ChatGPT

Chapter 9 – Conclusion

Summary of the Thesis Objectives

This thesis set out to explore the feasibility of classifying Parkinson’s Disease (PD) versus Healthy Control (HC) subjects using characteristics of their voice signals and classical machine learning methods. The primary objective was to develop and evaluate binary classification models based on acoustic features extracted from speech, in order to determine to what extent voice data can distinguish PD from healthy vocal patterns. A further goal was to compare different speech datasets and speaking tasks to understand how factors like sample size, feature set, and task type influence classification performance. Throughout, the emphasis remained on a rigorous, reproducible experimental approach rather than on achieving maximum accuracy, aiming to provide a cautious and methodologically sound assessment of voice-based PD detection.

Overview of the Experimental Framework

To address the objectives, two complementary datasets were utilized under a consistent evaluation framework. **Dataset A (MDVR-KCL)** consisted of raw audio recordings from approximately 36–37 individuals (PD patients and controls), each performing two speech tasks (one read text passage and one spontaneous dialogue). From these recordings, a set of 47 acoustic features was extracted per sample, capturing various properties of the voice (e.g. frequency measures, amplitude stability, etc.). Three classical classification algorithms – Logistic Regression, Support Vector Machine (with RBF kernel), and Random Forest – were trained and tested on Dataset A using a grouped **subject-level cross-validation** strategy. This means that folds were split by whole subjects rather than individual samples, ensuring that the same speaker’s voice did not appear in both training and testing sets. This procedure provided a more realistic evaluation at the cost of fewer effective training samples per fold.

In parallel, **Dataset B (PD_SPEECH_FEATURES)** was a larger corpus comprising 756 voice samples with 752 pre-computed features per sample (already extracted by the dataset creators). This dataset did not include explicit subject identifiers, so a stratified **sample-level cross-validation** (treating each recording as independent) was applied. The same three classifiers and performance metrics were employed on Dataset B to enable a comparative analysis. In both experiments, model performance was quantified using common classification metrics: accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (ROC-AUC). These metrics were reported as mean values with standard deviations across cross-validation folds. The experimental framework also included a comparison between the two speech tasks in Dataset A (read vs. spontaneous speech) to examine task-related effects on classification. This overall design allowed the thesis to systematically evaluate model behavior under different data conditions while adhering to rigorous validation protocols.

Synthesis of Main Findings

Voice Classification Feasibility and Variability (Dataset A): The results on Dataset A indicate that voice-based classification of PD is **feasible** but only achieved moderate accuracy under the given conditions, accompanied by high outcome variability. All three classifiers exhibited some ability to distinguish PD from healthy speech, yet the performance was limited and inconsistent across validation folds. In particular, the small number of subjects led to wide confidence intervals on all metrics – some folds yielded near-chance predictions while others were more accurate. This high variance (e.g. F1-score standard deviations on the order of 0.15–0.30) reflects the limited statistical power of a ~36-subject dataset and the strict subject-wise validation (which, by design, reduces the amount of training data per fold). An illustrative extreme case was observed with the SVM model: in several folds its ROC-AUC dropped below 0.5 (worse than random guessing), underscoring the **instability** of complex models in a small-sample, high-dimensional setting. These inconsistent outcomes should be interpreted cautiously – they suggest the presence of some disease-related vocal signal but are not definitive on their own.

Effect of Speech Task: Within Dataset A, a subtle trend emerged when comparing the two speech tasks. The spontaneous speech task yielded slightly higher classification metrics on average than the read text task, hinting at a higher discriminative potential in extemporaneous speech. This aligns with the intuitive expectation that spontaneous speaking might reveal more pronounced PD-related voice deficits (such as variations in prosody or fluency) than reading a fixed passage. However, the **overlapping confidence intervals** between the two task results mean this difference was not statistically significant. In practical terms, while spontaneous speech **suggests** a richer source of diagnostic features, the evidence remains inconclusive due to the small sample and variability. Thus, one cannot firmly claim that spontaneous speech is definitively more informative than read speech in this study – only that a promising trend was observed that would warrant further investigation with a larger cohort.

Performance on a Larger Feature-Rich Dataset (Dataset B): In contrast to Dataset A, the classification experiments on Dataset B showed **higher overall performance** across all models. With 756 samples and an extensive 752-dimensional feature set, the models achieved substantially better accuracy, F1, and ROC-AUC scores (e.g. on the order of ~0.85–0.94 ROC-AUC for the best models, compared to the 0.60–0.75 range in Dataset A). These results indicate that, under ideal conditions of abundant data and rich features, classical machine learning methods can capture PD-related vocal patterns with much greater success. The improvement is likely attributable to the **combined effects** of sample size and feature breadth: a larger dataset provides more examples of PD vs. healthy variations, and the diverse feature set (including detailed acoustic measures) offers the classifiers more informative cues to learn from. However, it is crucial to emphasize that these impressive metrics from Dataset B **must be interpreted with caution**. Because subject identities were not available, the cross-validation for Dataset B was done at the sample level – this raises the risk that some recordings from the same individual could appear in both training and testing splits. Such overlap (if present) could lead to overly optimistic performance, as models might inadvertently learn person-specific characteristics (e.g. a particular voice's timbre or recording nuances) rather than generalizable disease markers. In other words, the outstanding results on Dataset B represent a **benchmark under less stringent validation** rather than a guaranteed level of real-world performance. They demonstrate what the classifiers can achieve in a best-case scenario, but they likely **overestimate** true generalization to new, unseen patients. This finding underscores the importance of careful validation: higher accuracy in Dataset B does not necessarily equate to a truly better model, but rather to more favorable data conditions.

Cross-Dataset and Model Comparative Insights: Taken together, the findings from both datasets highlight how data characteristics influence outcomes. The gap in performance between Dataset A and Dataset B suggests that **larger sample sizes and richer feature representations substantially improve** the detectability of PD from voice, yet the methodological differences (especially the validation strategy) confound direct comparison. It would be misleading to conclude that the pre-extracted features of Dataset B are inherently superior to the handcrafted features of Dataset A without accounting for these factors. Additionally, the comparative evaluation of the three classifiers yielded some consistent patterns. Across both datasets, the Random Forest classifier emerged as the most **robust and reliable** performer, achieving high scores with relatively lower variance. This robustness can be attributed to the ensemble's inherent feature selection and its capacity to handle high-dimensional inputs, which likely helped it avoid overfitting in both the small and large dataset scenarios. Logistic Regression, while more modest in accuracy, provided stable and interpretable results, reflecting its simplicity and lower variance tendency. The SVM showed strong capability on the larger dataset but proved **sensitive** on the smaller dataset, where its performance became erratic. These observations imply that model choice should be guided by the data regime: simpler or ensemble methods may be preferable for limited data conditions, whereas more complex models can excel when sufficient data is available. Overall, the main findings of this thesis suggest that voice-based PD classification is **promising but challenging** – performance greatly depends on having adequate data and proper validation, and even then results should be interpreted within their methodological context.

Methodological Contributions

In addition to the empirical results, this research makes several methodological contributions to the study of voice-based medical classification. **First**, it demonstrates the critical importance of evaluation design by implementing rigorous **subject-level cross-validation** in Dataset A. This approach ensured that no individual's voice appeared in both training and test sets, thereby providing a more realistic assessment of model performance on truly unseen speakers. While this reduced the apparent accuracy (compared to a naive sample-level split), it delivered more trustworthy insights and highlights a best-practice for studies with multiple recordings per subject. **Second**, the thesis explored two distinct data processing pipelines – one starting from raw audio with manual acoustic feature extraction, and another leveraging a pre-featured dataset. By comparing these pipelines under a unified framework, the work illustrated how data preparation choices and feature representations can impact outcomes. This dual approach is a contribution in itself, showcasing the challenges and advantages of each pipeline: working from raw audio offers control and transparency in feature generation, whereas using a rich pre-extracted feature set can boost performance but may hide potential confounders. **Third**, the study provided a controlled, side-by-side comparison of multiple classical machine learning algorithms (logistic regression, SVM, random forest) on the same task. Such a comparative analysis, with all models evaluated under identical conditions and metrics, yields practical insight into their relative behavior for PD speech classification – for example, identifying the stability of Random Forest versus the sensitivity of SVM in a low-data regime. **Finally**, a strong emphasis was placed on **transparent feature extraction and reproducible methodology**. All feature definitions (e.g. jitter, pitch variation, MFCCs) were based on well-established acoustic measures documented in the literature, and the experimental procedures (from preprocessing to validation) were described in detail. This transparency facilitates replication of the results and builds trust in the findings. Taken together, these methodological elements strengthen the validity of the study's conclusions and serve as a reference for future researchers aiming to design robust voice classification experiments.

Limitations and Scope Boundaries

Despite its findings, this research has clear limitations that constrain the interpretation and generalizability of the results. **The foremost limitation is the small scale of Dataset A**, which included only on the order of tens of subjects. Such a limited sample size greatly increases uncertainty in the results – as evidenced by the large variance in performance across folds – and means that the reported metrics for Dataset A should be viewed as preliminary. The trends observed (e.g. the slight advantage of spontaneous speech, or the differences among models) are **indicative but not conclusive** given the wide confidence intervals. Moreover, the lack of an independent external test set for Dataset A implies that all evaluations were done via cross-validation on the same cohort; while we took care to avoid any data leakage through grouped CV, the true generalization of the models to entirely new patient populations remains unverified. In practical terms, the moderate accuracies achieved on Dataset A cannot be assumed to hold for all PD patients or all recording conditions – they pertain only to the specific group and tasks studied. **This thesis therefore does not establish a clinically actionable diagnostic system**, but rather a proof-of-concept within a controlled experimental scope.

For Dataset B, the key limitation lies in the **absence of subject identifiers and the potential confounding this introduces**. Because we could not group recordings by speaker, the evaluation may have been inadvertently easier than a real-world scenario where a model must generalize to new individuals. If multiple samples from the same subject were present in different folds, the classifier could partially leverage person-specific vocal qualities instead of purely detecting PD-related changes. This factor likely inflated the performance metrics on Dataset B, meaning the true error rate on new patients would be higher than suggested by our cross-validation. Thus, while Dataset B's results are useful as an upper bound, they **overstate the model's readiness for deployment**. More broadly, the two datasets and methodologies differ in ways that confine the scope of any direct comparisons – differences in how features were obtained, how subjects were sampled, and what speech was recorded all act as confounding variables. We explicitly refrain from claiming that any feature set or model is inherently superior based on this work, since the outcomes are context-dependent.

Finally, the scope of this thesis was deliberately focused on classical machine learning techniques and acoustic features from voice. It did not incorporate other potentially informative data modalities (for example, neurological exam scores, gait or handwriting data, etc.) nor did it explore modern deep learning approaches for feature learning. These choices were made to maintain interpretability and due to the dataset sizes available. Consequently, the findings are bounded to the scenario of classical supervised learning on acoustic features. There may well be patterns or improvements achievable outside this scope, but investigating those was beyond the aims of the current study. Recognizing these limitations and boundaries is important: the conclusions drawn here are **conditioned on the data and methods used** and should not be generalized beyond appropriate context. Any positive results are to be seen as encouraging signals rather than definitive evidence of a reliable diagnostic tool.

Future Work

Building on the insights and acknowledging the limitations of this thesis, several avenues for future research are suggested to advance voice-based PD classification:

- **Larger and More Diverse Datasets:** A top priority is to gather much larger speech datasets with PD and healthy subjects, ideally spanning diverse demographics and recording conditions. A greater number of subjects (with multiple recordings each) would improve statistical power and enable models to learn more generalized PD vocal characteristics. Larger cohorts would also allow exploration of more complex models (if warranted) without the severe overfitting risks encountered in small-sample settings.

- **Subject-Aware Evaluation Protocols:** Future studies should ensure **subject-level validation** whenever the data includes multiple samples per person. In cases similar to Dataset B, where samples arrive unlabeled by subject, efforts should be made to obtain that information or design experiments that approximate subject-independent testing. Adopting leave-one-subject-out or grouped cross-validation as a standard will yield more realistic performance estimates and prevent the optimistic bias observed under sample-wise splitting. This is crucial for developing models that truly generalize to new individuals.

- **Standardized Recording and Task Conditions:** It would be beneficial to record future speech data under more controlled and uniform conditions. Consistency in microphone quality, recording environment (noise levels), and elicitation tasks can reduce extraneous variability unrelated to PD. Additionally, incorporating carefully chosen speech tasks – including both scripted passages and free-form speech – in a balanced way could help identify which aspects of vocal performance are most sensitive to PD. A controlled protocol across subjects and sessions would strengthen the reliability of comparisons and feature measurements.

- **Expanded Feature Exploration:** There is ample room to broaden the spectrum of acoustic features and apply more advanced feature engineering techniques. For example, future work could include **dynamic speech features** such as delta and delta-delta MFCC coefficients (capturing changes in spectral features over time) or other time-series analyses of pitch and amplitude modulations. Non-linear signal features that quantify voice signal chaos or complexity (e.g. based on nonlinear dynamics or fractal measures) might capture subtle dysphonia characteristics not represented in standard features. At the same time, applying feature selection or dimensionality reduction methods (such as principal component analysis or genetic algorithms for feature subset selection) could help manage very high-dimensional feature sets, focusing models on the most informative attributes and potentially improving generalization.

- **Advanced Modeling Techniques:** While this thesis focused on classical algorithms, future research could investigate more complex machine learning models, including deep learning approaches, to automatically learn discriminative features from the raw audio. Neural network models (e.g. CNNs or RNNs applied to spectrograms or waveforms) may uncover patterns that handcrafted features miss. However, pursuing such models should be done cautiously: deep learning would require substantially more training data and careful regularization, and interpretability would become an important concern. Any gains in accuracy must be weighed against the need for transparency in medical decision support tools. Therefore, advanced models are a promising direction, but they should complement – not replace – rigorous validation and feature understanding.

- **External and Longitudinal Evaluation:** To move toward clinical relevance, future studies should validate findings on entirely independent datasets and, if possible, in longitudinal settings. Testing the developed models on a new cohort (collected at a different institution or later in time) would provide a stringent check on generalizability. Additionally, investigating how voice-based classifiers perform in tracking disease progression (for instance, analyzing the speech of PD patients over time or before/after treatments) could extend the utility of these methods beyond static diagnosis. Such longitudinal analysis was outside the scope of this work but represents a natural next step if reliable baseline classification is achieved.

By pursuing these future directions, researchers can build upon the foundation laid by this thesis. In particular, addressing data scale and validation rigor will **enhance the credibility** of voice-based PD classification models. Ultimately, the hope is that continued work in this area, with larger cohorts and refined methods, will clarify the true potential of non-invasive voice analysis as a tool for supporting PD diagnosis and monitoring – always keeping in mind the need for careful interpretation and thorough validation at each stage of development.