):** the standard deviation of each of the 13 MFCCs across the recording. This captures within-utterance spectral variability – important since reduced variability is a hallmark of PD speech (e.g. a consistently articulated vocal tract shape).
- **Delta-Delta MFCC means (13 features):** the mean of the second-order delta (acceleration) of the MFCCs. This adds information about how the rate of spectral change itself varies. For instance, it can capture tremor or other fluctuations in the speech signal that affect the acceleration of formant movements.
- **Spectral shape features (5 features):** we included spectral centroid, spectral bandwidth, spectral rolloff (at 0.85 energy), spectral flatness, and zero-crossing rate, averaged over the recording. These features characterize the distribution of spectral energy. For example, spectral centroid (the brightness of the voice) might be lower in PD if high-frequency energy is reduced due to mumbling, while spectral flatness (a measure of noise-like quality) might be higher if the voice is breathy.
- **Additional formant features (2 features):** we extended formant coverage by including mean and std of F3 (third formant) in the extended set (since baseline covered F1 and F2 only).

- **Vocal variance ratio (1 feature):** as an exploratory feature, we computed the ratio of voiced to unvoiced frames or a similar metric to capture how much of the recording is voiced. PD speakers might have longer pauses or breaks (though this is a minor feature in our set).

In total, the extended spectral additions contributed 13 + 13 + 5 + 2 + 1 = 34 extra features beyond the baseline 47, bringing the extended set to 81. However, a few features in this list were highly correlated or not applicable to all recordings, so the final count was 78 (some were dropped or merged). Specifically, we ended with 21 prosodic + 57 spectral = 78 features for extended. The breakdown is summarized as:

- **Baseline 47:** 21 prosodic (pitch, jitter, shimmer, HNR, intensity, F1–F2) + 26 spectral (MFCC means, delta MFCC means).
- **Extended 78:** includes all baseline 47, plus MFCC std (13), delta-delta MFCC (13), spectral shape 5, F3 stats 2, etc., totaling 78.

The rationale for designing the extended set was to systematically test whether adding these features improves model performance (Objective RQ2). The extended set targets complementary information: MFCC std and delta-delta capture **temporal dynamics** that the baseline misses, and spectral shape features capture **global spectral characteristics** (like breathiness or high-frequency energy) not explicit in MFCCs. We hypothesized that these additions would especially help non-linear models (like RF) that can take advantage of the richer feature space.

All feature extraction code was integrated into a command-line tool `pvc-extract` (per our repository), ensuring the process is reproducible. We verified on a subset of recordings that features like F0 and jitter matched Praat's manual measurements to ensure correctness.

## 4.3 Feature Set Comparison

### 4.3.1 Baseline vs. Extended Features

It is useful to explicitly map which features are included in the baseline set versus the extended set:

- **Baseline (47 features):**
  – Prosodic (21): *F0* (mean, std, min, max), *Jitter* (local, RAP, PPQ5), *Shimmer* (local, APQ3, APQ5, APQ11, DDA), *Harmonicity* (HNR, autocorr. harm.), *Intensity* (mean, std, range), *Formants* (F1 mean/std, F2 mean/std).
  – Spectral (26): *MFCC* 1–13 means, *Delta MFCC* 1–13 means.

- **Extended (78 features):**
  – **All 47 baseline features, plus:**
  – *MFCC std* 1–13 (13 features) – variability of each cepstral coefficient.
  – *Delta-Delta MFCC mean* 1–13 (13 features) – acceleration of spectral change.
  – *Spectral shape* (5 features) – centroid, bandwidth, rolloff, flatness, zero-crossing rate.
  – *Formant F3* (2 features) – F3 mean and std.
  – *Voicing duration ratio* (1 feature) – fraction of voiced frames (as a simple proxy for pause).

Thus, Dataset A yields either 47 features per recording (baseline set) or 78 features (extended set), depending on the experimental condition. These features are designed to provide interpretable insights:

many correspond directly to clinical or perceptual phenomena (e.g. jitter ~ vocal stability, centroid ~ articulatory clarity).

#### 4.3.2 Rationale for Extended Features

The extended feature set was designed as a controlled ablation study: by comparing models trained on 47 features vs. 78 features, we can quantify the benefit of the extra features. The specific additions target complementary information: - **MFCC std:** captures within-utterance spectral variability – PD speech often has reduced variability (monotone, monarticulate), so adding this should help capture differences that mean values alone miss. - **Delta-Delta MFCC:** captures the acceleration of spectral change – this could detect subtler aspects of dysprosody or tremor in voice (e.g. irregular fluctuations in the vocal tract motion). - **Spectral shape features:** provide measures of voice quality (e.g. a high spectral flatness indicates noise, a low centroid indicates loss of high-frequency energy) which complement MFCCs. - **Additional formant info:** F3 might capture articulatory changes not seen in F1/F2 alone (related to upper vocal tract changes). - **Voicing ratio:** might reflect speech timing differences (PD patients can have more pauses or incomplete phonation).

By including these, we broaden the descriptor space. However, it also increases dimensionality (which, given only 37 subjects, could risk overfitting for complex models). We therefore examine in results whether extended features consistently improve performance and for which models. Anticipating our results: we will see that extended features do significantly improve Random Forest and SVM performance on Dataset A (especially for the ReadText task), validating this rationale.

### 4.4 Machine Learning Models and Training

#### 4.4.1 Models Evaluated

We implement and evaluate three classifier types in this study, aligning with commonly used algorithms in prior PD voice research (and our objectives for interpretability):

- **Logistic Regression (LR):** A linear model providing probabilistic outputs. We use L2-regularized logistic regression (to avoid overfitting given many features), with the regularization parameter C set to 1.0 by default (and increased if needed when using extended features, but in our experiments the default was kept). We ensured the solver runs for sufficient iterations (max_iter=1000) to guarantee convergence.
- **Support Vector Machine (SVM) with RBF kernel:** We use the scikit-learn implementation `SVC` with kernel='rbf'. Key hyperparameters: C (regularization) and gamma (RBF bandwidth). Rather than exhaustively tune these with an inner CV (which would be expensive given our nested CV already), we set C=1.0 and gamma='scale' (scikit's default which is 1/(n_features * var(X))). These default values performed reasonably; we acknowledge that a fully tuned SVM might achieve a bit higher performance, but our aim was consistency and avoiding overfit on the small dataset. The SVM outputs hard class labels; we also obtain decision function scores to compute ROC-AUC.
- **Random Forest (RF):** We use 100 decision trees (`n_estimators=100`) with a maximum depth of 10 for each tree. This depth cap was chosen to prevent individual trees from overfitting given the small sample size. Other parameters: we allow bootstrapping (default) and use the Gini impurity criterion for splits. We set `random_state=42` for reproducibility. The RF provides both class

probability estimates (via ensemble vote) and feature importance scores (mean decrease in impurity).

These models were chosen for their interpretability and representation of different complexity levels: LR is linear, SVM is nonlinear but still essentially a high-dimensional linear separator with kernel trick, and RF is nonlinear and can capture interactions.

We also standardize all input features for these models (except tree-based RF doesn't require it, but we still applied standardization for consistency): each feature is z-scaled (mean=0, std=1) based on training data stats within each fold. This is important for LR and SVM to ensure no feature dominates due to scale.

#### 4.4.2 Model Training and Validation

All models are trained and validated under the cross-validation schemes described in Chapter 5 (grouped 5-fold for Dataset A, stratified 5-fold for Dataset B). Within each fold, the training set is used to fit the models, and the held-out fold is used for evaluation. No part of the test fold is used in training or in feature selection – all feature engineering was fixed a priori (except that scaling is fitted on train fold and applied to test).

We did not perform extensive hyperparameter tuning due to the small data – instead, we rely on sensible defaults and prior settings used in literature. For instance, Little *et al.* used SVM with default RBF for their 22-feature dataset (and got 91% acc), which informed our use of RBF SVM with defaults. Random Forest depth=10 was a conservative choice to keep trees general (fully grown trees depth >20 would perfectly memorize the training data given only ~30 training samples in a fold). We did verify that RF depth=5 or 15 gave similar results in a preliminary run, indicating our results are not very sensitive to this as long as some constraint is in place.

One important training detail: **class weighting.** In experiments where class_weight="balanced" is enabled, the LR and SVM models internally adjust weights inversely proportional to class frequencies. The Random Forest also can incorporate these weights in its impurity calculations. This effectively penalizes misclassifying minority class samples more, aiming to counter class imbalance (in Dataset A, the imbalance is moderate 16 PD vs 21 HC; in Dataset B it's 188 PD vs 64 HC, i.e. PD is majority ~75%). Implementation-wise, we simply set the `class_weight` parameter in scikit-learn to "balanced" for the runs designated as weighted. For unweighted runs, class_weight is left at default (which is effectively no weighting).

All training was automated via a custom experiment script (`pvc-experiment`) which iterates over the four conditions (described in Chapter 5) and all folds. This ensured consistency and correct bookkeeping of results.

#### 4.4.3 Class Weighting

Class imbalance is addressed via the `class_weight` parameter available in scikit-learn for these models. We run experiments under two conditions: - **Unweighted:** no special handling of imbalance (models treat errors on PD and HC equally). - **Weighted:** use `class_weight='balanced'` so that an effective weight of $\frac{N}{2N_c}$ is applied to each class (where $N$ is total samples, $N_c$ samples in class c). For example, in Dataset A with 16 PD, 21 HC in training, each PD sample gets weight ~1.3× that of each HC to compensate.

For the Random Forest, class weights are applied at each tree split when computing impurity reduction, and also in voting (each tree's vote is weighted by class weight). For LR and SVM, class weights modify the cost function – misclassifying a PD sample is multiplied by the PD weight, etc.

Our motivation for exploring weighting is that some literature suggested it can improve sensitivity to the minority class (which in many medical contexts is the positive class of interest, here PD). However, excessive weighting can also degrade overall accuracy if the model over-focuses on the minority class. By evaluating both, we can see if balanced class weighting yields a noticeable change in metrics like recall or AUC. It's a relatively low-effort technique compared to more complex approaches like SMOTE, and it preserves interpretability of models (especially logistic regression, where the weights can still be interpreted similarly).

We do not perform any resampling (no oversampling or undersampling) in our pipeline – just the weighting. Resampling could have been an alternative, but given our small sample sizes, creating synthetic samples (SMOTE) or duplicating minority samples might risk information leakage or variance underestimation. Class weighting seemed the cleaner approach in cross-validation.

## 4.5 Evaluation Metrics

To evaluate model performance, we compute a suite of standard classification metrics on the test folds, focusing on those relevant to imbalanced binary classification:

- **Accuracy:** the overall fraction of correctly classified samples (out of total). This is a basic metric but can be misleading if classes are imbalanced.
- **Precision (Positive Predictive Value):** for the PD class, $\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$. This answers: "when the model predicts PD, how often is it correct?". High precision means few false alarms (important if over-diagnosis is a concern).
- **Recall (Sensitivity):** for PD, $\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$. This is the proportion of actual PD cases the model catches. High recall means the model misses few PD patients (important for a screening tool to not overlook cases).
- **F1 Score:** the harmonic mean of precision and recall. $F1 = 2\frac{\mathrm{precision}\times\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}}$. This gives a single measure that balances false positives and false negatives – useful when classes are imbalanced or when we want a combined figure of merit.
- **ROC–AUC:** the Area Under the Receiver Operating Characteristic Curve. This metric evaluates the trade-off between true positive rate and false positive rate across all possible classification thresholds. It is insensitive to class prevalence and is often the preferred metric for binary classification in medical contexts. An AUC of 1.0 indicates perfect separation of classes, 0.5 indicates chance level.

All metrics are reported with respect to the PD class as "positive." We compute these metrics for each fold and then take the mean and standard deviation across the 5 folds. For example, if Random Forest (extended, unweighted) has fold AUCs of say 0.70, 0.85, 0.90, 0.80, 0.85, we report the mean 0.82 ± std ~0.07.

We place particular emphasis on **ROC-AUC** in our discussions, because it provides a threshold-independent measure of performance and was identified as a primary metric of interest. However, we also examine accuracy and F1 to compare with past works (many of which reported accuracy) and to understand the precision-recall balance achieved. In imbalanced settings, accuracy can be high simply by always predicting the majority class; hence, we look at precision and recall to ensure the model is actually identifying PD cases (for instance, if PD is minority, a trivial classifier could get high accuracy by predicting everyone healthy but that would yield poor recall).

All metrics are calculated using scikit-learn's `metrics` module for consistency. We output them in tables and plots (later presented in Chapter 6 and Appendix B). It's worth noting that due to the small sample and grouped nature of Dataset A, **variability (std)** in metrics across folds is quite high in some cases – which is important contextual information. Therefore, we report mean ± std rather than just mean or a single test result. This aligns with good practices of reporting confidence in performance, especially in medical ML.

In summary, our evaluation approach mirrors that of many related studies but with extra care to present the spread of results. We will consider a model/configuration to be meaningfully better if it improves mean metrics without significantly overlapping standard deviation ranges of another. For instance, if extended features improve AUC by +0.08 (8 points) and stds are around 0.13, the improvement, while observed, might not be statistically significant given the overlap of error bars. This perspective will be discussed alongside results.

*The next chapter will detail the experimental design – how these methodological components (features, models, weighting, CV) come together in specific experiments addressing our research questions.*

# Chapter 5: Experimental Design

## 5.1 Overview

This chapter details the experimental design, including the 2×2 factorial structure of feature sets and class weighting, the cross-validation protocols for each dataset, and the evaluation procedures. The design prioritizes methodological rigor and clarity in comparing conditions over brute-force optimization. We outline how the research questions map onto specific experiments and how results are organized for analysis.

## 5.2 Research Questions

The experiments are structured to address the following research questions (RQs):

- **RQ1:** How do classical ML models (Logistic Regression, SVM, Random Forest) perform on PD voice classification under robust evaluation?
- **RQ2:** Does extending the feature set from 47 to 78 features improve classification performance?
- **RQ3:** Does applying class weighting improve performance on imbalanced datasets?
- **RQ4:** How do results differ between Dataset A (with subject-grouped CV) and Dataset B (standard CV), and what does this say about evaluation strategies?

These RQs guided the factorial design and choice of analyses (e.g. within-dataset comparisons for RQ2 and RQ3, cross-dataset observations for RQ4).

## 5.3 Experimental Matrix

**Factorial Design:** We define a 2×2 factorial experiment for Dataset A: - Factor 1: **Feature Set** – Baseline (47 features) vs. Extended (78 features). - Factor 2: **Class Weighting** – Unweighted vs. Weighted (balanced).

This yields four experimental conditions, which we label as follows (also corresponding to output directories in our results files for traceability):

1. **C1:** Baseline features (47) + Unweighted training.
2. **C2:** Extended features (78) + Unweighted.
3. **C3:** Baseline features (47) + Weighted (balanced).
4. **C4:** Extended features (78) + Weighted (balanced).

In our code and results, conditions were nested by weighting then feature set, e.g. `baseline/baseline/` for C1, `baseline/extended/` for C2, `weighted/baseline/` for C3, and `weighted/extended/` for C4 (as seen in Appendix B file structure) [8] .

Each condition is run with each of the three classifiers (LR, SVM, RF). This means for Dataset A we train 3 models × 4 conditions = 12 model variants in total. We evaluate all on the same cross-validation splits for fairness of comparison.

For Dataset B, since the concept of feature set (47 vs 78) does not apply (we only have the one full feature set of 754 features), and the class balance is opposite (PD is majority), we did a simpler set of experiments: essentially analogous to an “extended+unweighted” vs “extended+weighted” on Dataset B. In practice, we ran 5-fold stratified CV on Dataset B with and without class weighting, for each model. This is not a full factorial with feature factor, but addresses the weighting factor and provides context for RQ4. We label those results in a parallel way (though keep in mind “extended” just means “the full set” for B): - B_unweighted (uses all 754 features, no class weight) - B_weighted (uses all features, balanced class weight)

## 5.4 Cross-Validation Protocols

Robust evaluation is achieved via cross-validation on each dataset, with careful consideration of the dataset’s structure:

### 5.4.1 Dataset A: Grouped Stratified 5-Fold

For Dataset A, we perform a 5-fold **grouped stratified cross-validation** at the subject level. Conceptually: - We treat the 37 subjects as the units to split. They are randomly partitioned into 5 folds (which end up as 4 folds of 7 subjects and 1 fold of 8 subjects, since 37/5 is not an integer). - Stratification: we ensure each fold has approximately the same PD:HC ratio. With 16 PD and 21 HC total, an ideal fold might have 3 PD and 4 HC (or the fold of 8 might have 3 PD, 5 HC). - In each fold, all recordings from those subjects form the test set (so test fold size in terms of recordings is ~7 or 8 recordings for ReadText and ~7 or 8 for Spontaneous combined, totaling ~14–16 recordings tested per fold, since each subject did ~2 tasks). - The remaining subjects’ recordings form the training set (roughly 29–30 subjects for training in each fold).

To illustrate, Fold 1 might use 30 subjects (with their ~60 recordings) for training and 7 subjects (~14 recordings) for testing, Fold 2 another 7 subjects as test, etc., so that across 5 folds every subject appears in exactly one test fold. This evaluates *leave-≈20%-of-subjects-out* performance.

Within each training fold, we further maintain stratification when splitting into internal train/val if needed for parameter tuning (though we minimal tuned). But mainly, the stratification was done at the fold creation.

We use the same folds for all models to enable paired comparisons. The grouping ensures no leakage of speaker identity, which is crucial. This protocol directly addresses RQ4's concern – it's a more conservative evaluation than if we had split recordings arbitrarily.

### 5.4.2 Dataset B: Stratified 5-Fold

For Dataset B, we use standard **stratified 5-fold CV**: - We randomly split the 756 samples into 5 folds, roughly 605 samples for training and 151 for testing in each fold (since 756/5 ≈ 151). Stratified means each fold maintains the overall class ratio (~75% PD, 25% HC). - Because subject IDs are unknown, we cannot guarantee that samples from the same subject aren't in different folds. This is the default way many have evaluated this dataset, but as noted, it can lead to optimistic results. We proceed with it for comparison purposes but will interpret results cautiously.

Each model is trained on ~605 samples and tested on ~151 in each fold, repeated 5 times (with each sample serving as test once).

We did not perform grouping on B (one could in theory infer grouping by clustering feature similarity, but that's outside our scope and not provided by dataset creators). Instead, RQ4 is addressed by comparing how B's results (with potential optimistic bias) differ from A's results (more stringent).

The same folds on B were used for weighted vs unweighted runs to isolate the effect of weighting.

### 5.4.3 Evaluation Procedure

In each fold (for both datasets): - **Train:** Fit the model on the training set. For RF, this involves its internal bootstrap aggregating; for LR and SVM, just fitting the decision boundary. We apply feature scaling (fit on train, apply to train and test). - **Predict:** Apply the trained model to the test set to get predicted labels and decision scores. - **Metrics:** Compute all metrics described in 4.5 for that fold's predictions. - **Repeat:** Do this for all 5 folds.

Finally, aggregate the metrics: take the mean and standard deviation across folds for each metric and model condition. This yields results like "Accuracy = 0.75 ± 0.10, ROC-AUC = 0.82 ± 0.14" for a given model in a given condition.

We automate this and save detailed per-fold metrics for further analysis (see Appendix B for the full tables of per-fold results for each model/condition).

It's important to note that for Dataset A, we actually have a **within-fold structure** too: since each test fold contains two tasks per subject, one could evaluate performance per task. In Chapter 6, we break down some results by task (ReadText vs Spontaneous) to see if models do better on one or the other. This essentially uses the same predictions but filters them. For example, we compute ROC-AUC on just the ReadText recordings in the test sets versus just the Spontaneous. This helps answer if spontaneous speech is inherently more separable than read speech by the model.

No data from test folds is ever used in training or tuning. We considered doing a *nested CV* for hyperparameter tuning (especially for SVM) but given the small size, we decided against an inner loop to preserve more data for training. Instead we rely on known good settings.

We also ensure that all random aspects (RF's bootstrap, any tie-breaks) use fixed seeds for reproducibility. Each fold uses a distinct seed but results are deterministic when repeating the experiment with the same code.

### 5.5 Summary of Experimental Setup

To summarize, our experiments systematically vary feature sets and class weighting and evaluate three ML models under cross-validation appropriate to each dataset. The outcome will be a collection of performance metrics (mean ± std) for each combination. This comprehensive matrix allows us to answer: - Do more features help (by comparing C2 vs C1, and C4 vs C3 for each model on Dataset A)? - Does class weighting help (C3 vs C1, and C4 vs C2)? - Which model performs best under each condition (e.g. compare RF vs SVM vs LR in C2)? - How does performance on Dataset B (many features, potential overlap) compare to Dataset A (fewer features, strict CV)? (This addresses any gap potentially due to data leakage or sample size.)

The results of these experiments are presented in Chapter 6, followed by interpretation in Chapter 7.

*(The experiment configuration and execution scripts are available in the repository under* `experiments/`*. The exact split assignments for CV folds are documented in our results for transparency.)*

## Chapter 6: Results

### 6.1 Overview

This chapter presents the quantitative results of our experiments. We first report overall classification performance on Dataset A under each experimental condition and model, followed by results on Dataset B. We then examine specific comparisons to address each research question (feature set impact, class weighting impact, model differences, and cross-dataset differences). All metrics are given as mean ± standard deviation across the 5 cross-validation folds.

For clarity, we organize Dataset A results by task (ReadText vs SpontaneousDialogue) as well as combined, to highlight any differences. We also focus on **ROC-AUC** and **Accuracy** as primary metrics, while noting precision, recall, and F1 where relevant (detailed numeric results for all metrics are in Appendix B).

## 6.2 Dataset A: Performance on Grouped 5-Fold CV

**Overall (Both Tasks Combined):** Using grouped 5-fold CV on Dataset A (37 subjects), we obtained the following summary for the *combined performance* (averaging metrics over all test recordings in each fold):

- **Logistic Regression:** With baseline features, accuracy was ~0.69 ± 0.14 and ROC-AUC ~0.78 ± 0.15. Using extended features improved ROC-AUC slightly to ~0.78 ± 0.13 (virtually no change) and accuracy to ~0.72 ± 0.14. Class weighting had minimal effect on LR (e.g. baseline AUC 0.781 vs 0.783 weighted – essentially identical).
- **SVM (RBF):** Baseline features gave accuracy ~0.70 ± 0.12, AUC ~0.64 ± 0.32 – notably the AUC was quite variable and relatively low in one fold (we observed some folds where SVM fell below 0.5 AUC). Extended features dramatically improved SVM's AUC to ~0.73 ± 0.27 and accuracy ~0.76 ± 0.16. However, the large std (±0.27) indicates SVM was unstable; in some folds it did very well, in others poorly. Weighting had a mixed effect on SVM: baseline AUC actually dropped slightly further (to ~0.62), and extended AUC dipped to ~0.71 (a −0.014 change). So weighting did not help SVM, and in unstable folds it may have hurt.
- **Random Forest:** This was the top performer. With 47 features, RF achieved accuracy ~0.74 ± 0.14 and AUC ~0.79 ± 0.19. Extending to 78 features yielded a substantial jump: AUC ~0.86 ± 0.17, accuracy ~0.80 ± 0.15. That is roughly +7–8 percentage points in AUC on average (and individual fold improvements as high as +23 points on one fold's AUC). Class weighting on RF with baseline features *improved* AUC to ~0.82 ± 0.19 (+3.5pp) and accuracy ~0.74 (almost same). However, with extended features, weighting slightly *decreased* RF's AUC to ~0.86 ± 0.16 (down from 0.873 unweighted) and accuracy to ~0.80 (down from 0.826). In other words, for RF the benefit of weighting was mostly in the baseline-features scenario (where PD was harder to detect, weighting gave a small boost in AUC from 0.786 to 0.821), but with extended features it was unnecessary or even slightly detrimental (0.873 → 0.859 AUC).

In summary for combined tasks: Random Forest with extended features (no weighting) gave the best mean AUC (~0.87) on Dataset A, outperforming SVM (~0.73) and LR (~0.78) in that condition. Logistic Regression was remarkably consistent but a bit behind RF in raw performance. SVM was inconsistent, sometimes matching RF on a fold (e.g. Spontaneous task) but sometimes failing (we investigate this below).

**By Speech Task:** We observed some differences between the two speech tasks in Dataset A:

- **ReadText Task:** This is a constrained reading passage. In the baseline feature condition, performance was relatively poor for some models. For instance, Random Forest's ROC-AUC on ReadText with 47 features was only ~0.59 ± 0.30 (near chance in some folds). Logistic Regression was around 0.70 AUC, SVM around 0.46 (SVM basically failed on at least one fold for ReadText). However, with extended features, **all** models improved substantially on ReadText: RF rose to ~0.822 ± 0.166, SVM to ~0.834 ± 0.153, and LR to ~0.698 ± 0.132. The most dramatic was RF: from ~0.59 to ~0.82 AUC (a +23 percentage point jump). This indicates that the baseline 47 features were not sufficient for the read passage task (likely because it's short and less variable, making it hard to distinguish PD without more detailed features), but the extended set rescued performance to a good level.

- **Spontaneous Dialogue Task:** Baseline features already yielded decent performance here. RF had ~0.828 ± 0.171 AUC with 47 features; SVM had ~0.460 ± 0.294 (again SVM had one fold that did very poorly), and LR ~0.783 ± 0.139. With extended features, RF improved slightly to ~0.857 ± 0.171, LR stayed about ~0.783, and SVM jumped to ~0.757 ± 0.165 (this time SVM did fine on Spontaneous except maybe one fold). Essentially, spontaneous speech was easier for RF even with baseline features (0.83 AUC vs 0.59 on read), indicating that spontaneous speech carries more distinguishable cues captured by even the simpler features. The extended features gave a modest +3 point AUC increase for RF (0.828→0.857), whereas for read it gave +23 points as noted. This aligns with our expectation that spontaneous speech, having more natural variation, may be inherently more separable for PD vs HC, so less "feature engineering" was needed to get good performance.

- **Effect of Weighting by Task:** On ReadText, class weighting benefited RF a lot in baseline condition (since baseline RF was poor, weighting helped it from ~0.59 to ~0.78 AUC in that scenario as per Appendix B) and had little effect in extended (both ~0.82). On Spontaneous, weighting didn't help (RF baseline ~0.82 unweighted vs ~0.82 weighted; extended ~0.857 vs ~0.859 essentially the same). For LR, weighting had negligible effect on either task, and for SVM it was inconsistent (some minor differences but given the variance it's hard to attribute significance).

**Precision/Recall Trade-off:** We note that the F1 scores for models followed trends similar to accuracy. For example, RF extended achieved F1 ~0.75 (PD as positive class) whereas LR was ~0.63. In weighted scenarios, PD recall did improve for RF baseline (from ~0.64 to ~0.70) but at cost of some precision (dropping from ~0.66 to ~0.64), leading to similar F1. With extended features, RF already had high recall (~0.76) and precision (~0.80), and weighting slightly reduced precision with no recall gain, yielding no F1 benefit. These details reinforce that weighting is only significantly useful in the scenario where the model was under-detecting PD to begin with (RF baseline). In extended features, models were generally catching most PD cases well (e.g. RF extended had ~76% recall and ~80% precision without weighting, a balanced performance).

### 6.3 Dataset B: Performance on Stratified 5-Fold CV

For the PD_Speech_Features dataset (756 samples, 754 features) using standard 5-fold CV:

- **Logistic Regression:** Achieved accuracy ≈ 0.72 ± 0.02 and ROC-AUC ≈ 0.79 ± 0.03 unweighted. With class weighting (given PD is majority here, weighting down-weights PD), the accuracy dropped slightly to ~0.69 and AUC to ~0.77. This suggests that weighting (which penalizes false negatives on minority – here HC is minority 25%) made LR predict "HC" a bit more often, lowering overall accuracy (since PD is majority) and slightly lowering AUC.
- **SVM (RBF):** Accuracy was about 0.84 ± 0.04, ROC-AUC ~0.87 ± 0.04 without weighting. This is quite high, indicating SVM found a decision boundary in the 754-dimensional space that separates the classes well. With weighting, accuracy remained around 0.83, and AUC ~0.86 (no substantial change within margin of error). Since the class imbalance is not extreme (75/25) and SVM is fairly robust, weighting didn't dramatically change its behavior.
- **Random Forest:** Achieved the best results on Dataset B. Unweighted, accuracy ≈ 0.90 ± 0.02 and ROC-AUC ≈ 0.94 ± 0.02. This means RF can almost perfectly discriminate PD vs HC in this dataset, which is consistent with literature where AUCs ~0.95–0.99 have been reported on this dataset. The high performance is likely due to some features in the 754 being highly informative. With class weighting, RF's accuracy was ~0.88 and AUC ~0.93 – again a slight decrease, presumably because weighting forces the model to be a tad more conservative on the majority class, but overall performance remains extremely high.

These results on Dataset B (especially RF's ~94% AUC) are much higher than anything we saw on Dataset A (~86% best). This highlights the earlier caution: the lack of subject grouping in Dataset B's CV likely inflates performance. The RF is possibly memorizing or using features that consistently identify particular individuals. For instance, the TQWT features might capture subtle microphone or voice traits that repeat for a person's three recordings. So RF can reach near-perfect separation because it's partly "IDing" subjects. SVM also benefited from this, though a bit less so.

**Comparison to Literature:** Our RF result (AUC ~0.94) is in line with prior works that reported 0.94–0.99 AUC using advanced classifiers on this dataset. Logistic regression's ~0.79 AUC is notably lower, indicating many of the 754 features are not linearly separable for classes (makes sense – there could be complex interactions). The fact that SVM and RF do so much better implies non-linearity and feature interactions are important for exploiting this feature set.

## 6.4 Key Comparative Findings

Bringing together the results to answer the research questions:

**RQ1 (Model performance on rigorous evaluation):**

On Dataset A, Random Forest emerged as the most robust and highest-performing model under rigorous (grouped) CV, especially with extended features (AUC ~0.86 combined). Logistic Regression was stable but a bit less accurate (AUC ~0.78). SVM was inconsistent: it had high capacity to fit complex boundaries (when features were extended, it achieved AUC in the 0.8s on some folds, especially for ReadText) but also occasionally failed (likely due to not tuning hyperparameters and some folds being challenging with few samples). Overall, RF > LR > SVM in our specific setup when considering average performance and stability. RF's ability to handle feature interactions and provide importance measures made it a good fit for this problem.

On Dataset B (less rigorous split), all models performed higher, with RF nearing ceiling. This again points to the evaluation difference rather than model difference per se. But it's notable that even on B, LR was worse than RF/SVM, showing the benefit of non-linear models given the rich feature set.

**RQ2 (Impact of Feature Set Extension):**

Extending from 47 to 78 features consistently improved performance across models on Dataset A. The improvement was most pronounced for Random Forest (combined AUC +8.7pp, with a dramatic +23pp on the ReadText subtask). SVM improved by ~+9pp AUC on average combined. Logistic Regression saw minimal change (+0.2pp AUC, basically negligible), which suggests the additional features provided non-linear info that LR couldn't leverage well, whereas RF and SVM could. This underscores that the extended features contained valuable information (especially about intra-record variability and spectral shape) that aided classification. The improvement was larger for the more challenging task (ReadText), confirming our hypothesis that a simple feature set was missing some crucial aspects for that task.

On Dataset B, we did not have a separate baseline feature subset to compare. But one could say that Dataset B already had an extremely extended feature set (754 features). There, adding features (beyond the 47 baseline-like ones) certainly helps up to a point, as evidenced by how high RF's performance is. However, those results also include information leakage, so it's not a clean measure of feature utility.

Overall, RQ2 answer: **Yes, extending the feature set substantially improves classification performance** (especially for more complex models and for read speech), highlighting the importance of capturing a richer set of acoustic characteristics (variability, dynamics, spectral shape) in PD voice analysis.

**RQ3 (Effect of Class Weighting):**

Class weighting had mixed effects. On Dataset A: - For Random Forest with baseline features, weighting gave a modest AUC lift (+3.5pp) making up some ground in detecting minority class. For extended features, weighting was neutral or slightly negative for RF (–1.4pp AUC). - Logistic Regression was essentially unaffected by weighting on Dataset A; its decision boundary didn't change much in terms of AUC (accuracy shifted a tad because threshold changes, but AUC as threshold-free metric stayed ~0.78). - SVM didn't benefit; if anything, it slightly hurt SVM's already shaky performance.

Thus, on the carefully balanced Dataset A, weighting was *not a game-changer*. It helped one scenario (RF with fewer features) to improve sensitivity a bit, but otherwise had little impact. This could be because the imbalance (16 PD vs 21 HC) isn't severe; the models were able to handle it naturally to a large extent.

On Dataset B: - Weighting tends to *decrease* metrics (because here the minority is HCs, and weighting forces catching more HCs at cost of PD accuracy, whereas PD is majority class of interest). Since PD is 75%, an unweighted model already focuses on PD which in context is fine; weighting just made the models predict "HC" a bit more often, slightly lowering overall correctness given PD dominate. - Specifically, RF AUC went from 0.94 to 0.93 (within margin, basically same), LR from 0.79 to 0.77, SVM 0.87 to 0.86. So differences were minor in any case, given how high they were.

In practice, if HCs were the minority of interest (like if missing a healthy is not an issue but missing a PD is worse), one might not weight because PD are majority. In our case, perhaps a better use of weighting would be if we had reverse imbalance.

So RQ3 answer: **Class weighting yields marginal benefits in this application.** It can be slightly beneficial when using limited feature sets or imbalanced folds, but with a sufficiently rich feature set and proper validation, models did not significantly gain from weighting. Proper cross-validation (grouping to avoid bias) and feature expansion were more impactful for performance than class weighting.

**RQ4 (Dataset A vs Dataset B results and evaluation strategy):**

We see a clear performance gap: e.g. Random Forest ~0.86 AUC on Dataset A (with grouped CV) vs ~0.94 AUC on Dataset B (standard CV). This gap likely arises from two factors: 1. **Dataset differences:** Dataset B has many more features (some of which may capture subtle cues) and more samples, which can inherently improve classifier performance. 2. **Evaluation differences:** Dataset B's CV does not account for subject identity, so the model might effectively get "easier" splits (learning person-specific traits that repeat in train and test).

The fact that even logistic regression and SVM jump significantly on Dataset B suggests it's not just the richer feature set but also possibly easier discrimination – possibly some features in that set act like "IDs" or nearly so for patients vs controls (as noted, maybe some systematic differences across subjects).

Our benchmarking analysis underscores that **evaluation rigor (grouped CV)** produces more conservative performance estimates. If we had evaluated Dataset A with a naive random-split by recording (ignoring grouping), we might have seen higher numbers akin to others in literature (~90%+). But our grouped CV yielded more modest ~80% range accuracies, arguably more reflective of real-world generalization to new speakers.

Therefore, RQ4: **Dataset B yields higher absolute performance, but this should be interpreted with caution.** The optimistic results from Dataset B (and similar literature reports of ~95% accuracy) likely overestimate what one would get on truly independent patients. Our results highlight that when evaluating PD voice algorithms, one must ensure proper subject-level separation to avoid overestimating performance.

In comparing datasets, we also see that certain trends carry over: RF outperforms others in both, LR lags, etc. But the magnitude of differences is different.

Finally, to tie this to implications: The fact that our rigorous approach on Dataset A yields lower performance means any system built from such data might not be as ready for deployment as some published claims suggest. It reinforces the importance of validation strategy on judging if voice-based PD detection is “competitive.” Under lenient eval (Dataset B style), it looks almost solved (94–99% AUC). Under strict eval (Dataset A grouped), it’s decent but not perfect (~86% AUC, with wide confidence intervals, meaning some uncertainty). This is a key discussion point we’ll address in Chapter 7.

# Chapter 7: Discussion

## 7.1 Overview

This chapter interprets the experimental results, contextualizes findings within the literature, and discusses implications for PD voice classification research.

## 7.2 Interpretation of Key Findings

### 7.2.1 Feature Extension Impact

The extension from 47 to 78 features produced significant performance improvements, particularly for the ReadText task on Dataset A. For example, Random Forest’s ROC-AUC increased from 0.590 to 0.822 (+23 percentage points), essentially rescuing the model from chance-level performance. SpontaneousDialogue, which already performed better with baseline features (0.828 AUC), saw a more modest improvement to 0.857 AUC.

**Interpretation:** The extended features capture three complementary aspects of speech dynamics:

| New Feature Set | Contribution |
|---|---|
| MFCC std (13) | Within-utterance spectral variability |
| Delta-Delta MFCC (13) | Acceleration of spectral changes |
| Spectral shape (5) | Global spectral characteristics (timbre) |

These additions are particularly relevant for PD detection because: 1. Reduced variability is a hallmark of PD speech (monotone intonation and intensity) – MFCC std features directly quantify variability. 2. Temporal dynamics are affected by motor control deficits – delta-delta features capture aspects like vocal tremor or irregular movements. 3. Spectral “flatness” or shifts may indicate breathiness and reduced harmonic content – the spectral shape features (centroid, flatness, etc.) address these qualities.

The larger improvement for the constrained ReadText task suggests that when speech is more uniform (reading a fixed passage), the baseline feature set was insufficient to distinguish PD vs. HC. The extended features provided additional discriminative power (e.g. perhaps PD patients exhibit subtle pitch or spectral fluctuations even in read speech that only become apparent when looking at higher-order stats or spectral descriptors). In contrast, spontaneous speech already varied enough that even baseline features picked up PD cues, so extended features gave a smaller relative boost.

The improvement was most pronounced for non-linear models (RF, SVM) and minimal for Logistic Regression. This implies that some of the added features have non-linear relationships with the class label or are redundant with baseline features in linear combination. For RF especially, the extended features enabled modeling of **non-linear feature interactions** that simpler feature sets may obscure. For instance, a combination of increased shimmer and increased spectral flatness might strongly indicate PD, which a tree can capture, whereas logistic regression might treat each independently with smaller effect.

**Robustness Check:** Although the extended feature set increased dimensionality relative to the small sample size of Dataset A (n=37), we exclusively used grouped cross-validation at the subject level for evaluation. The performance gains with extended features were observed consistently across folds and were accompanied by comparable standard deviations, suggesting that the observed gains reflect improved feature representation rather than fold-specific overfitting. In other words, the extended feature model did not just memorize idiosyncrasies of a particular fold’s subjects – it generalized better to unseen subjects in each fold, indicating true added value.

### 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 (no change) |
| SVM (RBF) | –1.3pp (baseline), –1.4pp (extended) |

**Interpretation:** The moderate imbalance in Dataset A (roughly 57% HC, 43% PD in our subject cohort) is not severe enough to substantially degrade unweighted classifiers. Weighting becomes more critical when: - Imbalance exceeds ~70:30 (either class dominating), - The minority class has a high cost of misclassification (e.g. missing a PD case in a screening test), - The overall sample size is very small (where every minority example is precious for learning).

In our case, PD and HC numbers were fairly close, and models like RF already achieved good sensitivity without weighting (RF unweighted recall for PD was ~0.76 with extended features). For Dataset B, where imbalance was reversed (25% HC minority), we saw that weighting would likely have a larger effect only if one's priority was catching the minority (HC), which is not typical in a PD screening context (usually PD is the class of interest). Indeed, for Dataset B some studies might weight to improve specificity, but that's a different use-case.

Empirically, we found that weighting *slightly* improved RF's ability to detect PD in the baseline-feature scenario (raising PD recall from ~0.64 to ~0.70, hence the AUC went up), but once features were extended and the model was stronger, weighting provided no further benefit and even caused a minor performance regression (likely because it started over-predicting PD in some folds that didn't need it).

For SVM, the variability in performance across folds meant any benefit of weighting was lost in noise – in fact, weighted SVM did a bit worse, possibly because the default hyperparameters were not re-tuned for the new effective class distribution. This underscores that weighting is not a panacea; it might require re-tuning C and γ for optimal results, which we didn't do (to avoid overfitting on our small data).

Overall, the takeaway is that **class imbalance was not the primary hurdle** in our PD classification problem – feature quality and validation methodology were more impactful.

#### 7.2.3 Model Performance Hierarchy

Across all conditions, Random Forest consistently outperformed the other models:

```
Random Forest > Logistic Regression ≈ SVM (RBF)
```

**Interpretation:**

Random Forest's advantages for this task include: 1. **Ensemble averaging** reduces variance on small datasets. Each tree may overfit to some extent, but averaging 100 trees smooths out idiosyncratic decisions. This gave RF more stable performance across folds (std of AUC for RF was typically smaller than that for SVM). 2. **Feature importance** provides interpretability. We leveraged this to identify which features were most useful (see Section 7.4). RF could focus on, say, shimmer and MFCC std and down-weight less useful features automatically. In contrast, SVM used all features via the kernel, potentially including noisy ones (unless explicitly filtered). 3. **Non-linear decision boundaries** capture complex patterns. PD vs HC separation likely involves logical conditions (e.g. "if jitter is high AND F0 variability is low, then PD"), which a decision tree handles well. SVM can also capture non-linear boundaries with RBF, but without careful tuning it might not align with the actual data distribution in high dimension. 4. **Robustness** to irrelevant features through feature subsampling. RF at each split considers a random subset of features; this is beneficial when many features are uninformative or correlated (as in our extended set). It essentially performs built-in feature selection. SVM and LR would be more affected by noise features unless regularized strongly.

Logistic Regression, while linear, proved quite competitive when given enough features – its performance was stable (did not suffer extreme drops like SVM did in some folds). This suggests that the PD vs HC separation is at least partially linearly separable in the enriched feature space. LR's disadvantage is it can't capture interactions: for example, if only the *combination* of low HNR *and* high jitter is a strong PD indicator, LR would need a feature representing that combination, whereas RF can branch on one then the other. The fact that LR's AUC stuck around ~0.78 and didn't improve with extended features indicates some interactions in extended features were key (which RF exploited to reach ~0.86).

SVM's underperformance here, relative to RF, might be due to the small sample + high feature regime being tricky for SVM without tuning or feature selection. SVM essentially attempted to maximize margin in a very high-dimensional space – with default parameters, it may have underfit (a wide margin that misclassifies some points, perhaps explaining lower AUC in some folds). Or, in other folds, it might have overfit (if a few support vectors ended up memorizing a quirk). The instability in SVM's results (some fold AUCs very low) points to high variance – perhaps needing nested CV to tune C, which we avoided due to small n. RF, by contrast, is more forgiving, as the ensemble can generalize well without much tuning (we set max_depth=10 somewhat arbitrarily, and it worked fine).

In literature, SVM was historically top on the smaller UCI dataset (195 samples, 22 features – where it achieved ~91% accuracy). But in our scenario with many features, RF's capability to handle feature redundancy and noise shines. Notably, a similar finding is reported in other bioinformatics contexts where tree ensembles often outperform SVM when features >> samples, unless SVM is carefully tuned or features are pre-selected.

## 7.3 Comparison with Literature

### 7.3.1 Performance Context

To contextualize our results, consider a few reference points from past studies:

| Study | Dataset | Best ROC-AUC | Method |
|---|---|---|---|
| Little et al. (2009) | UCI 22-feature | ~0.92 | SVM (10-fold CV) |
| Sakar et al. (2013) | Custom (multiple tasks) | ~0.86 (accuracy) | SVM (10-fold CV) |
| **This thesis** | **MDVR-KCL** (Dataset A) | **0.87** | **Random Forest** |

Our results are competitive with the literature, though direct comparison is limited due to: - Different datasets and features (e.g. our Dataset A vs Little's sustained vowels vs Sakar's mixed tasks). - Different validation strategies (many prior studies did not use grouped CV; some might have inadvertently allowed same-subject training/testing). - Different sample sizes (Little: 31 subjects; Sakar: 80 subjects; ours: 37 subjects).

For instance, Little et al.'s 0.92 AUC was on a dataset of sustained vowels in a controlled setting, with subject recordings likely mixed in CV. Our RF got 0.94 AUC on Dataset B which is similar sustained vowels with subject overlap, aligning with that – but on our Dataset A (more varied speech, proper grouping), we got 0.87. Sakar et al. reported ~86% accuracy on their dataset with an SVM, which is in the same ballpark as our RF's ~82–83% accuracy on Dataset A (taking into account their figure was accuracy, not AUC). This suggests that our methodology did not dramatically underperform conventional approaches – if anything, we demonstrated that with methodological rigor, one can achieve similar performance but with more confidence in generalization.

It’s worth noting that had we evaluated our models in a less strict manner (say, random splitting recordings), we likely would have seen inflated performance akin to 90%+ accuracies, which many papers tout. The fact that we enforce grouped CV means our numbers (e.g. 82.6% accuracy, 0.873 AUC for RF extended) are more conservative but more realistic.

### 7.3.2 Methodological Comparison

We highlight differences between typical literature approaches and ours:

| Aspect | Typical Literature | This Thesis |
|---|---|---|
| CV Strategy | Random split (record-level) | Grouped stratified (subject-level) |
| Subject handling | Often ignored (leakage possible) | Explicit grouping (no leakage) |
| Feature selection | Ad-hoc (sometimes none or simple filter) | Systematic ablation (baseline vs extended compared) |
| Reporting | Often only best result reported | All conditions (models, features, weighting) reported with variance |
| Emphasis | Max accuracy claims | Reproducibility and interpretability (feature importance, consistent CV) |

Our grouped CV approach provides **more conservative** but **more realistic** estimates of generalization performance. It guards against overly optimistic results that can arise from subject overlap. For example, in our experiments, treating each recording independently (if we had done so) would have made Task A performance approach Task B’s – but at the cost of realism. By showing the gap between rigorous and non-rigorous evaluation, we underscore a crucial point for the field: reported accuracies in the 90–99% range need scrutiny regarding evaluation methodology.

Additionally, by analyzing feature importance and including full result distributions, we aim to contribute transparency. Many previous works might report “we got X% accuracy with method Y,” but not the variability or the fact that a different split could yield a very different result. We found standard deviations of ~0.15 on AUC in some cases – meaning results could fluctuate by ±15 points just by fold composition. Literature rarely discusses such uncertainty, which could mislead one to think the method is consistently 95% accurate, rather than 80–95 depending on who is in test set.

Our approach to feature set comparison is also relatively novel. Instead of building a complicated hybrid model or doing extensive feature elimination, we framed it as an **ablation study** – demonstrating the value of adding certain feature groups. This kind of controlled experiment is important to understand where gains come from (rather than just throwing everything in and crediting the classifier).

In summary, our rigorous methodology prioritizes valid evaluation and interpretability, which we believe is essential for a field moving toward clinical translation. The downside is we didn’t squeeze out the absolute highest single-number metric – but the upside is confidence that our findings would hold on new data. This addresses a known concern in ML for healthcare: reproducibility and generalizability often lag behind headline performance in early studies.

## 7.4 Feature Importance Analysis

*(Detailed figures and tables for feature importance are provided in Appendix A. Here we discuss key insights.)*

### 7.4.1 Most Discriminative Features

The top features across models consistently include:

| Feature | Category | Relevance to PD |
|---|---|---|
| *f0_max* | Pitch | Reduced pitch range in PD (low max f0) [9] [10] |
| *delta_mfcc_2_mean* | Spectral dynamics | Temporal variability of formants |
| *autocorr_harmonicity* | Voice quality | Breathiness indicator (low in PD) |
| *shimmer_apq3* | Perturbation | Amplitude instability (high in PD) |
| *intensity_mean* | Prosody | Overall loudness (lower in PD on avg) |

These appeared in top-10 lists for Random Forest and/or Logistic Regression on both tasks [11] [12]. For instance, Random Forest ranked *f0_max* as the #1 feature for ReadText (importance 0.052) and also high for Spontaneous [9] [13], aligning with the observation that PD patients often have a lower maximum pitch (they don’t reach high pitches even when reading emotionally or in emphasis). Likewise, shimmer measures (like APQ3) featured prominently (e.g. #7 in ReadText RF importance [14], #2 in Spontaneous RF [15]), corroborating that vocal amplitude micro-variability is a strong PD marker (due to glottal inefficiency).

Interestingly, *delta_mfcc_2_mean* (the mean of delta-MFCC coefficient 2) was consistently important (RF rank #2 in ReadText, #5 in Spontaneous [11] [16]). MFCC_2 roughly correlates with formant spacing – a delta might capture how the first formant changes, perhaps reflecting articulation stability. PD speech might have less dynamic formant movement (due to monotony or slow articulation), so this feature being discriminative aligns with less variation being indicative of PD.

Harmonicity (autocorrelation-based) indicates periodicity of voice. Lower harmonicity (more noise) is expected in PD due to breathy/hoarse voice. Indeed, *autocorr_harmonicity* was in top 5 for ReadText RF [17] and in logistic regression’s top features. HNR was similarly valued in LR’s coefficients (e.g. HNR had a high positive weight meaning higher HNR (more harmonic, typical of healthy) pushes toward HC class).

Intensity mean being important likely ties to hypophonia – PD speakers often speak more quietly. But intensity can vary with recording conditions, so I suspect it was somewhat confounded (perhaps the dataset controlled recording setup, so lower intensity is genuine PD effect rather than mic distance). It showing up (RF rank #5) suggests PD voices indeed had lower overall energy.

In summary, the features our models found most informative align well with known clinical markers of PD speech: reduced pitch range, increased jitter/shimmer (especially shimmer APQ measures), increased noise (low harmonicity), and reduced loudness. It’s reassuring that the ML models “rediscovered” these – it lends face validity and also suggests that a focus on these features in simpler models might achieve nearly comparable results.

#### 7.4.2 Feature Category Contributions

We aggregated feature importances by category in Appendix A and found:

- **MFCC features** (cepstral coefficients and their derivatives) contribute the most to model performance overall. They dominated top ranks consistently. This highlights that spectral envelope characteristics (timbre) are highly affected by PD (perhaps due to changes in articulation and phonation).
- **Pitch features (F0)** are consistently important (ranked second overall category). This underscores monotony as a key separability factor – even simple features like std(F0) or max(F0) had impact.
- **Formant variability** (like F1_std, F2_std) shows moderate importance. In Random Forest, F1_std and F2_std had non-trivial importance (~rank 9 in ReadText RF) [18] . They reflect how clearly vowels are articulated; PD often reduces vowel space (so less formant variability).
- **Harmonicity and Shimmer** categories also appeared, though behind MFCC and pitch. Shimmer was more influential than jitter in our models (shimmer APQ3, APQ5 repeatedly out-ranked jitter local/RAP). Possibly because amplitude fluctuations (shimmer) were more consistently measurable across varied speech tasks than period fluctuations (jitter) which are classically measured on sustained phonation. In running speech, jitter might be harder to measure reliably.

From **Figure 7.1** (aggregated importance by category for ReadText) and **Figure 7.2** (for Spontaneous), we saw MFCC-based features contributing the largest portion of total importance (~40–50%), with Pitch features next (~20%), then Shimmer and Formant features splitting the remainder, etc. This quantification suggests that while prosodic and perturbation features are useful, the bulk of discriminative power in our extended set came from spectral/cepstral domain.

The analysis reveals: - **MFCC (spectral envelope)** is the #1 category. This implies that changes in the vocal tract output (due to slurred or reduced articulation and resonance changes) are the strongest indicators in our data. - **Pitch (F0)** being #2 category is logical, given PD's monotone speech trait. - **Shimmer (amplitude stability)** ranking #3 overall, slightly above **Delta MFCC** category in aggregated importance, is interesting – it implies the cycle-to-cycle amplitude perturbations are slightly more telling than the delta MFCC patterns for PD, within our features. Jitter was a bit lower (maybe rank 5 or 6 category). - **Harmonicity** ends up lower partly because we only had 2 features in that category, but those two were moderately weighted.

The analysis reveals what combination of characteristics yields the best differentiation: - Reduced spectral variability (MFCC std low, delta MFCC low), - Abnormal prosody (flat pitch, low loudness), - Unstable voice (high shimmer, low harmonicity).

These align well with the multi-faceted nature of hypokinetic dysarthria: PD affects phonation (breathiness -> harmonicity), articulation (vowel centralization -> MFCC patterns), and prosody (monotony -> F0, intensity changes).

#### 7.4.3 Cross-Task Stability

Comparing ReadText and SpontaneousDialogue tasks: - **Feature rankings are moderately consistent across tasks**, suggesting that the acoustic signatures of PD are **task-general** rather than task-specific.

For example, the top features for RF in both tasks had overlap: MFCCs, shimmer APQ, etc., albeit with some reordering. In Appendix A Figure A.3 vs A.7, we see many checkmarks for features appearing in top-10 for both tasks. Concretely, *f0_mean*, *delta_mfcc_2_mean*, *autocorr_harmonicity*, and a shimmer measure appeared in top lists for both ReadText and Spontaneous (as shown in A.5.1).

This cross-task consistency implies that our features are capturing fundamental voice characteristics of PD that manifest regardless of whether the person is reading or speaking freely. That is encouraging: it means a model trained on one type of speech could potentially work on another to some extent (though absolute performance differed, as we saw). It also justifies using a combined feature set for both tasks rather than entirely different features.

There were some differences too (e.g. intensity_min was top-10 in Spontaneous logistic regression features but not in ReadText, possibly because in spontaneous speech some PD patients trailed off at end of utterances leading to lower minima, whereas in reading a standardized text that might not happen as much). However, overall patterns persisted.

The practical implication is that a single system could possibly handle multiple speaking tasks as input, since the core features to monitor are similar. However, it also suggests diminishing returns in doing multi-task fusion (we did not explicitly combine tasks in training due to our CV strategy, but one could imagine training one model for all tasks vs separate per task – our analysis indicates the same features work for both).

## 7.5 Implications

### 7.5.1 For Feature Engineering

The success of extended features suggests that future work should: 1. **Include variability measures** (std, range) alongside means. Our results show that not just the average pitch or intensity matters, but how much they vary is crucial. So any PD voice feature set should incorporate measures of variability (pitch variability, intensity variability, formant variability). 2. **Capture temporal dynamics** beyond simple deltas. We added delta-deltas (acceleration) which helped somewhat. Future feature exploration might include measures of rhythm or timing patterns (e.g. speaking rate consistency, pause distribution) as additional dynamic features, given PD often affects timing. 3. **Provide spectral shape descriptors** that are not encapsulated by MFCCs. We added centroid, flatness, etc., which contributed unique info (like flatness for noise). The inclusion of such features, plus possibly others like spectral entropy or harmonics-to-noise ratio across different bands, could further improve detection.

Our controlled experiment highlights that **hand-crafted features still hold value**. Even though one might consider using end-to-end learning, classical features are interpretable and, as we saw, quite effective. By systematically expanding the feature set, we significantly improved model performance. It stands to reason that further expanding with other known features (e.g. non-linear vocal fold dynamics features used by Little 2007 like DFA or RPDE, or additional prosodic features like pause ratio) could capture remaining variance.

That said, more features is not always better without careful validation. We added 31 features and saw improvement with robust CV, indicating they were indeed useful. If adding some feature doesn't yield consistent fold-by-fold improvement, it might be spurious. So iterative feature engineering with CV feedback is recommended to avoid cluttering with noise.

### 7.5.2 For Model Selection

Random Forest is recommended for similar tasks due to: - **Robustness on small datasets:** as shown, RF handled 37 subjects with 78 features and still generalized better than SVM or overly simple LR. - **Built-in feature importance:** aiding interpretability, which is valuable for convincing clinicians of what the model is "looking at" (pitch, shimmer, etc., which relate to clinical concepts). - **Good handling of mixed feature types:** our features vary in scale and distribution (e.g. some are percentages, some frequencies in Hz). RF is invariant to monotonic transformations and not sensitive to feature scaling or distribution assumptions, whereas distance-based methods like SVM can be. - **Tuning simplicity:** we barely tuned RF (only depth limited) and got great results. SVM would have needed careful grid search to likely reach parity.

However, logistic regression shouldn't be dismissed. Its performance was not far off, and it offers maximum transparency (coefficients directly mapping to risk factors). One could envision using LR with a selected subset of the most important features as a baseline screening tool – it might yield slightly lower accuracy but easier clinical adoption.

If pursuing SVM or other non-linear models, one should invest in hyperparameter tuning or use techniques like cross-validation selection or even automatic methods (Gaussian Processes, etc.) to set kernel parameters. Otherwise, as we saw, default SVM might underperform.

Another consideration is ensembles or stacking: we only tried individual models. Possibly a simple ensemble of LR and RF (for example) could leverage LR's precision and RF's recall to boost F1 marginally. But given RF alone was strong, the gain might be small.

One might ask: what about deep learning? Our results show that classical ML can reach ~87% AUC with good features on limited data. A neural network might not outperform this without much more data (and would be less interpretable). In literature, some tried deep belief networks or CNNs on voice and got varied results; it often requires augmentations and large datasets to beat classical methods in this domain. So, our findings reinforce the viability of classical approaches as both effective and interpretable for PD voice classification at this stage.

### 7.5.3 For Evaluation Protocols

Grouped cross-validation should be **mandatory** when: - Multiple recordings exist per subject (as in most voice datasets). - Subject identifiers are available (which they should be whenever ethically possible). - Generalization to new subjects is the goal (which is usually the case in a screening scenario – you want the model to work on new people).

Our work clearly demonstrates how much optimistic bias can result from ignoring this. In Dataset B, if one mistakenly treated all 756 samples as independent, one might claim ~94% AUC (like a near-perfect test). But our Dataset A and discussion highlight that a more realistic evaluation yields ~87% AUC – a meaningful difference if we're talking about deploying a system (the former might imply only 6% error rate vs ~13% error rate in the latter; in screening large populations that difference is non-trivial).

Therefore, any PD voice study should incorporate subject-level CV or use an external test set where subjects are wholly separate from training. Without this, reported performance is suspect. This is not just theory: there are published works that later had to be tempered because they inadvertently exploited speaker-specific traits.

We also advise reporting variance (e.g. fold std) or confidence intervals. We found, for instance, RF extended had 0.873 ± 0.137 AUC – meaning in some folds it was as low as ~0.74 or as high as ~1.0. Knowing this spread is important for understanding model reliability. A single average can be misleadingly high if variance is large.

Finally, our approach of analyzing by subtask (read vs spontaneous) suggests that if a study uses only read speech or only monologue, results might differ. Ideally, one would test their model on multiple speech tasks to ensure it's capturing PD markers and not task-specific artifacts. For example, a model trained only on sustained vowels might do poorly on running speech without adaptation. In our case, we trained and tested within each task, but future work might consider training on one task and testing on another to truly examine generalization (our stable feature importance across tasks hints there might be generalization potential, but we haven't proven it by cross-task validation).

In summary, rigorous evaluation protocols (grouped CV, reporting uncertainty, testing across speech tasks) are crucial for progress in this field. We believe our thesis contributes an example of such rigor, hopefully encouraging its adoption.

# Chapter 8: Limitations and Threats to Validity

## 8.1 Overview

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

## 8.2 Sample Size Limitations

### 8.2.1 Dataset A: Small Subject Pool

| Metric | Value |
|---|---|
| Total subjects | 37 |
| Subjects per test fold (5-fold CV) | ~7 (one fold has 8) |
| PD subjects (minority) in dataset | 16 (approximately) |

**Implications:** - High variance in fold-level metrics (std > 0.15 for AUC and accuracy was common). With only ~7 PD and ~7 HC in each test fold, performance swings depending on which individuals fall in that fold. - Limited statistical power for detecting small effects. For example, a ~2% improvement due to a tweak might not be reliably observable given the noise – thus some potentially helpful modifications might go unnoticed. - Results may not generalize to broader populations. With 37 subjects from a specific cohort, we risk modeling quirks of that cohort (e.g. all our PD speakers might have similar severity, dialect, etc.). A different set of PD patients could yield somewhat different outcomes.

#### 8.2.2 Effect on Statistical Confidence

With 37 subjects and 5-fold CV: - Each fold has only ~7 test subjects, as noted. **A single misclassification shifts accuracy by ~14%** (1/7 ≈ 14%). Thus one subject being atypical can swing fold accuracy drastically. - Confidence intervals are wide by design. For instance, our Random Forest extended AUC 95% CI (using ±2 std/√n_folds) still spans tens of points due to high fold variance. So we cannot claim, say, "AUC is 0.873 significantly above 0.80" with high confidence. - This also affected comparing models: differences needed to be relatively large to be sure one model outperformed another consistently. For instance, SVM vs LR differences were within the variability range, so we had to interpret with caution.

**Mitigation:** Results focus on **relative comparisons** rather than absolute performance claims. We emphasize trends (extended > baseline, RF > LR ≈ SVM) consistent across folds rather than the exact numeric values. We also report std devs to remind readers of uncertainty.

However, ultimately the small N is a limitation we cannot fully overcome; future validation on larger cohorts is needed to firm up the findings.

### 8.3 Subject Identifier Limitations

#### 8.3.1 Dataset B: Missing Subject IDs

Dataset B (PD_SPEECH) provides no subject identifiers. This creates potential for: - **Subject leakage:** The same subject's recordings can appear in both training and test sets in CV, artificially inflating performance (model effectively "recognizes" the person rather than the disease). As discussed, this likely occurred in our cross-dataset comparison, where dataset B's high AUC partly stems from this issue. - **Optimistic bias:** Because of leakage, reported metrics (accuracy ~90%+, AUC ~0.94) are overly optimistic estimates of how the model would perform on truly independent subjects. - **Unknown generalization:** We can't assess how well models trained on Dataset B generalize to new individuals, since any CV mixing confounds that. The absence of IDs means one can only assume independence, which is a faulty assumption here (since we know there are 3 samples per person, albeit unlabeled as such).

**Caveat Statement:**

> *"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."*

We explicitly add such warnings in reporting any result that comes from Dataset B to avoid misinterpretation.

#### 8.3.2 Comparison Limitations

Direct comparison between Dataset A (grouped CV) and Dataset B (standard CV) is confounded by: - Different CV strategies (grouped vs not). - Different feature dimensionalities (78 vs 754). - Different sample sizes (37 vs 252 subjects).

Therefore, while we noted performance differences, attributing them solely to e.g. subject leakage vs feature count is not straightforward. A more controlled experiment would be needed (e.g. simulate subject splits within B, or reduce features in B to our 78 to isolate effect of CV vs features).

We caution against over-interpreting the cross-dataset gap as purely due to leakage; part of it is likely because Dataset B's feature set truly has more information (tqwt features etc.). Conversely, part is because some of that info is inadvertently personal.

Thus, one limitation is we did not obtain a dataset where we could apply grouped CV with 750+ features to see if AUC still remains ~0.94 or drops – that would have clarified this better. We were limited to available data as-is.

## 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:** There may be other acoustic features (e.g. particular frequency band energies, advanced nonlinear metrics) more discriminative than some we included. Because we stuck to interpretable, known features, we might have missed performance gains from exotic features. - **Fixed parameters:** We used Librosa/Praat default settings for window length, etc., without tuning. For example, MFCC count = 13 is standard, but maybe more coefficients could help. We didn't exhaustively optimize these parameters. - **No feature selection:** All 78 extended features were used regardless of correlation or redundancy. This could introduce noise. Perhaps a subset of ~10 highly informative features could achieve similar performance with less overfitting risk.

These choices were conscious (to avoid tailoring to this specific dataset too much), but it means our feature set might not be the absolute best possible.

### 8.4.2 Audio Quality Assumptions

Feature extraction assumes: - Reasonable signal-to-noise ratio. If recordings had heavy background noise, features like jitter/shimmer could be invalid. We assume our dataset's recordings are relatively clean (since from a clinical study), but in real-world conditions (home recordings via phone) noise might violate this assumption. - Consistent recording conditions. We assume similar microphone type, positioning, etc., across subjects. If some PD patients recorded differently than controls (say, different distances), features like intensity differences could be confounded by that rather than pathology. - No severe clipping or distortion. Our pipeline would not detect if audio was clipped. Clipping could artificially change features (e.g. reducing measured shimmer). We assume original data was monitored for quality.

The MDVR-KCL dataset's smartphone recordings may violate some conditions (different phones, etc.). We tried to mitigate by normalizing amplitude and trimming silences, but device frequency response differences could still affect spectral features (like MFCCs). We did not explicitly calibrate for device differences (if metadata existed).

## 8.5 Model Limitations

### 8.5.1 No Hyperparameter Tuning

All models used default or fixed hyperparameters:

| Model | Fixed Parameters |
|---|---|
| Logistic Regression | C=1.0, max_iter=1000 (no tuning) |
| SVM (RBF) | C=1.0, gamma='scale' (defaults) |
| Random Forest | n_estimators=100, max_depth=10 |

**Implications:** - Performance may be suboptimal. For example, SVM likely would improve with a different C or gamma; we saw it underperform, perhaps partly due to no tuning. - Results represent lower bounds in a sense. If we did tune (especially via nested CV), we might squeeze out higher AUC or accuracy. However, nested CV would further tax the small sample size and might overfit to folds. - Tuned models might change rankings. It's possible a well-tuned SVM could match or beat RF. Our conclusion that RF > SVM holds for default SVM; a fairer fight would tune SVM, which we omitted for aforementioned reasons.

We rationalized not tuning to preserve an un-biased evaluation on small data. Nonetheless, this is a limitation: our results do not guarantee that RF is inherently superior to an optimally tuned SVM on this task – just that in a reasonably simple configuration, RF performed best.

**Rationale for not tuning:** Nested CV on 37 subjects would lead to extreme variance and risk of overfitting the tuning process (the inner folds would be even smaller). We opted for simpler, robust settings (like an RF not too deep, an SVM not too aggressive) to ensure generality, at the cost of some accuracy.

### 8.5.2 Classical ML Only

This thesis explicitly excludes deep learning. Potential missed opportunities: - **End-to-end learning from spectrograms:** A CNN could, in theory, automatically learn discriminative features (maybe capturing subtle patterns our handcrafted features miss). Some research suggests spectrogram-based CNNs can equal classical methods when lots of data are available. - **Transfer learning from pre-trained speech models:** e.g. using `wav2vec` or other self-supervised models. We did not explore this; such models might have extracted features that correlate with PD (like vocal tract shape signatures) beyond our set. - **Attention mechanisms for temporal modeling:** A model like an LSTM or transformer that looks at the full time-series with attention might pick up longitudinal patterns (e.g. progressive slowing or fatigue within an utterance) that our summary features don't capture.

**Rationale:** Deep learning typically requires larger datasets (which we lack) and offers reduced interpretability – conflicting with our aim of interpretable, reproducible analysis. Prior works that tried deep nets on small PD datasets often risk overfitting or ended up not much better than classical methods. We prioritized reliable performance over potentially higher but less trustworthy performance.

Still, missing out on deep learning means we can't claim the absolute ceiling of performance was reached. It's possible a sophisticated model could, say, hit 95% AUC on Dataset A if it cleverly learned patterns – we just don't know from our study.

## 8.6 Methodological Limitations

### 8.6.1 No External Validation

All results use internal cross-validation. Limitations: - No held-out test set from a different source. We did not get to test our model on, say, an external dataset (like training on MDVR-KCL, testing on another voice dataset). Thus generalization across datasets (with different demographics or recording setups) remains unverified. - No multi-site validation. Our data likely come from a single site (KCL for dataset A, Istanbul for dataset B). PD voice characteristics might slightly vary by population (due to language, etc.). Without multi-site data, we can't ensure our model is robust to those differences. - Generalization to clinical settings unknown. We validated via cross-val, which, even grouped, may not capture all sources of variation present in real clinic or telemonitoring scenarios.

### 8.6.2 Binary Classification Only

The task is limited to PD vs HC classification. Not addressed: - Disease severity prediction (e.g. mild vs moderate vs severe PD). It's possible certain features correlate with severity (like more severe PD -> more monotone, etc.). Our model doesn't quantify severity. - Progression monitoring (tracking changes in a patient's voice over time). We just classify at one time point. - Differential diagnosis (PD vs other disorders that affect voice, like atypical parkinsonian syndromes or vocal pathologies). We only distinguished from healthy; a real screening test should also ideally distinguish PD from, say, stroke-related dysarthria, etc.

### 8.6.3 Single Speech Tasks

Each task analyzed separately. Not addressed: - Task fusion strategies: In a real evaluation, a patient might do multiple tasks (sustain vowel, read passage, etc.). We didn't investigate combining features from multiple tasks to improve accuracy. Perhaps using both tasks simultaneously could improve confidence (if model agrees on both tasks). - Multi-task learning: We trained separate models per task in CV. We didn't utilize the fact that we had paired ReadText and Spontaneous from each subject to do, say, a multi-output model that learns PD vs HC across both tasks jointly. Such an approach might leverage commonalities and improve robustness. - Optimal task selection: We didn't determine if one task is intrinsically better for PD detection than the other beyond noting performance differences. For instance, if Spontaneous is only marginally better, maybe the effort of recording spontaneous speech (which is less controlled) isn't needed if reading a passage yields nearly as good results. We haven't formally addressed which task provides more signal when controlling other factors.

## 8.7 Threats to Validity

We consider threats in terms of internal, external, and construct validity:

### 8.7.1 Internal Validity

| Threat | Status | Mitigation |
|---|---|---|
| Subject leakage | Controlled (Dataset A) | Grouped CV (no subject repeats in train/test) |
| Label noise | Unknown | Assumed correct (we trust the provided diagnoses; any mislabeled PD or HC would hurt performance in unpredictable ways) |
| Feature bugs | Possible | Unit tests, manual verification of extraction on a few samples (we spot-checked that features like F0 mean indeed reflected pitch visually from spectrogram) |

For feature bugs: we wrote our extraction code carefully and tested on example audio (comparing Praat outputs). But with many features, there's a chance of error (e.g. mis-scaling or using a wrong parameter). If a bug existed, it could affect results. We did not find evidence of glaring bugs (our intermediate checks like comparing computed jitter with Praat were within tolerance).

Label noise: if some "HC" in the dataset actually had undiagnosed PD (or vice versa a PD was misdiagnosed), that would confuse training. With clinical data, diagnostic confirmation is usually good, but it's a possibility. We had no way to detect that, so we assume labels are correct.

### 8.7.2 External Validity

| Threat | Status | Mitigation |
|---|---|---|
| Population bias | Likely | Document dataset demographics (if available). Our subjects might not represent all PD (e.g. maybe mostly moderate severity, English speakers). We acknowledge this and don't overclaim generality. |
| Recording variability | Present | Standardized extraction (we normalized amplitude, etc., to reduce device differences) |
| Temporal stability | Unknown | Single recording session used (we don't know if these voice markers fluctuate daily or with medication – our model sees one snapshot) |

Population bias: For example, if our PD subjects were all relatively young onset, their speech impairments might be milder than typical; a model trained on them might not generalize to older, more severely affected patients. We note this as a limitation. Ideally, multi-cohort training would address it.

Recording differences: We tried to mitigate by normalization, but subtle differences (like smartphone mic frequency response) we couldn't fully correct. This threatens how well the model would do on different hardware. We might mitigate by recommending using similar devices as in training or using calibration phrases, etc., but that's future work.

Temporal (in)stability: If a patient's voice improves after meds (levodopa) significantly, a recording at "on" vs "off" state could yield different classification results. Since we only had one recording per subject presumably at a similar condition, our model doesn't account for within-subject variability over time. This is a threat to using it in monitoring or even in screening at different times of day. We didn’t explicitly handle that beyond including jitter/shimmer which reflect momentary stability.

### 8.7.3 Construct Validity

| Threat | Status | Mitigation |
|---|---|---|
| Feature relevance | Assumed | Literature-based selection (we assumed features chosen do measure what we intend: e.g. shimmer reflects vocal stability, which relates to PD) |
| Metric appropriateness | Addressed | Multiple metrics reported (accuracy, precision, recall, F1, AUC) to give a full picture. Our primary focus on AUC is justified for class imbalance. |
| Class definition | Accepted | Binary PD/HC from source (we trust that the diagnostic criteria for PD were applied similarly across subjects; "PD" is a somewhat broad category including various severities, but we didn’t refine it) |

Construct validity wise: - One could argue whether acoustic differences are a direct proxy for PD. They are, but moderated by other factors (like age affects voice too). If our HC were younger on average than PD, features might partly capture age differences, not just disease (construct confounding). We don’t have details on matching of groups. So age or gender differences could confound some feature differences (e.g. PD group might have more males which have lower pitch, etc.). Ideally, groups are matched, but if not, our model might partly be picking up demographic differences. We didn’t explicitly control for that – that’s a threat to internal and construct validity.

- The metric selection: we emphasized AUC because we didn't want to choose a threshold and because class distribution might not reflect prevalence. That’s valid. But if one’s use-case cares about specific sensitivity or precision, we might not have directly optimized that. We show precision/recall though so one can see trade-offs.

- "PD vs HC" as a construct: We assume PD is a binary state (presence vs absence) which is fine. However, PD itself is heterogeneous. Two PD patients can have different speech symptom severity. Our model treats them the same class, effectively focusing on the average differences vs controls. It might not generalize to extremely mild PD that has almost normal speech (our PD sample likely had noticeable speech issues, else they wouldn't be in a voice study). So the construct of PD voice pathology severity isn't accounted for, which can limit where on the spectrum the model works. That falls under external validity too.

## 8.8 Reproducibility Considerations

### 8.8.1 Strengths

We took several steps to enhance reproducibility: - Fixed random seeds (42) for all stochastic processes (RF bootstrapping, etc.) so that results can be exactly reproduced given the same data splits. - Version-controlled code available (with documentation in our repository) – all data processing and training scripts are provided, allowing others to rerun or inspect steps. - CLI-based pipeline for feature extraction and experiments (ensures environment independence – one can run `pvc-extract`, `pvc-experiment` with provided settings and get the same outputs). - Documented dependencies (we list library versions for Librosa, scikit-learn, etc., so that subtle differences in implementations don't change results).

These measures mean someone else with the dataset should be able to replicate the core results. Indeed, we include unit tests and sample outputs to verify consistency.

#### 8.8.2 Limitations

However, some reproducibility constraints remain: - Library version drift may affect results. E.g., future versions of Praat or Librosa might compute jitter slightly differently. Our results are tied to versions used in 2026. We encapsulated our environment details in the documentation to help, but cannot guarantee absolute identical results if libraries change (though likely small differences). - Hardware differences in audio processing: If someone else tries to replicate feature extraction on differently formatted audio or via a different audio I/O library, minor numeric differences might occur (we saw tiny differences using Praat vs Parselmouth, for instance). - Dataset access may change. We used MDVR-KCL via Zenodo. If that dataset is updated or if audio files get re-encoded differently, results could shift. Also, Kaggle dataset might be updated or removed. We archived what we used, but future attempts might face data availability issues.

We've tried to offset these by packaging intermediate results (like listing out exactly the features we extracted and their definitions so others can ensure they compute the same).

### 8.9 Interpretation Guidelines

Given the limitations, results should be interpreted as follows:

#### 8.9.1 Appropriate Claims

"Extended features improved ROC-AUC on this dataset" – We have clear evidence that adding features X, Y, Z improved performance on our data. This is a specific, supported claim.

"Random Forest outperformed other models under these conditions" – Within our experimental setup, yes. We can claim RF was better than LR and SVM on MDVR-KCL data with our feature set.

"Grouped CV provides more conservative estimates than random splits" – We demonstrated that pretty convincingly by comparing Dataset A vs B outcomes. This is an important methodological point we can generalize (and indeed matches theory and other domains' findings).

These claims stick to what we directly tested and observed.

#### 8.9.2 Inappropriate Claims

"This system diagnoses Parkinson's Disease" – Our model, while promising, is far from a deployable diagnostic tool at this stage. We tested on 37 PD vs HC voices in lab conditions. Clinical diagnosis involves more than voice, and we haven't validated prospectively or against medical standards. Presenting it as a diagnostic could be misleading and irresponsible given current state.

“82.6% accuracy is clinically sufficient” – That’s an overclaim. A screening test might need different operating points (like very high sensitivity at some cost of specificity). And 82.6% accuracy might not translate to the same in a broader pop or early PD detection. We can’t assert sufficiency or readiness for clinical adoption from our limited retrospective study.

“These results will generalize to other populations” – as noted, our sample is specific (likely English speakers, certain recording conditions, moderate PD). Without testing on e.g. non-English speakers or those with comorbid voice issues, we cannot generalize. So we should not claim broad generality beyond our tested scenario.

In summary, while our research supports the feasibility of voice-based PD classification with classical ML, it should be seen as a *methodological demonstration* rather than a ready-to-use tool. The findings contribute to the understanding of what aspects of voice are most indicative of PD and how to evaluate models rigorously, rather than clinching an ultimate solution.

### 8.10 Future Work to Address Limitations

We outline potential solutions corresponding to the identified limitations:

| Limitation | Potential Solution |
|---|---|
| Small sample size | Multi-site data collection (increase N, diversity) |
| Missing subject IDs in dataset B | Insist on subject-labeled data; or require dataset creators to provide grouping info for proper CV |
| No hyperparameter tuning | Bayesian optimization or nested CV (with caution) to find optimal model settings |
| No external validation | Independent test cohort from different source (could be part of a challenge or collaboration) |
| Classical ML only | Carefully introduce deep learning with data augmentation or transfer learning (but verify interpretability and performance gain) |

Each of these points to clear next steps: - We are particularly keen on obtaining more data. One idea is combining multiple existing voice datasets (if comparable) to train a more general model, followed by external validation on an entirely separate set. - The subject ID issue is something we as a community should address – perhaps by advocating dataset releases to always include that information or by developing algorithms that can partially infer it (some have tried clustering by speaker in absence of IDs). - For hyperparameters, one could employ automated tools like AutoML on a held-out subset or synthetic data to get better values, then apply them. The risk is always overfitting to limited data, so techniques like leave-one-subject-out might help gauge if a tuned model still generalizes. - Testing on other languages: voice markers might differ cross-linguistically (tone languages vs non-tone, etc.). We could extend research by acquiring PD speech in other languages to see if the same features apply or if new ones are needed (like tonal range for tonal languages). - Another future direction: incremental learning or domain adaptation. Perhaps a model trained on one hospital’s data could adapt to another’s using a small calibration sample, since acoustic conditions might differ. We didn’t explore that.

By addressing these, we move closer to a robust, generalized PD voice screening tool. Each limitation surmounted will bolster confidence in the method's utility.

# 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. Also designed a systematic 2×2 factorial experiment (features × class weighting) to assess each factor's impact. All conditions and results (mean ± std) are transparently reported, providing a clear view of model performance variability.
2. **Feature Engineering Investigation** – Extended the acoustic feature set from 47 to 78 features, incorporating measures of spectral variability and shape. Demonstrated an +8.7 percentage point ROC-AUC improvement due to this extension. Identified most discriminative features (e.g. F0 range, shimmer, harmonicity, MFCC dynamics) that align with known PD speech characteristics.
3. **Class Weighting Analysis** – Evaluated `class_weight="balanced"` across all models. Found modest effects on moderately imbalanced data (e.g. RF AUC +3.5pp with baseline features, negligible with extended). Documented the interaction between feature richness and class weighting efficacy.
4. **Reproducible Pipeline** – Developed a CLI-based, open-source pipeline for feature extraction and modeling. Fixed random seeds and provided extensive documentation, enabling replication. The code repository includes configuration for experiments and unit tests, ensuring that others can reproduce our results exactly.

## 9.2 Key Findings

### 9.2.1 Primary Results

| Finding | Evidence |
|---|---|
| Best ROC-AUC: **0.873 ± 0.137** | Random Forest, Extended Features (Dataset A) |
| Feature extension improves performance | +8.7pp ROC-AUC (baseline → extended) for RF; notable gains for SVM as well. |
| Random Forest outperforms other models | Highest ROC-AUC across all conditions; more stable folds. |
| Grouped CV is essential for realism | Prevented ~6–8pp AUC inflation seen in non-grouped scenario. |

### 9.2.2 Best Configuration

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

*(This corresponds to Dataset A results on the combined tasks.)*

#### 9.2.3 Feature Importance Insights

The most discriminative features for PD detection include: 1. **f0_max** — Maximum fundamental frequency (pitch ceiling). PD speakers exhibit a lower f0_max (limited pitch range) [9]. 2. **delta_mfcc_2_mean** — Spectral dynamics (changes in formant structure). Lower in PD (less vocal modulation). 3. **autocorr_harmonicity** — Voice periodicity (harmonics-to-noise). Lower in PD (breathy, hoarse voice). 4. **shimmer_apq3** — Amplitude perturbation (short-term). Higher in PD (unstable volume control). 5. **intensity_mean** — Overall vocal intensity. Lower in PD (hypophonia).

These align with known clinical manifestations of PD: reduced pitch variability, increased hoarseness/noise, and reduced loudness. Our model's reliance on these features reinforces their importance as objective biomarkers. It also suggests that interventions targeting these aspects (e.g. voice training to increase loudness and pitch range) are addressing the right features, which our classification finds salient.

### 9.3 Research Questions Answered

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

Classical ML achieves **ROC-AUC up to 0.873** with Random Forest on the MDVR-KCL dataset using grouped cross-validation. This demonstrates that voice-based PD detection is feasible with good accuracy, though not perfect. Performance varies by model: - Random Forest > SVM > Logistic Regression in our experiments, with RF providing the best combination of sensitivity and specificity. For example, RF (extended) attained ~82.6% ± 12.2% accuracy, whereas Logistic Regression was ~72.4% ± 13.6%. - All models show substantial variance across folds (±10–15%), indicating that consistent generalization requires more data or model averaging.

In summary, classical ML (particularly ensemble methods) can detect PD from voice with competitive accuracy, but there is notable uncertainty due to small sample size. We highlight that these results are under rigorous validation; less strict evaluation in past studies reported higher performance (often >90% accuracy), but our approach provides more realistic estimates.

#### RQ2: Does feature set extension improve classification performance?

**Yes.** Extending from 47 baseline features to 78 features improved ROC-AUC by **+8.7 percentage points** for Random Forest (from ~0.79 to ~0.87). SVM saw a similar relative boost (~+9.1pp). The improvement is most pronounced for the more challenging read speech task (+23pp AUC for RF on ReadText), and more modest for spontaneous speech (+3pp). Extended features captured additional information: - MFCC std and delta-delta features provided insights into intra-utterance variability and acceleration that baseline features missed. - Spectral shape descriptors (centroid, flatness, etc.) added detection power by quantifying voice quality changes.

These results clearly show that a **richer feature representation enhances PD vs HC separability**. In practical terms, using an extended acoustic feature set (including prosodic, perturbation, and spectral features) is recommended over a minimal set. The gains are especially significant for subjects with mild impairments where subtle features make the difference between PD and HC classification.

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

**Modestly, and only in certain conditions.** On Dataset A (57% HC, 43% PD): - Class weighting improved Random Forest's AUC by +3.5pp with baseline features (0.786 → 0.821), primarily by improving PD recall, but had no benefit (and slight detriment –1.4pp) with extended features. - Logistic Regression was essentially unaffected by weighting (no change in AUC, minor changes in precision/recall). - SVM's performance did not improve; if anything, it decreased slightly with weighting.

Thus, for moderate imbalance, class weighting is **not a game changer**. The unweighted models already handled the imbalance reasonably. Weighting becomes more critical if imbalance is severe (e.g. 75/25) or if missing minority instances is especially costly. In our case, given PD and HC counts were relatively close and our features were informative, weighting did not significantly elevate performance.

On Dataset B (25% HC minority), applying class weighting actually *decreased* overall accuracy slightly (since it forces more false-positive PD predictions to catch a few more HCs). This underscores that weighting is context-dependent: our aim was PD detection (treating PD as positive), so unweighted models already favored PD sensitivity.

**Conclusion:** Class weighting can be a useful tool to tweak sensitivity vs specificity, but it is not a substitute for good features or proper validation. In PD voice classification, ensuring no data leakage and having robust features contributed far more to performance than class weighting.

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

Dataset B (standard 5-fold CV, subject overlap) showed higher absolute performance than Dataset A (grouped CV): - E.g., Random Forest ~0.94 AUC on Dataset B vs ~0.87 on Dataset A, and 90% accuracy vs 82.6%. - Similarly, SVM and LR scored higher on Dataset B.

**Interpretation:** The higher scores on Dataset B are likely due to optimistic bias from subject overlap and the larger, more complex feature set. We caution that these results **do not reflect true generalization to new subjects**.

When comparing, we find that: - **Grouped CV provides more conservative (and likely more accurate) performance estimates.** Many literature results that used random splits (like Little et al.'s 91% accuracy) correspond more closely to our Dataset B scenario, whereas our grouped CV results are lower but more trustworthy for deployment expectations. - Nonetheless, Dataset B's consistently high Random Forest AUC (~0.94) even in random splits suggests that if subject ID leakage were resolved (by grouping), we might still expect better performance than Dataset A's 0.87, likely due to the rich feature space (754 features including

TQWT coefficients). This indicates that there is additional signal in those advanced features – but one must handle them carefully to avoid overfitting to subject-specific patterns.

In essence, **evaluation strategy drastically affects reported performance**. Our work emphasizes that without guarding against subject leakage, one can be misled about a model's true accuracy. We advocate that future studies adopt grouped CV or independent test sets. The field should treat results from non-grouped evaluations as upper bounds (often unattainable on new data).

## 9.4 Implications

### 9.4.1 For Researchers

- **Use grouped CV** when multiple recordings per subject exist; failing to do so can overestimate performance by 5–15 percentage points (as we saw). This should become a standard practice in PD voice research to improve rigor.
- **Include variability features** in acoustic analyses. Our results highlight that features capturing variance (e.g. std of pitch, MFCC) were crucial. Researchers should ensure their feature sets are not limited to means.
- **Report all conditions**, not just the best result. We found value in examining baseline vs extended, weighted vs unweighted. Sharing such complete results (and code) improves the field's understanding. It helps avoid "over-tweaking" results for one scenario and encourages generalizable insights.
- **Acknowledge limitations** openly. By doing so (as we have in Chapter 8), we allow the community to build on this work with clear awareness of what needs improvement (e.g. testing on more data, trying deep learning, etc.).

### 9.4.2 For Practitioners

- Voice-based PD screening is feasible but **not yet clinical-grade**. Our best model had ~82–83% accuracy, meaning it misses some PD cases and flags some HCs. In a clinical or population screening context, this suggests it could be a useful *adjunct* tool but not a standalone diagnostic.
- Random Forest provides a robust baseline classifier that is relatively interpretable (via feature importance). Practitioners can examine which features are driving a prediction to ensure they make medical sense (e.g. the model flagged this person as PD mainly due to very low pitch variability and high jitter – which aligns with clinical impression).
- Feature interpretability supports clinical understanding. We can say "the algorithm focuses on monotone and breathy voice signs," which are terms a clinician understands. This is an advantage over black-box approaches and might increase trust in using such a tool for screening or monitoring.
- **Results require validation on independent cohorts**: Before any clinical deployment, our model (or one like it) should be tested on new patients from a different hospital or demographic to confirm performance. Practitioners should be cautious about any voice-based tool that hasn't been externally validated.

### 9.4.3 For Dataset Creators

- **Always include subject identifiers** in public datasets to enable proper evaluation. The differences we demonstrated between grouped and non-grouped evaluation underscore how crucial this is. If identifiers are omitted (as in some UCI datasets), researchers might inadvertently draw overly

optimistic conclusions. Including IDs (or a recommended train/test split by subject) helps ensure valid modeling.
- Document recording conditions and equipment. That way, if a model is developed on that data, users know the context (e.g. “this model expects audio recorded via smartphone at 44.1kHz in a quiet room” – if applied outside that, results may vary).
- Provide demographic information (age, gender, disease duration/severity) with the dataset. This allows researchers to control for confounding factors or to test the model’s robustness across subgroups. For example, we’d like to know if our model performs equally well for male vs female voices (pitch features behave differently across genders). Without that info, we assumed homogeneity.
- Consider longitudinal designs. A dataset with follow-up recordings could allow models that track progression. Currently, most datasets (including those used here) are cross-sectional single recordings, limiting models to snapshot diagnosis. Including time-series data of voice changes could open doors for progression monitoring algorithms.

## 9.5 Limitations Recap

Key limitations that bound the interpretation of results:

1. **Small sample size** (37 subjects) – leads to high variance and less stable estimates of model performance.
2. **No hyperparameter tuning** – our models might be under-optimized; tuning could potentially yield higher metrics.
3. **Single dataset source** – our training/testing all came from one dataset; thus, generalization to other recording conditions or populations is uncertain.
4. **Binary classification only** – we did not address PD severity or differentiation from other disorders.
5. **No external validation** – we validated via cross-val only; the model hasn’t been tested on a fully independent cohort, which is the acid test for any predictive model.

We advise readers and future researchers to keep these in mind. Our reported numbers should not be taken as universal – they apply to our dataset and conditions. Overcoming these limitations (especially via larger multi-center studies) is necessary to move toward clinical application.

## 9.6 Future Directions

### 9.6.1 Short-term Extensions

- **Hyperparameter optimization** with nested CV or Bayesian methods to see if SVM or other algorithms can catch up or surpass RF when properly tuned.
- **Feature selection** to reduce dimensionality and possibly improve generalization. For instance, using L1-regularized LR or tree-based selection to find a compact set of top features (maybe 10–20) that preserve performance. This could simplify models further and reveal the core minimal feature set needed.
- **Multi-task fusion** (ReadText + Spontaneous): Train a single model on combined feature vectors from both tasks per subject (where available). This might improve robustness by leveraging complementary information (read speech might highlight certain deficits, spontaneous others). One could concatenate features or use an ensemble that integrates both tasks’ predictions.

- **Additional acoustic features**: e.g. non-linear dynamic features (Lyapunov exponents, recurrence quantification) as used by Little et al. (2007), or newer features like Mel-spectrogram convolutional embeddings or voice tremor metrics. Incorporating these might further boost performance.
- **Data augmentation**: Create synthetic variations of audio (pitch shifted, noise added) to effectively increase training sample diversity and test model robustness to such variations.

### 9.6.2 Medium-term Research

- **External validation on independent datasets**: e.g. test our model on the standard 50-speaker Oxford dataset (Little's data) or the Italian Parkinson's vocal dataset, etc. Conversely, train on one, test on our data. This would assess generalization across demographic/language differences and identify features that hold up vs those that are dataset-specific.
- **Deep learning with appropriate regularization**: Explore if a 1D CNN or transformer on raw audio (or spectrogram) can outperform classical ML when trained carefully on a larger combined dataset. Use techniques like transfer learning (pre-train on large speech corpus, fine-tune on PD vs HC), and incorporate interpretability methods (like layer-wise relevance propagation) to retain insight into what the model uses.
- **Longitudinal tracking of disease progression**: If data can be collected from PD patients over time, see if voice features can predict changes in clinical scores (UPDRS speech item, etc.). This could extend our binary classifier into a regression or multi-class (mild/moderate/severe) problem.
- **Multi-class classification (severity levels)**: Using clinician-rated dysarthria severity as labels, train a model to classify severity from voice. This is challenging due to needing well-annotated data, but would be valuable for monitoring and for evaluating whether our features correlate with disease stage (some literature suggests they do, e.g. pitch variability decreases with progression).

### 9.6.3 Long-term Vision

- **Integration into smartphone applications**: Ultimately, the goal would be non-invasive, at-home PD screening or monitoring via voice. Our results are a step in that direction. For real deployment, one must handle a wide range of acoustic environments. Future work should test models on in-the-wild recordings (background noise, different devices) and possibly retrain models to be noise-robust (using methods like noise augmentation or training a denoising front-end).
- **Multi-modal biomarkers (voice + gait + tremor)**: Voice is one modality. Combining it with other modalities (e.g. analyzing handwriting or gait from sensors) could yield a more comprehensive digital biomarker. In practice, an app might collect voice and some simple tapping or drawing tasks – combining those outputs via a late fusion model might greatly improve accuracy in detecting PD.
- **Personalized baselines for individual tracking**: Each person has a unique voice, so differences from personal baseline might be more sensitive to PD changes than absolute values. In future, deploying voice monitoring in those at risk (e.g. REM sleep behavior disorder patients) and tracking deviations from their baseline could allow earlier detection. This requires longitudinal modeling for each individual (perhaps using anomaly detection techniques).
- **Clinical validation studies**: Ultimately, conducting studies where voice-based assessments are compared with neurologist diagnoses or gold-standard tests (like DaTscan), to quantify sensitivity/specificity in a real clinical workflow. This will be the proof of whether voice AI can become part of screening or telemedicine for PD.

## 9.7 Closing Remarks

This thesis demonstrates that **voice-based Parkinson's Disease classification is feasible** using classical machine learning with carefully engineered acoustic features. The **+8.7pp improvement** from feature extension highlights the importance of capturing speech dynamics beyond simple averages, reinforcing that PD's vocal imprint lies in subtle variability and quality measures.

However, the field faces significant challenges: - **Small datasets** necessitate rigorous methodology to avoid misleading results. We showed that subject-aware validation yields more realistic performance estimates than naive approaches. - **Subject identity** must be tracked for valid evaluation, and future datasets should heed this requirement to enable honest benchmarking. - **Clinical deployment** requires extensive validation – our work is a foundation, but translating it to a reliable tool entails testing on larger, diverse populations and ensuring robustness to recording conditions and comorbidities.

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

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

---

[1] [3] Lee Silverman voice treatment (LSVT) mitigates voice difficulties in mild Parkinson's disease - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC6504915/

[2] A comparison of regression methods for remote tracking of Parkinson's disease progression - ScienceDirect
https://www.sciencedirect.com/science/article/abs/pii/S0957417411016137

[4] [5] [7] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] GitHub
https://github.com/christos97/parkinsons-voice-classification/blob/508d0935ec5b84d623cc1d5bf8b2e1abb46ec429/docs/v2/APPENDIX_A_FEATURE_IMPORTANCE.md

[6] [8] CHAPTERS_1_2_3_4_5_6.pdf
file://file_00000000506c71f49181e562d33bd22b