ChatGPT

Chapter 1: Introduction

Parkinson’s Disease (PD) is a progressive neurodegenerative disorder that imposes a growing burden on public health. Recent epidemiological studies indicate PD is among the fastest-growing neurological conditions worldwide, with an estimated 6 million individuals affected as of 2016 [1]. Clinically, PD is characterized by motor symptoms such as tremor, rigidity, bradykinesia, and postural instability [2]. In addition to these cardinal signs, *voice and speech impairments are present in nearly 90% of people with PD*, reflecting the disorder’s impact on the vocal apparatus [3]. These vocal changes—ranging from reduced volume and monotony to dysarthria—motivate investigation into **voice analysis as a non-invasive biomarker** for PD. Early detection and objective monitoring of PD remain challenging; diagnosis typically relies on clinical examination and subjective rating scales, which can be time-consuming and prone to error. Notably, misdiagnosis rates for PD by general neurologists may reach up to 20% [4]. This situation creates an urgent need for accessible and accurate diagnostic support tools [5]. In this context, **voice-based analysis** offers an attractive approach: voice recordings are easy to obtain, inexpensive, and remote-friendly, making them a scalable medium for screening and monitoring PD in large populations [6] [7]. Over the past decade, numerous studies have reported promising accuracy in distinguishing PD from healthy controls using acoustic features of speech [7]. These works demonstrate that vocal biomarkers can potentially reflect the presence and progression of PD, encouraging further research into automated voice-based PD detection.

Despite encouraging results in prior voice-based PD classification studies, important gaps and challenges persist. Many earlier studies reporting high accuracies have used relatively small, homogeneous datasets collected under controlled conditions [6]. Such conditions can lead to overfitting or optimistic performance estimates that **do not generalize** well to broader populations. In real-world applications, variability in voice recordings (due to different microphones, environments, and speaker characteristics) can degrade model performance if not properly accounted for. There is thus a need for methodologically rigorous research that emphasizes generalizability, reproducibility, and interpretability. **Voice analysis for PD is still an emerging field**, and open questions remain about the best way to extract and utilize vocal features for robust classification. Addressing these questions requires careful experimental design – for example, ensuring that no subject’s recordings appear in both training and testing sets, to avoid “leakage” that inflates performance. Moreover, the selection of algorithms must consider the typical constraints of medical voice datasets: sample sizes tend to be small (on the order of tens of subjects), and feature sets may be high-dimensional, which can complicate model training.

In this thesis, we approach the problem of PD detection from voice recordings as a **binary classification task**: given acoustic data from a subject’s speech, the goal is to classify the subject as having Parkinson’s Disease (PD) or being a Healthy Control (HC). The central premise is to evaluate whether **classical machine learning methods** can effectively detect PD from voice features, and to do so in a manner that is interpretable and suitable for limited data. We deliberately focus on classical (non-deep) machine learning models – such as logistic regression, support vector machines (SVM), and random forests – rather than modern deep learning techniques. There are several motivations for this choice. First, classical models offer greater transparency; their decisions can often be explained by examining feature importance or coefficients, which is valuable in a medical context where interpretability is desired. Second, the datasets available for this research are relatively small ($N<100$ in our primary dataset), a scenario in which deep learning models would be prone to severe overfitting [8]. Classical models tend to be more sample-efficient in low-data regimes [9]. Third, by establishing the performance of interpretable baseline models, this work lays groundwork for future studies (for instance, using deep learning or larger datasets) while keeping the current project's scope modest and focused. It is worth noting that classical approaches have historically been quite successful in this domain: for example, early work by Little *et al.* achieved around 91% accuracy using an SVM on voice features [10], demonstrating the viability of non-neural methods for PD voice classification. Given these considerations, this thesis prioritizes *reproducibility and interpretability* over pushing absolute state-of-the-art performance. All experiments are conducted with careful control of random seeds, consistent evaluation protocols, and rigorous validation to ensure that results reflect genuine model generalization rather than artifacts of chance or data leakage.

Scope of Work: This research is confined to a well-defined scope, both in terms of the problem addressed and the approaches employed. The thesis investigates **binary classification** of PD vs. HC using voice data – no multiclass classification (such as differentiating varying severity levels of PD) is attempted. Two complementary datasets are utilized to broaden the analysis: (1) a *raw audio dataset* of voice recordings collected at King's College London Hospital (referred to as **Dataset A: MDVR-KCL**), and (2) a *pre-extracted feature dataset* from the UCI Machine Learning Repository (referred to as **Dataset B: PD_SPEECH_FEATURES**). Dataset A consists of recorded speech from 37 individuals (16 with PD and 21 healthy), each performing certain speech tasks (reading a standard text and engaging in spontaneous dialogue) under clinical conditions [11] [12]. These raw audio files allow us to implement a complete signal processing and feature extraction pipeline, yielding a tailored set of acoustic features per recording. In contrast, Dataset B is a public benchmark dataset containing **756 samples** (voice recordings) from a total of 252 subjects (188 PD, 64 healthy) [13]. Each sample in Dataset B is represented by a high-dimensional feature vector (752 features) extracted from sustained phonation of the vowel /a/ [14] [15]. By including both datasets, the thesis is able to **compare two paradigms**: a low-dimensional, interpretable feature set derived from raw audio (Dataset A) versus a large, rich feature set curated by prior researchers (Dataset B). This dual-dataset strategy provides a form of *comparative analysis* – any differences in classification outcomes between the two will shed light on the impact of feature representation, dataset size, and other factors on model performance.

The experimental work is structured around these two data sources. For **Dataset A (MDVR-KCL)**, a custom processing pipeline is developed: audio recordings are ingested and standardized, a set of acoustic features (e.g. Mel-frequency cepstral coefficients, jitter, shimmer, fundamental frequency statistics) is extracted from each recording, and then classical classifiers are trained to distinguish PD vs HC. A key design principle here is to avoid any form of subject overlap between training and testing data. The evaluation for Dataset A uses **grouped 5-fold cross-validation**, meaning that all recordings from a given subject are kept within the same fold (either in training or in testing, but never split) [16] [17]. This ensures the validation is truly subject-independent and reflects how a model would perform on entirely unseen individuals. For **Dataset B (PD_SPEECH_FEATURES)**, the workflow is simpler since features are already provided: the dataset is loaded as a table of features and labels, and the same types of classifiers are trained and evaluated. Because Dataset B does not include subject identifiers linking its 756 samples to specific individuals [18], a standard stratified 5-fold cross-validation (at the sample level) is applied [19]. It is acknowledged that this evaluation for Dataset B may inadvertently mix samples from the same subject across folds (an *unavoidable limitation* given the data), so results on this dataset must be interpreted with caution. Indeed, one **explicit constraint** of this thesis is that no claims of definitive clinical performance are made: the models are evaluated on these datasets to explore patterns and feasibility, but they are **not validated for real-world diagnosis**. In line with this, **no clinical trial or deployment** is within scope – the project does **not** include building a ready-to-use diagnostic system, and it makes **no medical recommendations or decisions**. All analyses are conducted offline on prerecorded data, and all findings are to be understood as research outcomes, *not* as a clinical tool at this stage [20] [21]. Furthermore, **no new data were collected** for this study; the work leverages existing publicly available datasets, avoiding any need for clinical data collection or ethics approval processes [22].

To evaluate classification performance, the thesis employs a set of **classical metrics common in binary classification**. These include *accuracy*, *precision*, *recall* (sensitivity), *F1-score*, and the *Receiver Operating Characteristic Area Under the Curve (ROC-AUC)* [23]. All metrics are reported as averages over cross-validation folds to provide an estimate of generalization performance along with variability. The use of multiple metrics is important in reflecting different aspects of model behavior (for example, accuracy alone can be misleading if classes are imbalanced, hence metrics like F1 and ROC-AUC offer additional insight). Consistent evaluation criteria are applied across both datasets and all models, ensuring a fair comparison. Another aspect of the methodology is an emphasis on **reproducibility**: the experiments are conducted with fixed random seeds and documented procedures [24], and the code is managed under version control, so that the results can be independently replicated. This rigorous approach strengthens the credibility of any conclusions drawn and aligns with the academic nature of the project.

Within this defined scope, the thesis addresses several **research questions** aimed at advancing understanding of voice-based PD detection using classical machine learning:

1. **Effectiveness of Voice-Based Classification:** *How accurately can classical machine learning models distinguish Parkinson's Disease patients from healthy controls using acoustic features of voice?* This question investigates the baseline feasibility of voice-based PD detection with interpretable models, using the features extracted from speech recordings. It examines whether the patterns in vocal biomarkers are sufficient for reliable classification in a controlled experimental setting.

2. **Impact of Feature Sets and Data Representation:** *To what extent does the choice of feature representation and dataset affect the classification performance?* Here we compare the two approaches – a limited, **interpretable feature set** derived from raw audio (Dataset A) versus a **high-dimensional, pre-extracted feature set** (Dataset B). The question encompasses whether a carefully engineered small feature set can perform comparably to a brute-force large feature set, and what trade-offs emerge (e.g., in terms of accuracy vs. overfitting risk, or interpretability vs. raw performance).

3. **Factors Influencing Model Performance:** *What are the main factors that influence the success or failure of classical classifiers in detecting PD from voice data?* This question seeks to interpret the outcomes by analyzing contributors such as **dataset size**, **feature dimensionality**, and **validation strategy**. For instance, if one approach yields higher accuracy, is it primarily due to having more training examples, a richer feature space, or perhaps a less stringent cross-validation method? Likewise, we consider whether certain algorithms are more robust than others on the small-sample data, and how the variability in results (e.g., fold-to-fold fluctuations) reflects underlying data limitations. Part of this inquiry also involves examining whether different speech tasks (reading a passage vs. spontaneous speech in Dataset A) exhibit different levels of discriminative power for PD detection, and why that might be the case from a clinical perspective.

These research questions define the investigative focus of the thesis. It must be emphasized that the work is **exploratory and foundational** in nature: the aim is not to deploy a clinically-ready system, but rather to deepen understanding of voice-based PD classification under realistic constraints. All findings will be discussed with appropriate caution. In particular, any positive results are not interpreted as clinical validation, and any comparisons between the two datasets or feature sets are made with awareness of confounding differences (such as the disparity in sample size and feature count). By staying within the aforementioned scope, the thesis ensures its contributions remain valid and interpretable without overstating their applicability.

Thesis Structure: The remainder of this thesis is organized as follows. **Chapter 2 – Literature Review** provides background on prior research in PD detection from voice, summarizing relevant findings from the last several years. This includes a review of clinical aspects of PD-related voice changes and various computational approaches (features and models) that have been explored in the literature. **Chapter 3 – Data Description** details the two datasets used in this study (MDVR-KCL and PD_SPEECH_FEATURES). It describes their collection protocols, contents, and any preprocessing or quality control steps taken. Special attention is given to class composition, the nature of the speech tasks, and known limitations of each dataset (for example, the absence of subject IDs in Dataset B). **Chapter 4 – Methodology** outlines the overall approach and experimental methodology. It covers the feature extraction process for raw audio, the design of the classification pipelines for both datasets, the selection of machine learning algorithms, and the cross-validation and evaluation procedures. All key decisions (such as focusing on classical ML and enforcing subject-independent validation) are explained in this chapter. **Chapter 5 – Experimental Design** defines the experiments conducted, including how the data was split and what comparisons were made. It may enumerate specific experiment conditions (for instance, training models on read-text vs spontaneous speech subsets, or comparing performance across datasets) and the rationale behind each experiment. **Chapter 6 – Results** presents the outcomes of the experiments. This chapter includes performance metrics for each model on each dataset (and each speech task for Dataset A), typically in tabular or graphical form, along with descriptive statistics. No interpretation is given in the Results chapter beyond factual observations of the numbers. **Chapter 7 – Discussion** provides an in-depth analysis of the results, linking them back to the research questions. In this chapter, the performance of the models is interpreted: for example, reasons for any performance gap between the two datasets are analyzed (considering sample size, feature richness, and possible data leakage issues), and the behavior of different algorithms is discussed (such as why the SVM might have struggled with the small dataset, or why random forests performed robustly). The discussion also relates the findings to existing literature and highlights practical implications (while reiterating that no clinical claims are made) [25] . **Chapter 8 – Limitations** explicitly addresses the constraints of the study. It candidly acknowledges the key limitations such as the small sample size of Dataset A [26] , the feature dimensionality mismatch between datasets [27] , the potential subject overlap issue in Dataset B [28] , and the restriction to classical models without exploring deep learning [29] . This chapter defines the boundaries within which the results should be interpreted and reinforces why certain choices were made (e.g. excluding deep learning due to data limitations). Finally, **Chapter 9 – Conclusion** closes the thesis with a brief summary of the work and its findings, and offers perspectives for future work. The conclusion reiterates the key contributions and observations of the thesis, while suggesting how subsequent research could build on these results – for instance, by incorporating more data, applying deep learning in a future phase, or conducting clinical validation studies to move closer to a practical PD diagnostic tool.

In summary, this thesis introduces and motivates the use of voice as a scalable, non-invasive biomarker source for Parkinson’s Disease detection, and investigates the efficacy of classical machine learning techniques for this task under realistic constraints. The work is carefully scoped to avoid overclaiming: it delivers comparative insights between two feature representation strategies and provides baseline results with interpretable models, without venturing into unvalidated clinical application. By doing so, the thesis aims to contribute a rigorous piece of evidence in the field of PD voice analysis, highlighting both the potential and the limitations of classical machine learning approaches in detecting Parkinson's Disease from speech. The following chapters will elaborate on each aspect in detail, beginning with a review of the relevant literature to place this study in context.

---

[1] [2] [4] [5] [6] [7] Assessing Parkinson's Disease at Scale Using Telephone-Recorded Speech: Insights from the Parkinson's Voice Initiative - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC8534584/

[3] Evidence-based treatment of voice and speech disorders in Parkinson disease - PubMed
https://pubmed.ncbi.nlm.nih.gov/25943966/

[8] [26] [27] [28] [29] CHAPTER_8_LIMITATIONS.md
file://file_00000000b24871f4b19e554caceec9a0

[9] [20] [21] [22] [24] SCOPE_AND_LIMITATIONS.md
file://file_00000000e76471f49b1b1381544f1662

[10] Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis | Scientific Reports
https://www.nature.com/articles/s41598-025-96575-6?error=cookies_not_supported&code=bc91e476-859a-435b-8160-11e378cf74a6

[11] [12] [14] [16] [18] CHAPTER_3_DATA_DESCRIPTION.md
file://file_00000000fff0720a870e80c843cc17ce

[13] [15] UCI Machine Learning Repository
https://archive.ics.uci.edu/dataset/470/parkinson+s+disease+classification

[17] [19] [23] CHAPTER_4_METHODOLOGY.md
file://file_00000000607c720a96363a37ae624382

[25] CHAPTER_7_DISCUSSION.md
file://file_00000000f5607246bb45cded6769b983