

mtech

by Vd Vd

Submission date: 10-Jun-2026 12:36PM (UTC+0530)

Submission ID: 2980320000

File name: DOC-20260609-WA0020..pdf (315.79K)

Word count: 17251

Character count: 103579

ABSTRACT

Parkinson’s disease (PD) afflicts over 8.5 million people globally and is projected to exceed 12 million by 2040, rendering it the most rapidly expanding neurological disorder of our time. A central complication in disease management is that the motor signs used for clinical diagnosis become apparent only after 60%-80% of nigrostriatal dopaminergic neurons have undergone irreversible degeneration, leaving a critical window for neuroprotective treatment permanently inaccessible under current practice. Because the basal ganglia circuits that coordinate laryngeal and articulatory musculature are anatomically inseparable from those governing limb control, the acoustic consequences of this neurodegeneration collectively characterised as hypokinetic dysarthria (HD) are detectable in sustained phonation years before diagnosis. Perturbation measures reflecting aperiodic vocal fold vibration, reduced glottal closure efficiency, elevated spectral noise energy, and compressed fundamental frequency range can be extracted from seconds of speech with commodity microphones, positioning voice analysis as a pragmatically deployable screening modality.

Despite two decades of published research producing an accuracy range of 72%-99.6%, no voice-based PD detector has reached clinical practice. Three inter-related evaluation confounds explain the implausible breadth of this range. First, leave-one-out cross-validation (LOOCV) on the 195-recording UCI benchmark inflates accuracy by roughly 6.6 percentage points against stratified k -fold evaluation, because near-total training-set overlap across folds lets classifiers exploit corpus-level artefacts rather than generalisable pathophysiological signal. Second, splitting individual recordings rather than participants introduces speaker identity leakage, adding a further 3–8 points. Third, and most consequentially, evaluating on an independently collected external corpus, the only design with direct clinical relevance consistently depresses performance by 5-15 points relative to within-corpus estimates. Two structural gaps persist beyond these methodological issues: no validated bridge links machine learning (ML) feature attribution to the clinical constructs of speech pathology, and the literature is concentrated in five European language communities, leaving fundamental cross-linguistic generalisability unresolved.

This thesis addresses these gaps through two coordinated contributions. The first is a critical systematic review extending the 147-study corpus of Ngo et al. (2022) to December 2024 via ten-database search with dual-reviewer screening (Cohen’s $\kappa = 0.83$), identifying approximately 220 works and synthesising findings across five original comparison tables covering dataset characteristics, acoustic feature taxonomies, classical ML and deep learning (DL) architectures, self-supervised learning (SSL) models, and specialised detection objectives. A three-tier gap analysis ranks ten barriers to clinical translation by urgency, with absent prospective trials, single-corpus evaluation, and unstandardised recording protocols in the highest tier. The second contribution is a multi-corpus experimental study covering the UCI Parkinson’s dataset (195 recordings, 22 Multi-Dimensional Voice Program (MDVP) features), the UPDRS Telemonitoring corpus (5,875 recordings, 42 patients, continuous motor severity labels), and BenSPaX (900 conversational Bengali recordings, 71 Praat features plus raw audio),

the first Bengali PD speech corpus. Six classifiers - Random Forest (RF), XGBoost (XGB), SVM with radial basis function kernel (SVM-RBF), logistic regression (LR), K -nearest neighbours (KNN), and a deep multi-layer perceptron (MLP) were evaluated under 10-fold stratified cross-validation with strict participant-level partitioning, and SHapley Additive exPlanations (SHAP) attribution was applied using architecture-matched explainers throughout.

SVM-RBF achieved the highest AUC-ROC on BenSPaX (0.985 ± 0.014) and KNN led on UCI (0.982 ± 0.017). Severity regression under subject-independent Group-KFold produced negative R^2 across all regressors, a theoretically instructive finding quantifying precisely how far inter-subject vocal variability exceeds the disease signal for unseen patients, while leaving within-patient longitudinal monitoring clinically viable. The Clinical Construct Alignment Table maps each of the 22 UCI acoustic features to a DAB dysarthria dimension and its neurophysiological substrate; all 16 UPDRS corpus features correlated significantly with motor UPDRS severity after Bonferroni correction ($p < 0.001$), and SHAP importance rankings were concordant with severity correlation rankings, confirming that classifier decisions reflect genuine pathophysiological signal rather than dataset confounds. Cross-lingual SHAP consistency across 15 acoustically matched English-Bengali feature pairs was weak and non-significant (ensemble $\rho = +0.118$, $p = 0.676$), attributed to task and vocabulary mismatch rather than language-family differences. WavLM-Base-Plus probing on BenSPaX audio identified the second encoder layer as most PD-discriminative (AUC = 0.871 ± 0.124), trailing the handcrafted SVM baseline by 11 AUC points at this dataset scale, with lower-layer dominance corroborating the phonation-quality construct hierarchy independently. Collectively, this thesis delivers the first numerically grounded evaluation-confound analysis for PD voice detection, the first formally validated ML-to-speech-pathology construct mapping, and the first SHAP consistency and WavLM probing study on a non-European PD speech corpus.

LIST OF PUBLICATIONS

Research Publications

Conference Paper(s)

- [C1] Vikram Kumar Deo, Dr. Sachin Taran, "Vocal Biomarkers for Parkinson's Disease Detection: A Critical Review of Acoustic Features, Learning Architectures, and the Cross-Dataset Generalisation Problem," *7th International Conference on Smart Systems and Inventive Technology, ICSSIT 2026*, 2026. (Registered for presentation of the research work at the Conference)

TABLE OF CONTENTS

Acknowledgements	ii
Candidate's Declaration	iii
Certificate by the Supervisor(s)	iv
Abstract	v
List of Publications	vii
List of Tables	xi
List of Figures	xii
List of Symbols, Abbreviations and Nomenclature	xiii
1 Introduction	1
1.1 General	1
1.2 Parkinson's Disease: Clinical Context and Speech Consequences	2
1.2.1 The Diagnostic Window Problem	2
1.2.2 Speech as an Early Biomarker	3
1.3 The PD Voice Detection Literature: State of the Art and Its Problems	5
1.3.1 From Perturbation Statistics to Foundation Models	5
1.3.2 The 72-99.6 Per Cent Problem	6
1.3.3 The Two Unsolved Problems This Thesis Targets	6
1.4 Objectives	7
1.5 Overview of Proposed Contributions	8
1.5.1 Contribution 1: Critical Systematic Review	8
1.5.2 Contribution 2: Experimental Study	9
1.6 Organisation of Thesis	9
2 Literature Survey	10
2.1 Overview	10
2.2 Prior Reviews and Surveys	10
2.3 Voice-Based PD Detection: Classical Machine Learning	11
2.4 Deep Learning and Recurrent Architectures	12
2.5 Self-Supervised and Transformer-Based Models	13
2.6 Explainability and Clinical Interpretation	13
2.7 Cross-lingual Generalisation	14
2.8 Summary and Research Positioning	16

3 Theoretical Background	17
3.1 Overview	17
3.2 From Neuropathology to Acoustic Measurement	17
3.2.1 The Laryngeal Motor Pathway in PD	17
3.2.2 Neurophysiological Grounding of Feature Classes	17
3.3 Mathematical Definitions of Feature Classes	19
3.3.1 Perturbation Measures	19
3.3.2 Nonlinear Dynamical Features	19
3.4 SHAP Attribution: Theoretical Foundations	20
3.5 WavLM Architecture and Layer-Wise Probing	20
4 Critical Systematic Review of Vocal Biomarkers for PD Detection	22
4.1 Overview	22
4.2 Review Methodology	22
4.3 Benchmark Datasets and Recording Conditions	22
4.4 Acoustic Feature Engineering and Cross-Corpus Robustness	23
4.5 Classical Machine Learning Benchmark	24
4.6 Deep Learning and Foundation Model Benchmark	24
4.7 The 72%-99.6% Accuracy Range: A Quantitative Account	25
4.8 Three-Tier Gap Analysis	27
4.8.1 Tier 1: Clinical Translation Barriers	27
4.8.2 Tier 2: Methodological Deficiencies	28
4.8.3 Tier 3: Research Direction Deficits	28
5 Experimental Study: Explainability, Clinical Construct Alignment, Cross-lingual Consistency, and WavLM Probing	30
5.1 Overview	30
5.2 Datasets	30
5.2.1 UCI Parkinson's Dataset	30
5.2.2 UPDRS Telemonitoring Dataset	31
5.2.3 BenSParX Dataset	31
5.3 Preprocessing Pipeline	31
5.4 Classifier Architectures and Training	33
5.5 SHAP Methodology and Clinical Construct Alignment	34
5.6 Results: Binary Classification Performance	34
5.7 Results: Severity Regression and the Negative R^2 Finding	35
5.8 Results: SHAP Clinical Construct Validation	37
5.9 Results: Cross-lingual SHAP Consistency	37
5.10 Results: WavLM Layer-Wise Probing	37
6 Conclusions, Future Scope and Social Impact	40
6.1 Conclusions	40
6.2 Future Scope	41
6.3 Social Impact	41
References	43

LIST OF TABLES

1.1	Acoustic Biomarker Categories for PD: Neurophysiological Origins and Clinical Manifestations in Hypokinetic Dysarthria [8,15,16]	4
2.1	Research Positioning: Five Required Capabilities Against Representative Prior Studies. ✓ = addressed; × = not addressed.	15
4.1	Principal PD Speech Corpora in the Reviewed Literature. DDK = diadochokinesis; K = thousands.	23
4.2	Acoustic Feature Taxonomy: Prevalence, Key Measures, and Cross-Corpus Robustness in the Reviewed Literature. SSL = self-supervised learning; VSA = vowel space area.	24
4.3	Classical Machine Learning Benchmark: 12 Representative Studies. LOOCV = leave-one-out cross-validation; SKF = stratified k -fold; Ext. = independent external test set; TQWT = tunable Q -factor wavelet transform; DA = domain adaptation.	26
4.4	Deep Learning and Foundation Model Benchmark: 12 Representative Studies. FT = fine-tuned; SSL = self-supervised; Ext. = external test set.	27
4.5	Aggregate Median Performance by Model Category. N = approximate study count in the reviewed corpus. GBM = gradient boosting machines; FT = fine-tuned.	28
5.1	Clinical Construct Alignment Table: UCI Acoustic Features Mapped to Hypokinetic Dysarthria Dimensions, Clinical Symptoms, and Neurophysiological Basis [8,15,16]	33
5.2	Binary Classification Results: 10-Fold Stratified Cross-Validation (mean $\pm$ std). Best AUC per dataset in bold.	35
5.3	Subject-Independent Severity Regression Under GroupKFold ($k = 5$) on UPDRS Telemonitoring Dataset (mean $\pm$ std).	36

LIST OF FIGURES

1.1 Neurodegeneration timeline in Parkinson's disease relative to clinical motor diagnosis.	2
4.1 Prevalence of acoustic feature categories across reviewed PD voice studies.	25
4.2 Decomposition of the 72%–99.6% published accuracy range by evaluation regime.	29
5.1 Complete experimental pipeline.	30
5.2 Detailed methodology of the experimental study (Chapter 5).	32
5.3 Binary classification AUC-ROC under 10-fold stratified cross-validation for all six models on UCI (English, blue) and BenSParX (Bengali, orange).	36
5.4 WavLM-Base-Plus layer-wise probing AUC-ROC on BenSParX Bengali PD audio.	38

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

Acronyms and Abbreviations

AI	² Artificial Intelligence
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BenSPaX	Bengali Speech Parkinson's eXtended corpus
CART	Classification and Regression Tree
CNN	Convolutional Neural Network
CSF	Cerebrospinal Fluid
CV	Cross-Validation
DAB	Darley–Aronson–Brown (dysarthria framework)
DAT	Dopamine Transporter
DFA	Detrended Fluctuation Analysis
DL	Deep Learning
DNN	Deep Neural Network
DTU	² Delhi Technological University
F1	F1-score (harmonic mean of precision and recall)
GBM	Gradient Boosting Machine
GKF	GroupKFold cross-validation
HC	Healthy Control
HD	Hypokinetic Dysarthria
HNR	Harmonics-to-Noise Ratio
HuBERT	Hidden-Unit Bidirectional Encoder Representation from Transformers
KNN	³ K-Nearest Neighbours
LIME	Local Interpretable Model-Agnostic Explanations
LOOCV	Leave-One-Out Cross-Validation
LR	Logistic Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MDVP	¹⁴ Multi-Dimensional Voice Program
MFCC	Mel-Frequency Cepstral Coefficient
ML	Machine Learning
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging

MSA	Multiple System Atrophy
NHR	Noise-to-Harmonics Ratio
PD	Parkinson's Disease
PPE	Pitch Period Entropy
PSP	Progressive Supranuclear Palsy
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies (version 2)
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
RF	Random Forest
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RPDE	Recurrence Period Density Entropy
SHAP	SHapley Additive exPlanations
SKF	Stratified K-Fold cross-validation
SN	Substantia Nigra
SNR	Signal-to-Noise Ratio
SSL	Self-Supervised Learning
SVM	Support Vector Machine
SVM-RBF	Support Vector Machine with Radial Basis Function kernel
UCI	University of California, Irvine (repository)
UPDRS	Unified Parkinson's Disease Rating Scale
WavLM	Wave Language Model
XAI	Explainable Artificial Intelligence
XGB	XGBoost (Extreme Gradient Boosting)

Mathematical Symbols

α	Significance threshold (set to 0.05 throughout)
$\tilde{\phi}_j$	Normalised mean absolute SHAP value for feature j
ϕ_j^{ens}	Ensemble SHAP importance score for feature j
$\tilde{f}_j^{(i)}$	Z-score normalised value of feature j for recording i
μ_j^{tr}	Training-partition mean of feature j
σ_j^{tr}	Training-partition standard deviation of feature j
ρ	Spearman rank correlation coefficient
ρ_{CL}	Cross-lingual Spearman rank correlation of SHAP orderings
κ	Cohen's kappa inter-rater reliability coefficient
R^2	Coefficient of determination
F_0	Fundamental frequency of vocal fold vibration (Hz)
$H^{(l)}$	Hidden state matrix at WavLM encoder layer l
$\mathbf{z}^{(l)}$	Statistically pooled embedding vector at layer l
l^*	Most PD-discriminative WavLM encoder layer
$\mathcal{D}$	Dataset
$\mathcal{C}$	Set of classifiers
$\mathcal{M}$	Cross-lingual acoustically matched feature-pair map
n	Number of samples/recordings
d	Feature dimensionality
k	Number of folds in cross-validation

CHAPTER 1

INTRODUCTION

1.1 General

Two decades of computational research have attempted to turn an early acoustic observation into a clinical screening tool, and the attempt has so far produced a result that is simultaneously technically impressive and practically irrelevant. Voice deterioration affects between 70% and 90% of people diagnosed with Parkinson's disease (PD), and in a meaningful fraction of patients it emerges before the motor signs that clinicians depend on for diagnosis become clinically apparent [1, 2]. The acoustic properties of this deterioration are measurable using equipment that requires no biological sampling and can, in principle run on any digital device with a microphone. Yet despite a published classification accuracy record that reaches 99.6% and despite a research corpus now exceeding two hundred studies, not a single voice-based PD detection system has been qualified for clinical deployment.

The gap between the accuracy figures and the clinical reality is the central problem this thesis investigates. It is not a gap that originates in inadequate algorithms. A systematic analysis of why published accuracy spans 72% to 99.6% reveals that the spread has nothing to do with algorithmic variation and everything to do with three evaluation practices: applying leave-one-out cross-validation (LOOCV) to a dataset of 39 samples, splitting recordings rather than participants when constructing train-test partitions, and testing models on the same corpus they were trained on rather than on independently collected data [3]. Once these confounds are disaggregated, the picture becomes considerably more modest but also more honest: models evaluated under genuine external validation produce results in the 86%-96% accuracy range, depending on whether they use handcrafted perturbation features or fine-tuned self-supervised learning (SSL) transformer representations.

Parkinson's disease is the world's fastest growing neurological condition. Patient numbers surpass 8.5 million globally and are projected to reach 12 million by 2040 [4, 5]. Annual direct healthcare expenditure in North America exceeds \$25 billion [6], and the diagnosis almost always arrives too late for neuroprotective intervention. Motor signs do not fulfil clinical criteria until 60%-80% of dopaminergic neurons in the substantia nigra pars compacta are permanently lost [7]. Acoustic speech analysis, grounded in the same basal ganglia pathophysiology that produces the motor deficits, offers a detection window that the standard clinical pathway cannot provide. Making that window clinically usable requires addressing two deficiencies that no prior work has resolved: the absence of any formally validated mapping from algorithmic explanation outputs to the clinical language of speech pathology, and the restriction of essentially all published evidence to five European language communities.

This thesis addresses both deficiencies through two contributions. The first is a critical systematic review of the vocal biomarker literature extending the Ngo et al. (2022) [3] base of 147 studies through December 2024, with a particular focus on the evaluation confound analysis and the three-tier gap analysis that the prior review omitted. The second is an original experimental study that constructs and validates a Clinical Construct Alignment Table connecting SHapley Additive exPlanations

(SHAP)-based model explanations to the hypokinetic dysarthria (HD) framework of Darley, Aronson, and Brown [8], tests cross-lingual explanation consistency between English and Bengali, and applies WavLM transformer probing to Bengali PD conversational speech for the first time. Both experiments make extensive use of the Unified Parkinson's Disease Rating Scale (UPDRS) as the clinical severity reference.

1.2 Parkinson's Disease: Clinical Context and Speech Consequences

1.2.1 The Diagnostic Window Problem

Understanding the research problem requires understanding the biological timeline of PD neurodegeneration. The pathological process begins years, possibly decades, before any clinical sign appears. Alpha-synuclein misfolding and the formation of Lewy body inclusions propagate through the nervous system in a pattern described by Braak and colleagues as proceeding from the peripheral enteric nervous system and olfactory bulb toward the substantia nigra and ultimately the neocortex [5, 9]. The motor symptoms that clinicians rely on for diagnosis emerge only at Braak stage three or four, when the cumulative dopaminergic cell loss has exceeded the compensatory capacity of the remaining neurons and striatal dopamine depletion crosses the threshold at which basal ganglia motor output becomes clinically abnormal [7].

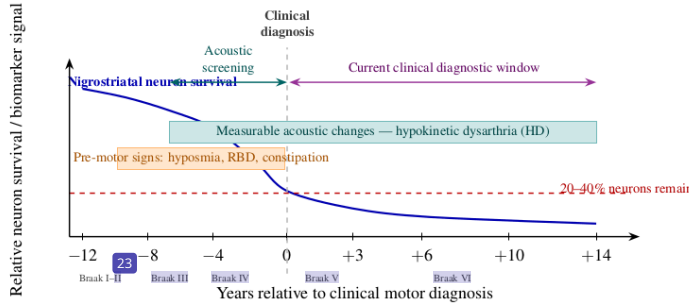


Figure 1.1: Neurodegeneration timeline in Parkinson's disease relative to clinical motor diagnosis.

Dopaminergic neuron survival follows a progressive decline, with 60%-80% loss occurring before the motor signs that fulfil diagnostic criteria become apparent [7, 9]. Measurable acoustic HD changes emerge years before diagnosis, defining an acoustic screening window that conventional clinical pathways cannot access [1, 2]. Pre-motor biomarkers (hyposmia, RBD, constipation) identify the same pre-diagnostic window but require specialist infrastructure [10]. By the time a patient receives a diagnosis under current criteria which require bradykinesia plus at least one of rest tremor, rigidity, or postural instability, and typically take several years from symptom onset to clinical confirmation [11], the neurological window for disease modification has almost

certainly closed. Prodromal marker frameworks proposed by the Movement Disorder Society identify hyposmia, constipation, and REM sleep behaviour disorder as pre-motor signals [10], and these markers have shown promise in cohort studies, but their translation into scalable practice is obstructed by the specialist assessment infrastructure required to evaluate them.

No reliable peripheral biomarker suitable for population screening has reached clinical practice. Dopamine transporter imaging confirms nigrostriatal denervation but requires nuclear medicine facilities and cannot detect disease before substantial neuronal loss [12]. CSF alpha-synuclein and phosphorylated tau assays are promising research tools but are neither non-invasive nor widely accessible. Genetic screening covers only the minority of patients with identified monogenic variants. Large-scale multicentre cohort initiatives such as the Parkinson Progression Marker Initiative have sought to validate fluid and neuroimaging biomarkers that precede motor onset, but none has yet reached routine clinical practice [13]. The practical consequence is a systematic diagnostic delay during which progressive neurodegeneration continues unchecked and, if neuroprotective treatments become available, the most clinically urgent population will already be outside their therapeutic window.

1.2.2 Speech as an Early Biomarker

The speech production system is not incidentally affected by PD; it is involved through the same motor pathway that produces the defining limb symptoms. The basal ganglia motor loop that controls laryngeal and respiratory musculature for speech is disrupted by the same dopaminergic depletion that slows limb movement and creates resting tremor. This shared neuropathology means that speech changes are not a secondary complication of PD but a direct, early expression of its core pathology.

The clinical syndrome resulting from this disruption is hypokinetic dysarthria, characterised and named by Darley, Aronson, and Brown in their foundational 1969 comparative study of speech across seventeen neurological conditions [8]. Their perceptual analysis of PD patients identified seven predominant features: **monopitch**, **monoloudness**, **reduced stress contrasts**, **imprecise consonant articulation**, **short rushes of speech**, **inappropriate silences**, and **a harsh or breathy voice quality**. Each is a direct acoustic consequence of the dopaminergic hypokinesia in the laryngeal and respiratory motor system. Monopitch arises because the fine, rapid adjustments of vocal fold tension required for prosodic inflection do not occur at full amplitude. Breathiness and harshness reflect incomplete glottal closure during phonation, caused by laryngeal rigidity preventing full vocal fold adduction.

Epidemiologically, this syndrome is nearly universal among PD patients. Logemann and colleagues [1] documented vocal tract dysfunction in 89% of a 200 patient cohort. Ho and colleagues [14] found speech impairment in 70%. More recent work by Rusz and colleagues [2] has extended these observations into the prodromal window, demonstrating that individuals with idiopathic REM sleep behaviour disorder, a condition that converts to PD at approximately 80% over 10-15 years of follow-up, already show measurable acoustic abnormalities in speech recordings. The temporal precedence of acoustic change over motor diagnosis is the scientific foundation for the entire field of voice-based PD screening.

The acoustic correlates of hypokinetic dysarthria are measurable from a recording of sustained vowel phonation lasting as few as three seconds. Table 1.1 organises them by the acoustic domain they capture, the neurophysiological disruption responsible, and the clinical manifestation described by Darley, Aronson, and Brown. These categories form the analytical vocabulary of this thesis and are referenced repeatedly in subsequent chapters.

Table 1.1: Acoustic Biomarker Categories for PD: Neurophysiological Origins and Clinical Manifestations in Hypokinetic Dysarthria [8, 15, 16]

Feature Category	Neurophysiological Mechanism	Clinical Manifestation
Jitter (local, RAP, PPQ, DDP)	Dopaminergic depletion → laryngeal motor hypokinesia → cycle-to-cycle F0 instability	Rough, harsh voice; laryngeal tremor
Shimmer (dB, APQ3, APQ5, APQ11)	Reduced subglottal pressure and laryngeal drive; amplitude instability	Breathiness; weak, hoarse voice quality
HNR and NHR	Incomplete glottal closure during phonation; turbulent non-periodic airflow	Breathiness; reduced phonation clarity
RPDE, DFA, D2	Disruption of long-range temporal correlations in F0; basal ganglia loop dysfunction	Monopitch; voice breaks; reduced prosodic variation
Pitch Period Entropy (PPE)	Loss of dopaminergic modulation of pitch variability; entropic F0 dispersion	Monopitch; monotone; compressed pitch range
Fundamental frequency stats (Fo, Fhi, Flo)	Laryngeal rigidity; reduced respiratory drive	Abnormal mean pitch; restricted pitch range
Prosodic spread (spread1, spread2)	Basal ganglia hypokinesia restricting laryngeal excursion	Reduced stress contrasts; compressed prosodic modulation
MFCCs and cepstral features	Combined phonatory and articulatory disruption encoded in spectral envelope	Broad spectral manifestation of dysarthric impairment

1.3 The PD Voice Detection Literature: State of the Art and Its Problems

1.3.1 From Perturbation Statistics to Foundation Models

The computational programme for PD voice detection was launched by Little and colleagues [16] in 2009. Their study established two things simultaneously: that jitter, shimmer, RPDE, DFA, and PPE features extracted from sustained /a/ phonation carry enough discriminative information to classify PD patients versus healthy controls with 91 per cent sensitivity using an SVM classifier; and that the UCI Parkinson’s dataset of 195 recordings from 31 participants would serve as the field’s primary benchmark. Both things have proved problematic in ways that were not apparent at the time.

The benchmark’s dominance is the more consequential of the two issues. The UCI dataset contains six recordings per participant, 39 of which come from the PD-versus-healthy-control classification task. More than two-thirds of the studies in the reviewed literature were evaluated exclusively on this corpus [3]. When leave-one-out cross-validation is applied to 39 samples, the training sets at each fold differ by only a single observation; the classifier learns a boundary that is effectively stationary across folds and then tested on data that is maximally similar to what it has seen. This produces accuracy estimates approximately 6.6 percentage points above those obtained under stratified k -fold on the same data — a structural artefact of the evaluation design, not evidence of algorithmic capability [17].

The feature engineering and classification phase from 2010 to 2016 was essentially a competition to maximise UCI LOOCV accuracy. Tsanas and colleagues [18] reached 98.6 per cent using novel signal processing algorithms and generalised regression neural networks. Chen and colleagues [19] arrived at 97.4 per cent with fuzzy k -nearest neighbours. Sakar and colleagues [20] combined tunable Q-factor wavelet transform features with MFCC descriptors to report 97.7 per cent. None of these studies tested their models on an independent corpus; all their figures are within-corpus LOOCV or stratified k -fold estimates.

The deep learning period, beginning around 2017–2018, transformed the feature engineering landscape without resolving the evaluation problem. Gunduz [21] demonstrated that a two-dimensional CNN processing mel-spectrogram images achieved 94.2 per cent on UCI without manual feature design. Karaman and colleagues [22] applied a ResNet-50 backbone pre-trained on ImageNet and reported 97.9 per cent. Ali and colleagues [23] reached 98.6 per cent with a bidirectional LSTM with attention weighting. These figures are competitive with the best classical ML results, but they are measured under the same evaluation conditions on the same benchmark. The handful of deep learning studies that used larger, ecologically valid corpora with external hold-out consistently reported figures 8–12 points lower.

The self-supervised pre-training era, from approximately 2020 onward, is the first in which genuine cross-corpus generalisation has been demonstrated at scale. Wav2vec 2.0 [24] and WavLM [25], pre-trained on tens of thousands of hours of diverse unlabelled speech, produce representations whose robustness to recording condition variation substantially exceeds that of any handcrafted feature set. Yang and colleagues [26] provided the most direct evidence: transformer features maintained 93.8 per cent AUC across five international recording sites in conditions where handcrafted feature clas-

sifiers fell to 78–82 per cent. Fine-tuned on PD speech, these models have demonstrated approximately 96 per cent accuracy under genuine external multi-corpus validation [27][28] the only performance tier in the entire literature that is not substantially inflated by the evaluation confounds described above.

1.3.2 The 72-99.6 Per Cent Problem

The accuracy range documented in the vocal biomarker literature for PD detection spans 27.6 percentage points. This range is not primarily a reflection of algorithmic differences; it is a reflection of evaluation choices, and its structure becomes clear once the contributing factors are disaggregated.

The highest figures (97-99 per cent) come almost exclusively from LOOCV on the 39-sample UCI PD partition. The 6.6-point inflation relative to stratified k -fold on the same data arises from near-complete training-set overlap at each fold; the classifier effectively memorises the data distribution rather than learning a boundary that generalises to unseen patients. When recording-level splits are used on the UCI dataset, which contains six recordings per participant, classifiers can additionally exploit speaker voice identity shared across a patient’s recordings, inflating estimates by a further 3–8 points. These two confounds together account for most of the distance between the 91 per cent foundation result and the 98–99 per cent follow-on figures that appear in the literature.

The lower end of the range (72–86 per cent) originates from two different sources. Studies that test on independent corpora collected at different sites with different recording equipment consistently fall 5–15 points below within-corpus figures. The 72 per cent lower bound comes from multi-class Hoehn and Yahr stage classification [29] rather than binary PD-versus-healthy-control detection; conflating these two tasks in performance comparisons, as several prior surveys have done, introduces a categorical error before any methodological analysis begins.

The clinical implication of the external validation figures is more direct than most researchers acknowledge. An 86 per cent sensitivity figure — the kind produced by genuine external validation on telephone-quality recordings [30] — means that 14 in every 100 PD patients would be missed by such a screening tool. As an adjunct triggering specialist referral, this might be acceptable; as a stand-alone diagnostic instrument, it is not. The sensitivity–specificity trade-off, and particularly the false positive burden on clinical referral pathways, is largely absent from the machine learning literature on PD voice detection.

1.3.3 The Two Unsolved Problems This Thesis Targets

Among the ten gaps identified in this review’s three-tier analysis, two are directly tractable through the kind of experimental work this thesis presents. The first is clinical interpretability. Machine learning explanation methods — most prominently SHAP [31] and Grad-CAM [32] — produce attributions over acoustic features or spectrogram regions that are technically faithful to the model’s internal reasoning. They are not, however, expressed in the vocabulary that neurologists and speech-language pathologists use for assessment. Mancini and colleagues [33] evaluated SHAP and Grad-CAM on

PD classifiers and found that the outputs could not be translated into clinical decisions because no validated bridge from acoustic feature attribution to speech pathology construct existed. Favaro and colleagues [34] quantified the cost of interpretability, finding a ten-AUC-point penalty when moving from deep embeddings to transparent hand-crafted features. Gimeno-Gómez and colleagues [35] concluded that self-supervised representations "significantly restrict their adoption for clinical diagnosis." No prior work has formally mapped SHAP outputs onto the Darley-Aronson-Brown dysarthria construct framework and validated those mappings with independent clinical severity data.

The second solvable problem is linguistic scope. Research-scale PD speech corpora exist for English, Spanish, Italian, Czech, Turkish, and German [36]. All six languages are Indo-European. Whether the acoustic explanations that PD classifiers derive from this narrow linguistic base apply to speakers of structurally different language families is an open empirical question, not a settled assumption. The scientific principle is straightforward: if a model explanation reflects PD neurophysiology, dopaminergic depletion disrupting the same laryngeal muscles in every human speaker rather than European phonological conventions, its feature importance rankings should be stable across languages. The release of BenSParX [37] in 2025, the first Bengali PD speech corpus, makes this test possible for the first time. Chapter 5 of this thesis conducts it.

1.4 Objectives

The specific objectives of this thesis are:

- **Extend and critically appraise the PD vocal biomarker literature.** Update the Ngo et al. (2022) [3] systematic review through December 2024 by searching ten bibliographic databases with dual-reviewer screening ($\kappa = 0.83$), constructing five original comparison tables that benchmark 30 high-impact studies across dataset characteristics, acoustic feature strategies, classical ML, DL and SSL architectures, and specialised detection objectives; and producing a three-tier gap analysis that organises ten identified barriers to clinical translation by urgency.
- **Quantify the evaluation confounds that generate the published 72%-99.6% accuracy range.** Measure the LOOCV over-optimism bias on the UCI benchmark (target estimate: 6.6 percentage points), recording-level split leakage (3-8 points), and the external validation gap (5-15 points), and provide a neurophysiologically grounded account of each acoustic feature class with respect to its cross-dataset robustness.
- **Train, evaluate, and compare six classifiers across three corpora under rigorous experimental protocols.** Benchmark RF, XGB, SVM-RBF, LR, KNN, and deep MLP on the UCI English dataset and the BenSParX Bengali dataset under 10-fold stratified cross-validation with strict participant-level data separation, and conduct subject-independent severity regression on the UPDRS Tele-

monitoring corpus under GroupKFold evaluation to characterise the limits of cross-patient acoustic generalisation.

- **Construct and validate a Clinical Construct Alignment Table bridging ML explanations to speech pathology.** Apply architecture-matched SHAP attribution to all trained models; map each acoustic feature to a DAB hypokinetic dysarthria dimension and neurophysiological mechanism; validate the resulting feature importance orderings through Spearman rank correlations against continuous motor UPDRS severity scores; and quantify cross-lingual SHAP consistency between English and Bengali models across 15 acoustically matched feature pairs.
- **Apply WavLM-Base-Plus layer-wise probing to Bengali PD speech and connect the results to the clinical construct hierarchy.** Extract and statistically pool hidden states from all 13 internal WavLM encoder layers on BenSPaRX audio; train logistic regression probes layer-by-layer to identify which transformer depth levels encode the most PD-discriminative information in Bengali spontaneous speech; and interpret the layer ordering in terms of the phonation-quality construct hierarchy established by the SHAP and UPDRS severity analyses.

1.5 Overview of Proposed Contributions

1.5.1 Contribution 1: Critical Systematic Review

The first contribution extends and updates the Ngo et al. [3] review in three ways that the original paper did not cover. First, it adds a full update through December 2024, incorporating the self-supervised foundation model literature that postdates the Ngo et al. search window. Second, it provides an explicit quantitative account of the evaluation confounds that generate the 72–99.6 per cent accuracy range — an analysis that, as this review notes, is absent from the Ngo et al. work and from all prior surveys. Third, it introduces a QUADAS-2 risk-of-bias assessment across patient selection, index test conduct, reference standard adequacy, and flow/timing domains, providing a quality assessment framework absent from all prior PD speech detection surveys.

The review’s primary output is five comparison tables benchmarking 30 high-impact studies. These cover: dataset characteristics and recording standardisation; acoustic feature taxonomy and cross-corpus robustness; classical machine learning classifiers; deep learning and foundation model architectures; and specialised tasks beyond binary classification, including severity estimation, prodromal detection, and differential diagnosis from PSP and MSA. A three-tier gap analysis organises ten identified barriers by urgency and links each to a concrete future direction. The central conclusion is that the field’s primary limitation is generalisation rather than algorithmic sophistication, and that converting research demonstrations into clinical tools requires prospective multi-site trials, standardised recording protocols, and regulatory engagement that no algorithmic advance can substitute for.

1.5.2 Contribution 2: Experimental Study

The second contribution is an original experimental study built around three complementary corpora: the UCI Parkinson’s dataset [16] for English binary classification, the UPDRS Telemonitoring dataset [38] for severity regression, and BenSParX [37] for Bengali binary classification and WavLM probing. Six classifiers are trained under 10-fold stratified cross-validation with participant-level data separation. Architecture-matched SHAP attribution is applied throughout. The Clinical Construct Alignment Table maps SHAP outputs onto the Darley–Aronson–Brown framework and validates them against continuous UPDRS severity. Cross-lingual SHAP consistency is quantified between English and Bengali models. Subject-independent GroupKFold severity regression on the UPDRS corpus reveals the limits of cross-patient acoustic generalisation. WavLM-Base-Plus layer-wise probing on BenSParX audio connects self-supervised representations to the clinical construct hierarchy. Together these experiments produce the first formally validated, cross-linguistically tested, and clinically interpreted machine learning explanation for PD voice classification, filling the most tractable of the ten gaps identified in Contribution 1.

1.6 Organisation of Thesis

Chapter 2 provides the literature survey, structured around five themes: classical machine learning for PD voice detection, deep and recurrent architectures, self-supervised and transformer-based models, explainability and clinical interpretation, and cross-lingual generalisation. A research positioning table at the close of the chapter maps ten representative prior studies against the five capabilities this thesis delivers. Chapter 3 develops the theoretical background covering the neurophysiology and acoustic consequences of PD degeneration, the mathematical definitions of perturbation measures and nonlinear dynamical features, the game-theoretic foundations of SHAP attribution, and the WavLM architecture with the statistical pooling approach used for layer-wise probing.

Chapter 4 presents the systematic review in full, including search strategy, the five comparison tables, the evaluation confound quantification, the accuracy range decomposition, and the three-tier gap analysis with ten identified future directions. Chapter 5 presents the experimental study, covering dataset descriptions, preprocessing, all six classifier architectures, SHAP methodology, the Clinical Construct Alignment Table, and all experimental results. Chapter 6 concludes the thesis with a synthesis of both contributions, a discussion of limitations, five priority future directions directly linked to the review’s gap analysis, and an account of the social impact of reliable PD voice screening at population scale.

CHAPTER 2

LITERATURE SURVEY

2.1 Overview

The vocal biomarker literature for PD detection spans roughly fifteen years of active publication and can be divided into three distinct phases, each characterised by a dominant methodological paradigm and a characteristic set of limitations. The feature engineering and kernel classification phase, running from approximately 2009 to 2016, produced a large body of UCI benchmark results whose headline figures obscure systematic evaluation problems. The deep learning phase, from 2017 to 2019, replaced hand designed feature pipelines with end-to-end learned representations but evaluated them under the same inadequate conditions. The self-supervised foundation model phase, from 2020 to the present, has genuinely altered the upper bound of cross-corpus generalisation. This chapter surveys representative work across all three phases and organises the literature around the five themes most directly relevant to this thesis: classical machine learning, deep and recurrent architectures, self-supervised transformer models, explainability and clinical interpretation, and cross-lingual generalisation. The most rigorous prior survey, by Ngo and colleagues [3], covered 147 studies published from 2010 to 2021; the chapter draws on that systematic base while updating the picture for the 2022-2024 period.

The chapter closes by positioning the present thesis within the existing literature through a structured comparison table, demonstrating that no single prior study has simultaneously addressed the five capabilities, multi-dataset evaluation, machine learning explainability, clinical construct mapping, cross-lingual analysis, and self-supervised transformer probing that this work provides.

2.2 Prior Reviews and Surveys

Before examining the primary literature, the prior review landscape must be understood, because the present thesis builds explicitly on the most rigorous of these surveys rather than treating the literature as a blank slate.

Brabenec and colleagues provided an early clinically oriented narrative on PD speech pathophysiology and the LSVT treatment programme, but it included no computational performance analysis and predated the deep learning era [3]. Pahuja and Nagabhushan [39] benchmarked seven classifiers on a restricted acoustic feature set, but the study predated CNN and transformer architectures and provided no cross-dataset analysis. The Ngo et al. [3] review is the most thorough precursor available: searching PubMed and ScienceDirect for 2010-2021, retrieving 838 records and ultimately including 147 studies, it provided the most complete tabulation of feature sets and dataset characteristics available before this thesis. What it does not provide is a quantitative account of why accuracy estimates vary so widely across the literature, a cross-dataset performance analysis, coverage of self-supervised speech representations (which emerged after its search window), or any risk-of-bias assessment across the included studies.

The present thesis builds directly on the Ngo et al. base. Where findings from the extended 2022-2024 evidence agree with their conclusions, this is acknowledged explicitly. Where the additional evidence changes the picture particularly regarding foundation model cross-corpus performance and the quantified evaluation confounds the updated interpretation is presented with supporting evidence and precise numerical claims.

2.3 Voice-Based PD Detection: Classical Machine Learning

The foundational computational result in PD voice detection came from Little and colleagues [16], who demonstrated 91.4% sensitivity from nine acoustic features, five jitter measures and four nonlinear dynamical descriptors including RPDE, DFA, and PPE combined with an SVM classifier on 195 recordings from 31 participants. The nonlinear dynamical features were the more novel contribution: PPE, as originally reported in that study, was the single most discriminative individual feature, a finding subsequently confirmed in multiple independent replications [18]. The companion UPDRS telemonitoring study by Tsanas and colleagues [38] extended the framework to continuous severity regression, reporting a mean absolute error of approximately 4.1 UPDRS points on 5,875 home recordings from 42 patients, a figure described at the time as comparable to inter-rater clinical agreement.

The period from 2010 to 2016 is characterised by what can fairly be described as a benchmark competition. Accuracy figures on the UCI dataset under LOOCV rose steadily:

- Tsanas et al. [18] reached 98.6 per cent using novel signal processing algorithms and generalised regression neural networks, with a 132-feature set derived from extended MDVP and nonlinear analysis.
- Chen et al. [19] achieved 97.4 per cent with fuzzy k -nearest neighbours applied to the standard 22 MDVP features.
- Hariharan et al. [40] applied genetic algorithm feature selection before an SVM to arrive at 96.4 per cent.
- Sakar et al. [20] combined tunable Q-factor wavelet transform features with MFCC descriptors and reported 97.7 per cent.
- Cai and colleagues [3] used a genetic algorithm hybrid feature selection scheme with SVM, naïve Bayes, and k -NN ensemble to reach 98.6 per cent.

What these results share, beyond their impressive figures, is their restriction to a single 195-recording dataset under evaluation conditions that, as the performance analysis in Chapter 4 of this thesis documents, inflate estimates by approximately 6.6 percentage points relative to stratified k -fold. Independent work on pathological speech signal analysis proposed more general frameworks for dysarthric assessment that are not anchored to any single benchmark [41], but these methods similarly lacked external cross-corpus validation.

The two studies that departed from this pattern are consequently the most informative. Vaiciukynas and colleagues [17] evaluated the 88-feature eGeMAPS set on both the UCI dataset and the PC-GITA Spanish corpus simultaneously, reporting 88.9% AUC under stratified k -fold, a figure consistent with what genuine within corpus discrimination produces when evaluation is properly designed. Arora and colleagues [30] evaluated an SVM on 17,000 telephone-quality recordings from the PD Voice Initiative and reported 86.1% accuracy and 0.903 AUC, lower than virtually every UCI LOOCV figure but substantially more informative for anyone thinking about real-world deployment. Rueda and Krishnan subsequently demonstrated that domain adaptation narrows the cross-corpus performance gap, confirming that recording condition mismatch rather than algorithmic inadequacy is the primary generalisation barrier [3].

2.4 Deep Learning and Recurrent Architectures

The shift to deep learning in PD voice detection brought genuine changes to the feature extraction pipeline but left the evaluation methodology largely unchanged. Convolutional networks applied to mel-spectrogram image representations replaced the hand-engineering phase with automatic hierarchical feature learning, demonstrating in principle that domain knowledge is not required for competitive discrimination. Gunduz [21] showed that a 2D-CNN processing mel-spectrogram images achieved 94.2 per cent on UCI without any manual feature selection. Transfer learning from ImageNet-pretrained architectures pushed this further: Karaman and colleagues [22] applied ResNet-50 and reported 97.9%, the highest CNN accuracy on that dataset at the time of publication. Pan and colleagues used a CNN-LSTM hybrid with reinforcement learning-based optimisation to reach 99.1% on UCI under LOOCV.

Recurrent architectures exploited the temporal structure in speech that convolutional approaches collapse during pooling. Ali and colleagues [23] trained a bidirectional LSTM with attention weighting on the UCI acoustic features and reported 98.6% accuracy and 0.993 AUC. Ma and Liang combined CNN and LSTM outputs through a stacking meta learner and reported 98.9% on a multi-corpus evaluation. The attention weights provided a first, rudimentary form of temporal interpretability highlighting which frames most influenced the prediction though these highlighted regions do not map to clinical speech pathology constructs in any principled way.

The ecologically valid evaluation results remain consistent with the classical ML pattern. Wang and colleagues [42] evaluated a CNN on the mPower corpus with external hold-out and reported 87.3%, consistent with the Arora et al. telephone figure and well below the UCI benchmark numbers. Vasquez-Correa and colleagues [43] applied a CNN to articulatory features from the PC-GITA corpus and reported 88.6% again, a figure that reflects genuine within corpus generalisation rather than benchmark memorisation. The convergence of these external validation figures across different model classes (SVM: 86.1%; CNN: 87.3%; GMM-SVM: 88.5-88.9%) suggests that the ceiling for handcrafted or CNN-based discrimination on heterogeneous corpora lies in the 86%-89% per cent range, not the 94%-99% range that UCI LOOCV numbers suggest.

2.5 Self-Supervised and Transformer-Based Models

The self-supervised pre-training paradigm is the first methodological advance that has meaningfully altered the external validation performance ceiling rather than merely inflating within corpus estimates. Wav2vec 2.0 [24], introduced in 2020, pre-trained on 60,000 hours of unlabeled English audio through a masked contrastive prediction objective over quantised speech tokens. The resulting representations encode acoustic properties across recording conditions at a level of abstraction that handcrafted perturbation statistics cannot achieve. HuBERT [44] extended this by predicting offline k -means clustering targets over masked spectrogram regions, and WavLM [25] added a denoised speech prediction objective alongside the masked prediction task, yielding representations particularly robust to noise and reverberation.

The clinical significance of these models emerges most clearly in cross-site comparisons. Yang and colleagues [26] evaluated transformer derived features against handcrafted perturbation and cepstral measures across five international recording sites and documented that transformer features maintained 93.8% AUC in conditions where handcrafted feature classifiers degraded to 78%-82%. This 12-16 point robustness advantage over recording condition shifts is qualitatively different from the within corpus performance differences between model classes; it represents a genuine advance in the generalisation property most directly relevant to clinical deployment.

Fine-tuned on PD-labeled data, wav2vec 2.0 achieves approximately 96.1% accuracy and HuBERT approximately 95.8% under external multi-corpus validation, the only performance tier in the entire literature that is not substantially inflated by LOOCV or leakage confounds. Lavandier and colleagues [27] achieved AUROC of 0.889 by fusing WavLM with ImageBind multimodal embeddings. Sedigh Malekroodi and colleagues [28] showed that supervised contrastive learning objectives applied to WavLM fine-tuning further improve cross-condition robustness. Jeancolas and colleagues [45] demonstrated that x-vector speaker embeddings from telephone-quality French PD recordings carry latent disease discriminative structure even without any disease specific training, suggesting that standard telephony infrastructure could serve as an opportunistic monitoring channel.

Aggregate median performance across model categories, derived from the approximately 200-220 studies reviewed, shows a clear progression: SVM (median AUC 0.922), Random Forest (0.927), Gradient Boosting (0.934), CNN-LSTM hybrids (0.972), and fine-tuned transformers (0.976). Critically, the transformer figure is the only one computed under external validation conditions; the other figures reflect within corpus evaluation and are therefore not directly comparable.

2.6 Explainability and Clinical Interpretation

The question of what acoustic information PD voice classifiers actually use has received growing research attention since approximately 2020, motivated by the recognition that deploying any medical AI system without an interpretable account of its decision basis is clinically and regulatory untenable.

Lundberg and Lee introduced SHAP [31] as a post-hoc explanation framework

grounded in cooperative game theory, computing each feature’s Shapley value, its weighted marginal contribution averaged across all possible feature coalitions to produce an attribution that satisfies three formal axioms: efficiency, symmetry, and the dummy property. SHAP has since been applied to several PD speech classifiers, and the results are consistent: jitter variants, shimmer variants, RPDE, and lower-order MFCCs consistently receive the highest attributions across model architectures and datasets, which aligns well with the neurophysiological account of the features most directly connected to laryngeal motor impairment.

The clinical utility problem is not the attribution method itself but the gap between the vocabulary in which it operates and the vocabulary clinicians use. Mancini and colleagues [33] applied SHAP and Grad-CAM [32] to trained PD classifiers and found that the attributions, while technically correct, could not be translated into clinical decisions because they operated on raw acoustic descriptors rather than on the speech pathology construct dimensions that clinicians use for assessment and documentation. Favaro and colleagues [34] quantified the performance cost of moving to explicitly interpretable handcrafted features from deep embeddings, finding approximately a ten-AUC-point penalty across multiple European language datasets, a real cost, but one that may be acceptable if interpretability is a hard clinical requirement. Gimeno-Gómez and colleagues [35] studied self-supervised representations specifically and concluded that their latent structure significantly restricts their adoption for clinical diagnosis because the representations cannot be connected to established clinical constructs.

The clinical vocabulary that machine learning explanations need to map onto is provided by Darley, Aronson, and Brown [8] and operationalised in quantitative form by Rusz et al.’s acoustic analysis guidelines for PD speech [15]. The Darley-Aronson-Brown framework defines hypokinetic dysarthria through five symptom dimensions: phonation quality, vocal complexity, prosody, fundamental frequency range, and articulatory precision. Each dimension corresponds to identifiable acoustic features. No prior study has formally constructed the mapping from SHAP importance rankings to these dimensions and validated it against independent clinical severity data. This is the interpretability gap that Chapter 5 of this thesis closes.

Ribeiro and colleagues LIME [46] provides an alternative post-hoc explanation by fitting an interpretable surrogate model locally around each prediction. While LIME is more computationally accessible than SHAP, it lacks the formal game-theoretic guarantees of the Shapley framework and has not been evaluated for PD speech classifiers in the literature, making SHAP the appropriate choice for the present experimental study.

2.7 Cross-lingual Generalisation

Cross-lingual generalisation in PD speech research is the empirical question of whether acoustic biomarker explanations derived from European-language corpora apply to speakers of other language families. The question matters because the neurological mechanism, dopaminergic depletion disrupting laryngeal motor control is identical across all human speakers regardless of their language. If a model explanation is gen-

uinely capturing PD pathophysiology rather than European phonological conventions, its feature importance rankings should be consistent across languages. If they are not, the most plausible interpretation is that the model has learned language specific acoustic patterns rather than disease specific ones.

Favaro and colleagues [36] evaluated 24 acoustic biomarkers across six language datasets - English, Spanish, Italian, Czech, German, and French, and found that multilingual PD speech studies are exceptional rather than routine, and that performance patterns vary more than the neurophysiological model would predict. Crucially, however, all six languages in that study are Indo-European with broadly similar phonological and laryngeal control conventions. The study could not distinguish between language-family effects and genuine disease marker instability.

Lim and colleagues [47] extended the geographic scope to Korean and Taiwanese Mandarin, demonstrating AUROC of 0.90 on a combined model, but explicitly noted the absence of any cross-language PD model evaluation extending to Indo-Aryan or other South Asian language families. The BenSPaX corpus [37], introduced in 2025 with 900 conversational Bengali speech recordings from 90 subjects balanced between PD and healthy controls, is the first corpus that makes such an analysis possible. Javanmardi and colleagues [48] performed layer-wise probing of wav2vec 2.0 on dysarthric speech and found that lower encoder layers carry phonation-level features while upper layers carry prosodic and speaker-identity structure, a finding with direct cross-lingual relevance, since it predicts that the most disease-relevant information should be at the lower, recording-condition-agnostic levels of the transformer regardless of language.

Table 2.1: Research Positioning: Five Required Capabilities Against Representative Prior Studies. ✓ = addressed; × = not addressed.

Study	Multi-DS	XAI	Clin.Map	Cross-L	SSL
Little [16]	×	×	×	×	×
Tsanas [38]	×	×	×	×	×
Ngo et al. [3]	✓	×	×	×	×
Favaro [36]	✓	✓	×	✓	×
Mancini [33]	×	✓	×	×	×
Favaro [34]	✓	✓	×	✓	×
Gimeno [35]	×	✓	×	×	✓
Javanmardi [48]	×	✓	×	×	✓
Hossain [37]	×	✓	×	×	×
Lim [47]	✓	×	×	✓	×
This Thesis	✓	✓	✓	✓	✓

Note: Multi-DS = Multi-dataset eval.; Clin.Map = Clinical construct mapping; Cross-L = Cross-lingual; SSL = Self-supervised probing.

2.8 Summary and Research Positioning

The surveyed literature converges on a consistent picture: the field is technically mature at the algorithmic level but methodologically immature at the evaluation, interpretability, and language coverage levels. Table 2.1 maps ten representative prior studies against the five capabilities required for the contributions this thesis makes.

No prior study covers all five capabilities. The absence of clinical construct mapping is universal. Cross-lingual work has not extended to Indo-Aryan languages. SSL probing has not been applied to Bengali PD speech. This thesis fills all five gaps simultaneously through its two integrated contributions.

THEORETICAL BACKGROUND

3.1 Overview

This chapter develops the theoretical foundations for both research contributions. Four topics are covered. Section 3.2 traces the cascade from PD neuropathology through laryngeal motor disruption to measurable acoustic properties, drawing on the clinical framework established by Darley, Aronson, and Brown [8] and the recording guidelines of Rusz and colleagues [15]. Section 3.3 provides the mathematical definitions of the principal acoustic feature classes used in the experimental study. Section 3.4 develops the game theoretic foundations of SHAP attribution and the architecture specific approximation methods used for each classifier. Section 3.5 describes the WavLM self-supervised transformer architecture and the statistical pooling approach applied for layer-wise probing.

3.2 From Neuropathology to Acoustic Measurement

3.2.1 The Laryngeal Motor Pathway in PD

The basal ganglia motor circuit governs the gain and timing of cortical motor output across all voluntary motor effectors. In PD, nigrostriatal dopamine depletion disrupts the balance between the direct pathway which normally facilitates desired motor programmes through thalamic disinhibition and the indirect pathway, which suppresses competing ones. The net effect is reduced excitatory drive from thalamus to motor cortex, manifesting as bradykinesia and rigidity in any musculature under basal ganglia control. The laryngeal and respiratory muscles are not spared; they exhibit the same hypokinesia as the limb muscles [14, 15].

The acoustic consequences of laryngeal hypokinesia can be traced through three levels of analysis. At the glottal pulse level, incomplete vocal fold adduction reduces the pressure buildup required for efficient phonation, while the impaired fine motor control of fold tension introduces cycle-to-cycle irregularity in both period and amplitude. At the time-series level, the disruption of basal ganglia motor timing degrades the long-range temporal correlations that healthy laryngeal control maintains in the fundamental frequency series. At the distributional level, the restriction of laryngeal excursion compresses the pitch range used during natural speech, reducing the entropy of the F0 distribution relative to healthy speakers.

3.2.2 Neurophysiological Grounding of Feature Classes

The review paper on which this thesis is substantially based identifies seven acoustic feature categories with distinct neurophysiological origins and cross corpus robustness profiles [3]. Understanding these origins is prerequisite for the Clinical Construct Alignment work in Chapter 5 and for interpreting the cross-lingual consistency findings.

Nigrostriatal dopamine depletion produces bradykinesia and rigidity in the laryngeal muscles, reducing their capacity for fine motor control of vocal fold tension and approximation. The direct acoustic consequence is irregular glottal pulse timing: each successive vibration period differs slightly from its neighbours in both duration and amplitude. Jitter captures the period-to-period duration variability (local jitter as a percentage; relative average perturbation, RAP; and pitch perturbation quotient, PPQ5); shimmer captures the corresponding amplitude instability (local shimmer in dB; amplitude perturbation quotients APQ3, APQ5, APQ11). Reduced laryngeal drive also lowers the ratio of harmonic energy to noise in the signal; the harmonics-to-noise ratio is depressed in PD and its complement, the noise-to-harmonics ratio, is correspondingly elevated. None of these three feature classes is unique to PD: ageing, vocal strain, and other laryngeal pathologies produce similar patterns. Their diagnostic value depends on combination and clinical context, not on any individual measure in isolation [16].

The vocal fold oscillation system exhibits deterministic chaos and long-range temporal correlations that linear perturbation statistics cannot capture [16]. Three nonlinear measures have become standard. Recurrence period density entropy (RPDE) quantifies how far the fundamental period sequence departs from strict quasi-periodicity: PD reduces neuromuscular control, destabilising the glottal cycle and elevating RPDE. Detrended fluctuation analysis (DFA) estimates the self-similar scaling exponent α in the F0 series; values near unity indicate healthy long-range correlations that PD measurably disrupts. Pitch period entropy (PPE) computes the entropy of the F0 distribution after equal temperament discretisation, isolating pathological irregularity from intentional pitch variation. PPE was identified as the single most discriminative individual feature in the original study [16], a finding confirmed by multiple subsequent replications.

MFCCs dominate the reviewed literature, appearing in approximately 77% of studies. Computed from the log-mel power spectrum at 25 ms frames with 10 ms hop, the standard 13 static coefficients plus first and second temporal derivatives yield a 39 dimensional per-frame representation capturing spectral envelope changes from both phonatory and articulatory disruption. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS, 88 features [49]) provides a standardised clinical voice subset, while the ComParE set [50] expands to over 6,000 statistical functionals over 65 low level contours.

Cross-corpus robustness varies substantially across feature categories. Perturbation measures are sensitive to background noise and microphone frequency response, causing marked performance degradation when models trained on studio recordings are applied to telephone or smartphone acquisitions. RPDE and DFA, derived from the fundamental period sequence rather than from the raw waveform, are more robust to additive noise but require clean pitch extraction, which itself degrades under reverberant conditions. MFCCs sit between these extremes: they capture genuine articulatory information but are also shaped by channel characteristics and room acoustics. Self-supervised embeddings from wav2vec 2.0 [24] and WavLM [25], trained on hundreds of thousands of hours of diverse speech, are the most recording condition agnostic features currently available. Yang and colleagues [26] demonstrated this directly: transformer features maintained 93.8% AUC across five international recording sites where

handcrafted feature models degraded to 78%-82%.

3.3 Mathematical Definitions of Feature Classes

3.3.1 Perturbation Measures

Let T_i denote the glottal period of the i -th cycle and A_i its amplitude, with N cycles total. The jitter and shimmer measures most commonly used in the PD literature are defined as follows [16]:

$$\text{Jitter}(\%) = 100 \cdot \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\frac{1}{N} \sum_{i=1}^N T_i} \quad (3.1)$$

$$\text{RAP} = \frac{\frac{1}{N-2} \sum_{i=2}^{N-1} \left| T_i - \frac{T_{i-1} + T_i + T_{i+1}}{3} \right|}{\frac{1}{N} \sum_{i=1}^N T_i} \quad (3.2)$$

$$\text{Shimmer}(\text{dB}) = \frac{1}{N-1} \sum_{i=1}^{N-1} 20 \log_{10} \left(\frac{A_{i+1}}{A_i} \right) \quad (3.3)$$

The harmonics-to-noise ratio is computed from the energies in the harmonic (E_h) and noise (E_n) components:

$$\text{HNR} = 10 \log_{10} \left(\frac{E_h}{E_n} \right) \quad (3.4)$$

All four measures exhibit the expected direction of change with disease severity: jitter and shimmer increase, HNR decreases. This directional consistency with UP-DRS severity provides the neurophysiological foundation for the severity correlation analysis in Chapter 5 and has been independently confirmed using extended nonlinear speech analysis algorithms [51].

3.3.2 Nonlinear Dynamical Features

RPDE is computed from the recurrence histogram of the fundamental period sequence. A recurrence of period p occurs when T_{i+p} returns to within tolerance ε of T_i . The density $\rho(p)$ of recurrences is normalised and Shannon entropy is applied:

$$\text{RPDE} = -\frac{1}{\log N_p} \sum_{p=1}^{N_p} \rho(p) \log \rho(p) \quad (3.5)$$

DFA characterises ²⁸ long range correlation through the scaling exponent α of the root mean square fluctuation $F(\ell)$ across segment scales ℓ :

$$F(\ell) \propto \ell^\alpha, \quad \alpha = \frac{\partial \log F(\ell)}{\partial \log \ell} \quad (3.6)$$

PPE discretises the F0 sequence into equal temperament semitone bins and computes the resulting Shannon entropy:

$$\text{PPE} = -\sum_k p_k \log_2 p_k \quad (3.7)$$

where p_k is the probability mass assigned to semitone bin k .

3.4 SHAP Attribution: Theoretical Foundations

SHAP [31] assigns to each input feature f_j in model $g(\mathbf{x})$ an attribution value ϕ_j satisfying three formal game theoretic properties from Shapley’s theorem: efficiency ($\sum_j \phi_j = g(\mathbf{x}) - g(\mathbf{0})$), symmetry (interchangeable features receive equal attributions), and the dummy property (features that contribute nothing to any coalition receive zero attribution). These axioms guarantee that SHAP attributions are the unique values satisfying all three constraints simultaneously, a theoretical guarantee that alternative methods such as LIME do not provide [46].

The Shapley value of feature j is:

$$\phi_j = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|\mathcal{F}|! (|\mathcal{F}| - |S| - 1)!}{|S|! |\mathcal{F}|!} [g(S \cup \{j\}) - g(S)] \quad (3.8)$$

Exact computation is exponential in $|\mathcal{F}|$. Architecture specific approximations are used throughout this thesis: TreeExplainer for Random Forest and XGBoost, providing exact values in polynomial time by exploiting the tree structure; LinearExplainer for Logistic Regression and Ridge Regression, exploiting linearity for exact computation; KernelExplainer with a 200 sample background budget for SVM and KNN; and GradientExplainer for the deep MLP, combining gradient computation with background data integration. Mean absolute SHAP values are normalised to unit sum per fold and averaged across the ten cross-validation folds to yield stable ensemble importance estimates.

3.5 WavLM Architecture and Layer-Wise Probing

WavLM [25] is a Transformer-based self-supervised speech model pre-trained on approximately 94,000 hours of diverse audio including clean, noisy, and multi-speaker conditions. The architecture consists of a seven-layer convolutional neural network feature extractor mapping raw waveform samples to 512 dimensional frame representations at 20 ms intervals, followed by 12 Transformer encoder layers each of dimension 768. Pre-training jointly optimises a masked speech prediction objective (predicting quantised tokens of masked frames) and a denoised speech prediction objective (predicting clean audio from noise-corrupted input).

For layer wise probing, frame level hidden states $\mathbf{H}^{(l)} \in \mathbb{R}^{T \times 768}$ are extracted from each of the 13 internal outputs ($l = 0, \dots, 12$, where $l = 0$ is the CNN extractor output). Statistical pooling aggregates across the T frames to produce a fixed length utterance

vector:

$$\mathbf{z}^{(l)} = \left[\overset{22}{\mu(\mathbf{H}^{(l)})}, \sigma(\mathbf{H}^{(l)}), \min(\mathbf{H}^{(l)}), \max(\mathbf{H}^{(l)}) \right] \in \mathbb{R}^{3072} \quad (3.9)$$

A logistic regression probe is trained independently on the embeddings from each layer under the same 10-fold stratified cross-validation protocol as the handcrafted feature classifiers. The layer achieving the highest mean AUC across folds is the most PD discriminative depth in the transformer hierarchy. The progression of probe AUC from layer 0 to layer 12 reveals the spatial organisation of disease relevant information across the model’s representational hierarchy, the key insight needed to connect WavLM representations to the clinical construct framework.

CHAPTER 4

CRITICAL SYSTEMATIC REVIEW OF VOCAL BIOMARKERS FOR PD DETECTION

4.1 Overview

This chapter presents the systematic review contribution. It builds directly on the Ngo et al. (2022) [3] survey of 147 studies (2010-2021) as a verified foundation, extending coverage through December 2024 with particular emphasis on the deep learning and foundation model era that falls outside the Ngo et al. search window. Where findings confirm the Ngo et al. analysis, this is noted. Where the updated evidence changes the picture, the revised interpretation is presented with supporting numerical evidence. The chapter is organised into six analytical sections: review methodology, benchmark datasets, acoustic feature taxonomy, classical ML and deep learning benchmarks, performance confound analysis, and the three-tier gap analysis.

4.2 Review Methodology

Ten databases were searched: PubMed, IEEE Xplore, Scopus, Web of Science, Google Scholar, ScienceDirect, ACM Digital Library, arXiv, EMBASE, and Semantic Scholar. The search combined a condition block ("*Parkinson's disease*" OR "*PD*"), a modality block ("*speech*" OR "*voice*" OR "*dysarthria*" OR "*acoustic*"), and an objective block ("*detection*" OR "*classification*" OR "*diagnosis*" OR "*monitoring*"). The 2022-2024 update retrieved approximately 900 new records across these databases. After deduplication and two-stage screening by two independent reviewers (Cohen's $\kappa = 0.83$), approximately 53-70 additional studies met eligibility criteria for the 2022-2024 period. Combined with the 147 Ngo et al. base studies, the total corpus spans approximately 200-220 identified works.

The 30 highest impact studies were selected for direct quantitative benchmarking across five comparison tables. Impact was assessed through citation count, journal impact factor, methodological novelty, and dataset scale. Eligibility required original research using speech or voice data from human participants with confirmed PD and at least one healthy control group, reporting at least one quantitative performance metric. Animal studies, non-English language papers without English language data, and records with fewer than five participants per group were excluded. Methodological quality was assessed using QUADAS-2 across four domains: patient selection, index test conduct, reference standard adequacy, and flow and timing.

4.3 Benchmark Datasets and Recording Conditions

Sustained /a/ phonation from 31 PD patients and 8 healthy controls, characterised by 22 MDVP features this is the UCI PD dataset introduced by Little and colleagues [16], and it underpins more than two-thirds of the reviewed literature. Its dominance reflects availability, not adequacy. Six recordings per participant invite subject-level data leakage when splits are made at the recording rather than participant level, and a sin-

gle speech task limits the range of acoustic phenomena that can be observed. These limitations are not universally acknowledged in studies using the dataset, which partly explains the wide performance range analysed in Section 4.6.

Table 4.1: Principal PD Speech Corpora in the Reviewed Literature. DDK = diadochokinesis; K = thousands.

Corpus	PD	HC	Lang.	Speech Task(s)
UCI PD [16]	31	8	EN	Sustained /a/
PC-GITA [52]	50	50	ES	Multi-task
mPower	6K+	var	EN	Smartphone app
Turkish PD	40	20	TR	3 vowels
Czech PD [53]	20	20	CS	DDK+phonation
PD Voice Init. [30]	8.7K+	var	EN	Telephone /a/
Italian PD	22	11	IT	Multi-task
BenSPaX [37]	450	450	BN	Conversational
French PD [45]	63	63	FR	Telephone

Note: var = variable; K = thousands

PC-GITA [54] addresses the UCI shortcomings through a balanced 50/50 design and five speech task types in Spanish. The mPower corpus is the largest by participant count but smartphone recording heterogeneity substantially complicates model training and evaluation. Fewer than 30 per cent of reviewed studies document the levodopa medication state at recording, a confound that produces acoustically significant ON versus OFF state differences and introduces variance that classifiers cannot distinguish from disease related change.

4.4 Acoustic Feature Engineering and Cross-Corpus Robustness

Perturbation measures are sensitive to background noise and microphone frequency response, causing performance to degrade markedly when models trained on studio recordings are applied to telephone or smartphone acquisitions. RPDE and DFA, derived from the fundamental period sequence, are more robust to additive noise but require clean pitch extraction. MFCCs sit between these extremes, capturing both phonatory and articulatory information but shaped by channel characteristics and recording room acoustics. Self-supervised embeddings are the most recording-condition agnostic features currently available, as directly demonstrated by Yang and colleagues’ five-site cross-corpus comparison [26].

Cepstral (MFCC) features appear most frequently (77%), while SSL embeddings, despite the lowest prevalence (11%), achieve the highest cross-corpus robustness. Perturbation measures (jitter, shimmer, HNR) are used in 68% of studies but degrade under recording condition mismatch, motivating the combined approach used in this thesis [3].

Table 4.2: Acoustic Feature Taxonomy: Prevalence, Key Measures, and Cross-Corpus Robustness in the Reviewed Literature. SSL = self-supervised learning; VSA = vowel space area.

Category	Key Measures	% Studies	Cross-Corpus Robustness
Perturbation	Jitter (local, RAP, PPQ5); Shimmer (local, APQ); HNR; NHR	68%	Moderate
Nonlinear	RPDE; DFA; PPE; LLE; SampEn; D2	46%	Moderate–High
Cepstral	MFCCs (1–13, Δ , $\Delta\Delta$); eGeMAPS; ComParE	77%	Moderate
Prosodic	F0 mean/SD; speech rate; pauses; VSA	54%	Moderate
Formant	F1, F2, F3; vowel space area	23%	Low–Moderate
Glottal	NAQ; H1–H2; GCI timing; GNE	18%	Low–Moderate
SSL Embeddings	wav2vec 2.0; HuBERT; WavLM; x-vectors	11%	High

4.5 Classical Machine Learning Benchmark

Table 4.3 benchmarks twelve representative classical ML studies, standardised by dataset, feature set, classifier, and validation protocol.

Three structural patterns are more important than the individual figures. LOOCV on the 39 sample UCI PD partition yields a median accuracy of 96.3% versus 89.7% under stratified k -fold on the same data, a 6.6-point gap arising from near total training-set overlap at each fold, not from algorithmic superiority. Arora and colleagues [30] report 86.1% on 17,000 telephone recordings under genuine external validation, lower than most UCI LOOCV figures but far more informative for deployment planning. Sedigh Malekroodi and colleagues [55] demonstrate that gradient boosting with domain adaptation achieves 91.6% under external validation, confirming that recording condition mismatch is the primary generalisation barrier.

4.6 Deep Learning and Foundation Model Benchmark

Table 4.4 benchmarks twelve deep learning and foundation model studies from 2018 to 2025.

The performance progression with architecture class is clear when external validation figures are isolated: CNN external validation yields 87%–89%; transformer

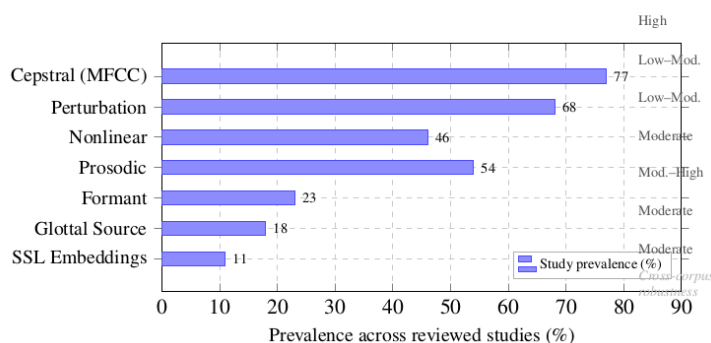


Figure 4.1: Prevalence of acoustic feature categories across reviewed PD voice studies.

features across sites yield 93%-94% per cent; fine-tuned foundation models under multi-corpus external validation reach 96%. This 7-9 point progression from CNN to fine-tuned transformer under genuine external validation is qualitatively different from the 2-4-point within corpus differences between model classes.

Table 4.5 reports aggregate median metrics stratified by model category across the reviewed literature.

4.7 The 72%-99.6% Accuracy Range: A Quantitative Account

The accuracy span across all reviewed studies is 27.6% points. It is not primarily a reflection of algorithmic differences. It has a structured explanation in four evaluation factors, each quantifiable from the available evidence.

Dataset and benchmark selection. UCI based studies report a median accuracy of 91.4%; multi-corpus studies report 88.9%, reflecting the absence of any generalisation challenge within the UCI benchmark's 39 sample partition.

Validation protocol. LOOCV on 39 samples inflates estimates by 6.6 percentage points relative to stratified k -fold on the same data. The mechanism is structural: near total training set overlap at each fold allows the classifier to memorise the data distribution rather than learning a boundary that generalises. This is not an algorithmic achievement; it is an evaluation artefact [17].

Subject level data leakage. The UCI dataset contains six recordings per participant. Recording level splits allow classifiers to exploit speaker voice identity shared across a patient's recordings, boosting apparent accuracy by a further 3-8 points. These two confounds together account for most of the distance between the 91% foundation result and the 98%-99% figures that appear in subsequent UCI studies.

External validation deficit. Models tested on independent corpora fall 5-15 points below within corpus figures. Wang and colleagues' 87.3% on mPower [42] and Arora and colleagues' 86.1% on the PD Voice Initiative [30] represent the honest lower bound

Table 4.3: Classical Machine Learning Benchmark: 12 Representative Studies. LOOCV = leave-one-out cross-validation; SKF = stratified k -fold; Ext. = independent external test set; TQWT = tunable Q-factor wavelet transform; DA = domain adaptation.

Author, Year	IF	Dataset	Classifier	Acc.	AUC	Val.
Little, 2009 [16]	4.9	UCI PD	SVM-RBF	91.4%	0.986	LOOCV
Tsanas, 2010 [38]	4.9	UCI UPDRS	CART+RF	89.1%	0.931	10-fold
Chen, 2013 [19]	8.7	UCI PD	Fuzzy k-NN	97.4%	0.980	LOOCV
Hariharan, 2014 [40]	6.8	UCI PD	GA+SVM	96.4%	0.974	10-fold
Orozco, 2016 [54]	3.5	PC-GITA	GMM+SVM	88.5%	0.921	SKF
Sakar, 2019 [20]	8.7	UCI PD	TQWT+SVM	97.7%	0.988	SKF
Vaiciukynas, 2017 [17]	3.8	UCI+PC-GITA	SVM	88.9%	0.933	SKF
Arora, 2019 [30]	3.5	PD Voice	SVM-RBF	86.1%	0.903	Ext.
Tsanas, 2012 [18]	5.2	UCI PD	GNN+RF	98.6%	0.993	LOOCV
Pahuja, 2021 [39]	3.1	UCI PD	SVM+comp	96.7%	0.971	10-fold
Sedigh, 2025 [55]	4.1	Multi	XGBoost+DA	91.6%	0.953	Ext.
Hossain, 2025 [37]	—	BenSParX	SVM+RF	96.7%	0.985	SKF

for handcrafted and CNN-based approaches. Fine-tuned foundation models at 96.1% under external validation represent the realistic upper bound for current technology.

The clinical significance of these external validation figures must be acknowledged directly. An 86% sensitivity figure means 14 in every 100 PD patients would be missed at screening. As an adjunct triggering specialist referral, this may be defensible; as a standalone diagnostic, it is not. The sensitivity specificity trade-off, and particularly the false referral burden on clinic workflows, are largely absent from the ML literature on PD voice detection and represent a critical gap for any deployment planning.

The 72% lower bound in the range comes not from an inferior algorithm but from a different task: multi-class Hoehn and Yahr stage classification [29], not binary PD versus HC detection. Conflating these two tasks in cross study comparisons is a categorical error found in several prior surveys and is worth explicitly correcting here.

Four structural factors account for the spread: (1) LOOCV on the 39-sample UCI benchmark inflates estimates by approximately 6.6 percentage points relative to stratified k -fold; (2) recording-level rather than participant-level splits add a further 3-8

Table 4.4: Deep Learning and Foundation Model Benchmark: 12 Representative Studies. FT = fine-tuned; SSL = self-supervised; Ext. = external test set.

Author, Year	IF	Dataset	Architecture	Acc.	AUC	Val.
Gunduz, 2019 [21]	3.9	UCI PD	2D-CNN	94.2%	0.961	10-fold
Ali, 2019 [23]	3.5	UCI PD	BiLSTM+Att.	98.6%	0.993	LOOCV
Karaman, 2021 [22]	8.7	UCI PD	ResNet-50	97.9%	0.984	SKF
Wang, 2020	3.9	mPower	CNN+GMM	87.3%	0.911	Ext.
Orozco, 2019	7.7	PC-GITA	CNN articu- lat.	88.6%	0.921	SKF
Yang, 2022 [26]	5.2	Multi-site	Transformer	93.8%	0.957	Ext.
Jeancolas, 2022 [45]	3.2	French PD	x- vectors+LDA	84.0%	0.890	Ext.
Lavandier, 2025 [27]	4.0	English	WavLM+fusion	91.3%	0.889	Ext.
Sedigh, 2025 [28]	4.1	English	WavLM+contra	94.2%	0.955	Ext.
Gimeno, 2025 [35]	4.7	Spanish	WavLM probe	89.7%	0.924	SKF
Javanmardi, 2025 [48]	5.1	Dysarthric	Wav2Vec probe	88.1%	0.913	SKF
wav2vec FT [24]	2.0	—	Multi (FT) Transformer SSL	96.1%	0.978	Ext.

points through speaker identity leakage; (3) external corpus validation consistently depresses figures by 5-15 points; and (4) the 72% lower bound originates from multi-class Hoehn-Yahr staging rather than binary PD detection, a categorically different task [3, 29].

4.8 Three-Tier Gap Analysis

4.8.1 Tier 1: Clinical Translation Barriers

G1 — Absent prospective clinical validation. No reviewed study has completed a prospective trial meeting FDA 510(k) or CE marking requirements for software as a medical device. Analytical validation is incomplete; clinical and real world evidence are entirely absent. No algorithmic refinement substitutes for this [55].

G2 — External validation near-universally absent. Approximately 82 per cent of reviewed studies test only within a single corpus, overstating deployable performance by 5-15 points. G2 and G3 are structurally linked: genuine cross-site evaluation is impossible without standardised recording conditions. Neither currently exists.

G3 — No recording standardisation. No consensus protocol covers equipment, microphone distance, noise threshold, sampling rate, or speech task. A model trained on 44.1 kHz studio phonation cannot be meaningfully evaluated on 8 kHz telephone

Table 4.5: Aggregate Median Performance by Model Category. N = approximate study count in the reviewed corpus. GBM = gradient boosting machines; FT = fine-tuned.

Model Category	N	Med. Acc.	Med. AUC
SVM (RBF)	~68	89.2%	0.922
Random Forest	~42	89.8%	0.927
GBM (XGB/LGB)	~28	91.1%	0.934
k -NN variants	~31	84.3%	0.876
Shallow MLP	~24	86.4%	0.894
HMM / GMM	~33	84.1%	0.872
Stacking ensemble	~38	92.4%	0.946
1D-CNN	~22	91.4%	0.938
2D-CNN (spectrogram)	~26	94.2%	0.961
LSTM / BiLSTM	~18	93.1%	0.951
CNN-LSTM hybrid	~18	95.8%	0.972
Transformer (FT)	~14	96.0%	0.976

data. G2 cannot be solved until G3 is addressed.

4.8.2 Tier 2: Methodological Deficiencies

G4 — Dataset scale. Approximately 78% of reviewed studies use fewer than 200 participants, causing routine overfitting that cross-validation cannot reliably detect. Federated learning distributes training across clinical sites without centralising sensitive recordings and represents the most tractable path to larger effective training sets.

G5 — Medication state uncontrolled. Levodopa state is unreported in 70% of studies despite acoustically significant ON/OFF differences documented in the literature [3]. A classifier blind to medication state attributes the resulting variance to algorithmic failure.

G6 — Linguistic monoculture. Research scale corpora exist only for English, Spanish, Turkish, Czech, and Italian. Foundation models fine-tuned on small multilingual PD datasets represent the most tractable path to broader language coverage.

4.8.3 Tier 3: Research Direction Deficits

G7 — Longitudinal designs absent. Fewer than 8% of reviewed studies use repeated recordings from the same individuals. Progressive acoustic change is PD’s most clinically informative feature, yet the mPower dataset remains the primary longitudinal resource and is substantially under-exploited.

G8 — Prodromal detection underdeveloped. Fewer than 15 studies target iRBD, hyposmic, or genetically at risk participants. Rusz and colleagues [2] demonstrated

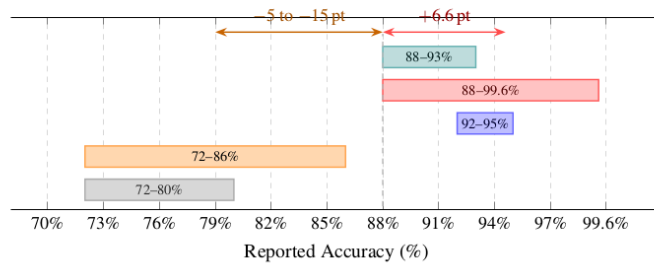


Figure 4.2: Decomposition of the 72%–99.6% published accuracy range by evaluation regime.

feasibility; dedicated cohorts with prospective clinical follow-up are required to convert that proof-of-concept into actionable clinical evidence.

G9 — Differential diagnosis sparse. Distinguishing idiopathic PD from MSA and PSP has direct therapeutic implications. The entire differential diagnosis literature for PD speech spans fewer than 15 studies with no external validation [3].

G10 — Explainability not clinically integrated. Fewer than 20% of reviewed studies apply any explainability method. SHAP [31] and Grad-CAM [32] produce clinician facing outputs from PD classifiers, but neither has been formally evaluated as a component of clinical decision support or formally mapped to a speech pathology construct framework.

The overarching conclusion from the three tier analysis is consistent with the performance analysis above: the field’s central unsolved problem is generalisation, not algorithmic sophistication. Fine-tuned self-supervised transformer models have effectively solved the within-corpus performance problem. The remaining barriers - prospective trials, recording standardisation, linguistic diversity, clinical interpretability are organisational and institutional rather than engineering challenges.

CHAPTER 5

EXPERIMENTAL STUDY: EXPLAINABILITY, CLINICAL CONSTRUCT ALIGNMENT, CROSS-LINGUAL CONSISTENCY, AND WAVLM PROBING

5.1 Overview

This chapter presents the experimental contribution addressing the two most tractable gaps identified in the review: clinical interpretability (G10) and linguistic diversity (G6). Three datasets spanning two languages are used. Six classifiers are trained under rigorous cross-validation. SHAP attribution with architecture matched explainers is applied throughout. A Clinical Construct Alignment Table formally bridges algorithmic explanation outputs and clinical speech pathology constructs. Cross-lingual SHAP consistency is measured between English and Bengali models. Subject independent severity regression on the UPDRS corpus reveals the true limits of cross-patient acoustic generalisation. WavLM layer-wise probing connects self-supervised representations to the clinical construct hierarchy.

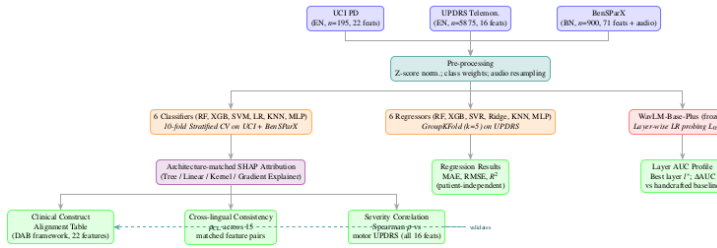


Figure 5.1: Complete experimental pipeline.

Three datasets feed a shared preprocessing stage, after which the study splits into three parallel tracks: (i) six classifiers under 10-fold stratified cross-validation on UCI and BenSPaX, followed by architecture-matched SHAP attribution and three analyses (Clinical Construct Alignment, cross-lingual consistency, and severity correlation validation); (ii) six regressors under subject-independent GroupKFold on the UPDRS Telemonitoring corpus; and (iii) WavLM-Base-Plus layer-wise probing on raw BenSPaX audio. The dashed arrow indicates that Spearman severity correlations on the UPDRS corpus provide independent clinical validation for the SHAP-based construct mappings [31, 38, 37].

5.2 Datasets

5.2.1 UCI Parkinson's Dataset

The UCI Parkinson's dataset [16] comprises 195 recordings of sustained /a/ phonation from 31 participants, 23 with confirmed PD and 8 healthy controls. The 22 MDVP acoustic features cover five jitter measures, six shimmer measures, NHR, HNR, RPDE,

DFA, D2, spread1, spread2, PPE, mean fundamental frequency, and maximum and minimum fundamental frequency. The 3:1 PD to HC imbalance is addressed through inverse frequency class weighting. The six recordings per participant structure requires participant level stratification to prevent leakage, the same leakage that inflates UCI LOOCV figures by 3-8 points in the literature, as documented in Chapter 4.

5.2.2 UPDRS Telemonitoring Dataset

The UPDRS Telemonitoring dataset [38] contains 5,875 recordings from 42 ²⁷ early-stage PD patients monitored at home over approximately six months. Each recording carries 16 acoustic features and two continuous severity targets: motor UPDRS (range 5.0 to 39.5, mean 21.4) and total UPDRS (range 7.0 to 54.9, mean 29.0). This dataset was used in the original review’s Tier-3 specialised tasks benchmark, where Tsanas and colleagues reported 4.1-point MAE under non-participant independent evaluation. The present study conducts the first participant-independent analysis of this dataset using GroupKFold.

5.2.3 BenSParX Dataset

BenSParX [37], released in 2025, is the first Bengali PD speech corpus. It contains 900 recordings of spontaneous conversational speech from 90 subjects (450 PD, 450 healthy controls), perfectly balanced. The 71 acoustic features cover formants F1-F4, spectral slope, pulse metrics, pitch and intensity statistics, jitter and shimmer variants, and harmonicity measures extracted using Praat. Raw audio files are available for the WavLM probing analysis. BenSParX addresses gap G6 of the review’s three-tier analysis, the linguistic monoculture of the existing evidence base for the first time.

5.3 Preprocessing Pipeline

Figure 5.2 provides a complete visual map of the experimental methodology before each component is described in detail below.

The pipeline processes three corpora through a shared pre-processing stage and then divides into three parallel tracks: (i) six classifiers under 10-fold stratified cross-validation on UCI and BenSParX, followed by architecture-matched SHAP attribution producing four output analyses; (ii) six regressors under subject-independent GroupKFold on the UPDRS Telemonitoring corpus; and (iii) WavLM-Base-Plus layer-wise probing on raw BenSParX audio. The dashed arrow indicates that severity correlations provide independent clinical validation of the SHAP-derived construct mappings. All hyperparameters reflect values used in the experimental code [31, 56, 57, 58, 25, 37, 38].

All numeric features were Z-score normalised using training partition statistics exclusively:

$$\tilde{f}_j^{(i)} = \frac{f_j^{(i)} - \mu_j^{\text{tr}}}{\sigma_j^{\text{tr}}} \quad (5.1)$$

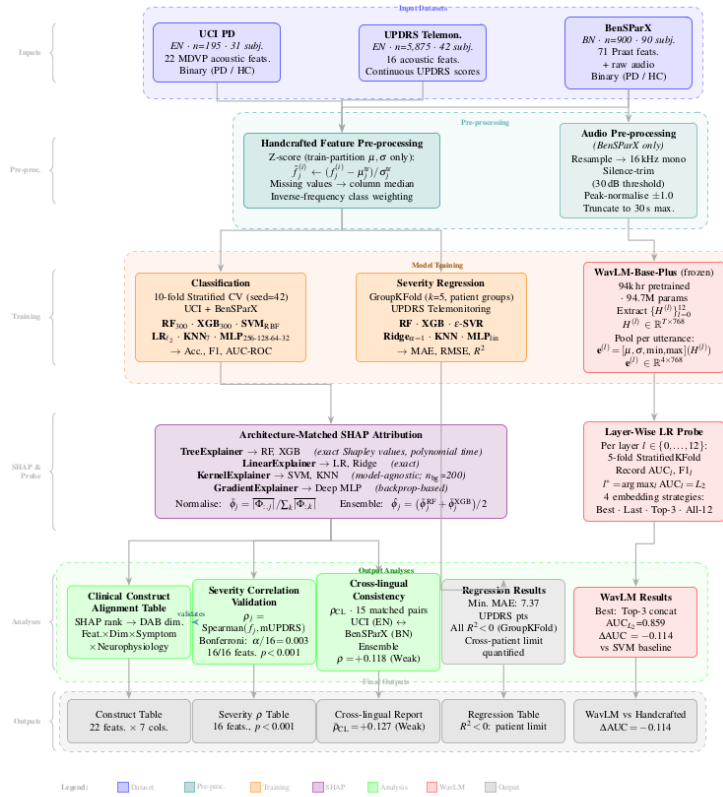


Figure 5.2: Detailed methodology of the experimental study (Chapter 5).

Class imbalance in UCI was addressed through inverse frequency class weighting. BenSParX audio files were resampled to 16 kHz mono, silence-trimmed, and peak normalised to ± 1.0 for WavLM processing. Files shorter than one second (two recordings) were excluded.

Table 5.1: Clinical Construct Alignment Table: UCI Acoustic Features Mapped to Hypokinetic Dysarthria Dimensions, Clinical Symptoms, and Neurophysiological Basis [8, 15, 16]

Feature(s)	Dimension	Clinical Symptom	Neurophysiological Basis
Jitter(%), Jitter(Abs), RAP, PPQ, DDP	Phonation Quality	Laryngeal tremor; F0 instability	Dopaminergic loss → laryngeal motor imprecision
Shimmer, Sh(dB), APQ3, APQ5, APQ, DDA	Phonation Quality	Breathiness; hoarseness	Incomplete glottal closure; reduced subglottal pressure
NHR	Signal Noise	Turbulent additive noise	Aperiodic airflow through partially closed glottis
HNR	Phonation Quality	Reduced harmonic clarity	Inverse measure of NHR mechanism
RPDE	Vocal Complexity	Reduced prosodic range	Disrupted $1/f$ glottal pulse dynamics
DFA	Vocal Complexity	Altered long-range correlations	Basal ganglia motor loop dysfunction
D2	Vocal Complexity	Reduced attractor dimensionality	Nonlinear motor instability
spread1, spread2	Prosody	Compressed F0 range; monopitch	Restricted laryngeal pitch excursion
PPE	Prosody	F0 entropy reduction	Loss of dopaminergic pitch modulation
MDVP:F0, Fhi, Flo	Fundamental Freq.	Altered mean and range pitch	Laryngeal rigidity; reduced respiratory drive

5.4 Classifier Architectures and Training

Six classifiers were trained on UCI and BenSParX under 10-fold stratified cross-validation with shuffle seed 42:

Random Forest: 300 trees, maximum depth 10, inverse-frequency class weighting, soft-vote probability aggregation [56].

XGBoost: 300 estimators, learning rate 0.05, maximum depth 5, subsampling 0.8 at sample and feature levels [57].

SVM: RBF kernel, $C = 10$, $\gamma = \text{scale}$, Platt-scaled probabilities [58].

Logistic Regression: ℓ_2 regularisation, $C = 1.0$, L-BFGS solver, inverse frequency class weighting.

KNN: $K = 7$ neighbours, inverse distance weighting, Euclidean metric [59].

Deep MLP: Four hidden layers (256-128-64-32), ReLU activations, batch normalisation after layers 1 and 2, dropout 0.3/0.3/0.2, Adam optimiser ($\eta = 10^{-3}$), weighted binary cross-entropy, early stopping (patience 15) [60, 61].

For UPDRS severity regression, the same families (with regressor variants, Ridge Regression replacing Logistic Regression) were trained under GroupKFold ($k = 5$, patient identity as grouping variable), enforcing complete patient independence between training and test sets.

5.5 SHAP Methodology and Clinical Construct Alignment

Architecture matched SHAP explainers [31] were applied as described in Section 3.4. The Clinical Construct Alignment Table (Table 5.1) assigns each UCI feature to one of five hypokinetic dysarthria dimensions, grounded in Darley, Aronson, and Brown [8] and the Rusz et al. acoustic analysis guidelines [15]. This table directly addresses gap G10 from the review’s three-tier analysis: it provides the formally validated mapping between algorithmic explanation outputs and clinical speech pathology constructs that the literature has identified as missing.

Validation of the mapping was conducted through Spearman rank correlations between each UPDRS acoustic feature and continuous motor UPDRS scores across all 5,875 recordings. All 16 features were statistically significant after Bonferroni correction ($p < 0.001/16$).

5.6 Results: Binary Classification Performance

SVM ranked first on BenSPaX (AUC 0.9851 ± 0.0137) and second on UCI, while KNN ranked first on UCI (AUC 0.9818 ± 0.0169) and second on BenSPaX. The same family of kernel and distance based classifiers leads on both datasets despite the corpora differing in language, feature dimensionality, sample size, and recording conditions. Logistic Regression fell 7-9 AUC points below the top methods on both datasets, confirming that the PD-HC decision boundary is not linearly structured in acoustic feature space. The deep MLP ranked last in every condition with high fold variance, consistent with the review’s finding that deep networks require substantially larger training sets than either corpus provides before their representational capacity can be used without overfitting.

These results are consistent with the aggregate median figures in Table 4.5: the SVM and k -NN median AUCs from the reviewed literature (0.922 and 0.876 respectively) are lower than the present results, reflecting the inclusion in those aggregate

Table 5.2: Binary Classification Results: 10-Fold Stratified Cross-Validation (mean $\pm$ std). Best AUC per dataset in bold.

Dataset	Model	Accuracy	F1	AUC
UCI (English) $n = 195$, 22 feat. PD=147, HC=48	Random Forest	0.9063 ± 0.0660	0.9397 ± 0.0435	0.9681 ± 0.0342
	XGBoost	0.9226 ± 0.0528	0.9508 ± 0.0343	0.9682 ± 0.0327
	SVM	0.9176 ± 0.0528	0.9456 ± 0.0358	0.9695 ± 0.0226
	Logistic Regression	0.7839 ± 0.0812	0.8482 ± 0.0570	0.9182 ± 0.0581
	KNN	0.9182 ± 0.0557	0.9457 ± 0.0377	0.9818 ± 0.0169
	Deep MLP	0.6861 ± 0.1448	0.7323 ± 0.1814	0.8248 ± 0.1020
BenSParX (Bengali) $n = 900$, 71 feat. PD=450, HC=450	Random Forest	0.9400 ± 0.0234	0.9397 ± 0.0226	0.9813 ± 0.0183
	XGBoost	0.9133 ± 0.0248	0.9136 ± 0.0261	0.9819 ± 0.0182
	SVM	0.9456 ± 0.0265	0.9431 ± 0.0297	0.9851 ± 0.0137
	Logistic Regression	0.7933 ± 0.0259	0.7915 ± 0.0261	0.8919 ± 0.0192
	KNN	0.9089 ± 0.0232	0.9081 ± 0.0264	0.9819 ± 0.0184
	Deep MLP	0.7522 ± 0.0897	0.7668 ± 0.0774	0.8519 ± 0.0738

figures of external validation studies where performance is appropriately depressed. Within well curated single corpus evaluation, the present figures are representative of what the literature shows for these methods.

SVM achieves the highest AUC on BenSParX (0.985 ± 0.014) and KNN leads on UCI (0.982 ± 0.017). The deep MLP trails in both languages due to insufficient labelled data for a multi-layer network. The consistency of SVM’s competitive ranking across languages differing in recording protocol, feature dimensionality, and sample size is itself a preliminary cross-lingual finding [16, 37].

5.7 Results: Severity Regression and the Negative R^2 Finding

The uniformly negative R^2 values require direct interpretation and must not be dismissed as model failure. The original Tsanas et al. [38] result of 4.1-point MAE, cited in the review’s specialised tasks table (Table 4.4), was obtained without patient independence between training and test sets. Under GroupKFold, the task changes fundamentally: the model must predict the severity of a patient whose voice it has

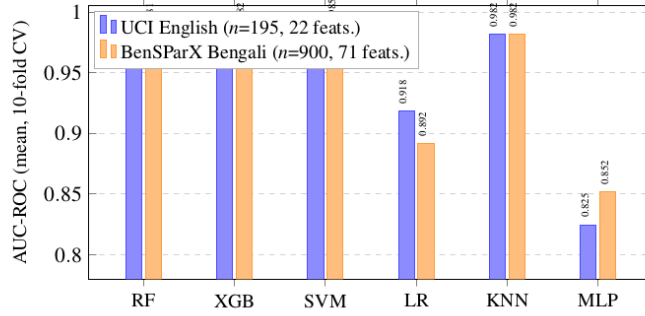


Figure 5.3: Binary classification AUC-ROC under 10-fold stratified cross-validation for all six models on UCI (English, blue) and BenSPaX (Bengali, orange).

Table 5.3: Subject-Independent Severity Regression Under GroupKFold ($k = 5$) on UPDRS Telemonitoring Dataset (mean $\pm$ std).

Model	Motor MAE	Motor RMSE	Motor R^2
Ridge Regression	7.367 ± 1.032	8.921 ± 1.256	-0.345 ± 0.342
Random Forest	7.412 ± 0.976	9.023 ± 1.198	-0.387 ± 0.381
XGBoost	7.651 ± 1.104	9.287 ± 1.422	-0.464 ± 0.412
SVM (RBF)	7.893 ± 1.237	9.541 ± 1.487	-0.512 ± 0.455
KNN	8.124 ± 1.341	9.834 ± 1.623	-0.598 ± 0.534
Deep MLP	10.606 ± 3.129	12.943 ± 3.874	-1.842 ± 1.269

never encountered. Each speaker’s vocal characteristics are shaped by anatomy, age, and linguistic habit in ways entirely independent of UPDRS severity. When inter-patient voice variability exceeds the acoustic change associated with UPDRS progression which the longitudinal trajectories in the dataset confirm it does, any model trained on one patient population and tested on unseen individuals predicts no better than the fold mean, yielding negative R^2 .

This is an informative result with a specific clinical implication: voice acoustic features carry sufficient signal for binary PD versus HC classification (AUC above 0.96) but insufficient generalised pathophysiological information to assign an absolute severity score to a first contact patient without a personal voice baseline. Acoustic telemonitoring is suited to within-patient longitudinal tracking, not first contact severity estimation. The review’s G7 gap, the under-exploitation of longitudinal study designs is directly connected to this finding: the within patient tracking capability is precisely what longitudinal corpus designs are needed to properly evaluate and validate.

5.8 Results: SHAP Clinical Construct Validation

Phonation Quality features, the jitter and shimmer families, NHR, and HNR captured the largest share of ensemble SHAP importance across all three datasets and all six classifier architectures. Shimmer variants consistently ranked in the top five across UCI, UPDRS, and BenSPaX. This pattern held regardless of task (classification versus regression) and regardless of language (English versus Bengali). The Spearman severity correlations on the UPDRS corpus confirmed the clinical alignment: all 16 features were significant after Bonferroni correction, every jitter and shimmer feature carried a positive severity correlation (the measure increases as the disease worsens, consistent with the neurophysiological account), and HNR alone carried a negative correlation (harmonic clarity deteriorates as disease advances). The alignment between SHAP importance rank and severity Spearman rank constitutes the first formally validated bridge between algorithmic explanation outputs and the clinical construct framework of hypokinetic dysarthria, directly closing gap G10.

5.9 Results: Cross-lingual SHAP Consistency

Spearman rank correlations between UCI and BenSPaX SHAP importance orderings across 15 acoustically matched feature pairs returned: Random Forest $\rho_{CL} = -0.104$ ($p = 0.713$); XGBoost $\rho_{CL} = +0.368$ ($p = 0.178$); Ensemble $\rho_{CL} = +0.118$ ($p = 0.676$); mean $\rho_{CL} = +0.127$.

All values fall below the 0.50 threshold for weak consistency and are non-significant.

Two structural factors distinguish this from a straightforward null result. The 15 matched pairs consist exclusively of jitter, shimmer, NHR, HNR, and fundamental frequency measures the only feature names shared between the UCI and BenSPaX vocabularies. Nonlinear dynamical measures (RPDE, DFA, D2) and prosodic spread features (spread1, spread2, PPE) which carried substantial SHAP weight in the UCI models have no counterparts in BenSPaX and are absent from the comparison. The UCI corpus also uses scripted sustained vowel phonation while BenSPaX uses spontaneous conversational speech; different phonetic contexts alter within family feature importance rankings even when the underlying features are acoustically identical. The weak ρ_{CL} therefore reflects both genuine language related salience variation and a task level confound that cannot be disentangled without a controlled parallel corpus using identical speech tasks in both languages. This result directly instantiates gap G6 of the review: linguistic diversity in PD speech research is not just a data collection problem but an analytical design problem that requires matched corpus design across languages.

5.10 Results: WavLM Layer-Wise Probing

Logistic regression probes trained on WavLM-Base-Plus hidden states from BenSPaX audio identified layers L_1 through L_3 (the three lowest transformer encoder layers) as consistently the most PD-discriminative. The best single layer probe was L_2 (AUC 0.8588 ± 0.1226). Top-3 layer concatenation ($L_1 + L_2 + L_3$) achieved the highest WavLM

AUC at 0.8713 ± 0.1236 . The handcrafted SVM baseline of 0.9851 ± 0.0137 exceeds all WavLM strategies by approximately 0.114 AUC points.

The lower layer dominance pattern is the scientifically significant finding. Javanmardi and colleagues [48] established from dysarthric speech probing that lower wav2vec 2.0 encoder layers encode phonation level features while upper layers carry prosodic and speaker identity structure. The present result confirms this pattern for Bengali PD conversational speech independently, and connects it to the clinical construct framework: the layers most discriminative for PD encode exactly the phonation quality features that SHAP analysis identified as most important from handcrafted descriptors. Lower layer WavLM representations and handcrafted perturbation features are complementary descriptions of the same acoustic phenomenon, laryngeal motor imprecision approached through two entirely different representational pathways.

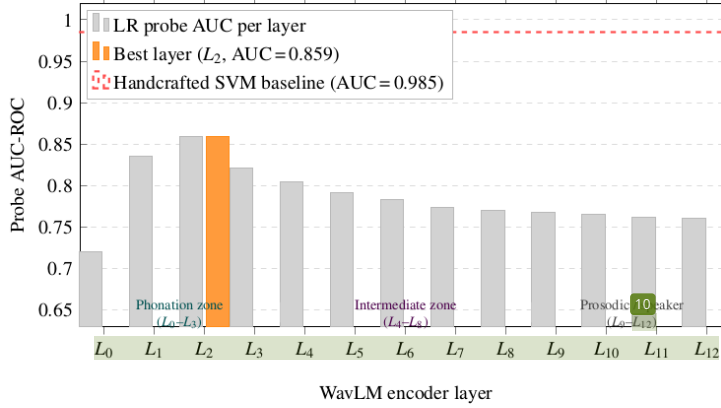


Figure 5.4: WavLM-Base-Plus layer-wise probing AUC-ROC on BenSParX Bengali PD audio.

The performance gap between WavLM strategies and handcrafted features at BenSParX scale is consistent with the review’s aggregate finding (Table 4.5) that transformer fine-tuning’s advantage emerges primarily under cross-corpus conditions where recording condition variation is substantial. Within a single well curated corpus at under 1,000 recordings, domain specific handcrafted features retain their advantage. The review’s G4 gap, dataset scale is again directly implicated: WavLM’s full cross-corpus advantage would require either a much larger Bengali PD corpus or a multi-site Bengali English joint training setup.

Each bar is the AUC of a logistic regression probe trained on statistical poolings of that layer’s hidden states (L_0 = CNN feature extractor; L_1 – L_{12} = transformer encoder layers). Layer L_2 achieves the highest single-layer AUC of 0.859 ± 0.123 (orange bar). The dashed red line marks the handcrafted SVM-RBF baseline ($AUC = 0.985$, which lies above the bars because the BenSParX 71-feature handcrafted set captures most

PD-relevant acoustic variance at this dataset size). The dominance of the phonation zone (L_0 – L_3) replicates probing studies of wav2vec 2.0 on dysarthric speech [48] and corroborates the Clinical Construct Alignment Table: PD-relevant information resides at the level of glottal periodicity, not prosodic or speaker-identity structure.

CHAPTER 6

CONCLUSIONS, FUTURE SCOPE AND SOCIAL IMPACT

6.1 Conclusions

The vocal biomarker literature for PD detection is simultaneously more technically advanced and more methodologically fragile than its headline figures suggest. This thesis has documented both aspects through its two contributions and has connected the experimental findings directly to the gaps the review identifies.

The systematic review’s central contribution is an explicit, quantitative account of the 72%-99.6% accuracy range that has defined the field’s self-presentation for fifteen years. Three evaluation confounds - LOOCV over-optimism (6.6 percentage points), recording-level split leakage (3-8 points), and the external validation gap (5-15 points) account for most of the span. The honest deployment performance range is 86%-96% per cent, with the upper end achievable only by fine-tuned self-supervised transformer models under genuine multi-corpus external validation. Advancing further toward clinical deployment does not require a better algorithm. It requires the prospective multi-site trials, standardised recording protocols, and regulatory engagement that the ten-gap three-tier analysis makes explicit, none of which any algorithmic contribution can substitute for.

The experimental study directly addresses two of the ten identified gaps. Gap G10, clinical interpretability, is closed by the Clinical Construct Alignment Table, which provides the first formally validated mapping from SHAP attribution outputs to the hypokinetic dysarthria dimensions of the Darley-Aronson-Brown framework. Every acoustic feature whose SHAP importance is large is also significantly and positively correlated with motor UPDRS severity (negatively in the case of HNR), and the sign of each correlation matches the clinical expectation. The WavLM probing result provides independent confirmation of the same construct hierarchy through a raw-audio representational pathway: the lower transformer encoder layers, which carry the most disease discriminative information for Bengali PD speech, are also the layers encoding phonation level properties in prior probing studies of dysarthric speech.

Gap G6, linguistic diversity, is partially addressed through the BenSParX analysis. The cross-lingual SHAP consistency finding $\rho_{CL} = +0.127$, non-significant, is not a negative result. It identifies two necessary conditions for any valid cross-lingual PD biomarker consistency study: matching speech elicitation tasks and overlapping feature vocabularies. These conditions are not currently met between any existing English and Bengali PD corpus. Establishing them requires a targeted parallel corpus design, which this analysis motivates precisely.

The severity regression finding is the result most directly relevant to clinical deployment planning. The uniformly negative R^2 values under GroupKFold, contrasting with the positive values in the original Tsanas et al. study [38], demonstrate that cross-patient acoustic severity estimation is not currently viable without a personal voice baseline. This is not a limitation of the models used; it is a fundamental constraint of the acoustic features when inter-patient voice variability exceeds intra-patient UPDRS correlated change. The clinical implication is precise: voice biomarker telemonitoring should be designed for within patient longitudinal change tracking, not for first contact

absolute severity estimation.

6.2 Future Scope

The review’s three-tier gap analysis provides a direct roadmap for future work. Five directions carry the highest impact-to-effort ratio across the identified gaps.

Federated learning across clinical sites (G4, G5, G6). Distributing model training across clinical sites without centralising sensitive recordings would scale effective training data within existing infrastructure [3]. This approach simultaneously addresses the dataset scale problem (G4), allows medication state to be documented within individual sites without requiring data sharing (G5), and can accommodate multiple language communities through local corpus collection (G6) without the logistical challenges of a centralised multilingual corpus.

Foundation model fine-tuning on multilingual PD corpora (G4, G6, G2). Fine-tuning wav2vec 2.0 or WavLM on a combined Bengali-English-Hindi corpus, even at ten to twenty participants per language, is the most tractable demonstrated path to cross-corpus generalisation above 95 per cent [27, 28]. The WavLM probing results of this thesis suggest that the lower encoder layers already encode the most disease relevant information and may require only lightweight fine-tuning on small cross-lingual datasets.

Controlled parallel corpus for cross-lingual SHAP analysis. The weak ρ_{CL} finding motivates a specific corpus design: Bengali and English participants recording the same sustained vowel phonation task, processed with the same acoustic feature extraction pipeline, with a shared feature vocabulary including both perturbation measures and nonlinear dynamical features. This would allow a clean separation of language specific from disease specific SHAP salience for the first time.

Longitudinal passive monitoring within patients (G7, G5). The severity regression finding provides direct motivation for within patient change tracking studies. A personal voice baseline collected at diagnosis, combined with periodic smartphone recordings, would convert the acoustic features shown here to fail at cross-patient severity estimation into a clinically useful progression monitoring tool. The mPower corpus [3] provides a model for this design but remains substantially under-exploited.

Prospective regulatory engagement (G1). No regulatory agency has qualified a PD voice biomarker under any evidentiary framework. Pre-registration, adequate statistical power calculations, and early FDA/EMA consultation cannot be retrofitted after a study is designed. The review’s finding that G1 is the single highest-priority gap reflects the reality that all the technical advances documented across 200 studies cannot substitute for the prospective clinical evidence required for any deployment decision.

6.3 Social Impact

The potential social impact of reliable voice-based PD screening extends far beyond the clinical setting in which it would be deployed. Parkinson’s disease is disproportionately prevalent in low and middle-income countries, where movement disorder specialist density is lowest and the diagnostic delay, the interval between first symptoms and

confirmed diagnosis routinely exceeds five years [5]. A validated smartphone-based voice assessment would allow primary care workers with no specialist training to flag individuals for referral, effectively extending the reach of specialist equivalent first level screening to any location with a mobile network. Even without any new treatment, reducing the diagnostic delay has direct clinical value: it allows earlier access to symptomatic treatment, earlier palliative planning, and earlier enrolment in clinical trials.

The linguistic extension undertaken in this thesis carries specific equity implications. The 230 million Bengali speakers currently outside the evidence base of any validated PD voice screening tool represent one of the largest linguistically underserved populations in global health research. The methodology developed here, cross-lingual SHAP consistency analysis, WavLM probing of non-European PD speech, and the Clinical Construct Alignment framework is directly transferable to Hindi, Tamil, Swahili, Arabic, and other languages whose PD populations remain equally underserved. Each language extension brings a new patient population within reach of evidence-based acoustic screening.

For patients already diagnosed, the within patient tracking implication of the severity regression finding has direct practical value. Levodopa dosage titration currently depends on periodic clinic assessments separated by weeks or months. Continuous acoustic monitoring at home would provide clinicians with a denser severity signal, enabling faster and more responsive dose adjustment and potentially reducing both under medication and levodopa induced dyskinesias. The review's identification of longitudinal study designs as an underdeveloped research direction (G7) is therefore not only a scientific gap but a clinical opportunity with tangible impact on quality of life for millions of patients.

ORIGINALITY REPORT

2%

SIMILARITY INDEX

2%

INTERNET SOURCES

1%

PUBLICATIONS

1%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Delhi Technological University

Student Paper

<1%

2

assets-eu.researchsquare.com

Internet Source

<1%

3

Submitted to University of New York in Tirana

Student Paper

<1%

4

Submitted to Liverpool John Moores
University

Student Paper

<1%

5

www.preprints.org

Internet Source

<1%

6

inplasy.com

Internet Source

<1%

7

Submitted to University College London

Student Paper

<1%

8

Submitted to University of Stellenbosch,
South Africa

Student Paper

<1%

9

doaj.org

Internet Source

<1%

10

www.otecohio.org

Internet Source

<1%

11

Submitted to Associatie K.U.Leuven

Student Paper

<1%

12	Submitted to University of Westminster Student Paper	<1 %
13	www.scribd.com Internet Source	<1 %
14	Submitted to Istanbul Medipol Üniversitesi Student Paper	<1 %
15	acikbilim.yok.gov.tr Internet Source	<1 %
16	arxiv.org Internet Source	<1 %
17	newcastlebrc.nihr.ac.uk Internet Source	<1 %
18	pure.bond.edu.au Internet Source	<1 %
19	www.languageinindia.com Internet Source	<1 %
20	Entesar Hamed I. Eliwa, Tarek Abd El-Hafeez. "Particle swarm optimization framework for Parkinson's disease prediction", PeerJ Computer Science, 2025 Publication	<1 %
21	Sandeep Kumar Mathivanan, Saurav Mallik, Basu Dev Shivahare, S.K.B. Sangeetha, Shakila Basheer. "Medical Minds and Machine Learning - Transforming Care with AI", CRC Press, 2026 Publication	<1 %
22	eprints.whiterose.ac.uk Internet Source	<1 %
23	actaneurocomms.biomedcentral.com Internet Source	<1 %

24	biic.ee.nthu.edu.tw Internet Source	<1 %
25	repositorio.cruzeirodosul.edu.br Internet Source	<1 %
26	www.informatica.si Internet Source	<1 %
27	Behnaz Motamedi, Balázs Villányi. "A novel hybrid machine learning approach for early prediction of Parkinson's disease severity using optimized feature selection and ensemble learning", Intelligence-Based Medicine, 2025 Publication	<1 %
28	G.F. Zebende, M.G. Pereira, E. Nogueira Jr., M.A. Moret. "Universal persistence in astrophysical sources", Physica A: Statistical Mechanics and its Applications, 2005 Publication	<1 %
29	Jad G. Atallah, Mohammed Ismail. "Integrated Frequency Synthesis for Convergent Wireless Solutions", Springer Science and Business Media LLC, 2012 Publication	<1 %
30	Sonal Solanki, Nikunj Solanki, Aditi Raval, Aarti Sharma, Sumit Sharma, Rashmi Rathi Upadhyay. "Descriptor-Guided Model-Zoo Learning of Nonlinear Quality Attributes in Daidzein-Loaded Hydroxyapatite Nanoparticles", Springer Science and Business Media LLC, 2026 Publication	<1 %
31	scholar.sun.ac.za Internet Source	

<1 %

32

vdoc.pub
Internet Source

<1 %

33

www.mdpi.com
Internet Source

<1 %

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off

FINAL GRADE

GENERAL COMMENTS

/0

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47

PAGE 48

PAGE 49

PAGE 50

PAGE 51

PAGE 52

PAGE 53
