http://dx.doi.org/10.62762/TSCC.2026.365257

RESEARCH ARTICLE

Non-Invasive Parkinson's Disease Screening from Vocal Biomarkers: A Subject-Level Machine Learning Framework with Bayesian Optimisation and Explainable AI

Ryan Wyton[1,*] **and Raza Hasan**[1]

[1] School of Technology and Maritime Industries, Southampton Solent University, Southampton SO14 0YN, United Kingdom

Abstract

An estimated ten million people live with Parkinson's, yet timely diagnosis remains constrained by specialist assessment, costly imaging, and unequal care access. Phonation analysis offers a low-cost alternative: dopaminergic degeneration causes vocal impairments years before motor symptoms emerge, enabling community screening. However, existing ML approaches are limited by recording-level data leakage and opacity. This paper presents a four-phase acoustic framework—spanning signal acquisition, feature processing, and interpretable decision support—applied to a multi-type Parkinson's speech dataset (40 training, 28 blind-test). The framework enforces subject-level partitioning via GroupKFold to eliminate leakage, derives 104 participant-level acoustic features through multi-statistic aggregation (mean, median, standard deviation, interquartile range), and employs Bayesian optimisation using Optuna. The optimised SVM achieved 90.00% cross-validated accuracy, a 12.50 percentage-point improvement over the 77.50% benchmark. Cross-validated sensitivity was 95.00%, specificity 85.00%, Youden Index $J = 0.80$, Brier Score $\mathrm{BS} = 0.1049$. External validation on a held-out vowel-only cohort achieved 85.71% sensitivity, surpassing the benchmark. A permutation test ($p = 0.349$) indicated no significant linear mapping between acoustic features and motor severity, motivating binary classification. Shapley Additive Explanations identified interquartile range of degree-of-voice-breaks and shimmer amplitude variability as primary diagnostic drivers, aligning with neuroacoustic evidence for aperiodic phonation. These results show that leakage prevention, variance-sensitive engineering, and transparent attribution together constitute a viable foundation for remote, telehealth-linked pre-screening instruments in resource-constrained settings.

Keywords: acoustic sensing, vocal biomarkers, multi-statistic feature fusion, remote health monitoring, intelligent decision support, explainable artificial intelligence, Bayesian optimisation, Parkinson's disease.

Submitted: 18 May 2026
Revised: 08 July 2026
Accepted: 13 August 2026
Published: 27 September 2026

Vol. 3, **No.** 3, 2026.
10.62762/TSCC.2026.365257

***Corresponding author:**
✉ Ryan Wyton
0wytor22@solent.ac.uk

Citation
Wyton, R., & Hasan, R. (2026). Non-Invasive Parkinson's Disease Screening from Vocal Biomarkers: A Subject-Level Machine Learning Framework with Bayesian Optimisation and Explainable AI. *ICCK Transactions on Sensing, Communication, and Control*, 3(3), 197–209.

1 Introduction

Parkinson's disease (PD) ranks among the most prevalent neurodegenerative disorders worldwide, second only to Alzheimer's disease in incidence, with an estimated ten million individuals affected and prevalence accelerating in an ageing population [1]. The condition arises from selective degeneration of dopaminergic neurons in the substantia nigra pars compacta, disrupting the basal ganglia motor circuit and producing the hallmark triad of bradykinesia, muscular rigidity, and resting tremor [2]. Despite substantial therapeutic progress, no disease-modifying treatment currently exists, placing a premium on early and accurate detection to maximise the window available for symptomatic management and future neuroprotective strategies.

Current diagnostic practice relies on specialist neurological examination grounded in Movement Disorder Society clinical criteria, supplemented by dopamine transporter imaging in equivocal presentations [3]. This pathway carries structural constraints: specialist access is geographically unequal, imaging costs are prohibitive across many healthcare systems, and subjective rating instruments introduce inter-rater variability. This combination of specialist dependency, imaging cost, and diagnostic subjectivity renders timely detection effectively inaccessible for a substantial proportion of at-risk individuals in community and resource-limited settings [3]. Critically, visible motor symptoms typically emerge only after 50 to 80 percent of nigral cell loss has already occurred, meaning the pre-diagnostic window, during which neuroprotective intervention would be most impactful, frequently passes undetected [4].

Vocal deterioration offers a compelling alternative screening signal. Hypokinetic dysarthria, the speech manifestation of PD, has been documented in over 89% of patients and often precedes motor symptoms severe enough to prompt specialist referral [5]. This produces characteristic acoustic changes including imprecise articulatory movements, reduced pitch range, elevated cycle-to-cycle frequency perturbation (jitter), amplitude perturbation (shimmer), and increased noise levels, all detectable through routine sustained vowel recordings requiring only a standard microphone [6]. Such assessments are compatible with remote administration, supporting integration into community screening and telehealth pathways consistent with evolving care guidelines [7]. Framed within a sensing–communication–control perspective, sustained-phonation screening constitutes an acoustic-sensing modality in which a commodity microphone serves as the sensor front-end, participant-level statistical aggregation performs signal processing and multi-statistic feature fusion, and a calibrated, explainable classifier drives the downstream triage decision. The low bandwidth and hardware demands of this pipeline make it well suited to telehealth-linked remote health monitoring in community and resource-limited settings.

The machine learning literature on vocal PD biomarkers has expanded substantially since the Oxford telemonitoring study [6], which achieved approximately 91% accuracy on sustained vowels using entropy-based dysphonia measures. Related telemonitoring research from the same research group subsequently mapped sustained-phonation features to longitudinal symptom-severity progression [8]. Subsequent studies have validated across multi-task speech corpora [9, 10] and explored architectures from ensemble learners to deep learning. Three persistent methodological weaknesses account for much of the performance variability in published studies: (i) speaker-identity data leakage arising when recordings from the same individual straddle train and test partitions; (ii) dependence on raw acoustic means that discard intra-subject phonatory variability; and (iii) opaque classifier decisions that challenge acceptance within regulated healthcare environments.

This paper addresses all three weaknesses through a structured four-phase pipeline. Subject-level boundary integrity is enforced through GroupKFold partitioning. Intra-subject variance is captured by engineering each participant profile from four statistical summaries per raw acoustic variable: mean, median, standard deviation, and interquartile range. Decision transparency is achieved through Shapley Additive Explanations (SHAP), mapping classifier decisions to individual acoustic attributes in a clinically actionable form. The framework is validated against the Sakar *et al.* [9] benchmark through internal cross-validation and blind external testing on a held-out cohort. Figure 1 provides an overview of the complete analytical pipeline.

This paper makes four contributions relative to existing voice-based PD classifiers. First, GroupKFold cross-validation enforces subject-level partitioning, eliminating the recording-level leakage present in the majority of published studies. Second, four-statistic aggregation (mean, median, standard deviation, and IQR) captures intra-subject phonatory

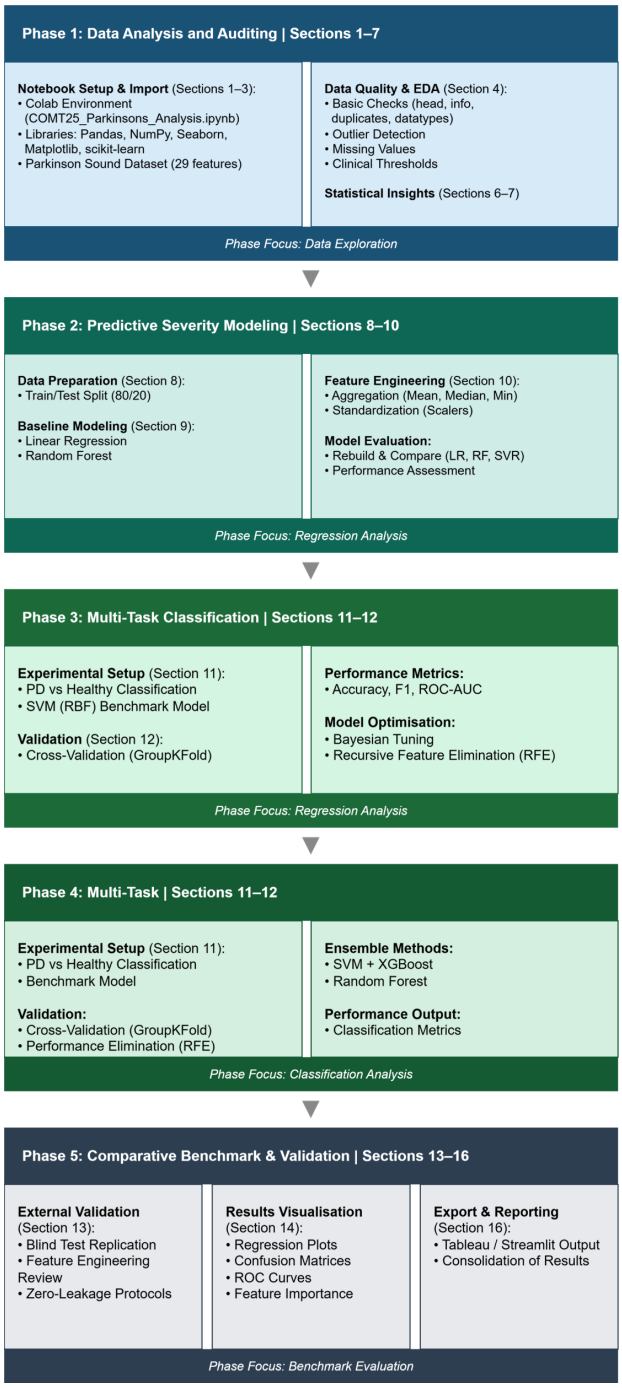


Figure 1. Overview of the four-phase analytical pipeline. Phase 1: data auditing and EDA. Phase 2: severity regression analysis. Phase 3: multi-task classification with Bayesian optimisation and SHAP. Phase 4 (external validation): vowel-isolation benchmark replication against Sakar *et al.* [9].

variability that mean-only representations discard. Third, Bayesian hyperparameter optimisation with Optuna TPE provides principled search over the SVM parameter space in place of manual tuning. Fourth, SHAP attribution maps classifier decisions to neuroacoustically grounded vocal features, satisfying the interpretability requirements of clinical decision support. Section 2 reviews acoustic biomarkers, machine learning methods, and explainability in PD detection. Section 3 details the dataset, feature engineering, classifier design, and optimisation workflow. Section 4 presents quantitative results with full metric reporting. Section 5 discusses the acoustic-sensing deployment and telehealth pathway, statistical interpretation, and limitations. Section 6 concludes.

2 Related Work

2.1 Acoustic Biomarkers in Neurological Screening

Sakar *et al.* [9] extended the scope of data collection to include running speech tasks alongside sustained vowels, assembling a multi-task dataset of 40 training and 28 test participants, with a vowel-specific sensitivity of 85.00% on the independent held-out cohort establishing the primary external benchmark for this study. A feature-expanded variant incorporating Tunable Q-factor Wavelet Transform coefficients was subsequently published by the same group [10], demonstrating incremental discriminative gains from the richer spectral representation.

2.2 Machine Learning Methods and Optimisation

Support vector machines with radial basis function kernels have consistently produced strong results on small-cohort vocal PD classification tasks, attributed to their margin-maximisation objective and capacity for non-linear boundary placement in sparse, high-dimensional acoustic feature spaces [11]. Bayesian optimisation directs search effort towards high-performing parameter regions through a probabilistic surrogate model, offering substantial efficiency gains over exhaustive grid search; Tree-structured Parzen Estimator implementations, such as the Optuna framework [12], represent the state-of-the-art realisation of this approach.

A recurring methodological flaw in published voice-based PD studies is the conflation of recording-level and subject-level evaluation. When a corpus contains multiple phonation attempts for each individual, standard random train–test splits acting on recordings allow the model to learn speaker-specific acoustic signatures rather than pathological generalisations, artificially inflating reported accuracy and producing classifiers that fail on genuinely unseen participants [13]. Subject-stratified partitioning with GroupKFold, which enforces that all recordings from a given individual remain within a single fold, addresses this directly, albeit at the cost of

reduced effective sample size.

Recent systematic evaluation of the field reinforces the critical importance of subject-level partitioning. Shokrpour *et al.* [14] surveyed 133 published PD voice classification studies and found that only 19% employed subject-independent evaluation protocols. The majority reported accuracy estimates derived from recording-level splits, which allow speaker-identity signals to substitute for pathological generalisation. This finding directly contextualises the GroupKFold approach adopted in the present study as addressing a methodological gap present in the substantial majority of prior work, and supports the interpretation that the 90.00% cross-validated accuracy reported here reflects genuine subject-level generalisation rather than recording-level memorisation.

2.3 Explainability in Clinical AI

Explainable AI techniques have also been applied directly to Parkinson's disease classification, demonstrating that feature attribution methods can bridge predictive performance with clinician-facing interpretability [15]. Growing recognition of the need for interpretable AI in healthcare has positioned post-hoc attribution methods as essential components of clinically deployable diagnostic models [16]. Shapley values, grounded in cooperative game theory, provide theoretically principled attribution of prediction contributions to individual input features, satisfying efficiency, symmetry, and linearity properties that alternative attribution methods do not guarantee [17]. For acoustic biomarker research, SHAP enables clinicians to verify that model decisions are driven by physiologically meaningful features rather than spurious data artefacts, a prerequisite for integration into regulated clinical decision support. The emergence of Kolmogorov–Arnold Networks as architectures for improved non-linear function approximation in PD voice analysis [18] represents a promising direction, though comparative evidence at the small-cohort scale common in this domain remains limited.

Recent work by Farfoura *et al.* [19] demonstrates that Self-Explaining Neural Networks (SENNs) can match post-hoc SHAP attribution performance in PD detection while providing structurally guaranteed explanations through intrinsic concept-based architecture. Unlike post-hoc methods such as SHAP, which approximate decision boundaries after training, SENNs enforce interpretability as an architectural constraint, ensuring that explanations faithfully reflect the model reasoning for each individual prediction. This distinction highlights an important direction for future clinical AI: while the SHAP-based approach adopted in the present study provides clinically actionable attribution at low data cost, intrinsic interpretability frameworks such as SENNs represent a promising complement where larger cohorts make their higher data requirements feasible.

3 Methods

3.1 Dataset

The Parkinson's Speech Dataset [9], retrieved from the UCI Machine Learning Repository, comprises recordings from 68 individuals. The pre-partitioned training set contains 1,040 recordings from 40 participants (20 with confirmed PD, 20 neurologically healthy controls), each contributing 26 recordings across three speech task modalities: sustained vowel phonations (/a/, /o/, /u/), digit and word lists, and connected sentence production. The independent test set contains 28 PD participants providing vowel-only phonations. Features comprise 26 acoustic variables for each recording: fundamental frequency statistics, five jitter measures, six shimmer measures, harmonics-to-noise ratio (HNR) and its inverse, the noise-to-harmonics ratio (NHR), autocorrelation coefficient, and four non-linear complexity measures (RPDE, DFA, PPE, and pitch period spread metrics). A continuous Unified Parkinson Disease Rating Scale (UPDRS) motor score and binary class label ($0 = \text{healthy}; 1 = \text{PD}$) accompany each recording.

The training cohort presents a balanced class distribution at the participant level (50% PD, 50% healthy controls). Figure 2 illustrates this balance at the recording level following participant aggregation. No missing values or duplicate records were identified during structural integrity auditing.

3.2 Data Quality and Clinical Plausibility

Prior to modelling, a domain-informed clinical plausibility audit was conducted to verify that acoustic feature values fell within physiologically plausible ranges. Jitter values were classified against clinical severity thresholds drawn from Teixeira *et al.* [20]: below 1% as healthy range, 1–2% as mild PD, 2–3% as moderate PD, and above 3% as severe PD. As shown in Figure 3, jitter_local and jitter_ddp exhibited substantial proportions in the moderate and severe PD severity ranges, corroborating the clinical composition of the cohort. Welch's independent samples t-tests confirmed statistically significant group differences for

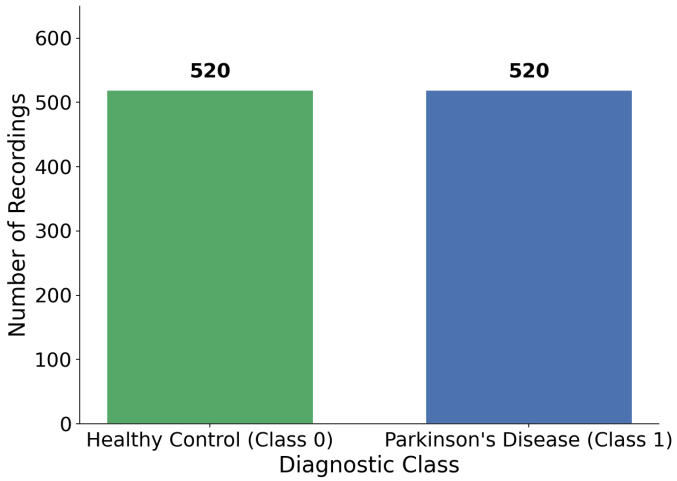


Figure 2. Class distribution across the 1,040 training recordings. Class 1 = PD; Class 0 = healthy control. The dataset presents a 50:50 balance at the participant level.

jitter_local ($t = 3.158$, $p = 0.002$) and mean_pitch ($t = -3.345$, $p < 0.001$) between PD and healthy participants, providing parametric validation that key acoustic features discriminate meaningfully between cohorts. One-way ANOVA across binned UPDRS severity groups further confirmed significant variation in jitter_local ($F = 15.756$, $p < 0.001$) and shimmer_local ($F = 4.087, p = 0.017$), consistent with the threshold-based severity classifications above.

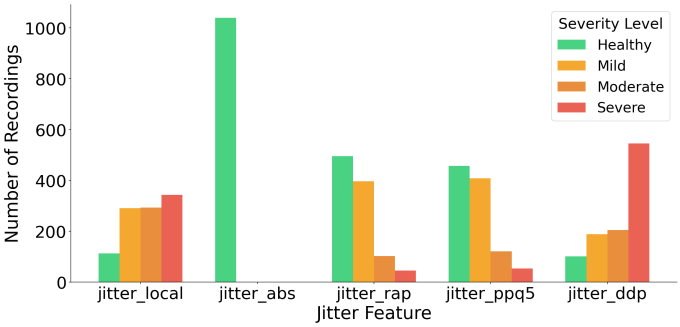


Figure 3. Jitter severity distribution across five jitter features, classified against clinical thresholds (healthy, mild, moderate, severe PD). The preponderance of moderate and severe classifications in jitter_local and jitter_ddp is consistent with the PD-dominant cohort composition.

3.3 Subject-Level Feature Engineering

A fundamental architectural decision is the transition from recording-level to participant-level representation. Training on individual recordings treats each phonation attempt as an independent observation, creating strong speaker-identity signals that classifiers can exploit without learning generalisable pathological patterns. To eliminate this confound, the 1,040 training records were aggregated into 40 subject profiles, each described by four statistical summaries per raw acoustic variable: arithmetic mean, median, standard deviation (SD), and interquartile range (IQR). This expands the 26-dimensional raw feature space to a 104-dimensional participant profile matrix.

$$\mathrm{IQR}(X_k) = Q_3(X_k) - Q_1(X_k) \tag{1}$$

where X_k is the n-dimensional vector of recordings for acoustic feature k, $Q_1(X_k)$ and $Q_3(X_k)$ are the 25th and 75th percentile values, and $\mathrm{IQR}(X_k)$ is the interquartile range. This four-statistic aggregation yields a 104-dimensional subject-profile matrix from the original 26-dimensional recording feature space.

The deliberate inclusion of SD and IQR encodes intra-subject phonatory variability directly into the feature representation. This reflects the clinical reality that Parkinsonian dysarthria manifests as aperiodic, intermittent impairment: dopaminergic tremor produces fluctuating rather than uniformly degraded acoustic quality across repeated phonation attempts. IQR is particularly valuable as it is resistant to individual outlying recordings and captures the stable spread of phonatory irregularity across a participant's session.

3.4 Multicollinearity Analysis

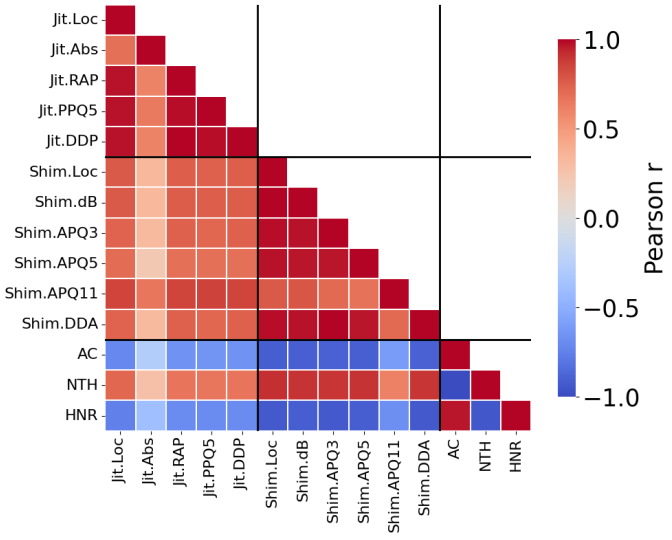


Figure 4. Pearson correlation heatmap across jitter, shimmer, and harmonics acoustic feature families (mean aggregation, $N = 40$ participants). Block-diagonal clustering confirms intra-family redundancy—jitter ($r > 0.95$ for four of five features), shimmer ($r > 0.90$)—with inverse correlations against harmonics-related measures, jointly motivating recursive feature elimination prior to classifier training.

Pearson correlation analysis across all features

characterised the collinearity structure of the acoustic space. Figure 4 shows pronounced block clustering within the jitter family (jitter_local, jitter_abs, jitter_rap, jitter_ppq5, jitter_ddp; inter-correlations $r > 0.95$ for four of the five features, with jitter_abs showing moderate intra-family correlation, $r \approx 0.60$–0.68) and the shimmer family (shimmer_local, shimmer_apq3, shimmer_apq5, shimmer_dda; $r > 0.90$). Harmonics-related measures exhibit inverse correlations with both perturbation families, consistent with the clinical expectation that increased instability degrades harmonic clarity. This collinearity structure empirically motivates feature pruning through recursive elimination prior to classifier training. The Pearson correlation between participant-level feature vectors f_i and f_j is

$$r_{ij} = \frac{\sum_s \left(f_{i,s} - \bar{f}_i\right)\left(f_{j,s} - \bar{f}_j\right)}{\sqrt{\sum_s \left(f_{i,s} - \bar{f}_i\right)^2}\sqrt{\sum_s \left(f_{j,s} - \bar{f}_j\right)^2}} \tag{2}$$

where f_i and f_j are participant-level feature value vectors, s indexes each participant, and r is the Pearson correlation coefficient. Feature pairs with $|r| > 0.90$ were flagged as highly collinear and candidates for recursive elimination. Figure 5 illustrates the discriminability distribution across all 104 aggregated features and the 16 features retained following recursive feature elimination.

3.5 UPDRS Regression Baseline

A UPDRS severity regression baseline was established using a Random Forest regressor to assess whether raw recording-level acoustic features carried reliable linear severity signal. A permutation test across 1,000 iterations returned $p = 0.349$, indicating no statistically significant mapping between the acoustic feature set and continuous motor severity scores at this level of representation. This null result confirmed that severity prediction from individual recordings is unreliable, motivating the architectural transition to binary PD versus healthy classification in Phase 3, where subject-level aggregated profiles replace recording-level inputs. Figure 6 shows the full null distribution.

3.6 Classifier Design and Optimisation

Classification used a Support Vector Machine (SVM) with a radial basis function (RBF) kernel, implemented with scikit-learn. The SVM was selected for its demonstrated efficacy in small-cohort, high-dimensional acoustic classification, and for its principled margin-maximisation objective. The SVM

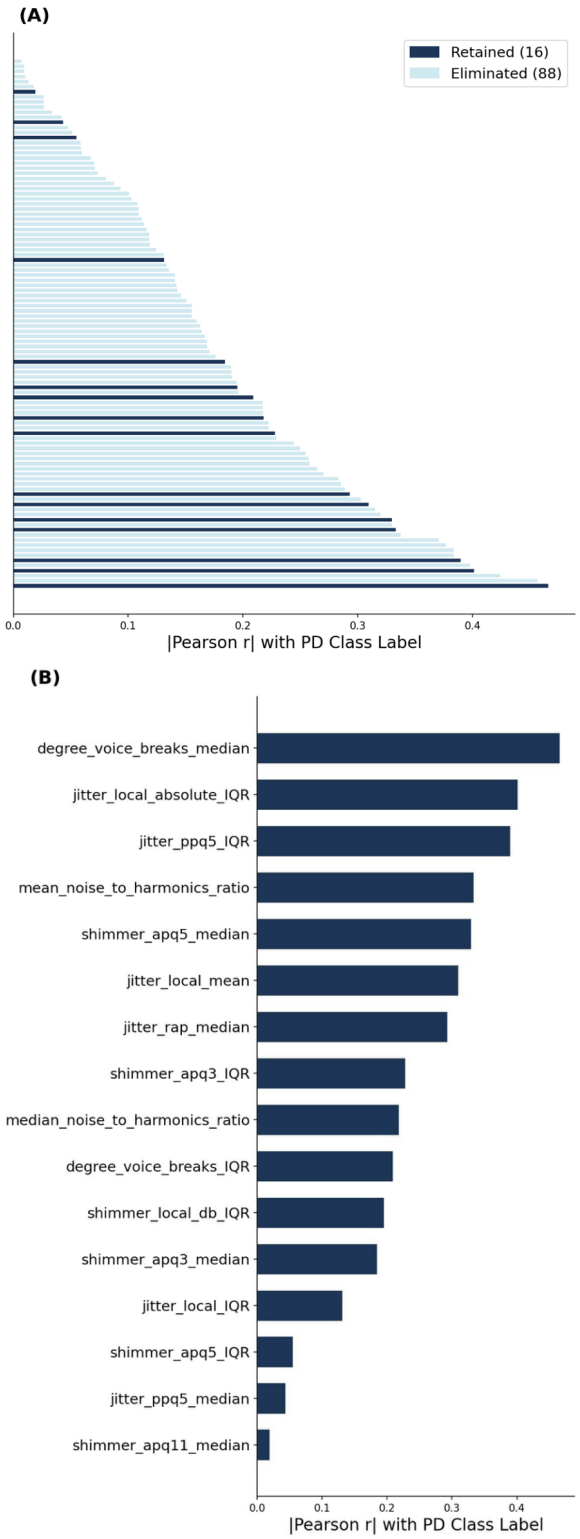


Figure 5. (A) All 104 aggregated features ranked by |Pearson r| with PD class label; dark bars = retained (16), light bars = eliminated (88). (B) The 16 features retained following recursive feature elimination, ranked by discriminability. IQR- and median-aggregated features dominate the retained set, providing empirical validation of the variance-sensitive aggregation design.

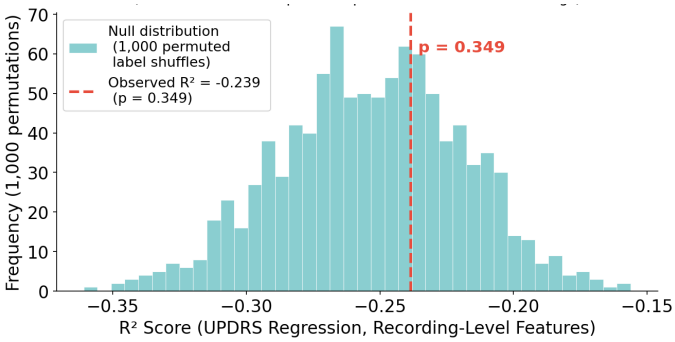


Figure 6. Permutation test null distribution for UPDRS regression (Random Forest, 1,000 iterations, GroupShuffleSplit 80/20, $n = 1{,}040$ recordings). The observed $R^2 = -0.239$ (dashed line) falls within the null distribution ($p = 0.349$), confirming no statistically significant acoustic-to-UPDRS mapping at the recording level.

decision function is

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{N} \alpha_i y_i \, K(x_i, x) + b\right) \tag{3}$$

where α_i are Lagrange multipliers, $y_i \in \{-1, +1\}$ are class labels, x_i are support vectors, b is the bias term, and K denotes the RBF kernel defined in (4). The cost parameter C and kernel width γ were jointly optimised through 30 trials of Bayesian search with the Optuna framework [12] with a Tree-structured Parzen Estimator acquisition function. The RBF kernel is

$$K(x_i, x_j) = \exp\left(-\gamma \, \|x_i - x_j\|^2\right) \tag{4}$$

where $\gamma > 0$ is the kernel width parameter optimised over $[0.0001, 1.0]$, yielding $\gamma = 0.008$, and $C = 44.97$. All training and preprocessing steps were conducted under GroupKFold cross-validation with five folds, stratified by participant identifier.

The Bayesian search selected the cost and kernel-width hyperparameters maximising the mean cross-validated accuracy across the K GroupKFold folds:

$$(C^*, \gamma^*) = \arg\max_{C,\gamma} \frac{1}{K} \sum_{k=1}^{K} \mathrm{Acc}(f_{C,\gamma}, V_k) \tag{5}$$

where C and γ are the SVM cost and RBF kernel-width hyperparameters, K is the number of cross-validation folds, $f_{C,\gamma}$ is the SVM trained with parameters (C, γ), V_k is the k-th held-out validation fold, and $\mathrm{Acc}(\cdot)$ denotes classification accuracy.

StandardScaler normalisation was fitted on training fold data only and applied to validation folds, preventing distributional leakage. The cost-sensitive configuration `class_weight = balanced` was applied to penalise false negatives more heavily, reflecting the asymmetric clinical cost of missing a PD diagnosis in a triage context.

An ensemble stacking architecture was also evaluated as an intermediate step: the optimised SVM, a Random Forest (100 estimators), and an XGBoost classifier were combined through a Logistic Regression meta-learner under the same GroupKFold scheme. Interpretable lightweight ensemble architectures have demonstrated generalisation across heterogeneous clinical datasets at modest cohort sizes [21], motivating the inclusion of a stacking ensemble alongside the optimised SVM as a complementary classifier.

3.7 Statistical Testing and Explainability

The Phase 2 permutation test confirmed no significant acoustic-to-UPDRS mapping ($p = 0.349$; see Section 3). Post-hoc model attribution used SHAP KernelExplainer with $k = 5$ k-means cluster centroids as the background dataset, generating approximate Shapley values through a weighted linear regression procedure. The Shapley value of feature i is

$$\phi_i(f) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S)\right] \tag{6}$$

where F is the complete feature set, S ranges over all subsets excluding feature i, $f(S)$ is the model output restricted to S, and $\phi_i(f)$ is feature i's average marginal contribution.

3.8 External Validation

External validation followed the Sakar *et al.* [9] protocol precisely. From the training set, three sustained vowel recordings per participant (/a/, /o/, /u/) were isolated and aggregated using mean and standard deviation only, mirroring the original study's constraints. The classifier was applied first using the Sakar *et al.* [9] default parameters ($C = 10$, $\gamma = 0.005$, RBF kernel) as a replication baseline, then under the Bayesian-optimised configuration. A label-shuffling integrity check confirmed near-chance sensitivity when training labels were randomised, validating the genuine pathological signal driving the main result.

4 Results

4.1 Baseline and Progressive Accuracy

The baseline SVM ($C = 1.0$, $\gamma = \text{scale}$) trained under GroupKFold cross-validation achieved 67.50% mean cross-validated accuracy and 75.00% on the held-out 8-subject partition. Figure 7 traces the accuracy progression across four modelling stages. The ensemble stacking architecture achieved 85.00% cross-validated accuracy following recursive feature elimination, representing an intermediate gain prior to hyperparameter search. Bayesian optimisation over 30 Optuna trials identified $C = 44.97$ and $\gamma = 0.008$ as the peak configuration, achieving 90.00% cross-validated accuracy at Trial 11. This represents a 12.50 percentage-point improvement over the Sakar *et al.* [9] multi-set benchmark of 77.50%.

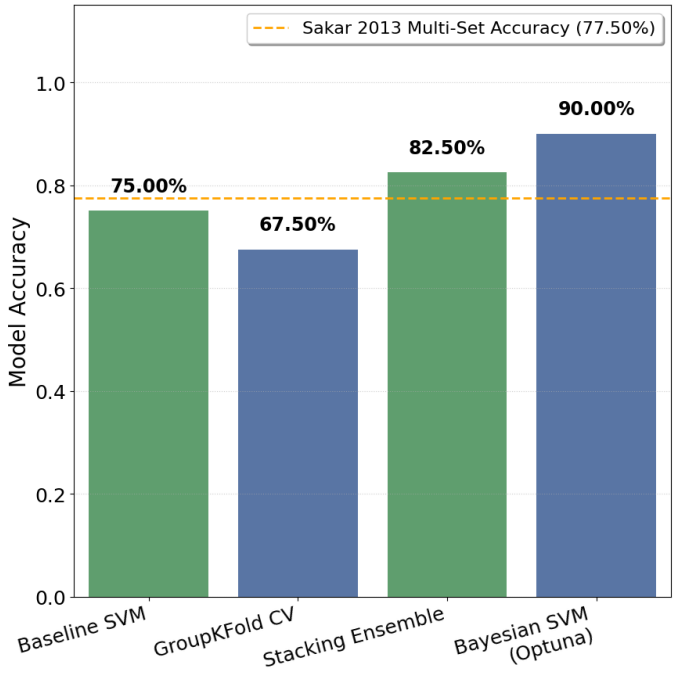


Figure 7. Progressive accuracy trajectory across four modelling stages. The dashed line marks the Sakar *et al.* [9] multi-set benchmark (77.50%). The Bayesian-optimised SVM at Trial 11 reached 90.00%, surpassing all prior configurations.

4.2 Generalisation Analysis

Figure 8 plots training and cross-validation accuracy against training cohort size ($n = 17$ to 32 participants). Training accuracy remained near-perfect (≥ 0.98) across all sample sizes, reflecting the expressive capacity of the RBF kernel. Cross-validation accuracy followed a consistent upward trajectory from approximately 48% at $n = 17$ to 90% at $n = 32$, with variance narrowing as cohort size grew. The progressive convergence of training and validation curves indicates genuine signal learning rather than memorisation of training instances and suggests that performance would continue to improve with an expanded participant base.

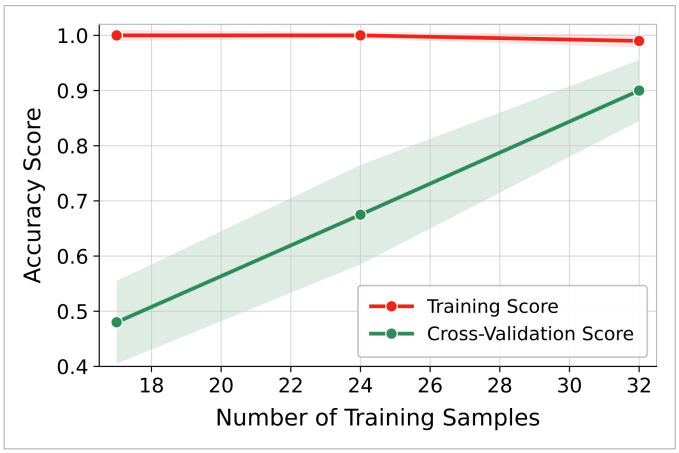


Figure 8. Learning curve showing training score and cross-validation score as a function of training cohort size. The narrowing band and upward CV trajectory indicate effective generalisation from $n = 17$ to $n = 32$ participants.

4.3 Performance Summary

Table 1 presents the benchmark comparison against Sakar *et al.* [9]; Table 2 reports the complete internal cross-validation metric suite. Notably, the optimised framework achieves a cross-validated area under the receiver operating characteristic curve (AUC-ROC) of 0.9200 (Table 2), indicating strong discriminative capacity. Both benchmark metrics show improvement, with the cross-validated accuracy gain being the most substantial.

Table 1. Benchmark comparison: Sakar *et al.* [9] versus the 2026 framework.

Metric	Sakar *et al.* [9]	2026 Framework	Δ
Multi-set CV Accuracy	77.50%	90.00%	+12.50 pp
Vowel-Task Blind PD Sensitivity	85.00%	85.71%	+0.71 pp
Permutation Test (p)	n/r	0.349	—

The optimised framework achieves a cross-validated Brier Score of 0.1049, representing a 58% improvement over the random classifier baseline of 0.25 on the balanced training cohort, confirming that probability estimates reflect genuine pathological signal (Table 2). The CV Youden Index (J) = 0.80 (CV Sensitivity = 95.00%, CV Specificity = 85.00%) confirms that the cost-sensitive class weighting correctly prioritises sensitivity over specificity, consistent with the clinical objective of minimising missed PD diagnoses. The external validation Brier Score of 0.0706 reflects

well-calibrated probability outputs on the 28-subject blind cohort; as this cohort is PD-only by the Sakar *et al.* [9] benchmark design, specificity and Youden Index cannot be computed from the external phase.

Table 2. Internal cross-validation metric suite for the optimised SVM (GroupKFold, $N = 40$, balanced cohort: 20 PD / 20 healthy controls).

Metric	Value
CV Accuracy	90.00%
CV AUC-ROC	0.9200
CV Sensitivity	95.00%
CV Specificity	85.00%
CV Youden Index J	0.80
CV Brier Score	0.1049

4.4 External Validation

Figure 9 presents the direct sensitivity comparison on the 28-subject blind test cohort. The optimised framework achieved 85.71% PD detection, surpassing the 85.00% RBF kernel peak reported in Sakar *et al.* [9]. The label-shuffling integrity check, which degraded sensitivity to near-chance levels, confirms the result reflects genuine pathological signal rather than systematic bias. The marginal 0.71 percentage-point gain over this benchmark is noted as practically meaningful within the constraints of the 28-subject test cohort.

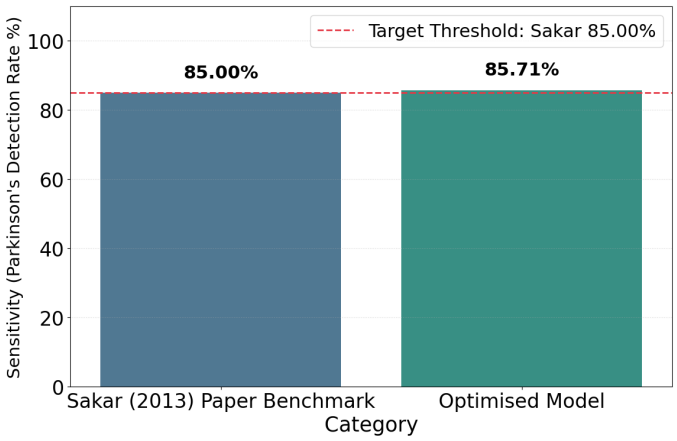


Figure 9. Sensitivity comparison on the 28-subject blind test cohort. The optimised framework (85.71%) surpasses the Sakar *et al.* [9] benchmark (85.00%). The dashed line marks the target threshold.

4.5 Diagnostic Confidence Distribution

Post-hoc error inspection examined the probability scores assigned to each of the 28 blind test participants. Figure 10 illustrates the distribution of PD likelihood scores. The majority of correctly identified PD cases were assigned scores exceeding 0.80, reflecting high-confidence detections well above the 0.50 decision boundary. The small cluster of false-negative cases received scores between 0.35 and 0.42, immediately below the threshold. This proximity suggests these participants exhibit subtler phonatory impairment, possibly indicative of earlier-stage disease or natural variability in vowel quality across recording attempts.

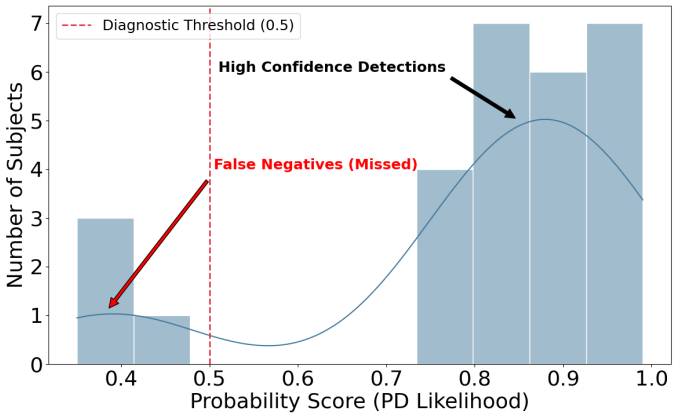


Figure 10. Histogram of PD probability scores for the 28-subject blind test cohort. High-confidence detections (score > 0.80) dominate the right-hand side of the distribution. False-negative cases cluster near the 0.50 decision threshold, suggesting subtle early-stage vocal pathology in missed participants.

4.6 Probability Calibration

Figure 11 presents the reliability diagram for the optimised SVM under GroupKFold cross-validation. The calibration curve shows the relationship between mean predicted probability and observed fraction of PD cases across five probability bins. Moderate overconfidence is evident at low-to-intermediate probability values, consistent with the documented behaviour of Platt-scaled SVMs on small cohorts. The CV Brier Score of 0.1049 represents a 58% improvement over the random classifier baseline of 0.25 on the balanced training cohort, confirming well-calibrated probability outputs overall.

4.7 SHAP Feature Attribution

Figure 12 shows the SHAP beeswarm plot for the optimised SVM trained on participant-level profiles. The three most influential features are *degree_voice_breaks_IQR*, *shimmer_apq11_median*, and *degree_voice_breaks_median*, all aggregation-derived statistics capturing the distribution of phonatory discontinuities and amplitude variability within an individual's recording session.

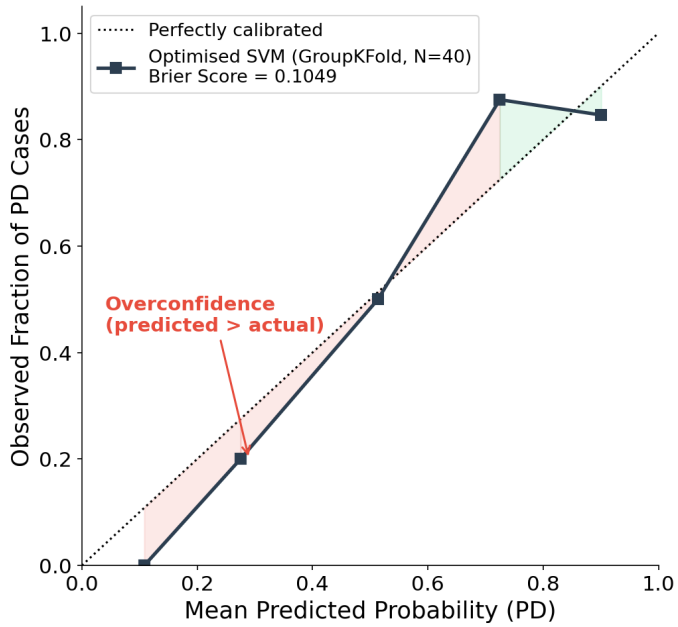


Figure 11. Reliability diagram: probability calibration of the optimised SVM (Platt scaling, $C = 44.97$, $\gamma = 0.008$, GroupKFold, $N = 40$). Moderate overconfidence at intermediate probability values is consistent with Platt-scaled SVM behaviour at small sample sizes. CV Brier Score $= 0.1049$.

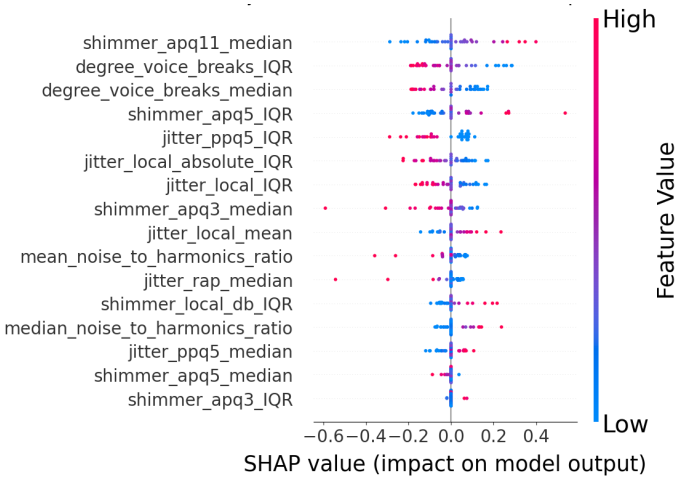


Figure 12. SHAP beeswarm plot for the optimised SVM. Each point represents one training participant; horizontal displacement indicates contribution magnitude and direction to PD classification probability. IQR and median aggregates of voice-break and shimmer measures dominate the attribution ranking.

IQR-derived features consistently outrank their mean-based equivalents in attribution magnitude, providing direct empirical support for the variance-sensitive aggregation design. The classifier discriminates PD from healthy controls primarily by detecting the spread and irregularity of voice break events across a participant's phonation attempts. This aligns with the neuroacoustic characterisation of Parkinsonian dysarthria as a disorder of phonatory consistency [22]: dopaminergic tremor generates unpredictable laryngeal motor fluctuations manifesting as irregular voice break episodes and inconsistent shimmer amplitude across trials. Mean statistics average over these fluctuations, suppressing the diagnostic signal; IQR preserves it. The identification of *mean_noise_to_harmonics_ratio* and *median_noise_to_harmonics_ratio* within the top ten SHAP features further aligns with established literature: elevated noise-to-harmonics ratio reflects turbulent phonation caused by incomplete glottal adduction, a known acoustic hallmark of Parkinsonian laryngeal hypokinesia [23].

5 Discussion

5.1 Interpreting the Performance Gains

The 90.00% cross-validated accuracy represents a practically significant advance over the 77.50% multi-set baseline of Sakar *et al.* [9]. Three architectural choices contribute to this gain. First, GroupKFold partitioning eliminates speaker-identity leakage and ensures the reported figure reflects generalisation to unseen participants rather than memorisation of vocal signatures. Removing this confound typically reduces apparent accuracy relative to naive recording-level splits, confirming that the observed 90% is a conservative, defensible estimate.

Second, participant-level feature aggregation yields an information-dense profile in which IQR and SD terms encode micro-tremor and phonatory irregularity signals absent from single-recording representations. SHAP analysis validates this empirically: the three leading predictors are IQR or median aggregates, confirming that the model exploits the distributional shape of a participant's recording session. This also highlights a limitation of the 2013 protocol, which aggregated using only mean and standard deviation: the IQR dimension, robust to outlying phonation attempts, provides additional discriminative power that simpler schemes omit.

Third, Bayesian optimisation located a configuration ($C = 44.97$, $\gamma = 0.008$) that substantially outperforms both the default SVM and the 2013 manual specification ($C = 10$, $\gamma = 0.005$). The non-monotonic optimisation trajectory in Figure 7 illustrates that performance gains require systematic search: some Optuna trials underperform the baseline before high-performing configurations are identified, underscoring the value of probabilistic sampling over

intuitive parameter choices.

5.2 Statistical Significance and Effect Size

The permutation test result ($p = 0.349$) pertains to the Phase 2 UPDRS severity-regression baseline (1,040 recordings) and does not bear on the Phase 3 classification performance. It indicates no statistically significant linear mapping between recording-level acoustic features and continuous motor scores, motivating the architectural shift to subject-level binary classification. Evidence that the classifier captures genuine pathological signal rather than systematic bias comes instead from the label-shuffling integrity check, which degraded external sensitivity to near-chance levels. Inferential testing remains constrained more broadly: the 28-subject blind test set sits well below the sample sizes required for reliable statistical testing, and the independence assumption is challenged by the single-site recording provenance. The 12.50 percentage-point internal accuracy gain represents a clinically meaningful improvement over the Sakar *et al.* [9] multi-set benchmark, while the 0.71 percentage-point sensitivity gain on the blind test confirms that the external benchmark is met and marginally surpassed. Consistency of improvement across both performance metrics in Table 1 supports a genuine performance advantage. Formal statistical confirmation awaits validation on multi-site cohorts of sufficient size.

5.3 Acoustic-Sensing Deployment and Telehealth Pathway

The SHAP attribution framework provides the interpretive bridge between model performance and clinical usability. Identifying *degree_voice_breaks_IQR* as the primary discriminative signal translates an abstract decision into a verifiable acoustic phenomenon: the variability of glottal closure failures across a patient's phonation session. This is a finding a neurologist can contextualise within clinical experience of hypokinetic dysarthria, enabling the kind of clinician trust that translatable AI systems require. The NICE [7] framework positions early recognition and timely specialist referral as quality indicators; a voice-based triage tool presenting SHAP decision support alongside a probability score—the decision layer of the acoustic-sensing pipeline—satisfies both interpretability and clinical governance requirements.

The bimodal diagnostic confidence distribution in Figure 10 maps naturally to a two-tier triage protocol: high-confidence cases (score > 0.70) trigger immediate specialist referral, while borderline cases (0.40 to 0.70) are scheduled for repeat assessment. This stratification reduces the operational burden of universal referral while preserving sensitivity for confident detections. Probability calibration analysis indicated slight overconfidence at intermediate values, a documented characteristic of Platt-scaled SVM outputs, which could be corrected through isotonic regression calibration prior to deployment.

5.4 Limitations

The training cohort comprises only 40 participants, limiting statistical power and the reliability of cross-validation confidence estimates. The 28-subject blind cohort comprises PD participants exclusively, precluding calculation of false positive rate and specificity in the external validation phase. Laboratory-quality microphone recordings may not transfer directly to smartphone or telephone-based capture environments typical of community screening. Medication status and dosage timing were not recorded in the Sakar *et al.* [9] dataset. Levodopa and dopamine agonist administration produces dose-dependent improvements in laryngeal motor precision, reducing the phonatory irregularity that the primary discriminative features, specifically the IQR of *degree_voice_breaks* and shimmer variability, are designed to detect. Recordings captured at peak medication effect may exhibit substantially lower voice-break variability in PD participants, potentially reducing model sensitivity. Future validation studies should record medication name, dose, and time since last dose for all participants and assess classification performance stratified by medication state. SHAP KernelExplainer produces approximate rather than exact Shapley values, introducing minor variance in feature importance rankings across runs. A further limitation concerns the external validation feature representation: to maintain strict replication of the benchmark protocol, the external test aggregation employed only mean and standard deviation, omitting the IQR and median aggregates identified as primary SHAP discriminators in the internal phase. This constrains the blind-test phase to a subset of the proposed feature space and precludes direct attribution of the sensitivity gain to the variance-sensitive IQR terms specifically.

6 Conclusion

This paper has presented and validated a structured machine learning framework for non-invasive PD

detection from acoustic vocal biomarkers, built on three design principles: subject-level leakage prevention through GroupKFold partitioning, variance-sensitive feature engineering through multi-statistic participant aggregation, and transparent decision attribution through SHAP decomposition. Applied to the experimental dataset, the framework achieved 90.00% cross-validated accuracy and 85.71% external sensitivity on a blind test cohort, improving on the benchmark on both reported metrics. SHAP analysis confirmed that IQR-encoded measures of phonatory irregularity and voice break frequency are the primary discriminative signals, consistent with the neuroacoustic profile of Parkinsonian hypokinetic dysarthria and directly interpretable by clinicians.

From a translational perspective, this framework is positioned as a pre-screening triage tool rather than a replacement for specialist clinical evaluation. Its low hardware requirements, participant-level probabilistic output, and SHAP-driven feature attribution make it suitable for embedding as an acoustic-sensing front-end within community healthcare and telehealth communication pipelines under current clinical guidelines. The framework's emphasis on combining ensemble prediction with post-hoc interpretability is consistent with recent advances in clinical XAI for PD detection, including intrinsic interpretability frameworks that structurally guarantee explanation fidelity. Future work will address three primary limitations: validation on multi-site datasets with recorded medication status; evaluation of generalisability under non-laboratory acoustic capture conditions; and exploration of deep learning architectures operating on raw mel-frequency cepstral coefficient spectrograms to bypass manual feature engineering entirely. The investigation of Kolmogorov–Arnold Networks as an alternative non-linear classifier at the small-cohort scale also warrants prospective evaluation.

Data Availability Statement

The Parkinson's Speech Dataset with Multiple Types of Sound Recordings is publicly available at the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/301/parkinson+speech+dataset+with+multiple+types+of+sound+recordings). Analysis code, the feature engineering pipeline, and supporting data files are openly available at https://github.com/RyanSolent95/COM725_Parkinson_Vocal_Biomarker_Analysis.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that Claude Opus 4.8 (Anthropic) was used for language editing of the manuscript. The authors have carefully reviewed, revised, and verified the AI-assisted output and take full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

This study constitutes a secondary analysis of a pre-existing, fully anonymised, publicly available dataset obtained from the UCI Machine Learning Repository. No primary data collection involving human participants was conducted. As the dataset contains no personally identifiable information and is freely accessible in the public domain, institutional ethical review was not required in accordance with applicable national guidelines for secondary analysis of anonymised public data.

References

[1] Dorsey, E. R., Constantinescu, R., Thompson, J. P., Biglan, K. M., Holloway, R. G., Kieburtz, K., ... & Tanner, C. M. (2007). Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology, 68*(5), 384–386. [CrossRef]

[2] Kalia, L. V., & Lang, A. E. (2015). Parkinson's disease. *The Lancet, 386*(9996), 896–912. [CrossRef]

[3] Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., ... & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Movement disorders, 30*(12), 1591-1601. [CrossRef]

[4] Fearnley, J. M., & Lees, A. J. (1991). Ageing and Parkinson's disease: substantia nigra regional selectivity. *Brain, 114*(5), 2283-2301. [CrossRef]

[5] Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1999). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural neurology, 11*(3), 131-137. [CrossRef]

[6] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering, 56*(4), 1010–1022. [CrossRef]

[7] National Institute for Health and Care Excellence. (2017). *Parkinson's disease in adults: Diagnosis and*

management (NICE Guideline NG71). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK447153/

[8] Tsanas, A., Little, M., McSharry, P., & Ramig, L. (2009). Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *Nature Precedings*, 1-1. [CrossRef]

[9] Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., ... & Kursun, O. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics, 17*(4), 828–834. [CrossRef]

[10] Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., ... & Apaydin, H. (2019). A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Applied Soft Computing, 74*, 255-263. [CrossRef]

[11] Elshewey, A. M., Shams, M. Y., El-Rashidy, N., Elhady, A. M., Shohieb, S. M., & Tarek, Z. (2023). Bayesian optimization with support vector machine model for Parkinson disease classification. *Sensors, 23*(4), 2085. [CrossRef]

[12] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631). ACM. [CrossRef]

[13] Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *Gigascience, 6*(5), gix019. [CrossRef]

[14] Shokrpour, S., MoghadamFarid, A., Bazzaz Abkenar, S., Haghi Kashani, M., Akbari, M., & Sarvizadeh, M. (2025). Machine learning for Parkinson's disease: a comprehensive review of datasets, algorithms, and challenges. *npj Parkinson's Disease, 11*(1), 187. [CrossRef]

[15] Raza, A., Ali, A., Kumar, A., Fatima, N., & Ali, M. (2026). Bridging Predictive Modeling and Clinical Interpretability: An Explainable AI Approach to Parkinson's Disease Detection. *Biomedical Informatics and Smart Healthcare, 2*(1), 20-37. [CrossRef]

[16] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery, 9*(4), e1312. [CrossRef]

[17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.

[18] Madany, M. E., Khalafallah, A., Ahmed, M. A., & Shoukry, A. (2024, December). Diagnosis of Parkinson's disease using Kolmogorov-Arnold Networks. In *2024 12th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)* (pp. 156-159). IEEE. [CrossRef]

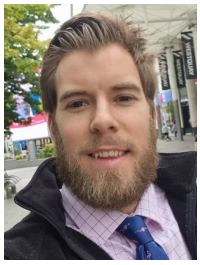
[19] Farfoura, M. E., Alkhatib, A. A., & Connie, T. (2026). Self-Explaining Neural Networks for Transparent Parkinson's Disease Screening. *Sensors, 26*(9), 2671. [CrossRef]

[20] Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia technology, 9*, 1112-1122. [CrossRef]

[21] Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023, June). Ensemble learning for disease prediction: A review. In *Healthcare* (Vol. 11, No. 12, p. 1808). MDPI. [CrossRef]

[22] Goberman, A. M., & Coelho, C. (2002). Acoustic analysis of Parkinsonian speech I: Speech characteristics and L-Dopa therapy. *NeuroRehabilitation, 17*(3), 237-246. [CrossRef]

[23] Rusz, J., Cmejla, R., Ruzickova, H., & Ruzicka, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *The journal of the Acoustical Society of America, 129*(1), 350-367. [CrossRef]



Ryan Wyton is a postgraduate student in Data Analytics and Visualisation at Southampton Solent University, Southampton, United Kingdom, specialising in machine learning and healthcare informatics. He received his bachelor's degree in music from the University of Chichester, where he first developed an interest in the analytical properties of acoustic signals and sound-based performance evaluation. His subsequent transition into data science has led to a focus on applying machine learning techniques to clinical and biomedical domains. He is currently focused on explainable artificial intelligence, acoustic signal processing, and non-invasive neurological screening, with particular emphasis on vocal biomarker classification for Parkinson's disease detection. His research interests include predictive modelling, interpretable AI, and feature engineering for clinical decision support systems. (Email: 0wytor22@solent.ac.uk)

Raza Hasan received his PhD in Informatics from Malaysia University of Science and Technology in 2021. He is currently a Senior Lecturer in Computing at Solent University, where he has worked since July 2023. Previously, he served as Deputy Head of Department, Programme Leader, and Associate Professor of Computing and IT at the Global College of Engineering and Technology in Oman. His research focuses on artificial intelligence, data science, learning analytics, educational data science, and data mining. (Email: raza.hasan@solent.ac.uk)