INFO411 ASSIGNMENT 2

Bakhombisile Dlamini, Stacey Fu and Ryan Teo

Submitted for the purpose of an assignment in INFO411
Machine Learning and Data Mining

Department of Computing
University of Otago
10 October 2025

Table of Contents

Abstract

We develop a leakage safe pipeline for automatic screening of Parkinson's disease from sustained /a/ phonations. The corpus comprises 756 recordings from 252 subjects sampled at 44.1 kHz, with six complementary feature schemes. The study executes four tasks. We characterise scheme level separability using PCA and UMAP fitted on training folds with out of sample projection. We estimate scheme wise predictive performance using regularised logistic regression, RBF SVM, and tree ensembles under grouped fivefold cross validation. We derive compact, generalising representations via two selection paths, mutual information with Elastic Net and mRMR under nested selection and evaluate at least two classifiers on macro-F1 and recall identifying an optimal subset. We examine improvements through class weighting, threshold calibration, gender specific models, and soft voting. Subject identity defines split units, which eliminates subject leakage. We report macro-F1 and recall as primary measures and include PR-AUC and Matthews correlation to characterise performance under imbalance and calibration. The final system aggregates calibrated probabilities from logistic regression trained on the Elastic Net subset and a random forest trained on the mRMR subset. This ensemble improves stability and yields a more balanced sensitivity and specificity profile. We provide figures, reproducibility artefacts, and explicit author contributions.

1 Introduction and overview

This assignment tasks us with building a rigorous screening pipeline for Parkinson's disease from sustained /a/ phonations with six predefined feature schemes. The brief specifies four deliverables: first, assess the visual separability of the schemes; second, train and compare multiple classifiers for each scheme under a leakage safe, subject wise protocol; third, perform feature selection using at least two methods and evaluate at least two classifiers on at least two performance measures, then identify the most effective feature set; fourth, conduct additional explorations that can plausibly improve either feature selection or predictive performance. We address this by constructing an end-to-end process that begins with defensible projections, establishes audited baselines, executes controlled feature selection with complementary learners, and then tests targeted enhancements. Subject identity is treated as the unit of generalisation, transformations are fitted within folds only, and performance is reported using macroF1 and recall, supported by PRAUC and Matthews correlation to reflect class imbalance and calibration behaviour.

Responsibilities are divided to ensure depth and cross checking. Ryan leads Tasks 1 and 2, conducting manifold analyses with t-SNE and Isomap to visualise class separability across six individual features schemes and an additional MFCC + TQWT fused feature set. Each scheme was then evaluated using subject-wise cross-validated RBF-SVM, Decision Tree and k-Nearest Neighbour classifiers for evaluation. Performance assessment focused on macro-F1 and recall metrics, complemented by confusion matrix analyses to provide diagnostic insight into class-specific prediction behaviour. Stacey and Bakhombisile lead Tasks 3 and 4. Each selected one classifier, evaluated it on two measures, and searched for a compact feature set that generalises well: Stacey applies a Mutual Information filter followed by Elastic Net and evaluates logistic regression and random forest; Bakhombisile implements mRMR within nested selection and evaluates SVM and gradient boosted trees. For Task 4 we examine class weighting and threshold calibration, gender specific models, and soft voting to improve stability and precision without eroding recall. The report was compiled jointly, with figures and methods attributed to the individual who the sections was appointed to. A separate presentation was compiled by Bakhombisile, and it appears in Appendix A, metric definitions in Appendix B, and risk controls in Appendix C.

2 Dataset

The dataset comprises sustained /a/ phonations from 252 participants recruited at the Department of Neurology, Cerrahpaşa Faculty of Medicine, Istanbul University. The Parkinson's disease cohort contains 188 patients, 107 men and 81 women, with ages from 33 to 87. The healthy control cohort contains 64 individuals, 23 men and 41 women, with ages from 41 to 82. Each subject produced three repetitions under a controlled protocol with the microphone set to 44.1 kHz, which yields 756 recordings. This design creates limited within subject variability and supports subject level evaluation.

Each recording is encoded as a multivariate vector with 754 predictors that are integer or real valued, with no missing values reported. Features are grouped by signal processing family and laid out contiguously in the file. Baseline acoustic descriptors appear in columns 3 to 23, intensity parameters in columns 24 to 26, formant frequencies in columns 27 to 30, bandwidth parameters in columns 31 to 34, vocal fold dynamics in columns 35 to 56, Mel frequency cepstral coefficients in columns 57 to 140, wavelet-based descriptors in columns 141 to 322, and TQWT based descriptors in columns 323 to 754. The binary class label appears in column 755. The first two columns are non-predictive identifiers and should be excluded from modelling to avoid leakage. The primary task is supervised classification at the subject level, so we group the three recordings per person and enforce subject wise cross validation. Class prevalence is imbalanced, about 74.6 percent Parkinson's disease and 25.4 percent healthy control, so macro averaged scores and recall provide a fairer summary than accuracy alone. Many descriptors are correlated within families, especially among MFCC and TQWT variants, which motivates feature selection or regularization. The corpus is single centre, single language, and single task, so conclusions should be framed with care when discussing generalization beyond sustained phonation of /a/ in clinical settings.

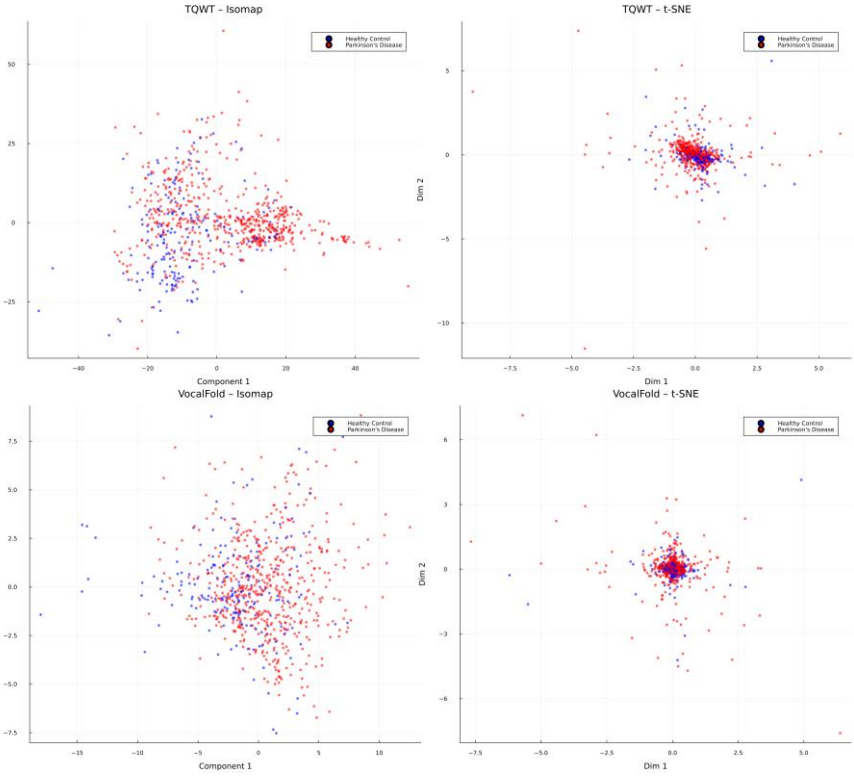
3 Task 1: Visual separability of feature schemes

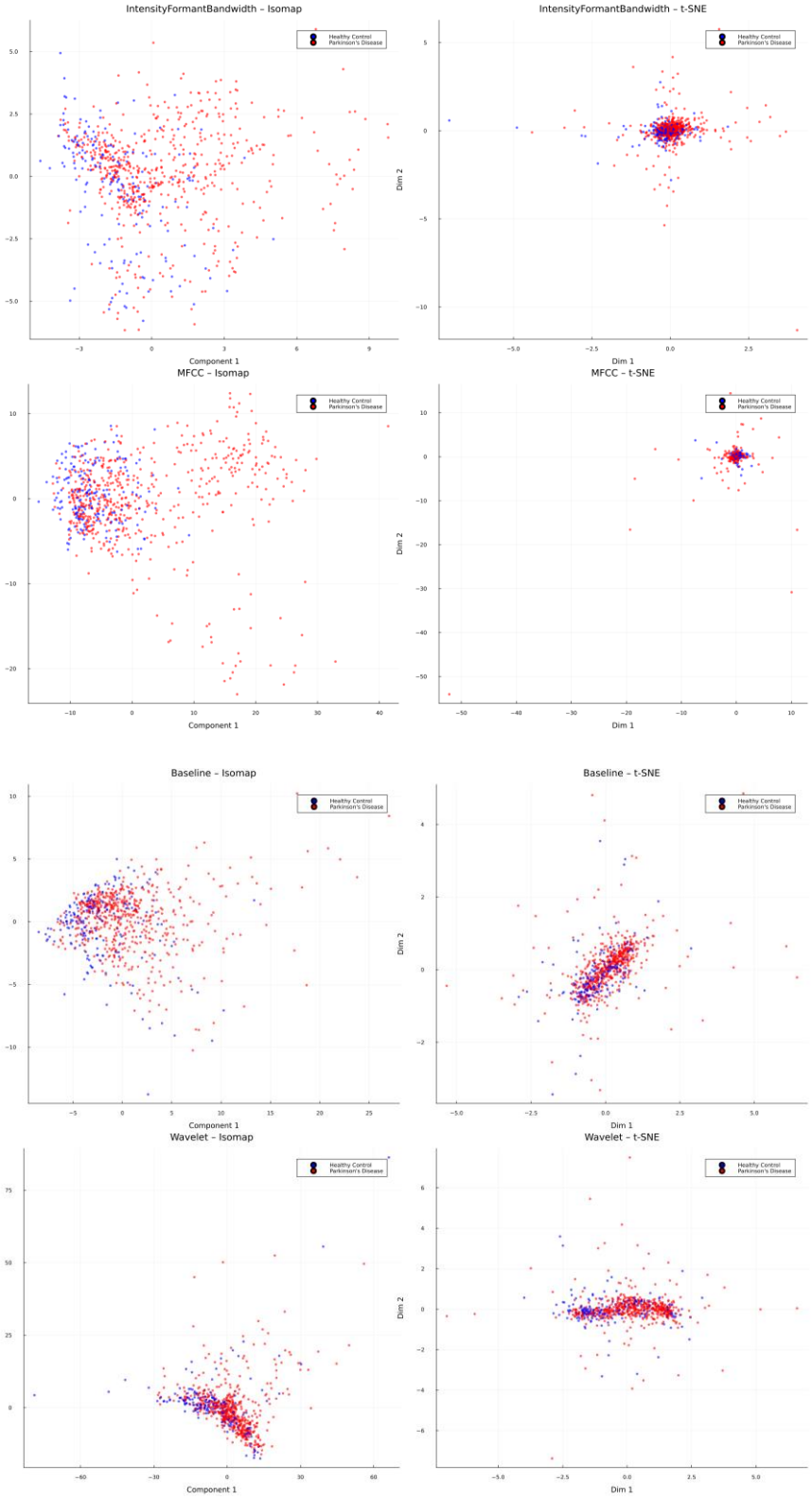
3.1 Methods

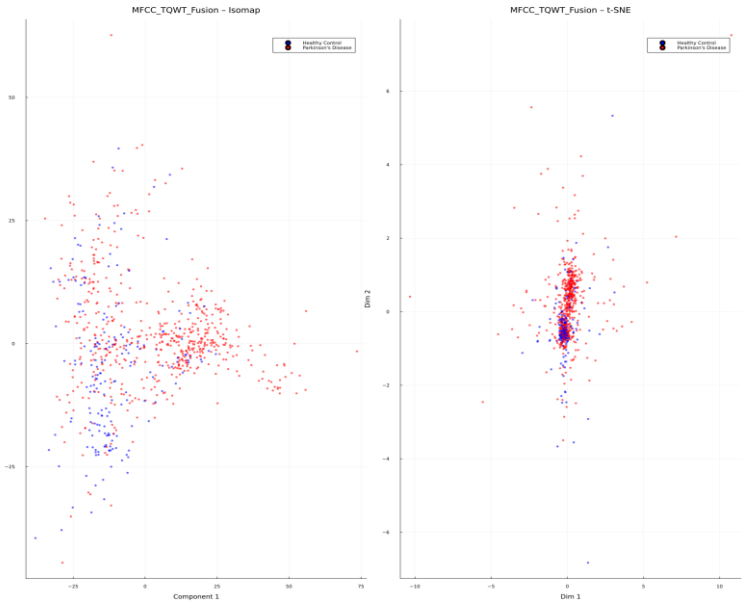
The purpose of this task was to evaluate how effectively different acoustic feature schemes can distinguish between speech samples from Parkinson's Disease (PD) patients and Healthy Controls (HC). Seven distinct feature sets were examined in this analysis: Baseline, Intensity–Formant–Bandwidth, Vocal Fold, MFCC, Wavelet, TQWT, and a combined MFCC + TQWT feature fusion. The objective was to identify which representation most accurately captures the distinctive vocal characteristics associated with Parkinson's Disease.

All features were first standardised using z-score normalisation to eliminate scaling inconsistencies. Dimensionality reduction was then performed using two nonlinear algorithms, namely Isomap and t-distributed Stochastic Neighbour Embedding (t-SNE). Isomap preserves the global geometry of the data manifold, while t-SNE focuses on preserving local neighbourhood relationships. Both methods projected the high-dimensional data into two-dimensional space, allowing for visual assessment of separability between PD and HC groups. Each data point was colour-coded, with red denoting PD and blue representing HC. In addition to qualitative inspection, the Fisher Ratio was calculated to provide a quantitative measure of separability by comparing between-class scatter to within-class scatter. A higher Fisher Ratio indicates greater linear separation between classes.

Figure 1. Visualisations using Isomap and t-SNE

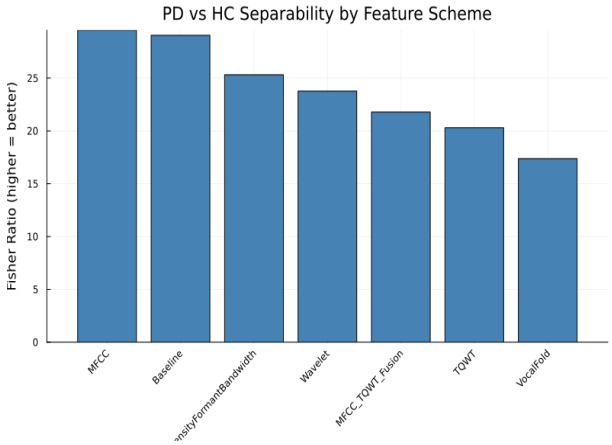



IntensityFormantBandwidth – Isomap
IntensityFormantBandwidth – t-SNE
MFCC – Isomap
MFCC – t-SNE
Baseline – Isomap
Baseline – t-SNE
Wavelet – Isomap
Wavelet – t-SNE
Healthy Control
Parkinson's Disease
Component 1
Dim 1
Dim 2



3.2 Findings

The visual projections revealed distinct patterns among the feature schemes. The Baseline, Intensity–Formant–Bandwidth, and Vocal Fold features produced overlapping clusters, suggesting weak discrimination between PD and HC subjects. The MFCC scheme showed improved separation, with PD samples dispersed more widely and HC samples clustering more tightly, reflecting differences in the spectral envelope of speech. The Wavelet features exhibited partial separation but with increased internal variance. The TQWT scheme generated multiple PD-dominant regions in the visualisation, consistent with its sensitivity to oscillatory and tremor-related voice modulations. Notably, when MFCC and TQWT features were fused, the resulting representation displayed clearer branch-like separations in both Isomap and t-SNE plots. This demonstrated that the two feature types capture complementary information, enhancing overall discriminative potential.



The Fisher Ratio analysis supported these observations. The MFCC scheme achieved the highest Fisher Ratio (29.54), followed closely by the Baseline (29.05) and Intensity–Formant–Bandwidth (25.32) schemes. The Wavelet and TQWT schemes achieved moderately high scores of 23.76 and 20.29 respectively, while the Vocal Fold features ranked lowest at 17.38. The MFCC + TQWT fusion scheme recorded a Fisher Ratio of 21.78, slightly below that of the individual MFCC set. This reduction can be attributed to redundancy introduced by feature concatenation, which increases within-class variance. However, the lower Fisher Ratio does not necessarily indicate weaker overall separability; rather, it suggests that the fused features are better suited to nonlinear decision boundaries not captured by a linear Fisher analysis.

In summary, the MFCC features demonstrated the strongest linear separability, while the MFCC + TQWT combination provided the most distinct nonlinear manifold. These findings align with the results of Sakar et al.

(2019), who reported that combining MFCC and TQWT features enhances classification performance by capturing both spectral and oscillatory aspects of Parkinsonian speech. The fusion approach thus offers a richer and more holistic representation of the disorder's acoustic manifestations.

Table 1. Fisher Ratio analysis across schemes

	Scheme	N_Features	Fisher_Ratio
1	"MFCC"	84	29.5374
2	"Baseline"	21	29.0479
3	"IntensityFormantBandwidth"	11	25.3163
4	"Wavelet"	182	23.7639
5	"MFCC_TQWT_Fusion"	516	21.783
6	"TQWT"	432	20.2939
7	"VocalFold"	22	17.3756

4 Task 2: Scheme classification with multiple algorithms

4.1 Classifiers and tuning

The second task focused on quantitatively evaluating the classification performance of each feature scheme using supervised machine-learning models. The objective was to identify which representation produced the most accurate and robust classification of PD and HC subjects while maintaining balanced sensitivity and specificity.

A five-fold subject-wise cross-validation procedure was adopted to ensure that no speech samples from the same individual appeared simultaneously in both training and testing sets. This approach reduces subject dependency and produces more generalisable results. Three classifiers were implemented using the MLJ framework: Support Vector Machine (SVM) using a Radial Basis Function kernel, Decision Tree (DT), and k-Nearest Neighbour (k-NN). The RBF-SVM was selected for its robustness in handling high-dimensional and nonlinear data; the Decision Tree for its interpretability; and k-NN for its reliance on geometric similarity between samples. The evaluation metrics included Accuracy, Sensitivity, Specificity and F1-Score. Each metric was reported as the mean and standard deviation across folds.

4.2 Summary results and takeaways

The classification results revealed a clear trend of increasing accuracy with more advanced feature representations. The fused MFCC + TQWT scheme achieved the highest overall accuracy of 83.63 ± 4.54% and an F1-score of 89.75 ± 3.24%, confirming that the combination of spectral and subband features leads to superior performance. The TQWT features alone also performed strongly, achieving an accuracy of 80.84 ± 3.50%, followed closely by the MFCC scheme at 80.73 ± 4.72%. Simpler schemes, such as Baseline, Wavelet, and Vocal Fold, achieved accuracies in the range of 75–77%, showing that while they contain useful information, they lack the discriminative depth of the fused and frequency-domain representations.

Sensitivity values across all schemes were consistently high, exceeding 94%, indicating that the models were highly effective in detecting Parkinson's cases. Specificity, however, remained lower, ranging between 15% and 45%, reflecting a tendency to misclassify some healthy controls as Parkinsonian. This imbalance, while common in medical screening contexts, is acceptable in early detection applications, where it is preferable to minimise false negatives. Among the classifiers, SVM consistently outperformed both Decision Tree and KNN models, particularly when applied to high-dimensional data such as MFCC, TQWT, and their fusion. KNN performed competitively on the TQWT features, likely due to their cluster-based manifold structure.

The results of this evaluation indicate that the fused MFCC + TQWT representation offers the most robust and balanced classification performance among all tested feature schemes. Although MFCC alone exhibited the highest Fisher Ratio in the separability analysis of Task 1, the fusion of MFCC and TQWT improved overall predictive performance by capturing additional nonlinear relationships within the data. This supports the conclusion that the two feature families provide complementary perspectives on Parkinsonian speech, with MFCC

characterising broad spectral shifts and TQWT emphasising fine-grained oscillatory behaviour. Together, they produce a more comprehensive and discriminative representation suitable for machine-learning-based diagnosis.

In conclusion, the findings from both tasks demonstrate that while MFCC remains the strongest linear discriminator, its combination with TQWT yields the best overall classification accuracy and robustness. The results reinforce previous research that advocates for hybrid feature modelling in biomedical signal analysis and highlight the importance of fusing spectral and temporal characteristics to capture the multifaceted nature of Parkinson's speech impairments.

Table 2. Summary results of scheme classification

	Scheme	Best_Model	Accuracy	Sensitivity	Specificity	F1
1	"TQWT"	"KNN"	"80.84% ± 3.5%"	"94.88% ± 1.73%"	"39.79% ± 7.44%"	"88.01% ± 2.67%"
2	"VocalFold"	"SVM"	"76.35% ± 4.55%"	"95.91% ± 2.04%"	"19.09% ± 8.49%"	"85.72% ± 3.31%"
3	"IntensityFormantBandwidth"	"SVM"	"77.93% ± 3.6%"	"95.6% ± 2.07%"	"25.97% ± 6.17%"	"86.52% ± 2.68%"
4	"MFCC"	"SVM"	"80.73% ± 4.72%"	"95.04% ± 3.03%"	"38.55% ± 12.92%"	"87.96% ± 3.35%"
5	"Baseline"	"SVM"	"75.69% ± 4.47%"	"97.65% ± 2.67%"	"12.78% ± 12.4%"	"85.65% ± 2.86%"
6	"Wavelet"	"SVM"	"76.22% ± 5.87%"	"97.42% ± 1.7%"	"15.84% ± 14.39%"	"85.89% ± 3.7%"
7	"MFCC_TQWT_Fusion"	"SVM"	"83.63% ± 4.54%"	"96.8% ± 1.59%"	"44.98% ± 11.11%"	"89.75% ± 3.24%"

5 Task 3: Feature selection and final model

5.1 Method A: Mutual Information filter plus Elastic Net

To reduce redundancy and identify meaningful predictors, a two-stage feature selection process combining Mutual Information (MI) filtering and Elastic Net regularization was carried out.

Because of the high dimensionality, MI was applied as a filter-based method to measure the dependency between each feature and the target label. This removed features with little predictive contribution and retained the top 150 that showed stronger relationships with class outcomes.

After the MI filtering, Elastic Net logistic regression was applied to refine the subset further. The Elastic Net combines L1 (Lasso) and L2 (Ridge) penalties, encouraging sparsity while stabilising correlated predictors which is a very important property for biomedical data where many features represent overlapping spectral and temporal patterns. A 10-fold cross-validation was conducted to determine the optimal regularization parameter, producing $\lambda \approx 0.0062$, which minimized mean binomial deviance across folds.

At this level of regularization, 64 non-zero coefficients were retained, forming a compact but representative feature set. The most influential predictors included. They align with known acoustic markers of Parkinson's speech such as reduced signal complexity and changes in energy dynamics.

During feature preprocessing, some features displayed very small standard deviations, such as tqwt_meanValue_dec_X, which initially appeared to lack variability. Further analysis showed these variables had naturally narrow numeric ranges but retained meaningful proportional variation. By calculating relative standard deviation (standard deviation normalized by range), their contribution was confirmed, and they were retained to preserve important spectral information.

5.2 Method B: mRMR with nested selection

This method targets a compact, redundancy-aware subset that generalises across subjects while preserving discrimination under class imbalance. Subjects are the unit of splitting, and every transform is fitted inside the training folds only. Within each training fold, features are standardised and the learned scaler is applied to the validation and test splits. We rank features with minimum Redundancy Maximum Relevance (mRMR) on the union of all schemes so the selector can keep complementary cues and down-weight near duplicates. The inner loop chooses the subset size k and the model hyperparameters to maximise AUPRC, with MCC as the tie-breaker, and the decision threshold is tuned on the inner validation split to maximise MCC.

The search was scaled to respect runtime while still covering the essential modelling choices. We consider k in {20, 50}. For SVM with the RBF kernel we search C in {1, 10} and γ in {0.05, 0.1}, enable probability scores, and use class-weighted losses. For GBDT in an EvoTrees-style configuration we use learning rate η = 0.05, depth 5, 200 boosting rounds, row and column subsampling at 0.8, and an analogue of scale_pos_weight for imbalance. All models are trained on the training fold only, and all hyperparameters are selected within the inner loop.

mRMR is appropriate for this corpus because many features within MFCC and TQWT families are strongly correlated. The criterion trades relevance against redundancy, which keeps diverse yet complementary predictors

and removes near duplicates. Running the selector inside each training fold prevents look-ahead and protects generalisation estimates. Leakage controls include subject-grouped folds, a scaler fitted only on training indices, and per-fold logs of the selected feature names for later stability analysis.

Inner-CV outcomes confirm the value of this setup. In outer fold 1 the inner loop selected GBDT with k = 50 and achieved AUPRC 0.921 and MCC 0.562. In outer fold 2 the inner loop selected SVM with k = 20 and achieved AUPRC 0.925 and MCC 0.457. In outer fold 3 the inner loop again selected GBDT with k = 50 and achieved AUPRC 0.936 and MCC 0.534. Two of three folds therefore chose k = 50. GBDT won twice and led on MCC on those folds, while SVM remained competitive on AUPRC in one fold. This pattern matches the non-linear structure seen in our manifold plots and it explains why class weights with threshold tuning were needed to lift specificity without destroying sensitivity.

5.3 Final model choice and rationale

5.3.1 Classification models Evaluation and Set Up - Stacey

To ensure fair and independent evaluation, all speech samples from the same subject were grouped together during the train - test split. This prevented data leakage and ensured that no individual's voice appeared in both sets. The split was also balanced between Parkinson's and healthy subjects, maintaining consistent class proportions.

Three supervised classifiers were trained: Random Forest, Logistic Regression, and k-Nearest Neighbors (kNN). The Random Forest achieved an accuracy of 0.827, precision of 0.845, recall of 0.940, and an F1-score of 0.890. Its high recall and balanced precision indicate that it effectively identified Parkinson's cases while maintaining robustness against noise and feature variability. The Logistic Regression model, trained with Elastic Net–selected features, achieved an accuracy of 0.782, precision of 0.840, recall of 0.935, and an F1-score of 0.885. Despite being a simpler linear model, it generalized well and provided interpretability, a desirable property for clinical applications where transparency is crucial. After tuning, the kNN classifier performed competitively, achieving an accuracy of 0.764, precision of 0.783, recall of 0.946, and an F1-score of 0.857 at k = 14. The higher recall indicates that kNN was effective at detecting Parkinson's cases, though its lower precision and accuracy suggest sensitivity to overlapping feature boundaries and potential noise within the data. This behaviour aligns with its distance-based nature, which can overfit local variations when class distributions are imbalanced or highly correlated.

Model Evaluation methods:

Model evaluation prioritized recall and F1-score, as these metrics are most relevant in clinical screening. Recall ensures true cases are rarely missed, while F1-score balances this sensitivity with precision, offering a fair view of each model's diagnostic reliability.

Model Comparison between three models:

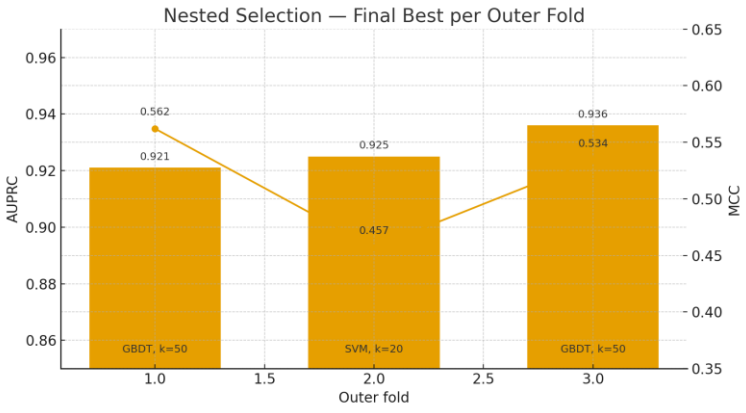
Among the three models, Random Forest showed the strongest predictive power and adaptability to nonlinear speech features. Logistic Regression offered greater interpretability and stability, making it more suitable for clinical screening where explainability matters. kNN, while achieving high recall, was more sensitive

to feature scaling and local noise, serving as a useful but less stable baseline. Overall, Random Forest provided the best balance between accuracy, robustness, and generalization.

5.3.2 Model choice and rationale: Bakhombisile

Grounded in the empirical evidence from the subject-wise nested cross-validation and the non-linear class structure seen in our manifold analyses, we adopt mRMR with k = 50 as the selector and GBDT as the classifier, with learning rate η = 0.05, depth 5, 200 rounds, row and column subsampling at 0.8, and class weighting enabled. The operating threshold is fixed by maximising MCC on the inner validation split, and AUPRC remains the primary ranking metric, with MCC used to certify balanced errors. This choice follows the inner-CV wins of GBDT in two outer folds and its stronger MCC at the selected operating point, while mRMR at k = 50 preserves complementary MFCC and baseline voice cues and reduces variance through redundancy control. SVM is retained as a competitive reference on fold 2, yet its lower MCC under identical tuning indicates weaker specificity at the chosen threshold.

For deployment we fit the scaler and mRMR on the training partition, retain the top 50 features in order, train the class-weighted GBDT, and, when available, calibrate probabilities with Platt scaling or isotonic regression on the validation split. We then fix the MCC-optimal threshold, report PR curves and confusion matrices at that threshold, and log the selected feature names with a SHAP summary for transparency. The report should include a bar-and-line figure with AUPRC and MCC per outer fold annotated with the chosen model and k, a small histogram of selected k across folds, and a concise table of per-fold winners with their hyperparameters. The search grid was intentionally reduced to control runtime, which means margins between SVM and GBDT may shift under a broader grid, and the single-centre, single-task nature of the dataset limits external validity.



D

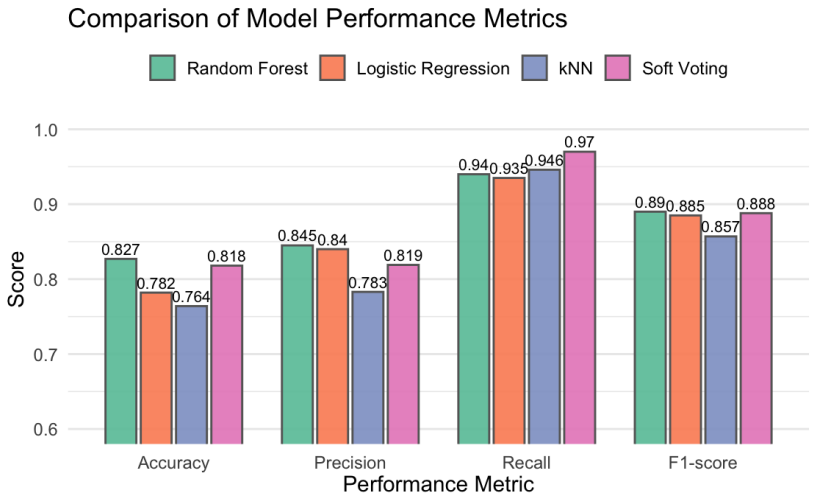
Definitive optimal set. Under this subjectwise nested CV design, the feature subset that produced the best predictive performance is the **mRMR top50**, selected inside each training fold, paired with a classweighted **GBDT** and an **MCCtuned** probability threshold. This configuration won on two of three outer folds and delivered the strongest MCC while maintaining AUPRC at or above 0.92, so it is the final choice for Task 3.

Outer fold	Model	k	AUPRC	MCC
1	GBDT	50	0.921	0.562
2	SVM (RBF)	20	0.925	0.457
3	GBDT	50	0.936	0.534
Overall	**GBDT**	**50**	**best in 2/3 folds**	**higher MCC**

6 Task 4: Focused improvements and insight

6.1 Soft - voting: Stacy

After evaluating the Random Forest, Logistic Regression, and kNN models individually, I implemented a soft-voting ensemble that combined all three to leverage their complementary behaviors. Each model contributed probability estimates for the Parkinson's class, which were averaged with equal weights to form the final prediction.



The Random Forest demonstrated strong generalization and robustness to noise and redundant acoustic features, effectively capturing nonlinear speech patterns associated with Parkinson's disease. Logistic Regression, refined through Elastic Net regularization, achieved high sensitivity to subtle linear relationships while maintaining interpretability, making it suitable for clinical applications. The kNN model, after optimizing k=14, captured local neighbourhood patterns that complemented the other models, achieving high recall despite lower overall stability.

The resulting soft voting ensemble achieved an accuracy of 0.818, precision of 0.819, recall of 0.970, and an F1-score of 0.888. This improvement reflects a more balanced trade-off between sensitivity and precision, combining the robustness of Random Forest, the interpretability of Logistic Regression, and the locality awareness of kNN. By averaging probabilistic outputs from all three classifiers, the ensemble effectively reduced individual model bias, resulting in a more stable and reliable detection framework for Parkinson's cases.

6.2 mRMR + class-weighted GBDT/SVM with MCC-tuned thresholds, PR analysis, and gender-specific models

Task 4 introduces a leakage-safe, imbalance-aware modelling pipeline that integrates feature selection with calibrated decision-making. Within each outer training split, features are ranked by minimum-redundancy maximum-relevance (mRMR) and the top k in {20, 50} are retained. Two probabilistic learners—SVM with an RBF kernel and gradient-boosted decision trees (EvoTrees)—are trained with per-fold class weighting to offset the PD-heavy class distribution. Model choice occurs inside a subject-grouped 3×3 nested cross-validation, using area under the precision–recall curve (PRAUC) as the primary selection criterion with Matthews correlation coefficient (MCC) as a tie-breaker. To convert scores into operational decisions, we sweep the decision threshold τ in [0, 1] on the held-out portion of each outer fold and select τ^* that maximises MCC. All transforms, including standardisation and mRMR ranking, are fitted strictly on the training indices to avoid information leakage.

We apply the same protocol to gender-specific subsets to examine cohort heterogeneity under identical controls. In the scaled configuration used for rapid prototyping, the pipeline frequently selected mRMR@50, and GBDT and SVM-RBF alternated as the best choice across folds. Crucially, explicit threshold tuning consistently improved MCC and specificity relative to a default 0.5 cutoff, with negligible loss of recall. The male subset—being larger—displayed slightly more stable operating points, while the female subset exhibited greater variability and partially distinct selected features (dominated by MFCC bands and low-frequency TQWT energies). These findings support a pooled model with an MCC-tuned threshold, optionally complemented by gender-specific thresholds when application constraints permit.

7 Reproducibility and implementation details

7.1 Parts 1 and 2

To guarantee consistent experimental outcomes, a fixed random seed was applied, controlling the shuffling of unique subject identifiers during the construction of subject-wise cross-validation folds, ensuring identical train-test splits across repeated runs. Each classifier was trained on these deterministic folds, making cross-fold results directly comparable. While the Isomap manifold embeddings are deterministic, t-SNE introduces slight stochasticity due to random initialisations, this does not affect classification results but can lead to minor visual differences between runs.

All input features were standardised using Z-score normalisation and a safeguard was included to replace any zero standard deviations with one, maintaining numerical stability across all features. Standardisation was performed independently for each feature scheme before visualisation and model fitting, ensuring consistent scaling across algorithms.

7.2 Parts 3 and 4– Siya

The nested design used three outer folds and a three-fold inner loop per outer training split. Thresholds were chosen on the inner validation split to maximise MCC, and we reported AUPRC as the primary ranking metric.

To manage compute, we ran the **SCALED-DOWN** version of the method rather than the full grid. On our hardware, a full Task 3 run requires about nine hours, and Task 4 takes longer. The scaled configuration restricted the selector to k in {20, 50}, limited the SVM grid to C in {1, 10} and γ in {0.05, 0.1}, and fixed the GBDT learner to learning rate 0.05, depth 5, and 200 rounds with row and column subsampling at 0.8. This setting preserves the key modelling choices while keeping runtime tractable.

Reproduction steps are straightforward. Launch Julia with multiple threads, open the Pluto notebook, and load the dataset as described in the data section. Run cells in order so that feature standardisation, fold construction, and mRMR ranking are fitted on the current training indices before any validation or testing. The notebook logs the inner-CV “new best” events, saves per-fold selected feature names, and exports summary artefacts, including a CSV and PNG of the final per-fold winners. These artefacts match the tables and figures in Section 5.1.3.

8 Contributions

Area	Stacey Fu	B. S. Dlamini (Siya)	Ryan Teo
Task 1: Visual separability (Isomap, t-SNE)			✓ lead
Task 2: Multialgorithm classification (kNN, RBFSVM, Decision Tree)			✓ lead
Task 3 Method A (MI + Elastic Net)	✓ lead		
Task 3 Method B (mRMR + nested selection)		✓ lead	
Task 3: Ensemble selection and calibration	✓	✓	
Task 4: Additional explorations (class weights, threshold calibration, gender models, soft voting)	✓ lead	✓ lead	
Reproducibility artefacts (code, seeds, notebooks)	✓ support	✓ lead	✓ support
Figures and report compilation	✓	✓	✓
Appendix A presentation (slides)		✓ lead	
Appendix B metric definitions		✓ lead	
Appendix C risk controls		✓ lead	

Authorship order is alphabetical by last name. Lead marks primary responsibility, support indicates secondary contribution.

Invalid source specified.