Vocal Biomarker Estimator for Parkinson's Disease

Final Technical Report

Author: Arifa Kokab (Group 11)

Program: M.Sc(Eng), Applied Artificial Intelligence — University of San Diego

AAI-590 Capstone Project

Abstract

This capstone presents the *Parkinsons-Vocal-Biomarker-App*, a two-part, voice-based system for Parkinson's disease (PD) screening and severity monitoring from short sustained phonations. Part 1 is a deployable screening classifier trained on engineered dysphonia features (e.g., jitter, shimmer, HNR, pitch statistics) from the Oxford/UCI PD Detection dataset. A Random Forest with five-fold cross-validated model selection achieved ROC AUC = 0.97 and test accuracy = 0.90 at threshold 0.50; an operating threshold of 0.63 was selected for public use to prioritize clean behavior (Healthy recall = 1.00; PD precision = 1.00). Part 2 addresses progression by forecasting visit-to-visit change in the motor component of the Unified Parkinson's Disease Rating Scale (Δmotor_UPDRS) using the Oxford Parkinson's Telemonitoring dataset. After visit-level aggregation and per-subject normalization, a compact BiGRU+1D-CNN predicts Δ, with the next absolute severity computed as last_score + Δ. On subject-held-out testing, mean absolute error ≈ 0.433; a naïve last-value baseline ≈ 0.427 and a ridge baseline ≈ 0.481. Despite modest directional accuracy for Δ, the model is well-calibrated (slope ≈ 1.01; intercept ≈ −0.35; R^2 ≈ 0.996) and supports clinically framed early-warning thresholds. The screening component is publicly deployed under the CarePath AI Foundation as a research/education prototype; the progression component remains research-only. Results indicate that feature-based voice screening is practical to deploy and that calibrated progression estimates from voice are promising for remote monitoring, pending larger cohorts and prospective validation.

Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by motor symptoms such as bradykinesia, rigidity, resting tremor, and postural instability, alongside non-motor features that affect voice, sleep, mood, and cognition. In clinical practice, diagnosis and follow-up rely on expert examination and structured instruments like the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS), which quantifies impairment and disability across multiple domains (Goetz et al., 2008). Although these methods are the clinical standard, they require clinic access and trained raters, which can be difficult for patients in underserved or remote settings.

Voice offers a practical digital window into PD because subtle changes in phonation often emerge early and reflect altered motor control of the vocal apparatus. Sustained phonation (e.g., an "aaah") captures frequency and amplitude perturbations—commonly summarized as jitter, shimmer, and harmonic-to-noise ratio—that have been linked to PD in prior work and explored for telemonitoring (Little et al., 2009; Tsanas et al., 2010). Because voice can be recorded with a smartphone or browser and processed locally or in the cloud, it is a promising modality for scalable screening and longitudinal tracking.

This capstone builds a two-part system around that idea. The first component is a deployable screening classifier that converts short voice samples into interpretable dysphonia features and predicts the likelihood of PD. The second component models progression by estimating how a diagnosed patient's motor severity will change from one visit to the next, producing calibrated signals that can flag sudden or unusual shifts for clinician review. The screening model is trained on the Oxford/UCI Parkinson's Disease Detection dataset of engineered

acoustic features with Healthy/PD labels (Little et al., 2009). The progression model uses the Oxford Parkinson's Telemonitoring dataset, which includes repeated phonations paired with clinician-rated UPDRS scores over time (Tsanas et al., 2010). In a live system, raw audio would be captured via web or mobile, standardized, transformed into the same feature family, and routed to the appropriate model head.

Part 1 — Classifier for Screening Model

The classifier uses a Random Forest trained on robust dysphonia features—pitch statistics, multiple jitter and shimmer variants, and noise-related measures—to provide a fast, explainable decision suitable for public use. The operating point is chosen to behave conservatively for a screening context. To move beyond the lab and demonstrate feasibility, this model has been deployed as a free public screening tool under the CarePath AI Foundation, where a user records a brief "aaah," the backend extracts features, and the app returns a likelihood-based screening result (Little et al., 2009).

Part 2 — Severity Progression Predictor Model

Screening is only the first step; people already living with PD need support between clinic visits. The second component estimates visit-to-visit change in the motor component of the UPDRS (Δmotor_UPDRS) from short sequences of visit-level features. A compact hybrid network—bidirectional gated recurrent units (BiGRU) alongside a 1D convolutional branch—captures longer-range temporal trends and local patterns while remaining small enough to train on modest longitudinal data. The output is translated into early-warning signals using conservative thresholds,

so that care teams can triage follow-ups without relying on raw regression values (Goetz et al., 2008; Tsanas et al., 2010).

The intent is practical: pair an interpretable, deployable screening head with a research-grade progression head, both grounded in published datasets and evaluated with clinically meaningful framing. The long-term vision is accessible screening and calibrated, at-home monitoring that complement—not replace—clinical judgment and formal assessment (Goetz et al., 2008; Little et al., 2009; Tsanas et al., 2010). In production, the same feature pipeline feeds both heads; the system routes to screening for new users and to progression for enrolled patients. Each model will be explored in more detail individually in this paper.

Classifier Screening Model for Parkinson's Disease (rf_model)

Dataset Summary

I trained the screening model on the Oxford Parkinson's Disease Detection dataset, a multivariate table of sustained-phonation voice recordings captured from 31 speakers (23 with Parkinson's disease), totaling approximately 197 instances and 22 engineered acoustic features plus a binary label, status (0 = healthy, 1 = PD). Each row corresponds to one short "aaah" recording; the first column is an identifier and is dropped prior to modeling. Predictors are real-valued (float64) dysphonia measures: fundamental frequency statistics (MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz)), frequency perturbation metrics (Jitter(%), Jitter(Abs), RAP, PPQ, DDP), amplitude perturbation metrics (Shimmer, Shimmer(dB), APQ3, APQ5, APQ11, DDA), noise and harmonic structure (NHR, HNR), and—depending on release—nonlinear/complexity features such as RPDE, DFA, spread1, spread2, D2, and PPE (Little et al., 2008).

I began by confirming class imbalance and basic distributions. The class histogram shows many more PD than healthy recordings, consistent with prior descriptions of this dataset (Figure A1: Class Distribution). Univariate density/histogram panels illustrate the characteristic right-skew of perturbation measures (jitter and shimmer families), a downward shift of HNR in PD recordings, and wider dispersion for pitch extrema (Fhi, Flo) (Figure A2: Per-Feature Distributions). A correlation heatmap reveals strong intra-family correlation among shimmer APQ variants and moderate inverse correlation between perturbation metrics and HNR (Figure A3: Feature Correlation Matrix). These patterns imply that redundant predictors are common;

accordingly, I favored a model class that is robust to correlated inputs without requiring aggressive manual decorrelation.

The raw UCI table has no missing cells; however, in real-world capture I expect occasional invalid acoustic measurements when the phonation is too short or the noise floor is high. To prepare for deployment, I standardized audio to mono 16 kHz WAV and computed features with Parselmouth (Praat), replacing rare invalids conservatively so the inference path matches the training feature space (Boersma & Weenink, 2024; Jadoul et al., 2018). The variables directly support the project goal: jitter and shimmer quantify cycle-to-cycle instability in frequency and amplitude; HNR summarizes harmonic energy relative to noise; and pitch statistics capture control of the vocal source—dimensions that commonly shift in PD hypophonia and dysarthria (Little et al., 2008; Tsanas et al., 2010). On this basis, I expected feature-based models to separate classes well and to produce clinically interpretable importance rankings.

Background Information

Voice-based PD screening has a substantial academic foundation using engineered dysphonia features and classical machine learning, with consistent evidence that perturbation and noise measures carry signal (Little et al., 2008; Tsanas et al., 2010). More recent work uses deep architectures on spectrograms or raw audio and reports gains in some settings (Vásquez-Correa et al., 2019). Commercial groups have operationalized voice biomarkers and smartphone tests at scale, underscoring practical feasibility (Aural Analytics, n.d.; Sonde Health, n.d.; Sage Bionetworks, n.d.).

I selected a Random Forest classifier for the screening head because it handles nonlinearities and interactions, tolerates correlated predictors, exposes feature importances for explainability, and offers fast, low-latency inference suitable for the web (Breiman, 2001). These properties align with the product requirement for a conservative, interpretable screener that can be communicated to non-technical users and clinicians while remaining inexpensive to serve.

Experimental Methods

I trained a RandomForestClassifier on the engineered feature set after removing the identifier column. A stratified train/test split preserved the class ratio; class imbalance was addressed with class weights during fitting. I performed grid search with five-fold stratified cross-validation, tuning n_estimators, max_depth, min_samples_split, min_samples_leaf, and using max_features='sqrt', optimizing weighted F1. The best configuration was n_estimators=100, max_depth=None, min_samples_split=2, min_samples_leaf=1, max_features='sqrt'. Model assessment included ROC and precision–recall curves, class-wise precision/recall/F1, and confusion matrices at multiple operating thresholds. I then selected an operating threshold of 0.63 for deployment, prioritizing perfect healthy recall and very high PD precision to reduce false alarms in a public screening context. I exported a bar chart of impurity-based importances to verify that top features agree with clinical expectations (Figure A4: RF Feature Importances) and plotted the ROC curve to summarize ranking quality (Figure A5: ROC Curve).

Results/Conclusion

At the default threshold of 0.50, test accuracy was 0.90 with ROC-AUC = 0.97. Class-wise metrics at 0.50 showed strong PD sensitivity and precision, with lower recall for healthy cases,

consistent with a PD-sensitive operating point. Shifting the threshold to 0.63 produced accuracy ≈ 0.897, Healthy recall = 1.00, and PD precision = 1.00 on this split, which suits a conservative public screener. The importance profile ranked MDVP:APQ and Shimmer:APQ5 highest, followed by Jitter:DDP, pitch statistics (Fo, Fhi, Flo), and other jitter/shimmer variants; HNR contributed as a countervailing signal. This ordering mirrors the dysphonia literature and strengthens trust that the model relies on physiologically meaningful cues (Little et al., 2008; Tsanas et al., 2010). Cross-validation tracked test performance closely, suggesting no material overfitting for this ensemble size.

The main surprise was how decisively a small threshold adjustment simplified user-facing behavior without materially affecting overall accuracy. If I continue this line, I will add external validation using browser-captured audio, conduct microphone/channel robustness checks, and evaluate calibrated gradient-boosted trees with uncertainty estimates. To prevent identity leakage in datasets with repeated measures per speaker, future splits can enforce speaker-grouped evaluation.

Model Deployment

To move beyond the lab, I deployed the trained classifier as a free public screening tool under the CarePath AI Foundation. The backend is a Flask API served by Gunicorn on Render.com; the frontend is React with Vite. The API converts uploaded audio to mono 16 kHz WAV, extracts the same dysphonia features using Parselmouth (Praat), imputes rare invalids conservatively, loads the serialized forest (rf_model.pkl), and returns both probability and label at the fixed 0.63 threshold. The frontend guides users through a 5-second sustained “aaah,” then displays the likelihood-based result with clear limitations and a non-diagnostic disclaimer. No

voice data or personally identifiable information are stored; processing is in-memory and transient (Boersma & Weenink, 2024; Jadoul et al., 2018; Grinberg, 2018; Gunicorn, n.d.; Meta, n.d.; Vite, n.d.; Render, n.d.). This architecture keeps the compute-heavy step—feature extraction—server-side for consistency with training and provides a stable, low-latency path from browser capture to inference.

Operationally, I monitor threshold behavior and user feedback to balance false reassurance with alarm fatigue. Also, I monitor production for dataset shift using Population Stability Index (PSI) on deployed feature distributions, weekly AUROC and calibration (ECE/Brier) on a small, labeled holdout cohort, and synthetic probe clips spanning microphones/noise conditions; alerts fire when PSI > 0.2 or when AUROC/calibration/probe scores deviate beyond set bounds, triggering threshold re-calibration or a retraining cycle. Productionizing beyond an educational prototype will require dataset shift monitoring across devices, periodic calibration refreshes, and IRB/HIPAA-aligned data governance if real patient data are collected (Goetz et al., 2008).

Severity Progression Predictor Model for Parkinson's Disease

Dataset Summary

I modeled progression using the Oxford Parkinson's Telemonitoring dataset, which contains longitudinal voice recordings paired with clinician-rated UPDRS scores, including the motor_UPDRS component (Tsanas et al., 2010). Each clinic visit includes several sustained-phonation recordings captured during the same session. Because clinical decisions are made at the visit level, I aggregated multiple phonations per visit into a single row (visit-level table) by robust averaging, preserving subject ID and visit day.

The supervised target for this head is the next visit's change in motor severity; I compute that as the one-step difference Δmotor_UPDRS: $\Delta\text{motor\_UPDRS}_{t+1} = \text{motor\_UPDRS}_{t+1} - \text{motor\_UPDRS}_{t}$

Absolute predictions for the next visit are reconstructed from the last observed level plus the predicted change $\hat{y}^{\text{abs}}_{t+1} = \text{motor\_UPDRS}_{t} + \hat{\Delta}_{t+1}$.

I started with exploratory analysis to understand variance, temporal structure, and feature behavior. The distribution of Δmotor_UPDRS is sharply centered near zero with modest tails, which already suggests that sign prediction will be challenging at short horizons (Figure B1, ΔUPDRS histogram). At visit level, motor_UPDRS spans a clinically wide range, supporting regression modeling on the absolute scale (Figure B10). Per-subject statistics confirm substantial between-person differences—means and standard deviations vary widely—so I applied per-subject z-scoring before sequence modeling to remove personal baselines and stabilize optimization $x'_{i,t} = \frac{x_{i,t} - \mu_i}{\sigma_i}$ (Figures B3–B4).

Visits are typically spaced a few days apart, with occasional longer gaps, so sequence windows must tolerate irregular sampling; the gap distribution is summarized in Figure B5. After aggregation, subjects retain roughly two dozen visits on average, which is sufficient for short windows while supporting a subject-held-out evaluation (Figure B6). At the row level**,** motor_UPDRS spans a clinically wide range (Figure B8).

Feature behavior matches expectations from the screening head. At visit level, shimmer/jitter families and HNR/NHR show strong intra-family structure (Spearman heatmap, Figure B9) and plausible relationships to motor_UPDRS in pair plots (Figure B10). To probe predictive value for change, I estimated mutual information between features at time *t* and ΔUPDRS (Figure B11) and Spearman correlations with the same target (Figure B12). Information scores are moderate and correlations are small in magnitude, consistent with the narrow Δ distribution; this reinforces the need for models that emphasize calibration and robust early-warning behavior rather than crisp directional classification. As a defensive preprocessing step, I clipped HNR at the [0.1%, 99.9%] quantiles to stabilize training (Figures B2a–B2b). Subject trajectories show diverse dynamics—steady decline, gradual improvement, and non-monotonic courses—which motivates a hybrid model that captures both slow trends and local transients (Figure B13). The average lag-correlation plot makes the temporal signal explicit: absolute levels are highly autocorrelated and decay with lag, while Δ shows much weaker persistence (Figure B14). Likely data issues include microphone/environmental variability and small day-to-day motor changes (Goetz et al., 2008; Tsanas et al., 2010).

Background Information

The goal here is monitoring, not diagnosis: estimate how severity will change by the next visit so a care team can triage follow-ups. Speech-based telemonitoring for PD has been explored using engineered features with classical regression as well as deep approaches on spectrograms or raw audio (Little et al., 2008; Tsanas et al., 2010; Vásquez-Correa et al., 2019). I chose a compact hybrid of a bidirectional gated recurrent unit (BiGRU) and a shallow 1D-CNN. GRUs summarize temporal dependencies with fewer parameters than LSTMs while still capturing long-range patterns (Cho et al., 2014). A 1D-CNN is effective at detecting short local motifs and adds a complementary inductive bias (Kiranyaz et al., 2016). This hybrid fits sequences that mix slow drifts and short-term fluctuations, which is exactly what the per-visit motor_UPDRS trajectories show (Figure B14). Because the dataset is modest and Δ is small, I emphasized regularization and calibration over capacity.

Experimental Methods

I framed the task as sequence-to-one regression on visit-level features. For each subject, I built fixed-length windows of the last L visits as input and the next visit's Δmotor_UPDRS as target. I normalized features per subject using training-fold statistics only $x'_{i,t} = \frac{x_{i,t} - \mu_i}{\sigma_i}$.

To avoid leakage, I used GroupKFold with subject ID for cross-validation and then evaluated on a subject-held-out test split.

Architecture

I designed a two-branch architecture: a small BiGRU with dropout and L2 weight decay, and a 1D-CNN with a short kernel and global pooling. I concatenated the branch outputs and passed them through a linear head to predict the next-visit delta (the change in motor_UPDRS). I

trained with mean squared error (MSE) loss and the Adam optimizer, used ReduceLROnPlateau for learning-rate scheduling, enabled mixed precision when available, and monitored validation loss with early stopping. The training objective is mean squared error $\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2$.

Optimization

I ran a grid over L ∈ {5, 7, 9} and L2 ∈ {1×10^{-3}, 3×10^{-4}}; the best grouped-CV setting was L = 5, L2 = 1×10^{-3}.

For clinical framing, I translated the predicted delta into conservative alerts using two thresholds: a single-visit absolute change of 0.5: $|\Delta\text{motor\_UPDRS}_{t+1}| \geq 0.5$ and a rolling three-visit absolute-change sum of 3.0: $S_t^{(3)} = |\Delta_{t-2}| + |\Delta_{t-1}| + |\Delta_t|$ trigger if $S_t^{(3)} \geq 3.0$ roughly aligned with commonly cited minimally important differences for UPDRS-III (Goetz et al., 2008).

The table below summarizes the final hyperparameters and the rationale for each choice.

Component	Hyperparameters	Values	Notes / Rationale
Windowing & normalization	Sequence length LL; target; scaling	L = 5 visits; target = Δmotor_{t+1}; per-subject z-score	Short windows fit the visit cadence and available history; subject-level scaling removes baselines and stabilizes training.
BiGRU branch	Hidden size; layers; dropout	GRU hidden = 32, 1 bidirectional layer; dropout = 0.20	Small recurrent state to avoid overfitting on modest data; dropout combats co-adaptation.
1D-CNN branch	Filters; kernel; pooling; dropout	filters = 32, kernel = 3, stride = 1; global avg pool; dropout = 0.20	Short kernels capture local motifs; global pooling yields a compact summary vector.
Optimization	Optimizer; LR; β's; batch; epochs; LR schedule	Adam, lr = 1e-3, β=(0.9, 0.999); batch = 32; max epochs = 200; ReduceLROnPlateau (patience 10, min-lr 1e-5)	Stable defaults for small/medium models; on-plateau LR drop helps escape flat regions.
Regularization & early stop	L2 weight decay; early-stopping; selection	L2 = 1e-3; ES patience = 20 on val MSE; select lowest val loss (seed = 42)	Matches the best CV setting L=5, L2=1e-3); patience balances fit vs. overtraining.

Results/Conclusion

On the subject-held-out test split, the model achieved MAE ≈ 0.433 on absolute motor_UPDRS when reconstructing next-visit level as last + predicted delta $\hat{y}^{abs}_{t+1} = \text{motor\_UPDRS}_t + \hat{\Delta}_{t+1}$

A naïve last-value baseline was ≈ 0.427 MAE on this split; a ridge baseline was ≈ 0.481 MAE. Directional accuracy for the sign of Δ was ≈ 46.5%**,** consistent with the narrow Δ distribution (Figure B1). Critically for monitoring, the absolute predictions were well-calibrated: the calibration scatter shows slope ≈ 1.01, $R^2 \approx 0.996$, and a small negative intercept (Figure B15). Bland–Altman analysis shows bias ≈ −0.19 with limits of agreement around −1.15 and 0.77 (Figure B16), indicating tight agreement in the clinically relevant range (Bland & Altman, 1986).

These results support the design: predicting small per-visit deltas is intrinsically hard, but the hybrid produces calibrated absolute levels that can drive early-warnin**g** flags using conservative thresholds, reducing alarm fatigue. The behavior matches the temporal statistics in Figure B14: strong level persistence but weak **Δ** persistence. Next steps include expanding the longitudinal cohort to strengthen Δ supervision, adding richer acoustic embeddings, and layering uncertainty quantification (e.g., conformal prediction). For productionization, I will add device/channel drift monitoring and scheduled calibration refreshes; workflow integration will emphasize trend summaries and threshold-based flags.

Final Capstone Project Conclusion and Next Steps

The screening model achieves strong discrimination on the test set. At the default threshold of 0.50, the model attains accuracy of 0.90 and a ROC-AUC of 0.97. Precision, recall, and F1 are balanced for the PD class, while Healthy recall is lower because the model is tuned for sensitivity to PD at that threshold. Adjusting the operating point to 0.63 yields a very clean operating behavior for a public screener: Healthy recall is 1.00 and PD precision is 1.00, with overall accuracy of 0.897 on this test split. Feature importance ranks MDVP:APQ and Shimmer:APQ5 at the top, followed by jitter-family features and HNR. This matches clinical expectations that amplitude and frequency perturbations are prominent markers in dysphonia for PD and supports trust in the model's decisions. The artifact exported for deployment reproduces this behavior when the same feature pipeline and threshold are used.

The progression model's goal is different: provide a calibrated early-warning signal rather than a perfect forecast. On held-out subjects, the model reaches a mean absolute error of about 0.433 when predicting the next visit's absolute motor_UPDRS via "last score + predicted Δ." A naïve baseline that simply carries the last score forward has MAE near 0.427 on this split, while a ridge regression baseline achieves approximately 0.481. Directional accuracy for the sign of Δ is modest at roughly 46.5 percent, reflecting the narrow distribution of small changes between visits. Crucially, the progression model is well-calibrated: the calibration slope is about 1.01 with minimal intercept bias, and R^2 approaches 1.0 on the absolute scale. Bland–Altman analysis shows small mean bias with limits of agreement well inside a clinically interpretable band. Framing change detection with two thresholds—a per-visit Δ threshold of 0.5 and a rolling three-visit sum

of 3.0—helps translate regression outputs into actionable flags that can cue check-ins without overwhelming clinicians with noise.

Several outcomes follow from these findings. First, feature-based screening is effective and can be deployed with predictable behavior at a chosen operating point. Second, forecasting ΔUPDRS is challenging because most short-horizon changes are small, but a compact hybrid network can provide calibrated estimates that, when summarized with conservative thresholds, act as a practical early-warning layer for ongoing care. Third, success criteria for real-world use should emphasize calibration, stability, and interpretable thresholds more than raw direction accuracy. If the project continues, priorities include expanding the longitudinal cohort, incorporating richer acoustic representations alongside hand-crafted features, formal uncertainty quantification, and prospective validation with clinician-defined alert policies. Productionizing the progression head would also require data-drift monitoring across microphones and environments and scheduled calibration refreshes.

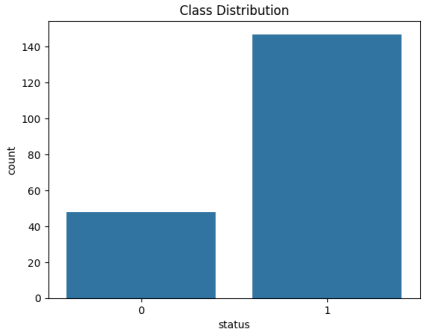
Ethics & Limitations

This system is an educational prototype and not a diagnostic device; it must not replace clinical evaluation. Voice datasets can encode demographic and linguistic biases, so performance may vary by age, sex, accent, language, and comorbidities, as well as by device and environment (microphone type, room acoustics). To mitigate risk, I standardize capture, document limitations, and plan stratified evaluation and bias testing before any clinical use. If real patient audio is collected, I will obtain IRB approval and informed consent and apply privacy-by-design practices (data minimization, encryption in transit/at rest, limited retention, and user deletion rights). The

current deployment processes audio in-memory only and does not store recordings; any future data collection would follow HIPAA-aligned safeguards and clinician oversight.

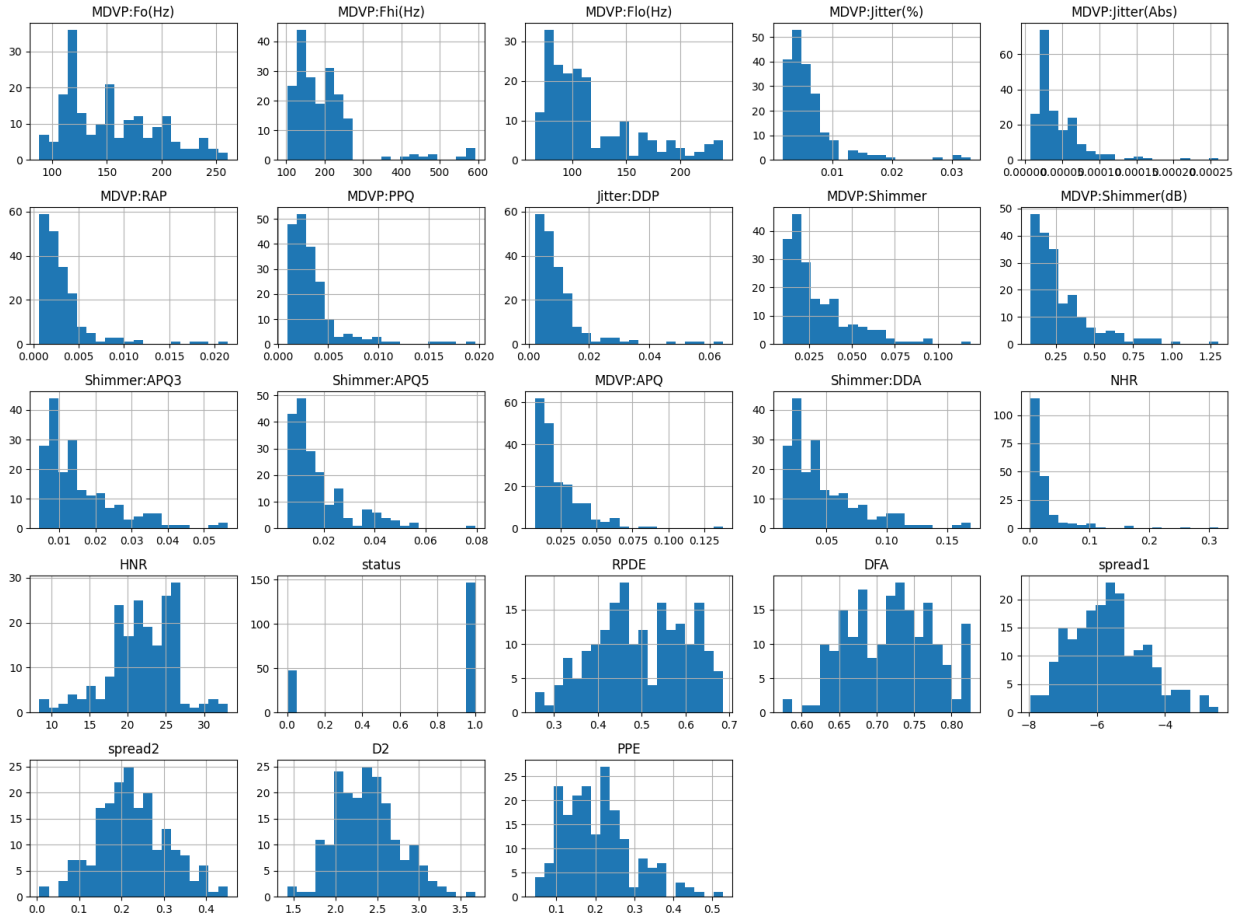
Appendices

Figure A1. Class Distribution



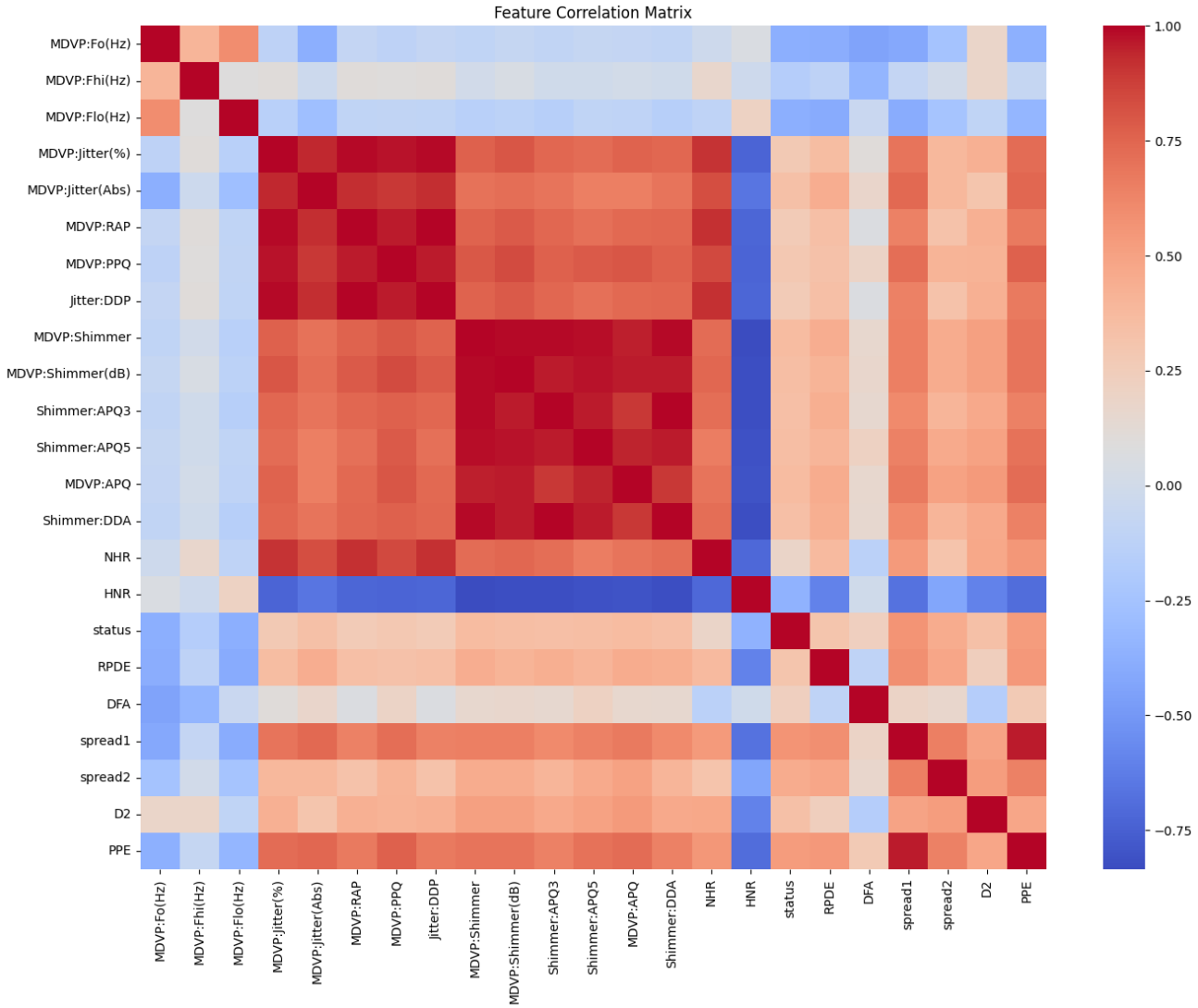
Bar chart showing the imbalance between healthy and PD recordings (status=0 vs. status=1)

Figure A2. Per-Feature Distributions



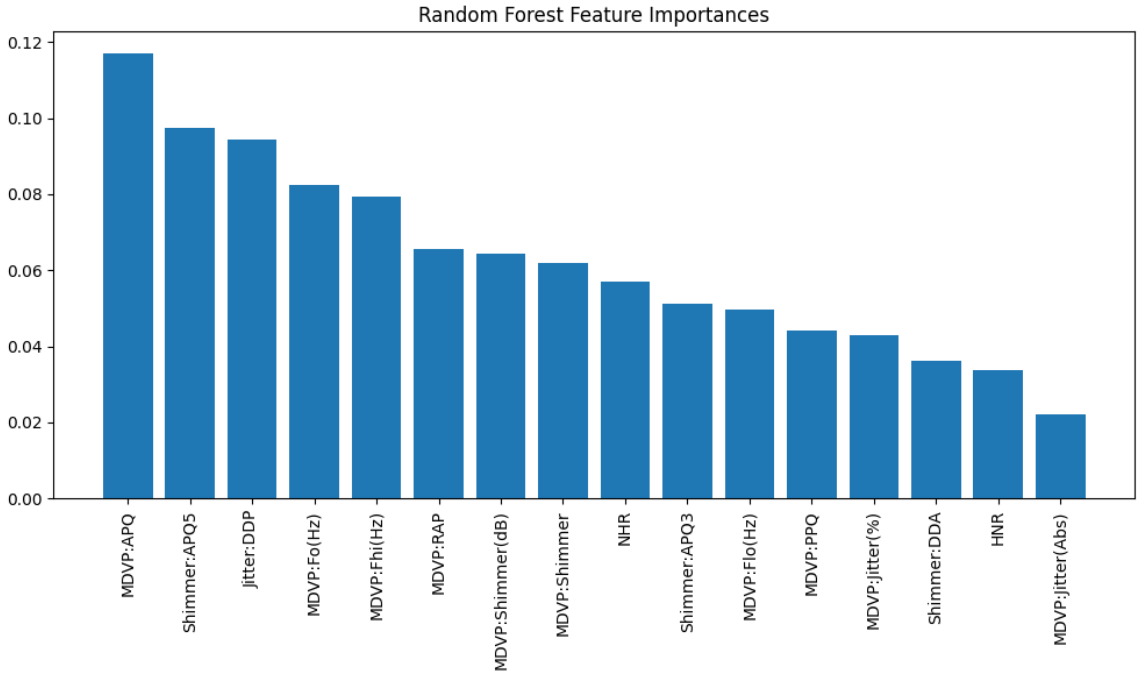
Histogram grid for all predictors, illustrating right-skew in perturbation metrics, dispersion of pitch extremes, and the downward shift in HNR among PD recordings

Figure A3. Feature Correlation Matrix



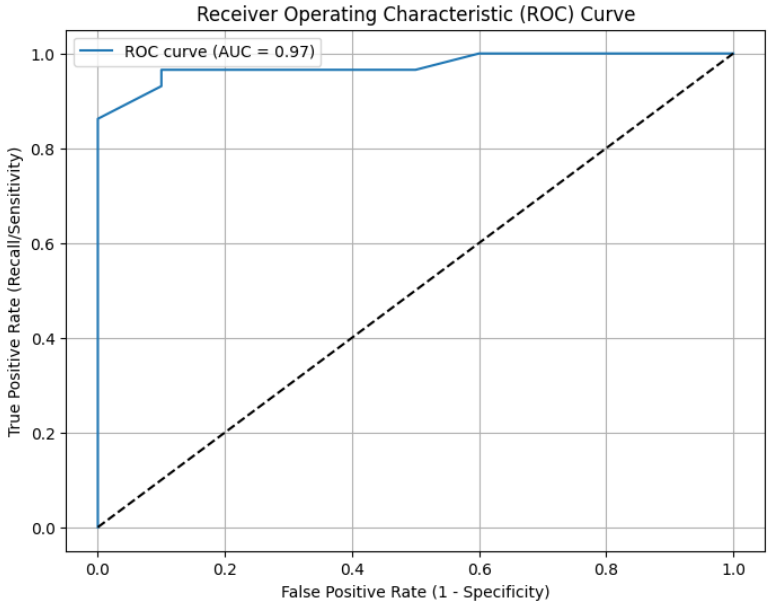
Heatmap showing strong intra-family correlations among shimmer APQ variants and inverse association between perturbation metrics and HNR

Figure A4. Random Forest Feature Importances



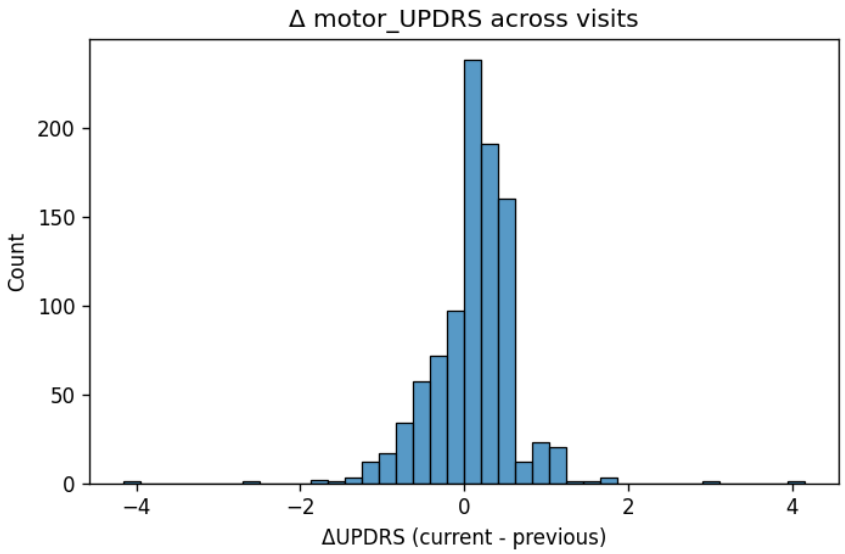
Bar plot of impurity-based importances; APQ and jitter-family features dominate, with pitch statistics and HNR contributing

Figure A5. ROC Curve (AUC = 0.97)



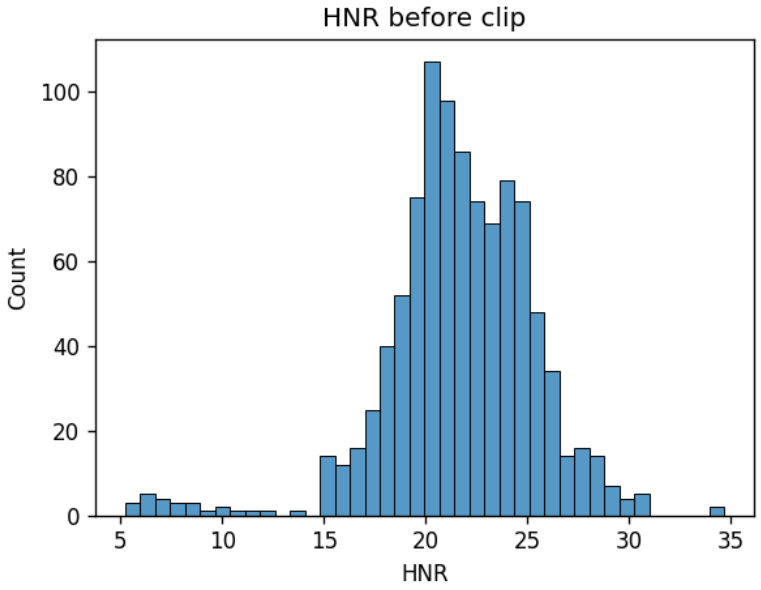
ROC curve summarizing ranking quality across thresholds and supporting the choice of a conservative deployment threshold

Figure B1. Δmotor_UPDRS across visits



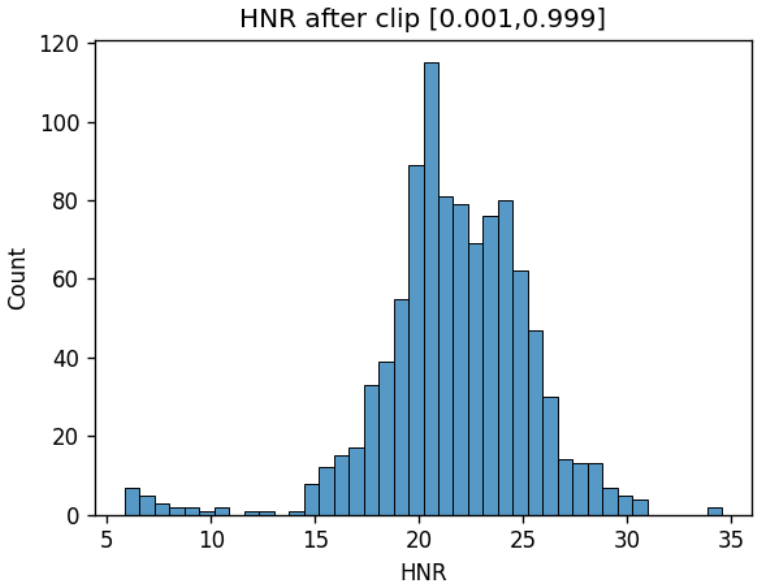
Histogram of Δmotor_UPDRS showing a sharp peak near zero and modest tails; highlights the intrinsic difficulty of sign prediction at short horizons

Figure B2a. HNR before clipping



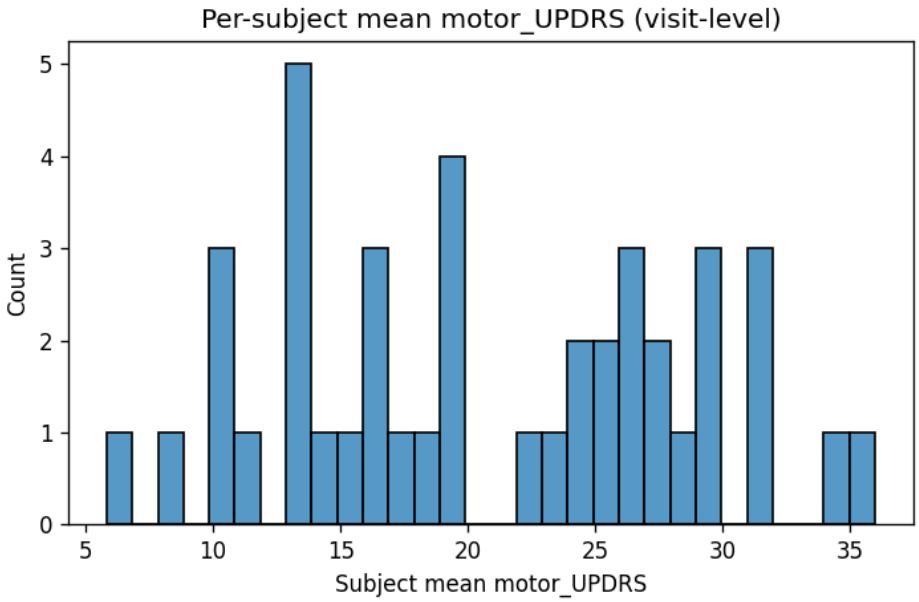
Visit-level HNR distribution with mild heavy tails; motivates defensive preprocessing

Figure B2b. HNR after clipping [0.1%, 99.9%]



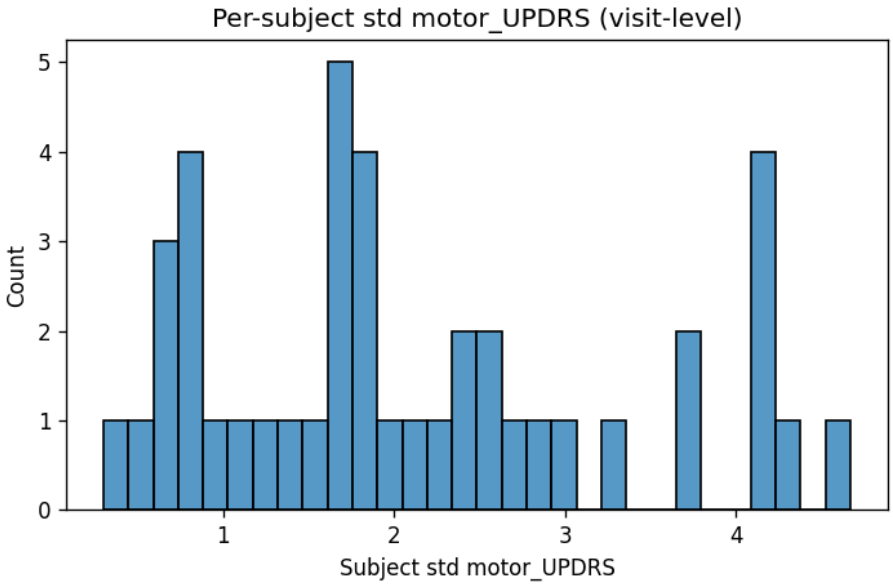
Same distribution after quantile clipping; stabilizes training without materially changing central tendency

Figure B3. Per-subject mean motor_UPDRS



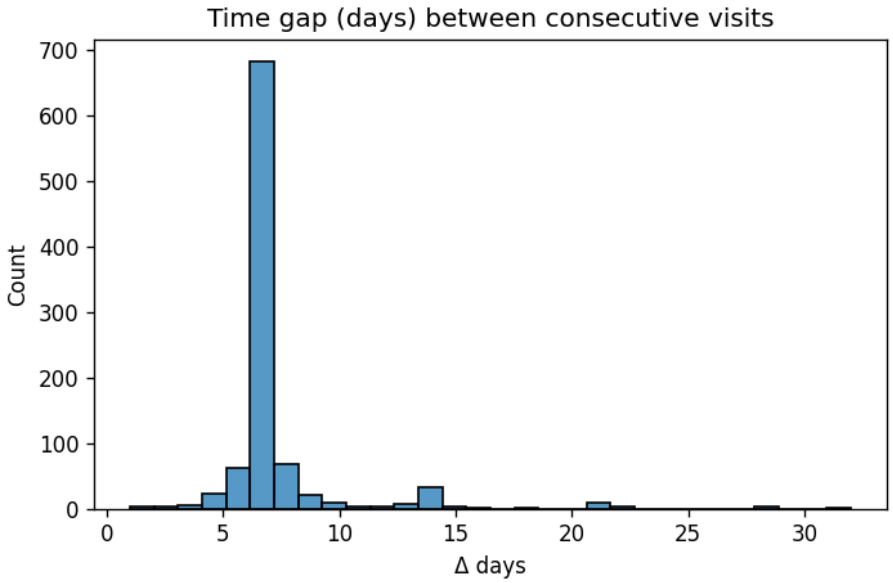
Distribution of subject-level means; justifies per-subject normalization to remove baseline differences

Figure B4. Per-subject standard deviation of motor_UPDRS



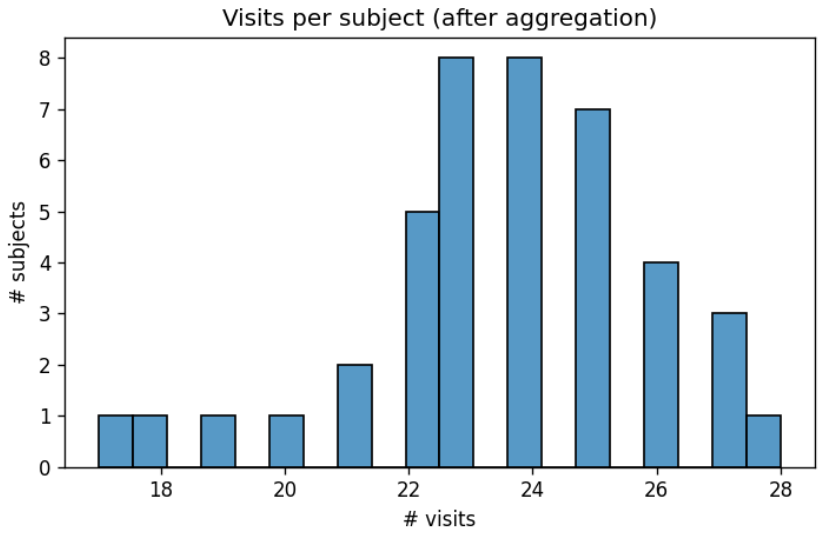
Variation of within-subject volatility; indicates heterogeneous dynamics across people

Figure B5. Time gap between consecutive visits



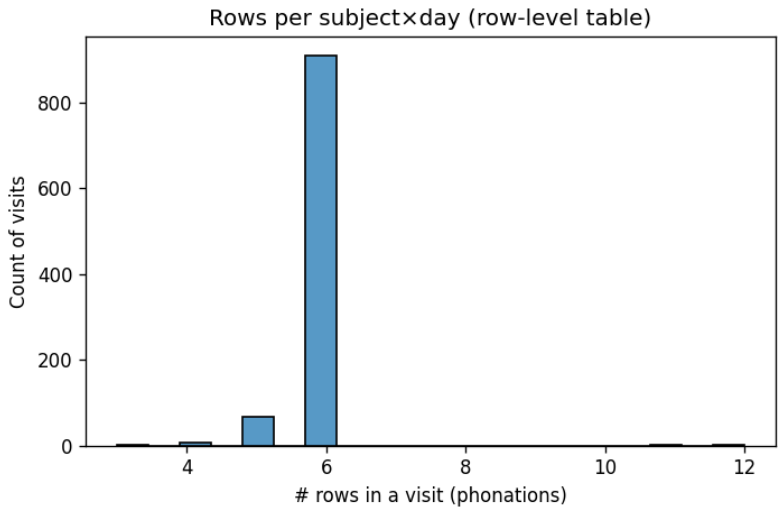
Histogram of day gaps; shows mostly short intervals with occasional longer gaps, relevant for sequence windowing

Figure B6. Visits per subject (post-aggregation)



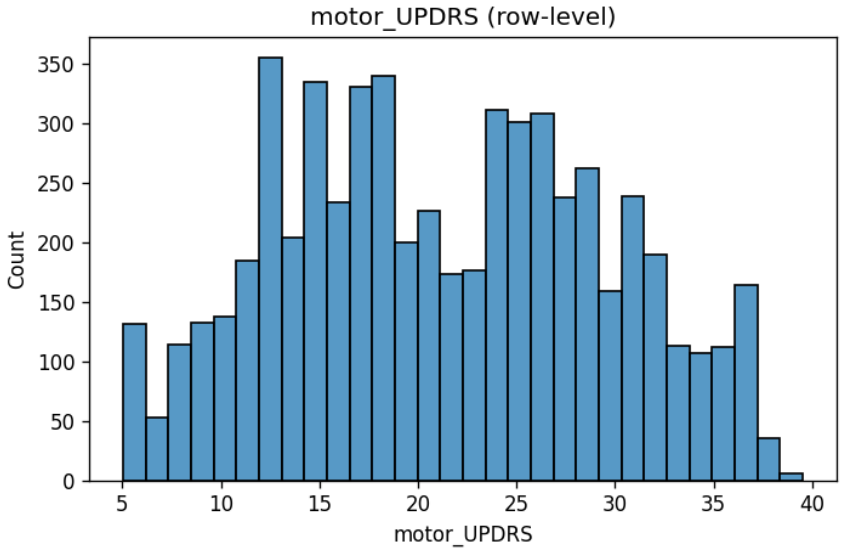
Counts of available visits by subject; supports short, fixed-length windows

Figure B7. Rows per visit (phonations)



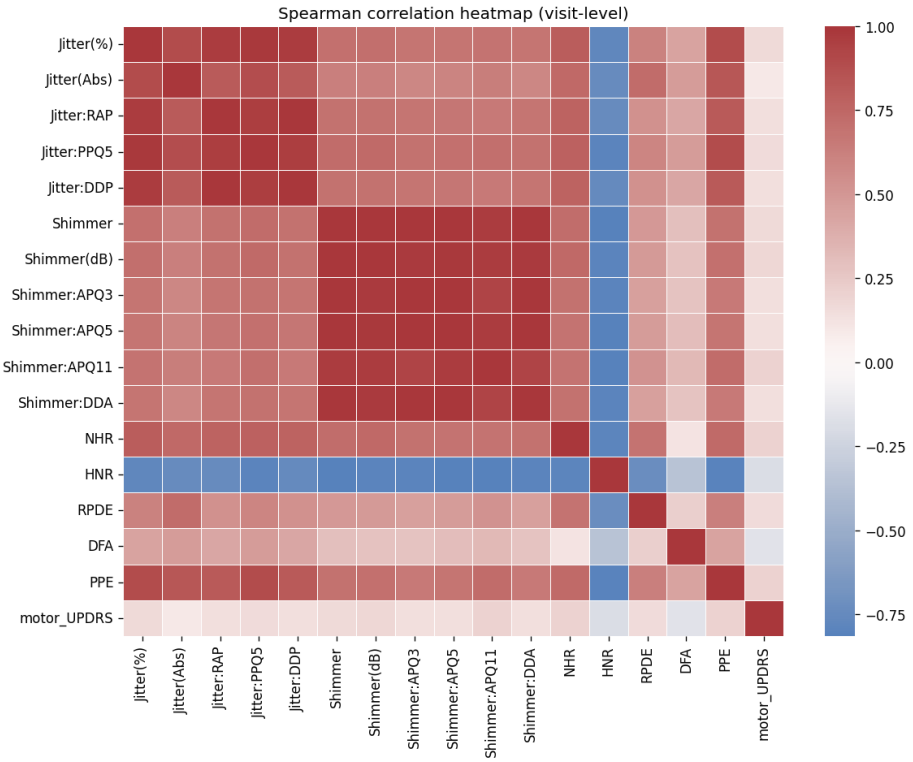
Distribution of phonations per visit in the row-level table; confirms aggregation choice

Figure B8. motor_UPDRS histogram (row-level)



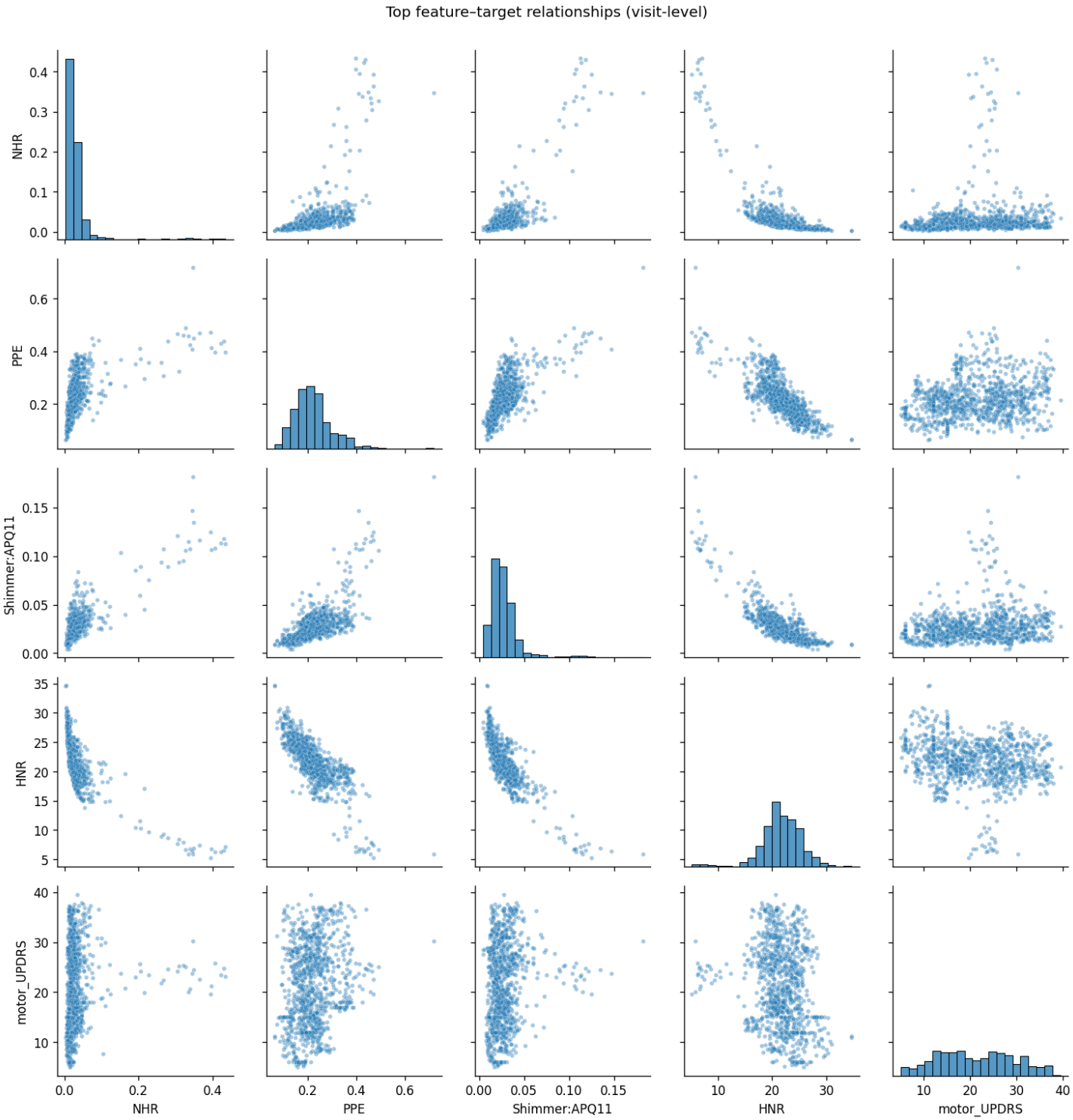
Broad distribution across phonations supports modeling on the absolute scale

Figure B9. Spearman correlation heatmap (visit-level)



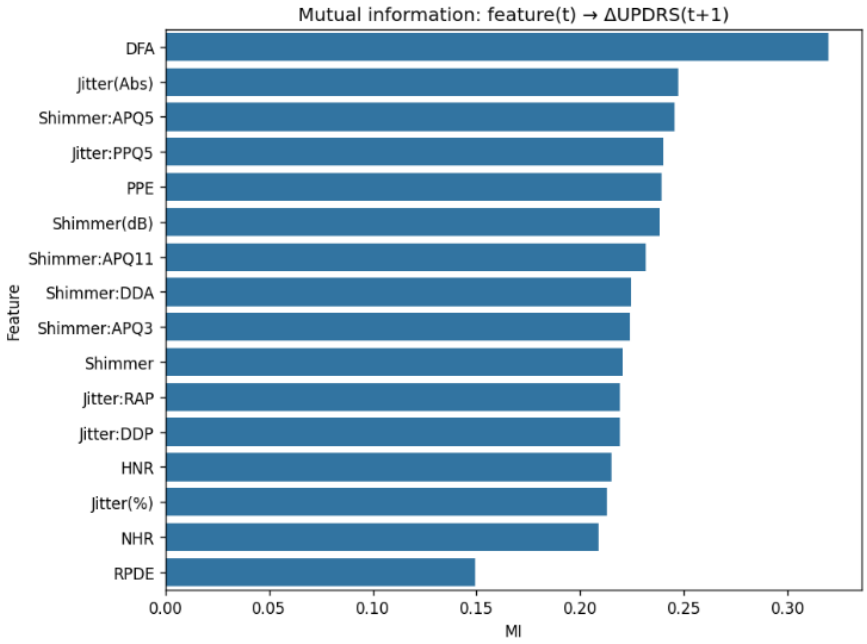
Feature–feature and feature–target relationships; shows intra-family structure and inverse relation with HNR

Figure B10. Pair plot of top feature–target relationships



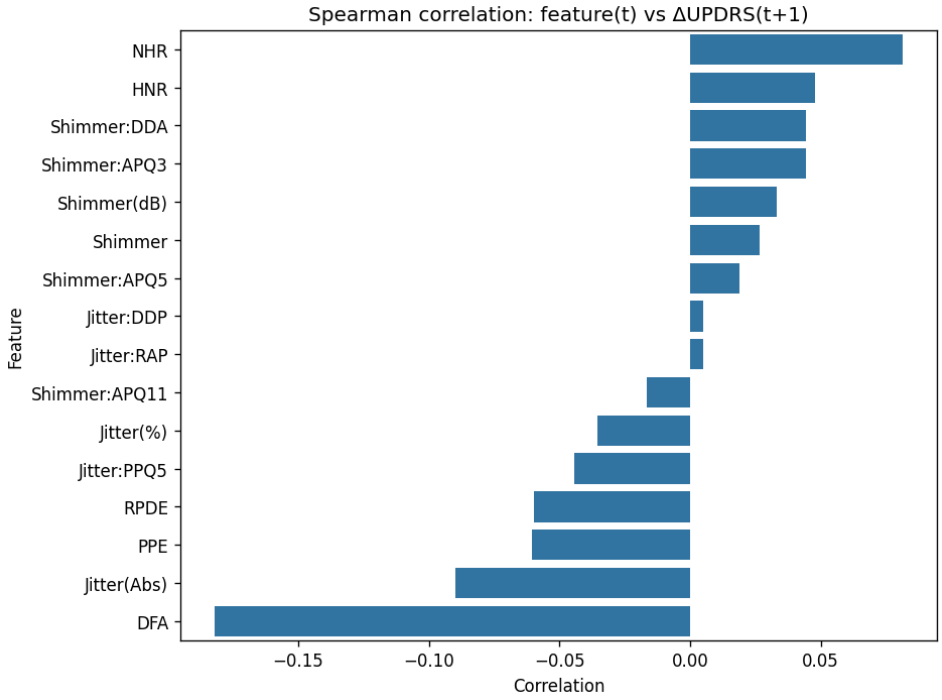
Scatter and marginal distributions for selected features versus motor_UPDRS; visually confirms plausible monotone trends

Figure B11. Mutual information, feature(t) → ΔUPDRS(t+1)



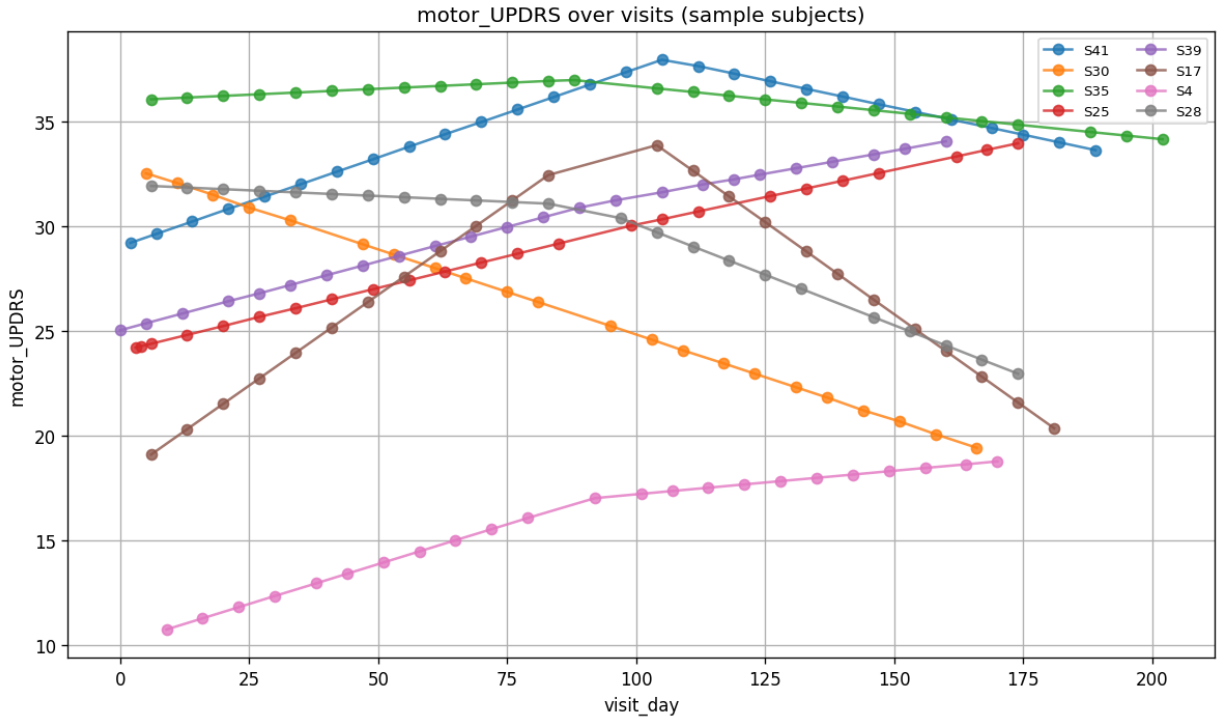
Bars for MI scores indicating moderate, non-dominant contributions across families

Figure B12. Spearman correlation, feature(t) vs ΔUPDRS(t+1)



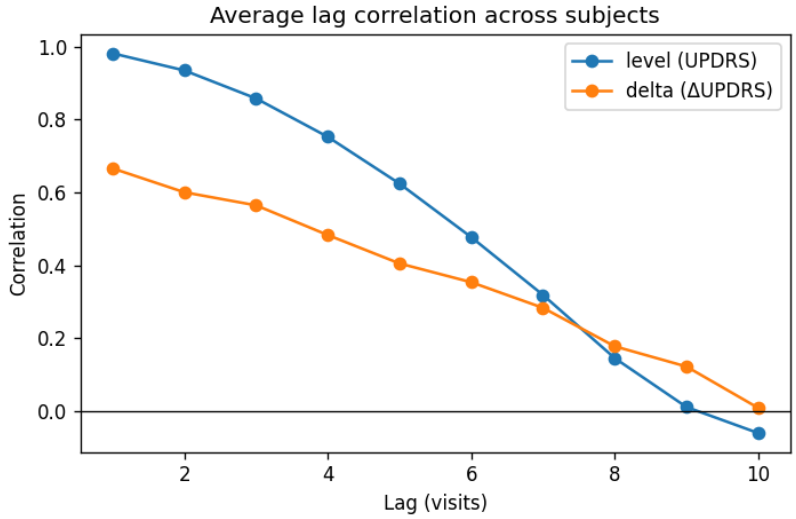
Correlations are uniformly small (|ρ| < 0.2): DFA most negative (~−0.18), Jitter(Abs) ~−0.09, PPE −0.06; NHR/HNR mildly positive (+0.08/+0.05). Small effects reflect the narrow ΔUPDRS and favor a calibration-first approach.

Figure B13. motor_UPDRS over visits (sample subjects)



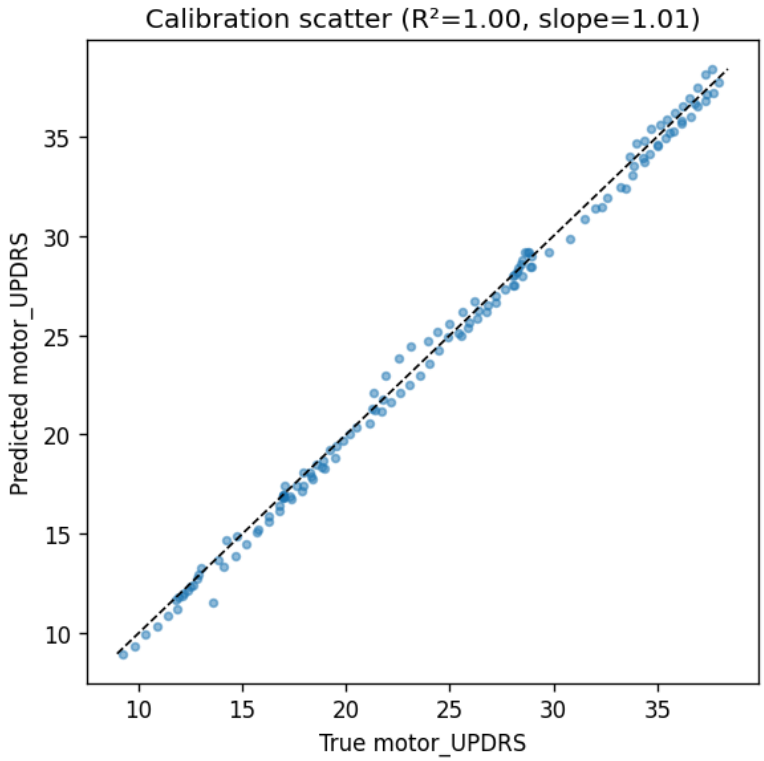
Trajectories showing improving, worsening, and non-monotonic courses; motivates a hybrid temporal model

Figure B14. Average lag correlation across subjects



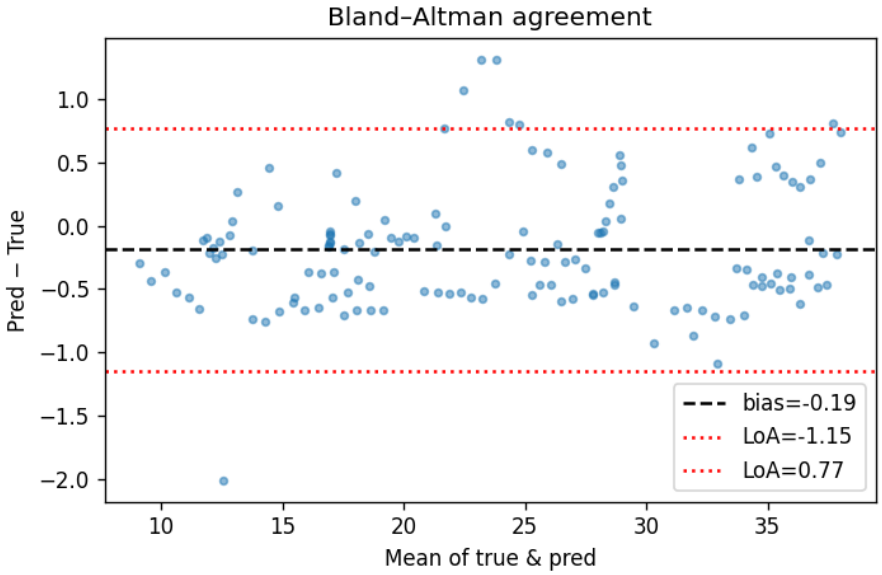
Level (UPDRS) shows strong autocorrelation at short lags and decays to ~0 by lag 9–10; the one-step change (ΔUPDRS) displays only modest short-lag persistence and approaches zero by lag ~9, reinforcing the calibration-first framing

Figure B15. Calibration scatter



Predicted versus true motor_UPDRS with regression line; slope ≈ 1.01 and R^2 ≈ 0.996 indicate strong calibration on the absolute scale

Figure B16. Bland–Altman agreement



Predicted–true differences versus mean, with bias and limits of agreement; shows tight agreement and small negative bias

Additional Capstone Project Assets

Capstone Project Github Repository:

https://github.com/arifakokab/parkinsons-vocal-biomarker-app

Deployed screening demo repository:

https://github.com/arifakokab/classification-for-parkinsons

Deployed Public PD Classifier screening app:

https://www.parkinsonsaiscreening-carepathai-foundation.care

References

Aural Analytics. (n.d.). Aural Analytics for neurological disease monitoring. https://auralanalytics.com/

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), 307–310.

Boersma, P., & Weenink, D. (2024). Praat: Doing phonetics by computer (Version 6.x) [Computer software]. http://www.praat.org/

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., … LaPelle, N. (2008). Movement Disorder Society–sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Movement Disorders, 23(15), 2129–2170. https://doi.org/10.1002/mds.22340

Grinberg, M. (2018). Flask web development (2nd ed.). O'Reilly.

Gunicorn. (n.d.). Green Unicorn: WSGI server. https://gunicorn.org/

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. In Proceedings of Interspeech 2018 (pp. 2748–2752). https://doi.org/10.21437/Interspeech.2018-1424

Kiranyaz, S., Ince, T., & Gabbouj, M. (2016). Real-time patient-specific ECG classification by 1-D convolutional neural networks. IEEE Transactions on Biomedical Engineering, 63(3), 664–675. https://doi.org/10.1109/TBME.2015.2468589

Little, M. A., McSharry, P. E., Hunter, E. J., & Ramig, L. O. (2008). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 55(10), 2591–2595. https://doi.org/10.1109/TBME.2008.2005954

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 56(4), 1015–1022. https://doi.org/10.1109/TBME.2008.2005954

Meta. (n.d.). React documentation. https://react.dev/

Render. (n.d.). Render docs. https://render.com/docs

Sage Bionetworks. (n.d.). mPower: Mobile Parkinson disease study. https://parkinsonmpower.org/

Sonde Health. (n.d.). Sonde Health. https://sondehealth.com/

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. IEEE Transactions on Biomedical Engineering, 57(4), 884–893. https://doi.org/10.1109/TBME.2009.2036000

Vásquez-Correa, J. C., Klumpp, P., Orozco-Arroyave, J. R., & Nöth, E. (2019). Convolutional neural networks and a transfer learning strategy to classify Parkinson's disease from speech in three different languages. Computers in Biology and Medicine, 104, 36–43. https://doi.org/10.1016/j.compbiomed.2018.10.017

Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Lorraine O. Ramig (2009), 'Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests', IEEE Transactions on Biomedical Engineering. https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring

Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering. https://archive.ics.uci.edu/dataset/174/parkinsons

Vite. (n.d.). Vite documentation. https://vitejs.dev/