ChatGPT

Chapter 7: Discussion

7.1 Overview

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

7.2 Interpretation of Key Findings

7.2.1 Feature Extension Impact

The extension from 47 to 78 features produced significant performance improvements, particularly for the ReadText task where Random Forest ROC-AUC increased from 0.590 to 0.822 (+23 percentage points), essentially rescuing the model from chance-level performance. **SpontaneousDialogue**, which already performed well (0.828), saw a more modest improvement to 0.857.

Interpretation:

The extended features capture three complementary aspects of speech dynamics:

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

These additions are particularly relevant for PD detection because:

1. **Reduced variability** is a hallmark of PD speech (monotone)
2. **Temporal dynamics** are affected by motor control deficits
3. **Spectral flatness** may indicate breathiness/reduced harmonic content

The larger improvement for non-linear models suggests that the extended features enable modeling of **non-linear feature interactions** that simpler feature sets may obscure.

Robustness Check:
Although the extended feature set increased dimensionality relative to the small sample size of Dataset A (n=37), performance was evaluated exclusively using grouped cross-validation at the subject level. Improvements were observed consistent across folds and were accompanied by comparable standard deviations, suggesting that the observed gains reflect improved feature representation rather than fold-specific overfitting.

7.2.2 Class Weighting Effects

Class weighting showed **modest and inconsistent effects** on Dataset A:

Model	Δ ROC-AUC (weighted vs unweighted)
Random Forest	+3.5pp (baseline), -1.4pp (extended)
Logistic Regression	0.0pp
SVM (RBF)	-1.3pp (baseline), -1.4pp (extended)

Interpretation:

The moderate imbalance in Dataset A (57:43 HC:PD) is not severe enough to substantially degrade unweighted classifiers. Class weighting becomes more critical when:

- Imbalance exceeds 70:30
- Minority class has high cost of misclassification
- Sample size is very small

For Dataset B (25:75 imbalance), class weighting would likely have a larger effect, though this remains to be tested with subject-grouped CV.

7.2.3 Model Performance Hierarchy

Across all conditions, Random Forest consistently outperformed other models:

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

Interpretation:

Random Forest's advantages for this task include:

1. **Ensemble averaging** reduces variance on small datasets
2. **Feature importance** provides interpretability
3. **Non-linear decision boundaries** capture complex patterns
4. **Robustness** to irrelevant features through feature subsampling

7.2.4 High Variance Across Folds

Standard deviations frequently exceeded 0.15 (15%), indicating substantial fold-to-fold variability.

Causes:

1. **Small sample size** (37 subjects → ~7 subjects per test fold)
2. **Subject heterogeneity** in disease severity

3. **Recording variability** (smartphone recordings)

Implications:

- Absolute performance numbers should be interpreted cautiously
- Relative comparisons across conditions are more reliable
- Confidence intervals overlap for many comparisons

7.3 Comparison with Literature

7.3.1 Performance Context

Study	Dataset	Best ROC-AUC	Method
Little et al. (2009)	UCI	0.92	SVM
Sakar et al. (2013)	Custom	0.86	SVM
This thesis	**MDVR-KCL**	**0.87**	**RF**

Our results are competitive with the literature, though direct comparison is limited due to:

- Different datasets and features
- Different CV strategies (many studies do not use grouped CV)
- Different sample sizes

7.3.2 Methodological Comparison

Aspect	Typical Literature	This Thesis
CV Strategy	Random split	Grouped stratified
Subject handling	Often ignored	Explicit grouping
Feature selection	Ad-hoc	Systematic ablation
Reporting	Best result only	All conditions

Our grouped CV approach provides **more conservative** but **more realistic** estimates of generalization performance.

7.4 Feature Importance Analysis

7.4.1 Most Discriminative Features

The top features across models consistently include:

Feature	Category	Relevance to PD
f0_max	Pitch	Reduced pitch range in PD
delta_mfcc_2_mean	Spectral dynamics	Temporal variability
autocorr_harmonicity	Voice quality	Breathiness indicator
shimmer_apq3	Perturbation	Amplitude instability
intensity_mean	Prosody	Hypophonia marker

7.4.2 Feature Category Contributions

Feature Importance by Category - ReadText

Figure 7.1: Aggregated importance by feature category (Random Forest, ReadText).

The analysis reveals:

- **MFCC features** contribute most to classification
- **Pitch features** (F0) are consistently important
- **Formant variability** (F1–F3 std) shows moderate importance

7.4.3 Cross-Task Stability

Comparing ReadText and SpontaneousDialogue tasks:

Feature Importance - Spontaneous

Figure 7.2: Feature importance by category for SpontaneousDialogue task.

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

7.5 Implications

7.5.1 For Feature Engineering

The success of extended features suggests that future work should:

1. **Include variability measures** (std, range) alongside means
2. **Capture temporal dynamics** (delta, delta-delta)
3. **Provide spectral shape descriptors** (centroid, rolloff)

7.5.2 For Model Selection

Random Forest is recommended for similar tasks due to:

- Robustness on small datasets
- Built-in feature importance
- Good handling of mixed feature types

7.5.3 For Evaluation Protocols

Grouped cross-validation should be **mandatory** when:

- Multiple recordings exist per subject
- Subject identifiers are available
- Generalization to new subjects is the goal

7.6 Addressing Research Questions

7.6.1 RQ1: ML Model Performance

> **How do classical ML models perform on PD voice classification?**

Classical ML achieves ROC-AUC up to 0.873, demonstrating feasibility of voice-based PD detection. Random Forest outperforms linear models.

7.6.2 RQ2: Feature Extension Impact

> **Does feature set extension improve classification performance?**

Yes. Extending from 47 to 78 features improved ROC-AUC by **+8.7 percentage points** (Random Forest). The improvement is most pronounced for non-linear models.

7.6.3 RQ3: Class Weighting Impact

> **Does class weighting improve performance on imbalanced datasets?**

Marginally. On Dataset A (moderate imbalance), class weighting improved Random Forest ROC-AUC by **+3.5 percentage points** with baseline features but showed inconsistent effects elsewhere.

7.6.4 RQ4: Cross-Dataset Comparison

> **How do results compare between Dataset A and Dataset B?**

Dataset B typically shows higher performance, likely due to:
- Larger sample size
- Potential subject overlap (unmeasurable)
- Different feature sets

Direct comparison is limited by these confounds.

7.7 Unexpected Findings

7.7.1 SVM Performance Variability

SVM (RBF) showed high variance and occasional fold-level failures (ROC-AUC < 0.5 in some folds). This suggests:

- Sensitivity to hyperparameters (not tuned in this study)
- Potential kernel mismatch for this feature space
- Need for larger training sets

7.7.2 Limited Benefit of Weighting with Extended Features

When using extended features, class weighting provided **no additional benefit** (and sometimes slightly reduced performance). This suggests that the richer feature representation already captures minority class characteristics effectively.

7.8 Summary

Key discussion points:

1. **Feature extension is the primary driver of improvement** (+8.7pp ROC-AUC)
2. **Random Forest is the most robust model** for this task
3. **Grouped CV provides conservative estimates** but ensures validity
4. **Class weighting has modest effects** on moderately imbalanced data
5. **High variance** necessitates cautious interpretation of absolute numbers

Chapter 8: Limitations and Threats to Validity

8.1 Overview

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

8.2 Sample Size Limitations

8.2.1 Dataset A: Small Subject Pool

Metric	Value
Total subjects	37

Metric	Value
Subjects per test fold	~7
PD subjects (minority)	15–16

Implications:

- High variance in fold-level metrics (std > 0.15 common)
- Limited statistical power for detecting small effects
- Results may not generalize to broader populations

8.2.2 Effect on Statistical Confidence

With 37 subjects and 5-fold CV:

- Each fold has only ~7 test subjects
- A single misclassification shifts accuracy by ~14%
- Confidence intervals are wide by design

Mitigation: Results focus on **relative comparisons** rather than absolute performance claims.

8.3 Subject Identifier Limitations

8.3.1 Dataset B: Missing Subject IDs

Dataset B (PD_SPEECH) provides no subject identifiers. This creates potential for:

- **Subject leakage:** Same subject in train and test sets
- **Optimistic bias:** Inflated performance estimates
- **Unknown generalization:** Cannot assess new-subject performance

Caveat Statement:

> "Results on Dataset B may be optimistic due to unknown subject overlap across folds. The absence of subject identifiers prevents validation of true out-of-subject generalization."

8.3.2 Comparison Limitations

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

- Different CV strategies
- Different feature dimensionalities (78 vs 752)
- Different sample sizes (37 vs 756)

8.4 Feature Extraction Limitations

8.4.1 Deterministic Feature Set

The feature set was designed *a priori* based on literature review, not data-driven optimization. Limitations include:

- **Potentially suboptimal features:** Other features may be more discriminative
- **Fixed parameters:** Librosa/Parselmouth defaults used without tuning
- **No feature selection:** All 78 features used without reduction

8.4.2 Audio Quality Assumptions

Feature extraction assumes:

- Reasonable signal-to-noise ratio
- Consistent recording conditions
- No severe clipping or distortion

The MDVR-KCL dataset's smartphone recordings may violate these assumptions.

8.5 Model Limitations

8.5.1 No Hyperparameter Tuning

All models used default or fixed hyperparameters:

Model	Fixed Parameters
Logistic Regression	C=1.0, max_iter=1000
SVM (RBF)	C=1.0, gamma='scale'
Random Forest	n_estimators=100, max_depth=10

Implications:

- Performance may be suboptimal
- Results represent lower bounds
- Tuned models might change rankings

Rationale for not tuning: Nested CV on 37 subjects would lead to extreme variance; fixed parameters ensure reproducibility.

8.5.2 Classical ML Only

This thesis explicitly excludes deep learning. Potential missed opportunities:

- End-to-end learning from spectrograms
- Transfer learning from speech models
- Attention mechanisms for temporal modeling

Rationale: Deep learning typically requires larger datasets and offers reduced interpretability.

8.6 Methodological Limitations

8.6.1 No External Validation

All results use internal cross-validation. Limitations:

- No held-out test set from a different source
- No multi-site validation
- Generalization to clinical settings unknown

8.6.2 Binary Classification Only

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

- Disease severity prediction
- Progression monitoring
- Differential diagnosis (PD vs other conditions)

8.6.3 Single Speech Tasks

Each task was analyzed separately. Not addressed:

- Task fusion strategies
- Multi-task learning
- Optimal task selection

8.7 Threats to Validity

8.7.1 Internal Validity

Threat	Status	Mitigation
Subject leakage	Controlled (Dataset A)	Grouped CV
Label noise	Unknown	Assumed correct
Feature bugs	Possible	Unit tests, manual verification

8.7.2 External Validity

Threat	Status	Mitigation
Population bias	Likely	Document dataset demographics
Recording variability	Present	Standardized extraction
Temporal stability	Unknown	Single recording session

8.7.3 Construct Validity

Threat	Status	Mitigation
Feature relevance	Assumed	Literature-based selection
Metric appropriateness	Addressed	Multiple metrics reported
Class definition	Accepted	Binary PD/HC from source

8.8 Reproducibility Considerations

8.8.1 Strengths

- Fixed random seeds (42)
- Version-controlled code
- CLI-based pipeline
- Documented dependencies

8.8.2 Limitations

- Library version drift may affect results
- Hardware differences in audio processing
- Dataset access may change

8.9 Interpretation Guidelines

Given the limitations, results should be interpreted as follows:

8.9.1 Appropriate Claims

"Extended features improved ROC-AUC on this dataset"
"Random Forest outperformed other models under these conditions"
"Grouped CV provides more conservative estimates than random splits"

8.9.2 Inappropriate Claims

"This system diagnoses Parkinson's Disease"
"82.6% accuracy is clinically sufficient"
"These results will generalize to other populations"

8.10 Future Work to Address Limitations

Limitation	Potential Solution
Small sample size	Multi-site data collection
Missing subject IDs	Require IDs in future datasets
No hyperparameter tuning	Bayesian optimization with nested CV
No external validation	Independent test cohort
Classical ML only	Careful deep learning with augmentation

8.11 Summary

This research has significant limitations including:

1. **Small sample size** (37 subjects) leading to high variance
2. **Missing subject IDs** in Dataset B preventing leakage control
3. **No hyperparameter tuning** potentially limiting performance
4. **No external validation** limiting generalization claims
5. **Binary classification only** excluding severity/progression

These limitations are acknowledged to ensure appropriate interpretation of results. Despite these constraints, the methodology prioritizes **validity over optimization**, providing a rigorous foundation for future work.

Chapter 9: Conclusion

9.1 Summary of Work

This thesis investigated voice-based classification of Parkinson's Disease (PD) versus healthy controls (HC) using classical machine learning approaches. The work addressed key methodological challenges in the field, including subject-level data leakage, class imbalance, and feature representation.

9.1.1 Contributions

1. **Rigorous Evaluation Framework**
2. Implemented grouped stratified cross-validation to prevent subject leakage
3. Systematic 2×2 factorial design (features × class weighting)

4. Transparent reporting of all conditions with confidence intervals

5. **Feature Engineering Investigation**

6. Extended feature set from 47 to 78 acoustic features
7. Demonstrated +8.7 percentage point ROC-AUC improvement

8. Identified most discriminative features (F0, MFCCs, harmonicity)

9. **Class Weighting Analysis**

10. Evaluated `class_weight="balanced"` across all models
11. Found modest effects on moderately imbalanced data

12. Documented interaction between features and weighting

13. **Reproducible Pipeline**

14. CLI-based tools for feature extraction and experiments
15. Fixed random seeds and documented parameters
16. Complete code repository with documentation

9.2 Key Findings

9.2.1 Primary Results

Finding	Evidence
Best ROC-AUC: 0.873 ± 0.137	Random Forest, Extended Features
Feature extension improves performance	+8.7pp ROC-AUC (baseline → extended)
Random Forest outperforms other models	Highest ROC-AUC across all conditions
Grouped CV is essential	Prevents optimistic bias from subject leakage

9.2.2 Best Configuration

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

9.2.3 Feature Importance Insights

The most discriminative features for PD detection include:

1. **f0_max** — Maximum fundamental frequency (pitch ceiling)
2. **delta_mfcc_2_mean** — Spectral dynamics
3. **autocorr_harmonicity** — Voice quality measure
4. **shimmer_apq3** — Amplitude perturbation
5. **intensity_mean** — Overall vocal intensity

These align with known clinical manifestations of PD: reduced pitch range, monotonous speech, and hypophonia.

9.3 Research Questions Answered

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

Classical ML achieves **ROC-AUC up to 0.873** with Random Forest on the MDVR-KCL dataset using grouped cross-validation. This demonstrates the feasibility of voice-based PD screening, though performance varies substantially across folds due to small sample size.

RQ2: Does feature set extension improve classification performance?

Yes. Extending from 47 baseline features to 78 features improved ROC-AUC by **+8.7 percentage points** for Random Forest. The additional features capturing spectral variability (MFCC std), temporal dynamics (delta-delta MFCC), and spectral shape contributed to this improvement.

RQ3: Does class weighting improve performance on imbalanced datasets?

Modestly. On Dataset A (57:43 imbalance), class weighting improved Random Forest ROC-AUC by **+3.5 percentage points** with baseline features. However, effects were inconsistent across models, and no benefit was observed when combined with extended features.

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

Dataset B (standard CV, no subject IDs) showed higher absolute performance than Dataset A (grouped CV), consistent with potential optimistic bias from subject leakage. **Grouped CV provides more conservative but more realistic estimates** of out-of-subject generalization.

9.4 Implications

9.4.1 For Researchers

- **Use grouped CV** when multiple recordings per subject exist
- **Include variability features** (std, delta-delta) in feature sets
- **Report all conditions** rather than cherry-picking best results
- **Acknowledge limitations** transparently

9.4.2 For Practitioners

- Voice-based PD screening is feasible but not yet clinical-grade
- Random Forest provides a robust baseline for similar tasks
- Feature interpretability supports clinical understanding
- Results require validation on independent cohorts

9.4.3 For Dataset Creators

- **Always include subject identifiers** to enable proper CV
- Document recording conditions and equipment
- Provide demographic information
- Consider longitudinal designs

9.5 Limitations Recap

Key limitations that bound the interpretation of results:

1. **Small sample size** (37 subjects) creates high variance
2. **No hyperparameter tuning** may underestimate potential
3. **Single dataset source** limits generalization claims
4. **Binary classification only** — no severity prediction
5. **No external validation** on independent test set

9.6 Future Directions

9.6.1 Short-term Extensions

- Hyperparameter optimization with nested CV
- Feature selection to reduce dimensionality
- Multi-task fusion (ReadText + SpontaneousDialogue)
- Additional acoustic features (wavelets, TQWT)

9.6.2 Medium-term Research

- External validation on independent datasets
- Deep learning with appropriate regularization
- Longitudinal tracking of disease progression
- Multi-class classification (severity levels)

9.6.3 Long-term Vision

- Integration into smartphone applications
- Multi-modal biomarkers (voice + gait + tremor)
- Personalized baselines for individual tracking
- Clinical validation studies

9.7 Closing Remarks

This thesis demonstrates that **voice-based Parkinson’s Disease classification is feasible** using classical machine learning with carefully engineered acoustic features. The **+8.7 percentage point improvement** from feature extension highlights the importance of capturing speech dynamics beyond simple statistical summaries.

However, the field faces significant challenges:

- Small datasets require rigorous methodology
- Subject identity must be tracked for valid evaluation
- Clinical deployment requires extensive validation

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

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

Appendix A: Feature Importance Tables

A.1 Overview

This appendix presents the top-20 most important features for each experimental condition, as determined by model-native importance measures:

- **Logistic Regression:** Absolute coefficient values
- **Random Forest:** Gini importance (mean decrease in impurity)

A.2 Dataset A — ReadText Task

A.2.1 Random Forest — Top 20 Features

Rank	Feature	Importance	Std
1	f0_max	0.052	0.019
2	delta_mfcc_2_mean	0.039	0.018
3	f3_std	0.038	0.011
4	autocorr_harmonicity	0.038	0.017
5	intensity_mean	0.035	0.021

Rank	Feature	Importance	Std
6	f0_mean	0.032	0.012
7	shimmer_apq3	0.032	0.013
8	mfcc_12_mean	0.031	0.007
9	f1_std	0.031	0.022
10	mfcc_6_mean	0.030	0.024

Visualization:

Feature Importance - ReadText - Random Forest

Figure A.1: Top-20 feature importances for Random Forest on ReadText task.

A.2.2 Logistic Regression — Top 20 Features

Rank	Feature	Coefficient	Std
1	f0_max	0.754	0.203
2	hnr_mean	0.649	0.178
3	shimmer_apq11	0.553	0.145
4	delta_mfcc_4_mean	0.496	0.163
5	delta_mfcc_2_mean	0.492	0.103
6	delta_mfcc_1_mean	0.474	0.273
7	mfcc_5_mean	0.470	0.127
8	mfcc_4_mean	0.426	0.233
9	mfcc_10_mean	0.418	0.221
10	mfcc_11_mean	0.388	0.192

Visualization:

Feature Importance - ReadText - Logistic Regression

Figure A.2: Top-20 feature importances for Logistic Regression on ReadText task.

A.2.3 Feature Importance by Category

Feature Importance by Category - ReadText

Figure A.3: Aggregated feature importance by category for ReadText task.

A.2.4 Cross-Model Heatmap

Feature Importance Heatmap - ReadText

Figure A.4: Normalized feature importance heatmap comparing models on ReadText task.

A.3 Dataset A — SpontaneousDialogue Task

A.3.1 Random Forest — Top 20 Features

Rank	Feature	Importance	Std
1	mfcc_5_mean	0.080	0.022
2	shimmer_apq11	0.069	0.007
3	delta_mfcc_8_mean	0.051	0.015
4	jitter_local	0.041	0.012
5	delta_mfcc_2_mean	0.040	0.018
6	autocorr_harmonicity	0.037	0.011
7	shimmer_local	0.036	0.017
8	mfcc_1_mean	0.034	0.010
9	f0_std	0.032	0.022
10	f0_mean	0.031	0.013

Visualization:

Feature Importance - Spontaneous - Random Forest

Figure A.5: Top-20 feature importances for Random Forest on SpontaneousDialogue task.

A.3.2 Logistic Regression — Top 20 Features

Rank	Feature	Coefficient	Std
1	mfcc_5_mean	0.722	0.039
2	delta_mfcc_8_mean	0.615	0.121
3	shimmer_apq11	0.559	0.136

Rank	Feature	Coefficient	Std
4	delta_mfcc_2_mean	0.493	0.196
5	intensity_min	0.459	0.170
6	mfcc_3_mean	0.388	0.271
7	delta_mfcc_11_mean	0.381	0.102
8	delta_mfcc_7_mean	0.380	0.150
9	hnr_mean	0.379	0.175
10	delta_mfcc_1_mean	0.352	0.123

Visualization:

Feature Importance - Spontaneous - Logistic Regression

Figure A.6: Top-20 feature importances for Logistic Regression on SpontaneousDialogue task.

A.3.3 Feature Importance by Category

Feature Importance by Category - Spontaneous

Figure A.7: Aggregated feature importance by category for SpontaneousDialogue task.

A.3.4 Cross-Model Heatmap

Feature Importance Heatmap - Spontaneous

Figure A.8: Normalized feature importance heatmap comparing models on SpontaneousDialogue task.

A.4 Dataset B — PD Speech Features

A.4.1 Random Forest — Top 20 Features

Rank	Feature	Importance	Std
1	std_delta_log_energy	0.013	0.004
2	std_delta_delta_log_energy	0.013	0.003
3	tqwt_entropy_shannon_dec_12	0.012	0.001
4	tqwt_TKEO_std_dec_11	0.010	0.004

Rank	Feature	Importance	Std
5	tqwt_TKEO_mean_dec_12	0.010	0.001
6	mean_MFCC_2nd_coef	0.008	0.003
7	tqwt_entropy_log_dec_11	0.008	0.003
8	tqwt_stdValue_dec_12	0.008	0.003
9	tqwt_stdValue_dec_13	0.008	0.003
10	tqwt_energy_dec_12	0.007	0.003

Note: Dataset B uses 752 pre-extracted features including TQWT (Tunable Q-factor Wavelet Transform) coefficients not present in Dataset A.

Visualization:

Feature Importance - PD Speech - Random Forest

Figure A.9: Top-20 feature importances for Random Forest on Dataset B.

A.4.2 Logistic Regression — Top 20 Features

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

Visualization:

Feature Importance - PD Speech - Logistic Regression

Figure A.10: Top-20 feature importances for Logistic Regression on Dataset B.

A.5 Cross-Task Feature Consistency

A.5.1 Features Appearing in Top-10 Across Multiple Tasks

Feature	ReadText RF	Spontaneous RF	Consistent
f0_mean	Rank 6	Rank 10	
delta_mfcc_2_mean	Rank 2	Rank 5	
autocorr_harmonicity	Rank 4	Rank 6	
shimmer_apq3/local	Rank 7	Rank 7	

A.5.2 Interpretation

The consistency of certain features (F0, delta MFCCs, harmonicity, shimmer) across tasks suggests these capture **task-general** acoustic signatures of Parkinson's Disease rather than task-specific artifacts.

A.6 Feature Category Summary

A.6.1 Category Rankings by Aggregated Importance

Category	ReadText	Spontaneous	Overall
MFCC	1	1	**1**
Pitch (F0)	2	3	**2**
Shimmer	4	2	**3**
Delta MFCC	3	4	**4**
Formants	5	6	**5**
Harmonicity	6	5	**6**

A.6.2 Key Observation

MFCC-based features (mean, std, delta) consistently dominate across all tasks and models, indicating the importance of spectral envelope characteristics for PD voice classification.

Appendix B: Detailed Results Tables

B.1 Overview

This appendix provides complete numerical results for all experimental conditions, including per-fold breakdowns and task-level performance.

B.2 Condition 1: Baseline Features (47) + Unweighted

Output directory: `outputs/results/baseline/baseline/`

B.2.1 Summary Statistics

Model	Accuracy	Precision	Recall	F1	ROC-AUC
LogisticRegression	0.696 ± 0.133	0.657 ± 0.262	0.702 ± 0.284	0.655 ± 0.246	0.781 ± 0.152
SVM_RBF	0.703 ± 0.143	0.603 ± 0.357	0.545 ± 0.390	0.547 ± 0.347	0.635 ± 0.311
RandomForest	0.744 ± 0.173	0.653 ± 0.373	0.638 ± 0.421	0.615 ± 0.369	0.786 ± 0.235

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

Output directory: `outputs/results/baseline/extended/`

B.3.1 Summary Statistics

Model	Accuracy	Precision	Recall	F1	ROC-AUC
LogisticRegression	0.699 ± 0.152	0.660 ± 0.289	0.641 ± 0.314	0.630 ± 0.281	0.783 ± 0.126
SVM_RBF	0.757 ± 0.143	0.703 ± 0.330	0.651 ± 0.354	0.657 ± 0.321	0.726 ± 0.265
RandomForest	0.826 ± 0.122	0.814 ± 0.255	0.760 ± 0.327	0.759 ± 0.271	0.873 ± 0.137

B.3.2 Improvement over Baseline

Model	Δ Accuracy	Δ ROC-AUC
LogisticRegression	+0.3pp	+0.2pp
SVM_RBF	+5.4pp	**+9.1pp**
RandomForest	+8.2pp	**+8.7pp**

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

Output directory: `outputs/results/weighted/baseline/`

B.4.1 Summary Statistics

Model	Accuracy	Precision	Recall	F1	ROC-AUC
LogisticRegression	0.687 ± 0.141	0.654 ± 0.271	0.696 ± 0.280	0.649 ± 0.247	0.781 ± 0.152
SVM_RBF	0.748 ± 0.115	0.690 ± 0.314	0.670 ± 0.333	0.659 ± 0.299	0.622 ± 0.316
RandomForest	0.736 ± 0.141	0.664 ± 0.315	0.660 ± 0.393	0.628 ± 0.322	0.821 ± 0.191

B.4.2 Effect of Weighting (vs Condition 1)

Model	Δ Accuracy	Δ ROC-AUC
LogisticRegression	-0.9pp	0.0pp
SVM_RBF	+4.5pp	-1.3pp
RandomForest	-0.8pp	**+3.5pp**

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

Output directory: `outputs/results/weighted/extended/`

B.5.1 Summary Statistics

Model	Accuracy	Precision	Recall	F1	ROC-AUC
LogisticRegression	0.724 ± 0.136	0.687 ± 0.283	0.696 ± 0.307	0.670 ± 0.270	0.783 ± 0.126
SVM_RBF	0.757 ± 0.165	0.718 ± 0.338	0.693 ± 0.319	0.693 ± 0.309	0.712 ± 0.305
RandomForest	0.801 ± 0.146	0.798 ± 0.259	0.760 ± 0.327	0.748 ± 0.268	0.859 ± 0.162

B.5.2 Effect of Weighting (vs Condition 2)

Model	Δ Accuracy	Δ ROC-AUC
LogisticRegression	+2.5pp	0.0pp
SVM_RBF	0.0pp	-1.4pp
RandomForest	-2.5pp	-1.4pp

B.6 Cross-Condition Comparison Matrix

B.6.1 Random Forest ROC-AUC

	Baseline Features	Extended Features
Unweighted	0.786 ± 0.235	**0.873 ± 0.137**
Weighted	0.821 ± 0.191	0.859 ± 0.162

B.6.2 Random Forest Accuracy

	Baseline Features	Extended Features
Unweighted	74.4% ± 17.3%	**82.6% ± 12.2%**
Weighted	73.6% ± 14.1%	80.1% ± 14.6%

B.7 Statistical Significance Notes

B.7.1 Confidence Interval Overlap

Due to high standard deviations (often > 0.15), confidence intervals overlap across many comparisons. This limits the ability to make strong statistical claims about differences between conditions.

B.7.2 Practical Significance

Despite overlapping CIs, the consistent pattern of:
- Extended > Baseline features
- Random Forest > other models

...suggests **practically meaningful** differences even if not statistically significant at conventional thresholds.

B.8 Raw Data Files

All results are available in CSV format:

```
outputs/results/
├── baseline/
│   ├── baseline/
│   │   ├── all_results.csv      # Per-fold, per-metric details
│   │   └── summary.csv          # Aggregated statistics
```

```
│   ├── extended/
│   │   ├── all_results.csv
│   │   └── summary.csv
│   ├── importance_readtext.csv  # Feature importance (baseline)
│   ├── importance_spontaneous.csv
│   └── importance_pd_speech.csv
│
└── weighted/
    ├── baseline/
    │   ├── all_results.csv
    │   └── summary.csv
    └── extended/
        ├── all_results.csv
        └── summary.csv
```

B.8.1 CSV Column Descriptions

all_results.csv:
| Column | Description | |--------|------------------------------| | model | Classifier name | | fold | CV fold number (1-5) | | metric | Evaluation metric | | value | Metric value | | dataset | Dataset name | | task | Speech task (ReadText/SpontaneousDialogue) |

summary.csv:
| Column | Description | |--------|--| | model | Classifier name | | metric | Evaluation metric | | mean | Mean across folds | | std | Standard deviation across folds | | mean_std | Formatted string (mean ± std) |
