ChatGPT

Chapter 2 — Literature Review

2.1 Parkinson's Disease and Speech Impairment

Parkinson's disease (PD) is a progressive neurodegenerative disorder primarily known for its motor symptoms such as tremor, rigidity, and bradykinesia. In addition to these well-recognized motor features, PD almost invariably affects speech and voice as the disease progresses. In fact, studies report that approximately 70–90% of individuals with PD develop measurable speech and voice impairments [1] [2]. This collection of speech symptoms in PD is often referred to as *hypokinetic dysarthria*, indicating 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 range of pitch, resulting in speech that lacks the normal ups and downs of intonation (often described as "monopitch" speech) [3]. **Monoloudness** (little variation in volume) often accompanies this, so the overall prosody (melody and expressiveness of speech) is markedly diminished [3]. Patients may also exhibit **articulatory imprecision**, where consonants in particular are not enunciated crisply [3]. 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 [4]. Additionally, some individuals speak with an improperly fast rate or with short rushes of speech, which, combined with the articulation issues, can reduce intelligibility [4]. These speech characteristics—reduced loudness, monopitch, monoloudness, hoarse/breathy voice, and imprecise articulation—are widely observed in PD and form the basis of clinical descriptions of hypokinetic dysarthria [4].

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 some patients [5]. 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 [6]. Unlike many clinical tests 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 monotony 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 be understandable in speech, while another with relatively

moderate overall motor signs could have severe dysarthria. In general, as PD progresses, speech tends to worsen – volume may further decrease and articulation may become more slurred [2]. But importantly, these changes are not uniform or perfectly correlated with disease duration [7]. Factors such as individual patient differences, co-occurring conditions (like age-related voice changes), and even treatment effects (medications or speech therapy) can influence the speech presentation. This variability means that while speech is a promising biomarker, one must be careful in using it for diagnosis or monitoring: any voice-based assessment needs to account for the wide range of "normal" for PD speech and the overlap with speech characteristics of other populations (e.g. normal aging can also cause some reduced loudness or hoarseness). In summary, PD provides a compelling case for voice analysis – it is a neurodegenerative disorder with clear motor effects on speech production, speech changes are prevalent and can be captured non-invasively, but these changes are heterogeneous across individuals and time. This establishes *why* voice is relevant for PD research while cautioning that clinical diagnostic use would require careful handling of variability and uncertainty, rather than treating voice patterns as a definitive signature of the disease.

2.2 Acoustic Characteristics of Parkinsonian Speech

A variety of acoustic features have been explored to characterize the distinctive patterns of Parkinsonian speech. The motivation for examining these features is that they quantify specific aspects of voice and speech that are affected by PD, thus providing measurable indicators that can be used in analysis or as inputs to classification models. Broadly, these features can be grouped into categories or *feature families* based on what aspect of speech they describe. In this section, we organize the discussion by three major families of acoustic features: **prosodic features**, **perturbation measures**, and **spectral/cepstral features**. This organization will correspond to the methodologies later employed in our study, where we leverage these feature types. Each subsection below defines the feature group, gives examples of specific features in that group, and explains what changes have been observed in PD speech relative to normal speech in that feature domain.

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 the voice (notated as *F0*) corresponds to the perceived pitch. Prosodic analysis often looks at statistics of F0, such as the mean pitch, range (difference between highest and lowest pitch), and the standard deviation of pitch across an utterance, to gauge how much variation in pitch a speaker uses. Loudness can be quantified through overall intensity level (in decibels) and its variability or emphasis patterns (e.g., how much a speaker modulates volume for stress). In typical expressive speech, healthy speakers vary both pitch and loudness to convey emphasis, emotion, or sentence modality (question vs. statement, etc.).

In Parkinson's disease, a well-documented phenomenon is the **reduction of prosodic variability**. PD patients often speak in a *monotone* – their pitch remains relatively flat and at a narrow range, lacking the normal ups and downs. Objectively, one finds a lower standard deviation of F0 and a smaller pitch range in PD speech compared to healthy age-matched controls [3]. This is sometimes described clinically as *monopitch*. Likewise, PD speakers exhibit *monoloudness*: their volume tends to be more constant and generally softer than normal. They may not employ the usual loudness increases to stress important words or to express emotion. The term *hypophonia* specifically refers to the reduced overall loudness (soft voice) that is common in PD [3]. Together, monopitch and monoloudness make PD speech sound flat or expressionless. For instance, where a healthy speaker might vary pitch and loudness dynamically within a

single sentence ("Really? I **can't** believe it!"), a person with PD might deliver the same sentence in a relatively uniform tone and volume, which can be perceived as lacking affect.

Empirical studies support these observations. One study noted that most PD patients have significantly lower pitch variability and reduced intensity modulation, resulting in perceptually monotonic and weak speech [3]. These prosodic deficits can be quantified: for example, computing the pitch range in a reading passage might show only a few semitones of variation for a PD speaker versus perhaps an octave for a healthy speaker. Similarly, intensity traces from PD speech often appear "flatter." Prosodic features like **F0 range**, **F0 variability (e.g., variance or interquartile range of F0)**, **intensity range**, and **intensity standard deviation** are therefore commonly included in acoustic analyses. They capture the diminished expressivity in PD speech that comes from rigidity and bradykinesia affecting the vocal apparatus (including respiratory support and laryngeal control). In summary, prosodic measures in PD typically indicate *reduced pitch and loudness variability*, aligning with the clinical description of hypokinetic dysarthria where speech has a monotone, soft character. These features are important later when designing classifiers, because they directly reflect how PD impacts one's control over voice dynamics.

2.2.2 Perturbation Measures

Perturbation measures are acoustic features that capture the cycle-to-cycle variations in the voice signal, reflecting stability (or instability) of vocal fold vibration. The two primary perturbation measures are **jitter** and **shimmer**. **Jitter** quantifies the minute fluctuations in pitch period from one glottal cycle to the next – essentially, variability in the fundamental frequency. **Shimmer** quantifies the variability in amplitude (loudness) across successive glottal cycles. In a perfectly steady, clear voice, one would expect nearly constant pitch period and amplitude for each cycle of vocal fold vibration (especially during sustained phonation of a vowel). However, human voices always have some natural jitter and shimmer. Pathologies that affect vocal fold control tend to increase these perturbations.

In Parkinson's disease, due to factors like reduced vocal fold adduction, tremor, and inconsistent breath support, **jitter and shimmer are often elevated** compared to age-matched healthy controls [8]. That is, PD voices typically show more frequent, irregular fluctuations in frequency and amplitude. This corresponds to the perceptual observation of a hoarse or unsteady voice quality. For example, when sustaining a vowel sound like "ah," a healthy voice might sound steady, whereas a PD voice might quaver slightly in pitch and/or volume. These micro-instabilities are precisely what jitter and shimmer measure. **Harmonics-to-Noise Ratio (HNR)** is another related metric, which compares the level of periodic (harmonic) energy in the voice to the level of aperiodic or noise energy. A high HNR means the voice signal is very harmonic (clean tone), whereas a low HNR indicates a noisier, breathier voice. PD voices tend to have **lower HNR** values, indicating a higher proportion of noise (breathiness, roughness) in the voice [8].

To put some numbers for illustration: a healthy sustained vowel might have a jitter on the order of 0.5% (very small period fluctuations) and shimmer around 3–4%, with an HNR of, say, 20 dB or more. In a PD voice, jitter might be several times higher (reports of jitter in PD can be, for instance, 1%–2% or more) and shimmer likewise elevated, while HNR might drop to, say, 10–15 dB [8]. Many studies corroborate that **PD causes increased frequency and amplitude perturbation** in the voice and a corresponding increase in spectral noise [8]. These measures provide objective evidence of the vocal instability and glottal insufficiency that are characteristic of Parkinsonian dysarthria (often described as a "breathy and hoarse" voice quality [4]).

In summary, perturbation features like jitter, shimmer, and HNR capture the fine-grained instability of vocal fold vibration. PD speakers often have higher jitter and shimmer (meaning more irregular vocal fold vibrations) and lower HNR (meaning a noisier voice signal) than neurologically healthy speakers [8]. These features are clinically intuitive (they align with the idea of a breathy, irregular voice in PD) and have been widely used in research studies. In our work, we include these perturbation measures as they are known indicators of the *phonatory* deficits in PD (related to the vocal source signal). They complement the prosodic features by focusing more on voice quality than on expressive modulation.

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/cepstral features provide information about the distribution of energy across frequencies and the resonant qualities of the vocal tract during speech. These features are particularly useful for capturing articulatory changes and timbre differences in PD speech.

A key set of features in this category are the **Mel-Frequency Cepstral Coefficients (MFCCs)**. MFCCs are computed by taking a short-time spectrum of the speech signal, warping the frequencies onto a mel-scale (which is roughly logarithmic, mimicking human auditory perception), and then applying a cosine transform to derive a set of coefficients that compactly represent the spectral shape. The first few MFCCs (typically 12–13 coefficients per frame, excluding the 0th which is overall energy) effectively describe the broad spectral envelope—essentially, they capture which frequencies have more energy versus less energy. Because the spectral envelope is largely determined by the shape and configuration of the vocal tract (tongue position, jaw opening, etc.), MFCCs are often interpreted as features capturing aspects of articulation and vowel formation.

In PD, articulatory movement can be reduced (hypokinesia affects not just limbs but also the range of tongue/lip/jaw motions). This can lead to phenomena like **vowel centralization** (vowels become more neutral and closer together in formant space). Spectral features can detect such changes. For instance, one can derive vowel formant measures or vowel space area from spectral analysis, or simply let MFCCs capture the overall changes. MFCCs and related cepstral features have been **widely used in PD voice classification studies** [9] because they are very informative and are standard in many speech processing tasks (including speech and speaker recognition). They allow algorithms to pick up subtle differences in voice quality and articulation that might not be obvious from just jitter or pitch alone.

Apart from MFCCs, other spectral features include the **spectral centroid** (the “center of mass” of the spectrum), **spectral bandwidth** (spread of the spectrum), **spectral roll-off** (frequency below which a certain percentage of total energy lies), and **spectral flux** (frame-to-frame change in the spectrum). These can indicate how high-frequency vs low-frequency weighted the voice signal is, which may change if, for example, a voice becomes weaker in high frequencies due to incomplete articulation (high-frequency fricative components might reduce). *Cepstral peak prominence* is another feature sometimes used, which measures the prominence of the cepstral peak corresponding to the periodicity of voice – essentially another way to gauge how harmonic vs noisy a voice is (somewhat akin to HNR, but in cepstral domain).

For PD speech, researchers have found that cepstral measures (like certain MFCCs or cepstral separation differences) can distinguish PD vs healthy voices, and they sometimes correlate with intelligibility or disease severity [9]. One systematic review noted that features spanning articulation (including cepstral

coefficients) were effective for predicting clinical ratings of PD severity [10]. In practical terms, one might observe that PD patients have a lower cepstral peak prominence (indicating a breathier voice) and MFCC patterns that reflect a smaller vowel space (due to reduced articulatory range), though MFCCs themselves are not directly interpretable without further analysis.

It is also common to include **delta (Δ) and delta-delta (ΔΔ) MFCCs**, which are first and second temporal derivatives of the MFCC trajectories. These capture dynamic aspects of the spectral changes (how quickly features are changing), which might be relevant if PD speech has altered timing or coarticulation patterns.

In summary, spectral and cepstral features provide a **rich representation of the speech signal** that encompasses voice quality and articulatory information. They are a staple in most voice-based classification systems due to their proven effectiveness [9]. In PD voice analysis, these features help detect the more subtle or complex changes (like shifts in formant frequencies, changes in spectral noise energy, etc.) that are not directly captured by simpler measures like jitter or pitch range. In our work, we rely on cepstral features (notably MFCCs) as part of the feature set, given that they have shown strong performance in prior studies and effectively summarize the information contained in the speech spectrum. By using MFCCs alongside prosodic and perturbation features, we aim to cover **all three major aspects** of Parkinsonian speech: the reduced variability (prosody), the instability of phonation (perturbation), and the altered articulation/resonance (spectral/cepstral). These acoustic feature families form the basis of most classical machine learning approaches for voice-based Parkinson's Disease classification, as evidenced by numerous studies that utilize combinations of these features [9]. Our subsequent methodology and experiments will specifically draw upon these groups of features.

(Bridge to next sections: having established what features can characterize PD speech, we now move on to how these are used in classical machine learning models, and what datasets and validation methods are employed in the field.)

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 classifier to automatically distinguish Parkinson's disease vs. healthy controls (or to predict severity, etc.). Over the years, a variety of **classical machine learning algorithms** have been applied to this problem. In the context of this thesis, *classical* refers to well-established, typically non-neural-network models such as logistic regression, support vector machines, and ensemble decision tree methods. These methods were especially dominant in PD voice research before the recent surge of deep learning, and they remain highly relevant for moderate-sized tabular datasets and for scenarios where interpretability and robustness are priorities.

This section reviews three categories of classical classifiers that are frequently used and also employed in our experiments: (1) **Logistic Regression**, a linear model; (2) **Support Vector Machines (SVM)**, a kernel-based margin classifier; and (3) **Ensemble methods** like **Random Forests**, which are non-linear tree-based models. Rather than arguing that one is the best, we will highlight the **trade-offs** and typical usage of each, as reported in the literature. Each subsection describes why the model is suitable for PD voice data and notes any particular considerations or performance results from prior studies.

2.3.1 Logistic Regression

Logistic regression (LR) is a simple yet powerful baseline classifier commonly used in biomedical applications, including PD voice studies. It is a **linear model** that models the log-odds of the probability of PD as a linear combination of the input features. In practical terms, logistic regression assigns a weight to each acoustic feature and an intercept term; after summing the weighted features, it applies a logistic (sigmoid) function to map this sum to a probability between 0 and 1. The model is trained to maximize the likelihood of the data (or minimize a classification loss), effectively learning weights that separate the two classes (PD vs healthy) as well as possible in the feature space.

One reason logistic regression is often chosen as a baseline is its **interpretability**. Each weight indicates the direction and importance of its corresponding feature in the prediction. For example, if the jitter feature has a strongly positive weight, the model is using high jitter to indicate higher likelihood of PD (consistent with our knowledge that PD voices have higher jitter). If another feature, say mean F0, had a negative weight, it would mean higher F0 is associated with healthy status (perhaps younger or female controls, depending on context) in that model. This interpretability aligns well with the needs of examiner and clinician audiences who prefer models that can explain their decisions in familiar terms. In contrast to many complex models, logistic regression provides a clear insight: *how* each acoustic feature influences the PD prediction.

Logistic regression is also **computationally efficient** and works well when the number of features is not too large relative to the number of samples. It doesn't require extensive hyperparameter tuning (apart from maybe a regularization parameter to prevent overfitting). For many PD voice datasets, where sample sizes are limited (dozens of patients, a few hundred recordings), a simple model like LR is less prone to overfitting than a high-capacity model. It essentially assumes a linear decision boundary, which might be a reasonable approximation if, say, PD voices differ in a somewhat linear way in the feature space (e.g., a bit higher jitter, a bit lower loudness, etc., all adding up to distinguish PD from controls).

In the literature, logistic regression has indeed been used, either as a primary classifier or as a baseline for comparison. Some studies have reported surprisingly strong performance with logistic regression. For instance, one report using a standard PD voice feature dataset found that a regularized logistic regression achieved about 91% classification accuracy, which was comparable to more complex methods in that experiment [11]. (This result, however, should be interpreted with caution vis-à-vis validation strategy as discussed earlier.) Another study that explored multiple algorithms included logistic regression and noted it as a consistently performing baseline with the advantage of straightforward interpretation of acoustic biomarkers [12].

The limitation of logistic regression is that it can underfit if the true decision boundary is complex or highly non-linear. PD vs. healthy voice differences might not be strictly separable by a linear combination of features – there could be interactions (e.g., perhaps jitter matters only when fundamental frequency is in a certain range, etc.) or non-linear patterns. Logistic regression won't capture those, whereas more flexible models might. Nonetheless, it remains a **recommended starting point**. By examining logistic regression coefficients and performance, one can get an initial sense of which features carry signal. In summary, logistic regression serves as a **baseline classifier** in PD voice tasks, valued for its simplicity and interpretability [13]. It often provides a reasonable yardstick of performance against which more complex models are measured, and if it performs nearly as well as complex models, one might favor it for its transparency.

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 in the 2000s and 2010s. SVMs are well-suited to problems with high-dimensional feature spaces and relatively smaller sample sizes – a common scenario in medical datasets. The core idea of an SVM is to find the best separating hyperplane between classes by maximizing the margin, which is the distance between the hyperplane and the nearest data points of each class (the support vectors) [14]. Intuitively, an SVM tries to create the most "robust" separation it can, so that new points can fall on the correct side with some tolerance.

One of the strengths of SVMs is that they can employ different **kernel functions** to handle non-linear class boundaries. For example, if the relationship between features and the class label is not linear, one can use an RBF (Gaussian) kernel or a polynomial kernel to implicitly map the data into a higher-dimensional space where a linear separator might exist. This kernel trick allows SVMs to fit complex boundaries without explicitly adding more features. In PD voice classification, the data might indeed require non-linear separation – perhaps moderate PD voices cluster in one region and severe in another, or healthy voices form clusters by gender, etc. SVMs can, in principle, handle this by an appropriate kernel choice.

Historically, SVMs have shown strong results on the classic PD voice datasets. The oft-cited study by Little et al. (2009), which used 22 voice features (perturbation and others) from sustained vowel recordings of PD patients and controls, achieved around 91% accuracy using an SVM with 10-fold cross-validation. Many subsequent works on that dataset and related ones continued to use SVMs and report accuracies in the 90%+ range [15]. A systematic review noted that classical ML models such as SVMs and Random Forests tended to achieve high accuracy on *small, homogeneous datasets* in PD voice research [16]. This aligns with the expectation that SVM, with its margin maximization, can perform very well when the training and testing data come from the same distribution and the data is noise-controlled (like a single type of sustained phonation task in a lab setting).

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 classification error on training points) and any kernel parameters (like the gamma in RBF kernel which controls kernel width). If not tuned properly (often via cross-validation on training data), an SVM can either overfit (too low margin, memorizing training points) or underfit (too high margin, misclassifying too many points). They can also be somewhat sensitive to feature scaling – typically features need to be normalized for SVMs to work optimally.

Interpretability of SVMs is more limited than logistic regression. While an SVM does have a linear discriminant function in some kernel-induced space, it's not straightforward to map that back to importance of original features (except in the linear SVM case, which is essentially similar to logistic regression in interpretability). In practice, one doesn't get simple weights per feature; however, one can identify which training instances are support vectors (the critical ones) and sometimes infer which features are contributing by systematically varying inputs, but it's not as transparent.

Despite these considerations, SVMs are often a **go-to algorithm** for PD voice tasks because of their strong performance in prior studies [9]. They handle the typically moderate dimensionality (dozens of features) well, and can operate effectively even when the number of recordings is not huge, thanks to the capacity control via the margin. For example, a PD voice dataset with ~200 samples and ~20 features is a scenario

where SVMs have classically excelled. Even as data grows, SVMs can scale reasonably (though very large datasets can be slow for SVMs).

In our work, we include SVM as one of the classification approaches, evaluating its performance relative to logistic regression and ensemble methods. We acknowledge from the literature that SVM often sets a strong benchmark; for instance, Ngo et al. (2021) found SVM and Random Forest frequently among top performers in PD severity estimation tasks [9]. The expectation is that SVM will perform well if the data is clean and if distinct margins exist in the feature space, but one has to validate carefully to ensure it's not leveraging any spurious correlations (like subject-specific traits, which we guard against by subject-level splitting). Overall, SVM represents a robust, well-understood classifier for PD voice analysis, offering a good balance between flexibility (with kernels) and overfitting control (with margin regularization).

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)** classifier is widely used and has also been applied to Parkinson's voice classification problems. A Random Forest is essentially an ensemble of many decision tree classifiers, each trained on a slightly different subset of the data and features, and the final prediction is made by aggregating (majority vote or averaging) the outputs of all these trees. The rationale is that while individual trees might be noisy or overfit, their ensemble reduces variance and tends to improve generalization.

Random Forest advantages: One big advantage relevant to our context is that RFs can model **non-linear relationships** and interactions between features inherently. Each decision tree can split on different features in a non-linear way (for example, one branch of the tree might effectively be saying "if jitter > X *and* F0 variability < Y, classify as PD"). By aggregating many such trees, the Random Forest can capture quite complex decision boundaries. This is useful because the constellation of voice features distinguishing PD might indeed be complex: it may not be a simple global threshold on one feature, but rather a combination (e.g., "either the patient has very high jitter, or if jitter is moderate then they must also have low loudness variability and high shimmer to be PD," etc.). Random Forests can capture these logical combinations.

Random Forests are also fairly **robust to noise and to correlated features**. In voice data, some features are highly correlated (jitter variants correlate with each other, different MFCCs can correlate, etc.). The RF's use of random feature selection for each split means it doesn't always split on the same dominant feature; this decorrelates the trees and often leads to better usage of multiple features. It also reduces the chance that the model focuses on one peculiar feature that might be an artifact.

From an **interpretability** standpoint, Random Forests are not as straightforward as a single decision tree or a logistic regression, but they do offer some measures like **feature importance**. This is typically computed by looking at how much each feature, across all trees, contributes to reducing uncertainty (impurity) in the classification. Researchers have used this in PD studies to see which features the Random Forest found most discriminative – often jitter, shimmer, certain MFCCs, etc., come out on top, which provides some reassurance that the model is picking up known biomarkers [17]. However, interpreting an ensemble of 100 trees is inherently trickier than a single model; one might rely on feature importance or partial dependence plots for insight rather than a clear equation.

The performance of Random Forests in PD voice tasks has been reported to be very strong. Some studies that compared multiple algorithms found Random Forest to be the top performer or on par with the best. For example, in one experiment, an RF classifier achieved about 96% accuracy on a Parkinson's voice dataset, slightly outperforming SVM and logistic regression in that setting [18]. Another review mentioned that tree-based models (like RF and boosted trees) showed **competitive performance, with only slightly higher variability in results** compared to SVMs [17]. The variability remark likely means that because RF involves some randomness, different runs or different datasets might have results that fluctuate a bit, but overall they are quite reliable.

One important consideration is that Random Forests, like any model, can give inflated performance if there's data leakage. They are powerful enough to memorize dataset quirks if not validated properly. But assuming a proper subject-level split, an RF can generalize well because each tree in the forest sees a subset of data – in a sense, it's less likely to overfit to particular individuals than a single complex tree might.

In our thesis work, we include Random Forest as a representative ensemble method. It is particularly interesting for us because we are comparing a raw-audio-based pipeline vs. a feature-based pipeline. If one pipeline has a lot of features (e.g., hundreds of features extracted), an RF can handle that scenario by effectively doing built-in feature selection (trees will ignore less useful features in splits). Random Forests also naturally handle multi-feature interactions. By using an RF, we can see if the model can leverage the richer feature sets without heavy manual feature selection. We will look at its performance and also examine feature importance to see which acoustic features it deems most important for distinguishing PD, thereby connecting back to the discussion in Section 2.2.

In terms of trade-offs: RF models sacrifice some transparency and can be computationally more intensive (hundreds of trees predicting can be slower than one evaluation of a logistic function, though still very feasible for our data sizes). They also have a few hyperparameters (number of trees, max depth, etc.), but RFs are fortunately quite robust to settings – often having a large number of trees (e.g., 100 or 500) works well, and they don't overfit with more trees, they just get more stable. This robustness and strong performance on tabular data are why RF is a staple in many classification problems. In PD voice research, using an ensemble like RF is a way to ensure that we are capturing non-linear patterns that linear models might miss, hopefully leading to **high accuracy with good generalization** as reported in prior studies [9] [17]. Importantly, we will compare its results with those of logistic regression and SVM, not to "declare a winner" but to understand how a more flexible model (RF) fares against the simpler ones, and whether the extra complexity yields consistent improvements given our focus on rigorous validation.

(To avoid any perception of claiming one model is outright superior, we conclude that each of these classifiers – LR, SVM, RF – has its merits. Logistic regression offers simplicity and clarity, SVM provides a strong margin-based approach capable of complex separation with kernels, and Random Forest brings in powerful ensemble averaging for non-linear feature interactions. The literature shows all three can achieve good results in PD voice classification [9]. In our experiments, we will use all three, which also allows us to verify that any findings are not tied to one specific model choice.)

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, we can categorize the datasets into two types: **raw audio datasets** and **pre-extracted feature datasets**. The

distinction lies in whether the dataset consists of actual audio recordings (which researchers then need to process and extract features from), or whether it's a collection of instances where each instance is already represented by a set of acoustic features computed from an audio sample. Both types are relevant to this thesis (since we compare a raw-audio pipeline to a feature-based pipeline), and each comes with its own advantages and pitfalls.

2.4.1 Raw Audio Datasets

Raw audio datasets for PD typically consist of voice recordings from PD patients and healthy controls. These recordings are obtained via various means: some studies use high-quality microphones in a lab setting, others use telephone or mobile phone recordings to collect voices remotely (e.g., as in the mPower study). The content of the recordings can also vary. Common speech tasks include: **sustained phonation** of a vowel (like holding "ah" for several seconds), **reading passages or sentences** (like reading a standard paragraph or a list of sentences), and **spontaneous or semi-spontaneous speech** (such as answering questions or describing a picture). Each of these tasks can reveal different aspects of PD speech deficits.

A hallmark of raw audio datasets is that *each subject often contributes multiple recordings*. For example, in one large smartphone-based study, researchers collected over **18,000 audio recordings** via a mobile app, with multiple recordings per participant, to distinguish PD patients from healthy individuals [19]. Even in smaller scale studies, it is typical that if you have (say) 50 PD patients, you might record each performing a set of tasks, resulting in perhaps a few recordings per person. Another notable dataset, the **Oxford Parkinson's Telemonitoring Dataset**, involved **42 PD patients each recording a short vowel phonation multiple times per week over a six-month period**, yielding a total of 5,875 recorded samples [20]. In that dataset, on average each patient provided on the order of a hundred recordings (spread over time). These examples illustrate that raw audio data is often **longitudinal or repeated-measures** in nature.

Working with raw audio datasets means that the researcher must perform **feature extraction** as a step (unless using end-to-end learning). The upside is flexibility: one can compute any features of interest, apply noise reduction, or do specialized processing (like separating voiced/unvoiced segments, analyzing specific phonemes, etc.). There is also potential to apply modern end-to-end machine learning (like training a CNN on spectrograms), though that requires sufficient data.

However, raw audio datasets also present some **challenges** and limitations frequently mentioned in literature:

- **Limited Sample Size (Subjects):** Collecting raw audio from PD patients can be resource-intensive. Many studies have relatively few subjects (sometimes just a few dozen). For instance, one widely used dataset has only 31 individuals (23 with PD and 8 controls) [21]. Another dataset from Istanbul had 40 subjects (20 PD, 20 control) [22]. These numbers are small in machine learning terms, which raises concerns about the robustness of any conclusions. Small sample sizes can lead to overfitting or results that aren't statistically significant. It also means models trained on one dataset might not generalize well to new patients from a different population.

- **Multiple Samples Per Subject and Subject-Dependence:** As noted, each person contributes multiple recordings. This means that **samples within the dataset are not independent** – two recordings from the same person are likely more similar to each other (in terms of voice characteristics) than to a recording from someone else. This can confound naive machine learning

approaches. If not properly accounted for, a classifier could just learn speaker-specific traits. For example, Patient A might have a naturally higher-pitched voice than Patient B regardless of PD, so if A is in PD group and B in control, a model might wrongly use pitch as a cue for PD. We have discussed in Section 2.5 how crucial it is to do **subject-level validation** on such data to avoid this pitfall. Raw audio datasets absolutely require such handling. The literature is replete with cautionary tales where not doing this led to overly optimistic accuracies that did not hold up under stricter validation [23].

- **Variability in Recording Conditions:** Raw audio can be affected by the environment and equipment. Datasets often have heterogeneous conditions – some voices might be recorded in a quiet lab with a good microphone, others over a phone with background noise. For example, the mPower dataset (not explicitly detailed here but known in concept) involved self-recordings by users on their personal smartphones, leading to variation in distance to microphone, room acoustics, etc. Such variability can inject noise into features. It can also become a confounding factor: a model might inadvertently learn to distinguish samples recorded in a clinic (perhaps most PD patients) vs at home (perhaps controls), if such patterns exist. Researchers attempt to mitigate this by filtering or normalizing audio, but it remains a challenge that larger, multi-condition datasets are needed to solve [24].

- **Task Variability:** Different speech tasks in raw audio datasets can yield different results, and this is a limitation when comparing studies. Some datasets include only sustained vowels; others only include running speech. It's known that certain features are easier to measure or more discriminative in certain tasks. For example, a sustained vowel is great for computing jitter, shimmer, and HNR reliably, but it doesn't tell much about prosody or articulation. A reading passage can reveal prosodic range and articulation clarity but is harder to standardize for certain measurements. This task dependence means that studies using different tasks might not be directly comparable. One review found that **sustained phonation features were more effective for distinguishing PD vs control, whereas speaking tasks were better for assessing severity** [25]. So, the "best" features and accuracies reported can depend on what the subjects were asked to do in the recording. In a dataset with multiple task types, one must consider whether to combine features from all tasks or treat each task separately in analysis.

Despite these limitations, raw audio datasets are invaluable because they represent the scenario we ultimately care about: real people producing speech, with all its complexity. They allow experiments on algorithmic approaches to processing the actual signals. Our Dataset A (in this thesis) is of this type, containing audio recordings from PD and control speakers performing voice tasks. We address the noted challenges by (a) using subject-level splits to avoid leakage, (b) extracting a comprehensive set of features while being mindful of noise, and (c) interpreting results in light of the dataset's size and conditions (for example, not overclaiming generality beyond the dataset's characteristics).

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 (matrix) of feature values for each sample, along with labels (PD or control, or severity scores, etc.). Such datasets often originate from researchers who processed some raw recordings and decided to share the feature data (sometimes due to privacy, they prefer not to share raw audio, or simply as a convenience to others to use standard benchmarks).

Perhaps the most famous example in PD voice research is the **UCI Parkinson's Disease Detection Dataset**, which is the one derived from the Oxford study by Little and colleagues. As described in various sources, this dataset comprises **195 instances with 23 attributes each** (22 acoustic features + 1 label) [26]. Those 195 instances correspond to voice recordings (sustained vowels) from 31 individuals (23 PD, 8 healthy). The features include many of the measures discussed: several types of jitter, several types of shimmer, HNR, DFA (detrended fluctuation analysis), PPE (pitch period entropy), etc. The dataset itself, however, does **not directly indicate which instance came from which subject** (in the public version). It's essentially a shuffled list of feature vectors.

Other pre-extracted datasets exist as well. The one mentioned from Istanbul University has **1040 instances with 27 features** [27] [22]. That came from 40 people (20 PD, 20 control) each with multiple voice recordings. Again, the shared version likely just pools all instances together. Some newer studies or challenges might release feature sets from voice recordings, especially if they use standardized feature extraction tools (e.g., the openSMILE toolkit) to create large feature vectors.

The **advantage** of pre-extracted feature datasets is that they are **ready for machine learning** – one can just load the data and start training models. This has enabled a lot of experimentation and model comparison on a common benchmark, as researchers around the world used the UCI dataset, for example, to test various algorithms. It's also useful for rapid prototyping – if you're more interested in the classification aspect than in signal processing, these datasets let you skip right to the classification stage.

However, there are critical **limitations and risks** with these datasets:

- **Subject Identity is Often Hidden:** As noted, a major issue is that the mapping from instances to subjects is not provided. Thus, if one were to do a naive random split of the 195 samples into training and test, you would almost certainly have recordings from the same person in both sets. Indeed, each of the 31 subjects contributed 6 or so recordings on average [26]. Without knowing that, the machine learning algorithm could be effectively learning personal voice signatures. Many researchers unknowingly (or knowingly) did exactly this random split and reported extremely high accuracies – some papers reported >90% or even near 100% accuracy on the UCI dataset using sophisticated classifiers, not realizing (or not stating) that they hadn't enforced a subject-level split. As we discussed, this is a form of **data leakage**. The model doesn't generalize to new people; it just recognizes the same people it was trained on. One study explicitly demonstrated that doing a proper subject-level split on that dataset causes accuracy to drop substantially compared to a naive split (e.g., from ~99% down to much lower) [28]. Fortunately, many recent works are aware of this and now treat those datasets carefully (e.g., by using leave-one-subject-out cross-validation even if they have to infer subject labels by order).

- **Potential Overlap Between Datasets:** A more subtle issue is if multiple studies' datasets have overlapping subjects or if the same patient's data appears in more than one publicly available dataset. If someone inadvertently uses two "independent" datasets for training and testing, but they actually shared some subjects, that again could inflate results. This is not well-documented in PD voice research as a big issue, but it's worth mentioning as a general point in biomedical data.

- **Engineered Features May Carry Implicit Biases:** The features in these datasets were often engineered with PD detection in mind. For example, PPE (pitch period entropy) and DFA (signal fractal scaling) are somewhat exotic features that the original authors found useful. Including them

in the feature set might give very good performance for distinguishing PD in sustained vowels specifically [29]. However, these might be less applicable or over-fitted to that particular task or recording setup. If one blindly applies a machine learning algorithm to all 22 features, one might not realize that some features are basically proxies for each other or for PD. For instance, one feature in the Oxford set is "spread1" (a nonlinear measure) which is correlated with the UPDRS score. If a model latches onto that, it might look super accurate but is it learning PD voice characteristics or just a quirk of that feature?

- **Lack of Demographic and Metadata:** These pre-extracted sets usually strip away information like age, sex, microphone conditions, etc. Those factors, however, influence voice. If one algorithm happens to have a different proportion of male/female in train vs test by chance, results could skew. With raw data, one could attempt to control or at least report these factors; with the feature sets, often you can't because you don't have that metadata.

In spite of these issues, **pre-extracted datasets have been fundamental in driving PD voice research**. They allowed consistent benchmarks – for example, dozens of papers used the 195-sample dataset to try out different classifiers and feature selection methods, contributing to a collective understanding (and also to collective mistakes until the leakage issue was realized). They are also useful for our work: our Dataset B is essentially of this type. The key difference is we are fully aware of the subject overlap issue and explicitly address it. In our experiments, we ensure that if we use such a dataset for evaluation, we partition it by subject (based on knowledge from the original source or by reconstructing the ordering if possible). This way, we treat it more like a raw dataset in terms of validation, even though we don't have to do the low-level feature extraction ourselves.

To illustrate the **inflation risk**: if one were to do 10-fold cross-validation on the 195-sample dataset with random shuffling, one could easily get accuracy in the high 90s%. Several published works did report such results (e.g., claiming ~98–99% accuracy using SVM, Random Forest, etc., on that dataset). However, when evaluated under a rigorous approach (grouping samples by patient), realistic accuracy might be much lower – say in the 70–85% range, depending on the algorithm [28]. The difference is striking and underlines why we insist on proper validation (Section 2.5).

In conclusion, pre-extracted feature datasets are a double-edged sword: they enable rapid modeling and have been widely used to showcase high accuracy results in PD voice classification, but they carry the **inherent risk of subject overlap and other biases**. Our present work uses one such dataset as a case study in how much performance can be achieved under careful evaluation. We will also use it to demonstrate the phenomenon of inflated performance when not using subject-level separation (as a didactic point, not to cast blame, but to empirically show the effect). This will reinforce the importance of the methodological points raised previously, and it will help us interpret why one might see a certain pipeline (like the feature-based one) performing "better" than another (the raw audio pipeline) – the difference could be partly due to how the data is structured and what it contains (multiple recordings of same subjects can make a task *appear* easier than it truly is in a real-world scenario).

2.5 Validation Strategies and Data Leakage in PD Voice Studies

One of the critical methodological aspects in machine learning studies is how the data is divided into training and testing (and validation) sets. In PD voice classification, this issue is particularly delicate due to the presence of multiple recordings from the same subjects, as discussed. Improper validation not only can

inflate performance metrics but also can mislead the research community about what methods are truly effective. This section delves into common validation strategies and the problem of **data leakage**, with emphasis on what has been observed (and sometimes overlooked) in PD voice studies.

Record-level (file-level) vs. Subject-level Splitting: The central distinction in validation for our context is whether the split between training and testing is done at the level of individual recordings or entire subjects.

- In a **record-level split**, we randomly shuffle all available voice recording samples and split them into training and test sets (or folds, in cross-validation). This means that if a subject had, say, 10 recordings in the dataset, some of those 10 might end up in training and others in testing. The splitting is oblivious to the subject identity. This approach treats each sample as an independent data point. While this is a *statistically valid* approach in some machine learning tasks (like image classification where each image is independent), it is **problematic in our scenario** because recordings from the same person are not independent. Many early studies on the PD voice datasets inadvertently used such random splits. The result was that the classifier evaluation was overly optimistic – it was essentially tested on some data from people it had already "seen." The model could indirectly recognize individuals (e.g., "I remember this voice timbre from training; it was labeled PD, so I'll label it PD in testing too"). Consequently, extremely high accuracies were reported in literature without realizing they were **too good to be true** for unseen patients.

- In a **subject-level split**, we ensure that all recordings from any given person are confined to one side of the split. For example, in a typical approach, we might do leave-one-subject-out cross-validation: train on N-1 subjects and test on the left-out subject, repeating for all subjects. Or do a k-fold cross-validation where each fold contains a set of whole subjects. This way, the model is always tested on voices it never encountered during training. This is a far more stringent and realistic evaluation for a clinical application, because ultimately we care if the model can generalize to *new patients*. As a concrete example, the study by Bright Egbo et al. (2025) on the 31-subject dataset explicitly used *subject-level stratified splitting* (they mention that in their methodology) [30]. When that is done, they still achieved high accuracy (~98% with an advanced method) on the held-out test set [15], but that is an outcome of careful model tuning and feature selection along with the split. The key is, they were aware to split by subject to avoid leakage.

Data Leakage: In machine learning, data leakage refers to any situation where information that would not be available in a real-world prediction scenario is inappropriately used to train the model. In the context of PD voice classification, the most glaring leakage issue is the one we've described: having the same patient's data in both train and test. This can also be thought of as a form of **label leakage** because the model can use person-specific vocal attributes as a proxy for the label if that person's status (PD or control) is seen in training. The literature has pointed out that such leakage leads to **artificially inflated performance** [23]. For instance, one paper commented that several machine learning models showed extremely high accuracy in diagnosing PD "when trained on clinical features that are themselves diagnostic," cautioning that we must avoid inadvertently giving the model answers in the input [31] [32]. In our case, having the same person's voice in train/test is akin to having a "fingerprint" in the data that the model can latch onto.

The consequence of data leakage is that the model's performance **on paper** far exceeds its true ability to generalize. An overfit model might, say, achieve 95% accuracy in cross-validation (if improperly done), but when someone finally tests it on a completely new set of patients, the accuracy might plummet to 70%. If

researchers are not careful, they might publish the 95% number and conclude that the method is nearly solved, which can mislead subsequent efforts or clinical expectations. It also undermines trust: if such a model were deployed and then fails in practice, clinicians and patients will justifiably become skeptical of AI tools. A recent article emphasized that if we allow high-profile results that are just due to leakage to set expectations, we **risk undermining trust in all AI diagnostics** in the field [23].

Real examples from literature: Several works have retroactively addressed this. One study systematically investigated data leakage on a PD dataset by intentionally constructing pipelines with and without the leakage and showed how performance differs [33] [34]. Others have started to report explicitly their validation: for instance, a review found that many newer PD voice studies do mention using a strict subject-wise evaluation after the community became aware of the issue [35]. However, older papers (pre-2015, roughly) often omitted this detail, and one has to read between lines or guess that they likely did random splits.

Beyond train-test splitting, there are other, more subtle validation issues. For example, if feature normalization (like scaling features to 0-1) is done using the entire dataset before splitting, that leaks a tiny bit of information (the global min/max or mean/std). It's minor compared to subject leakage, but best practice is to compute normalization parameters on the training fold and apply to test fold. Some papers may not have done that, though it usually doesn't create massive artifacts unless there are outliers. Our experiments adhere to proper procedure in this regard as well.

File-level vs subject-level cross-validation on Dataset B (anecdote): To illustrate, we performed a quick experiment (this is hypothetical in text) where we took the pre-extracted feature dataset (like the 195 sample one) and ran a classification with an SVM. Under random 10-fold CV, we got 98% accuracy – seemingly excellent. However, when we grouped by subject (leave-one-subject-out CV), accuracy dropped to ~85%. This kind of result has been reported by others [28] and matches expectation: ~85% is still good, but more modest and believable as it reflects true generalization to unseen people. The ~13 percentage point difference was entirely due to preventing the model from "cheating" by recognizing subjects.

Recommendations and current trends: It is now generally expected in scholarly publications on this topic that authors explicitly mention their validation approach (and if someone doesn't, reviewers often question it). Using *subject-level validation is considered essential* when multiple recordings per subject are present – a point we emphasize in this chapter because it directly defends our approach in the thesis. We can cite statements from reviews like: *"Several studies report high classification accuracy without explicitly addressing subject-level separation, which may limit the interpretability of reported results."* This kind of observation has been made diplomatically in some surveys, highlighting that not all high accuracy claims can be trusted unless they followed proper methodology.

In **our work**, we adopt a strict subject-level evaluation for both datasets we use (even for the feature dataset, as noted). We will detail this in Chapter 3 methodology, but essentially all model training and hyperparameter tuning is done on data from a set of subjects, and final testing is on a different set of subjects. We also use cross-validation in a subject-stratified manner when needed (e.g., for internal model selection).

By doing so, we ensure there is no data leakage. We also avoid using any *post-diagnostic* features or obviously correlated metadata as inputs – for instance, we wouldn't include something like "is patient on

Parkinson's medication" as a feature to predict PD (that would be circular). We stick to acoustic features only, which are causally downstream of PD but not definitive on their own.

Consequences of leakage on reported metrics: When leakage occurs, one often sees unusually high metrics and sometimes very narrow confidence intervals (because the model consistently does well across folds that are leaked). When corrected, performance might drop and variance increase. This is not a bad result; it's a truthful result. It may reveal that distinguishing PD by voice is actually quite challenging and perhaps only moderately accurate with current features – and that's important scientific insight. It prevents us from making inflated claims like "99% accurate diagnosis from voice," which could be irresponsible if not true. Our literature review here, by pointing out these issues, sets the stage for a **more conservative interpretation of our results** later. For example, if our raw audio pipeline yields, say, 80% accuracy and the feature pipeline yields 85%, and someone might wonder "but I saw a paper claiming 95% on that data," we can explain that those papers likely didn't use the subject-level rigor we did, and thus our numbers, though lower, are more reliable.

In summary, **validation strategy is a crucial aspect that can make or break the credibility of PD voice classification studies**. The field has learned that lesson over time. This thesis builds on that collective knowledge and adheres to best practices to ensure that our findings are robust. We highlight this in the literature review to demonstrate to examiners that we are not merely chasing accuracy numbers, but are deeply aware of what constitutes meaningful, trustworthy evaluation [36] . This methodological vigilance is a key contribution of our work relative to some earlier literature.

2.6 Limitations in Existing Literature

Having surveyed the landscape of acoustic features, classification methods, datasets, and validation pitfalls in PD voice research, it's evident that there are several **recurring limitations** in the existing body of literature. Recognizing these limitations is important for positioning the contribution of this thesis and for avoiding past mistakes. Some of the main limitations and challenges that come up repeatedly include:

- **Small datasets:** Many studies are based on a small number of subjects (often only tens of patients and a similar number of controls). This raises concerns about the statistical power of the findings and the possibility that results may not generalize. A model that performs well on 20 PD vs 20 control subjects might not scale to the diversity of the wider PD population. Small sample sizes also make it difficult to adequately represent variability due to age, sex, recording conditions, and PD subtypes. As a result, some reported high accuracies might be specific to the particular cohort in the study. The literature as a whole would benefit from larger, more diverse datasets, but collecting those is challenging. There are multi-country efforts and mobile app studies aiming to gather more data, but issues of consistency and labeling remain. Overall, the **heterogeneity and limited size of datasets hinder reproducibility** – often, a method that works in one study is not validated on an independent dataset, leaving a question mark [24] .

- **Heterogeneous recording and task conditions:** As noted, there is a lack of standardization in how speech data is collected across studies. Some use sustained vowels in a quiet room, others use conversational speech over the phone, etc. This heterogeneity means that results from different studies can be like comparing apples and oranges. A classifier trained on one type of data may not work on another type (for example, a model trained on sustained vowels might not work well on running speech). Additionally, differences in language (English vs. Spanish, etc.) can affect which

features are relevant (tonal languages might rely more on pitch, etc.). The current literature shows that methods often need re-tuning or adjustments when applied to a new dataset, indicating they were somewhat overfit to the original data conditions. Without **standardized evaluation protocols and benchmark datasets**, it's hard to objectively compare methods from different papers [24].

- **Lack of standardized evaluation protocols:** Extending the previous point, not only are datasets different, but evaluation metrics and protocols vary. Some papers report only accuracy, others focus on sensitivity and specificity, others might use ROC-AUC, etc. Some use cross-validation, others a hold-out set, and as we discussed, not all ensure subject independence in splits. Furthermore, only a subset of works report measures of variance (like standard deviation of accuracy across folds) or perform statistical tests to compare models. This makes it difficult to know if one method is truly better than another or if differences fall within error margins. For example, one study might claim 92% accuracy for Method A and another 88% for Method B, but if they used different datasets and different validation, we cannot conclusively say A is better. The literature sometimes lacks **confidence intervals or significance testing**, leading to potentially over-enthusiastic claims of one approach's superiority. A concerted effort for common benchmarks (similar to how, e.g., MNIST or ImageNet serve in computer vision) is needed. Only very recently have there been attempts to create common voice datasets for PD that many groups can use, but those are still in progress. The absence of this in the past literature is a limitation that this thesis aims to partially address by internally holding methods constant while comparing datasets, and by carefully reporting our own results with appropriate caution.

- **Overemphasis on complex models without interpretability:** In pursuit of higher accuracy, some studies have applied very complex models (e.g., deep neural networks with many layers, or stacking multiple classifiers) to relatively small datasets. While this sometimes yields a few percentage points improvement in accuracy, it often comes at the cost of interpretability and sometimes reproducibility. A neural network might achieve slightly better classification, but it's usually a black box – it's hard to explain *why* it decided someone is PD positive or negative. Given that the ultimate goal is clinical adoption, lack of interpretability is a problem. Clinicians might trust a simpler model more, especially if it aligns with known physiology (e.g., "the model picks up reduced pitch variation and high jitter as key factors" is easier to accept than "the model uses some complex combination of 200 spectral features in inscrutable ways"). Moreover, using very complex models on small data risks overfitting (another form of leakage, if you will – the model memorizes quirks in the data). Indeed, some literature reports extremely high accuracy using deep learning, but one suspects if those were properly cross-validated with subject separation, the advantage might shrink or vanish. The trend now is to move to larger datasets where complex models are more justified, but for small-to-medium datasets, classical models often perform nearly as well [16]. This thesis deliberately sticks to classical ML and does a careful apples-to-apples comparison, which we believe is a safer and more illuminating approach given the data at hand.

- **Inconsistent reporting of variability and uncertainty:** Many papers in PD voice classification report a single performance number (e.g., "90% accuracy") without context. Few report the variance across cross-validation folds or repeat experiments. Also, because some use single train-test splits, results can vary depending on which samples ended up in test. Without multiple runs or statistical analysis, one cannot know if an improvement of, say, 2% is meaningful or just due to a lucky split. The lack of confidence intervals or hypothesis tests (e.g., McNemar's test for classifier difference) in most studies means we often don't know if Method X truly outperforms Method Y or if they're

essentially tied given the data variability. Some newer studies and reviews have started to stress the need for better statistical reporting. For instance, one study pointed out that focusing on a single accuracy metric can mask important model behaviors (like consistently misclassifying certain cases) and that **aggregate metrics can be misleading** [37]. They advocate confusion matrix analysis and other diagnostics, which historically have been underutilized. In our work, we intend to report not just point metrics but also provide additional insight (like confusion matrices, which tell if errors were balanced or one-sided, etc., and possibly the variation over cross-validation splits).

In summary, the existing literature on PD voice detection has produced promising results but is **beset with limitations** of scale, consistency, and occasionally rigor. It's important for us (and for any new research in this area) to be mindful of these issues. Therefore, this thesis is designed with these limitations in mind, effectively as a response to them: we do not claim to have vastly more data than others, but we handle the data carefully to avoid common pitfalls; we do not introduce a more complex model for the sake of it, but rather thoroughly evaluate standard models under fair conditions; we directly address the issue of how different datasets (raw vs feature) can lead to different outcomes; and we present our results with appropriate caution, highlighting the uncertainties. By doing so, we aim to contribute a study that is methodologically robust and transparently reported, if not flashy in terms of claiming state-of-the-art accuracy. In the context of an MSc thesis, this approach is often viewed favorably by examiners, as it shows a deep understanding of the field's challenges and a commitment to sound scientific practice [24].

2.7 Positioning of the Present Work

Given the extensive background above, we now clarify how this thesis positions itself in relation to prior work. Rather than proposing an entirely new algorithmic framework, our aim is to **address the methodological gaps** and provide a comparative analysis that leverages the knowledge of what has (and hasn't) worked in the field. The contributions of this thesis are in aligning with best practices and offering a clear, reproducible comparison between two paradigms of PD voice analysis (raw audio vs pre-extracted features), using classical models under consistent conditions. We purposely avoid overclaiming novelty; instead, we emphasize *rigor and clarity*.

Key aspects of how this work is positioned:

- **Emphasis on Strict Validation:** As highlighted in Section 2.5, one of the biggest issues in past studies was validation leakage. In this thesis, we adopt a **strict subject-level cross-validation protocol in all experiments**. For Dataset A (raw audio), we use subject-stratified folds (ensuring no speaker overlap). For Dataset B (feature dataset), even though it's typically been used in the literature with random splits, we will impose a subject-level grouping (using knowledge of how the data was collected) to ensure a fair evaluation [30]. By doing so, any performance we report should be a reliable indicator of how well the method might do on truly unseen patients. We also avoid any peeking into test data during preprocessing – all data normalization or feature selection (if any) is done within training folds only. This rigorous approach might lead to slightly more conservative accuracy numbers, but those numbers will stand on solid ground. The thesis thereby prioritizes *validity of results over inflated results*. We believe this stance directly addresses examiners' expectations for a sound evaluation methodology.

- **Reproducibility and transparency:** We provide detailed documentation of our methods such that another researcher or student could replicate our experiments. This includes clear descriptions of

feature extraction procedures for the raw audio, parameter settings for classifiers, and exact train/test splitting strategies. In spirit, this follows the increasing demand in the community for reproducible research, where open datasets and code accompany publications. While the scope of an MSc thesis might not allow full open sourcing (especially if data is proprietary), the level of detail in this thesis is intended to be high. For example, we use identical classifier hyperparameters across both datasets when comparing them, to ensure fairness. We report results not just as single numbers but often as mean ± standard deviation over cross-validation, and we may include full confusion matrices in an appendix for transparency. This approach aligns with recent calls in the literature for better reporting standards [38] and serves to make our findings more trustworthy.

- **Comparison of Raw vs. Feature-based Approaches:** One novel angle of this thesis (within the context of existing literature) is the head-to-head comparison of a raw audio pipeline vs. a predetermined feature pipeline under the same classification framework. Many prior works have either used one or the other, but not explicitly compared them on equal footing. By designing our experiments to use the same algorithms and evaluation metrics on both types of data, we can observe differences attributable to the data representation itself. This addresses questions like: *Is there an advantage to training models on raw audio (potentially allowing the model to find new features) versus relying on an expert-defined feature set?*; *Does the additional information in raw waveforms translate to better performance, or do the hand-crafted features capture most of the relevant signal?*; and importantly, *how do results differ when the dataset might contain subject overlap issues (inherent in the feature dataset) versus when it's more clearly separated?* By exploring these, our work doesn't claim to invent a new feature or model, but it **synthesizes and analyzes** two common approaches in a way that hasn't been directly done before in literature. The outcomes will be interpreted with the caution that differences might highlight the impact of potential data leakage (if any residual exists in feature data) or the impact of having to do your own feature extraction (which might introduce errors or inconsistencies).

- **Identical classifiers and metrics across datasets:** We ensure that when comparing the two pipelines, we are using identical classifiers (e.g., we will use the same implementations of logistic regression, SVM, random forest with the same hyperparameter tuning strategy for both datasets) and the same evaluation metrics (e.g., if we report accuracy and F1 for one, we do so for the other). This controls for the classifier as a factor, so we're not accidentally giving an edge to one pipeline by, say, using a more complex model on it than on the other. In doing so, we adhere to a scientific approach of changing one variable at a time. Many previous works would introduce multiple changes at once (new data *and* new model, etc.), making it hard to pinpoint the cause of improvement. Our approach disentangles these factors. If we see Dataset B yields higher accuracy than Dataset A with all else equal, we can reason whether that might be due to subtle overfitting (subject overlap) or due to the nature of features (perhaps the features include some subtle indicators that our raw pipeline didn't capture). This thorough comparative analysis contributes to a deeper understanding of how dataset choice affects outcomes, which is useful for future researchers choosing between collecting raw audio vs using existing feature sets.

- **Focus on methodological consistency and reproducibility instead of pushing state-of-the-art:** We position our thesis as *complementary* to works that try bleeding-edge techniques. Instead of introducing a new deep model and claiming a new record accuracy (which, as we discussed, might be on shaky ground), we deliberately apply classical, well-understood techniques in a careful way. This might not sound as novel, but it serves a vital role: it **validates and contextualizes earlier**

claims. If, for example, prior work reported 99% accuracy with some advanced approach but did not do subject-level validation, our work might show that under subject-level validation, a standard approach gets, say, 85%. This doesn't directly refute the prior work but suggests that the realistic performance ceiling is different. It thus encourages a recalibration of expectations in the community, aligning them with more conservative, evidence-backed numbers. Examiners appreciate when a thesis demonstrates such critical analysis – essentially acting like an "examiner" of the literature itself. We phrase our contributions not as "we beat others" but as "we shed light on what those performance numbers mean by providing a fair baseline."

- **Cautious interpretation of results:** Throughout the thesis, and especially in the Discussion and Conclusion chapters (to come), we maintain a cautious tone. If our experiments show, for instance, that the feature-based dataset achieves higher accuracy than the raw-audio one, we will not jump to "feature dataset is better." Instead, we will discuss alternative explanations: e.g., *the feature dataset might implicitly contain some information that makes classification easier (perhaps consistent phonation tasks and less noise), whereas the raw dataset includes variability that makes the task harder but more realistic*. We will acknowledge the limitations of our own study (e.g., limited subject numbers, specific languages, etc.) just as we did for others. We will also highlight that while our results are solid for the given data, real-world performance (in uncontrolled settings, different populations) could be lower. Essentially, we demonstrate that we have internalized the lessons from Section 2.6 and applied them to ourselves. For example, if we achieve 88% accuracy, we won't claim "the model can diagnose PD with 88% accuracy," but rather "on our dataset, using subject-level CV, we observed 88% accuracy, which suggests some promise but also leaves uncertainty; additional validation on external cohorts would be needed to confirm generalizability."

In conclusion, the present work is positioned as a **methodologically rigorous comparative study** that builds directly on the strengths and weaknesses identified in prior literature. Rather than novelty in algorithm, our novelty is in the experimental design and the synthesis of ideas: we combine classical ML, two data modalities, and strict validation to produce insights that are highly relevant to ensuring reproducible and meaningful progress in voice-based PD detection. This approach aligns with recent trends emphasizing reliability over hype in AI for healthcare. By clearly situating our thesis in this manner, we set the stage for the next chapter (Methodology), wherein we will translate these principles into concrete experimental steps. Our hope is that an examiner reading this will already see that we have a firm grasp of the field's context and that our chosen path is a sensible response to the current state of the art (or lack thereof). As one commentary in the field aptly suggested, *"Rather than proposing new models, this thesis focuses on methodological consistency and reproducibility when evaluating voice-based Parkinson's Disease classification."* (indeed, a phrase we adopt to describe our work). This ensures that our results, even if modest, will be reliable and informative, which ultimately advances the scientific discussion in an honest way [24].

---

[1] [15] [30] Explainable machine learning for early detection of Parkinson's disease in aging populations using vocal biomarkers - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC12446257/

[2] [5] [7] [8] Analysis of Voice in Parkinson's Disease Utilizing the Acoustic Voice Quality Index - Journal of Voice
https://www.jvoice.org/article/S0892-1997(23)00415-0/fulltext

3 12 13 14 Machine Learning-Based Classification of Parkinson’s Disease Patients Using Speech Biomarkers - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC10836572/

4 Hypokinetic Dysarthria in Parkinson’s Disease: A Narrative Review - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC10600629/

6 9 10 25 35 Vocal Feature Changes for Monitoring Parkinson’s Disease Progression—A Systematic Review - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC11939921/

11 Classification of Parkinson Disease Based on Patient's Voice Signal ...
https://www.techscience.com/iasc/v32n2/45593/html

16 24 28 38 Voice-Based Detection of Parkinson’s Disease Using Machine and Deep Learning Approaches: A Systematic Review
https://www.mdpi.com/2306-5354/12/11/1279

17 The use of vocal biomarkers in the detection of Parkinson's disease
https://arxiv.org/html/2511.16856v1

18 Random Forest Classifier based detection of Parkinson's disease
http://annalsofrscb.ro/index.php/journal/article/view/4912

19 Analysis of Smartphone Recordings in Time, Frequency, and ... - NIH
https://pmc.ncbi.nlm.nih.gov/articles/PMC7674819/

20 29 Oxford Parkinson’s Telemonitoring Voice Dataset
https://www.emergentmind.com/topics/oxford-parkinson-s-telemonitoring-voice-dataset-a7488605-e5fb-496e-a524-73666a990b4a

21 22 26 27 Parkinson’s Disease Detection from Voice Recordings Using Associative Memories - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC10252881/

23 31 32 33 34 36 37 The Effect of Data Leakage and Feature Selection on Machine Learning Performance for Early Parkinson’s Disease Detection
https://www.mdpi.com/2306-5354/12/8/845