International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

A Subject-Aware Mathematical Framework for Parkinson's Disease Severity Modeling from Longitudinal Voice Biomarkers

Sparsh Arya
Software Engineer, C+AI, Microsoft R&D India

August 2026

Abstract - This report re-analyzes the Oxford Parkinson's Telemonitoring dataset as a longitudinal, repeated-measures problem rather than as an ordinary independent-and-identically-distributed regression dataset. The central result is that the observed voice–UPDRS association is dominated by persistent differences between subjects. Approximately 91.9% of motor-UPDRS variance and 93.5% of total-UPDRS variance are between subjects. Consequently, random row-level validation substantially overstates generalization: a supplied random-forest analysis gives approximately $R^2 = 0.36$–0.37, whereas subject-grouped validation gives negative R^2 values of approximately -0.39 and -0.37. Recomputed regularized linear baselines show the same structural failure, with grouped R^2 between about -0.35 and -0.23.

We propose a hierarchical longitudinal formulation combining subject-specific intercepts, subject-specific slopes, time effects, and a within/between decomposition of voice biomarkers. This formulation distinguishes three scientifically different questions: (i) cross-sectional severity differences between people, (ii) longitudinal changes within a person, and (iii) forecasting future severity for an already-calibrated person. The resulting framework is mathematically interpretable, explicitly handles repeated measurements, and is aligned with recent literature showing that linear mixed-effects and generalized additive mixed-effects models can outperform high-capacity neural models on this small-N longitudinal dataset.

1 EXECUTIVE SUMMARY

This dataset contains **5,875 recordings from 42 subjects**, with repeated measurements for each participant. The supplied CSV is complete, with **no missing values and no duplicate rows**. Each subject contributes between **101 and 168 recordings**.

The most important analytical finding is that the apparent relationship between voice measurements and UPDRS is driven much more by **differences between subjects** than by short-term changes within the same subject. This distinction is critical for prediction. Random row-level validation gives optimistic results because recordings from the same person appear in both training and test sets. Subject-level validation is much more difficult and produces poor out-of-subject performance.

Dataset quality summary

The direct analysis of the supplied CSV confirms 5,875 rows, 22 columns, 42 subjects, zero missing values, and zero duplicate rows. The repeated-measures structure is therefore not a data-quality problem; it is the central statistical structure of the dataset.

Distributional findings

The voice variables are highly skewed, particularly jitter, shimmer, and NHR. Several measures are mathematically redundant. In the supplied data,

- jitter_ddp/jitter_rap has median 3.000 and mean approximately 3.000;
- shimmer_dda/shimmer_apq3 has median 3.000 and mean approximately 3.000.

Thus, treating all 16 voice variables as independent predictors introduces substantial multicollinearity. Regularization, dimension reduction, or feature grouping is preferable to unconstrained coefficient interpretation.

Main target relationship

Motor and total UPDRS are extremely strongly related. In the supplied data,

$$r_P = 0.947, \qquad r_S = 0.958,$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

and the empirical regression is

$$\widehat{\text{Total UPDRS}} = 2.467 + 1.247\,\text{Motor UPDRS}.$$

This one-variable relationship explains approximately $R^2 = 0.897$ of total-UPDRS variation. The strength of this relationship is expected because motor impairment is a major component of the broader UPDRS construct.

Voice measures associated with UPDRS

At the recording level, the strongest positive rank associations with motor UPDRS are shimmer_apq11 ($\rho = 0.164$), ppe (0.163), shimmer_db (0.140), shimmer (0.136), nhr (0.136), and jitter (0.128). The strongest negative associations are hnr (-0.158) and dfa (-0.131).

These patterns are directionally consistent with the established literature on Parkinsonian dysphonia: jitter and shimmer quantify perturbations in frequency and amplitude, HNR/NHR quantify harmonic-versus-noise structure, and RPDE/DFA/PPE capture nonlinear or irregular aspects of vocal dynamics (**little2007?**; **little2009?**; **tsanas2010?**). However, the correlation magnitudes are modest, so no individual voice feature is a strong standalone estimator of UPDRS.

Between-subject versus within-subject variation

The variance decomposition is especially informative. Approximately 91.9% of motor-UPDRS variance and 93.5% of total-UPDRS variance are between subjects. Among voice variables, DFA has approximately 76.9% between-subject variance, HNR 60.6%, NHR 56.2%, shimmer 55.2%, and several other shimmer measures roughly 50–57%.

When correlations are computed separately within each subject, median Spearman correlations are close to zero. This yields the key empirical conclusion:

> *The cross-sectional association between voice measures and UPDRS does not reliably translate into a within-person longitudinal association.*

A feature can therefore distinguish a person with a high average UPDRS from a person with a low average UPDRS without reliably increasing when the first person deteriorates.

Progression over time

Subject-specific linear slopes using test_time have means of approximately 0.012 points/day for motor UPDRS and 0.019 points/day for total UPDRS. The corresponding ranges are approximately $[-0.081, 0.086]$ and $[-0.094, 0.092]$, respectively. The population trend is therefore positive but highly heterogeneous.

Predictive modeling and leakage

The supplied random-forest analysis gives approximately $R^2 = 0.36$ for motor UPDRS and 0.37 for total UPDRS under random row-level splits, but approximately -0.39 and -0.37 under subject-grouped splits. Negative grouped R^2 means that the model is worse than the training-set mean predictor on unseen subjects.

Independent recomputation of simple linear baselines on the supplied CSV produces the same qualitative result: random-split R^2 is only about 0.10, while grouped R^2 is negative for OLS, Ridge, Lasso, Elastic Net, and PLS. Thus, the leakage phenomenon is not specific to random forests.

Recommended modeling strategy

For symptom monitoring, the preferred formulation is not a universal model that predicts raw UPDRS from a single voice recording. The preferred model is a **personalized longitudinal model** that learns a person's baseline voice phenotype and estimates deviations from that baseline. The recommended hierarchy is:

1. split evaluation by subject;
2. estimate preprocessing parameters on training subjects only;
3. reduce redundant jitter and shimmer measures;
4. include time explicitly;
5. use random intercepts and random slopes for subject heterogeneity;

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

6. separate between-subject and within-subject voice effects;
7. for deployment, calibrate each new subject using an initial baseline period;
8. evaluate future-recording prediction, not random-row prediction.

2 DATASET AND LITERATURE CONTEXT

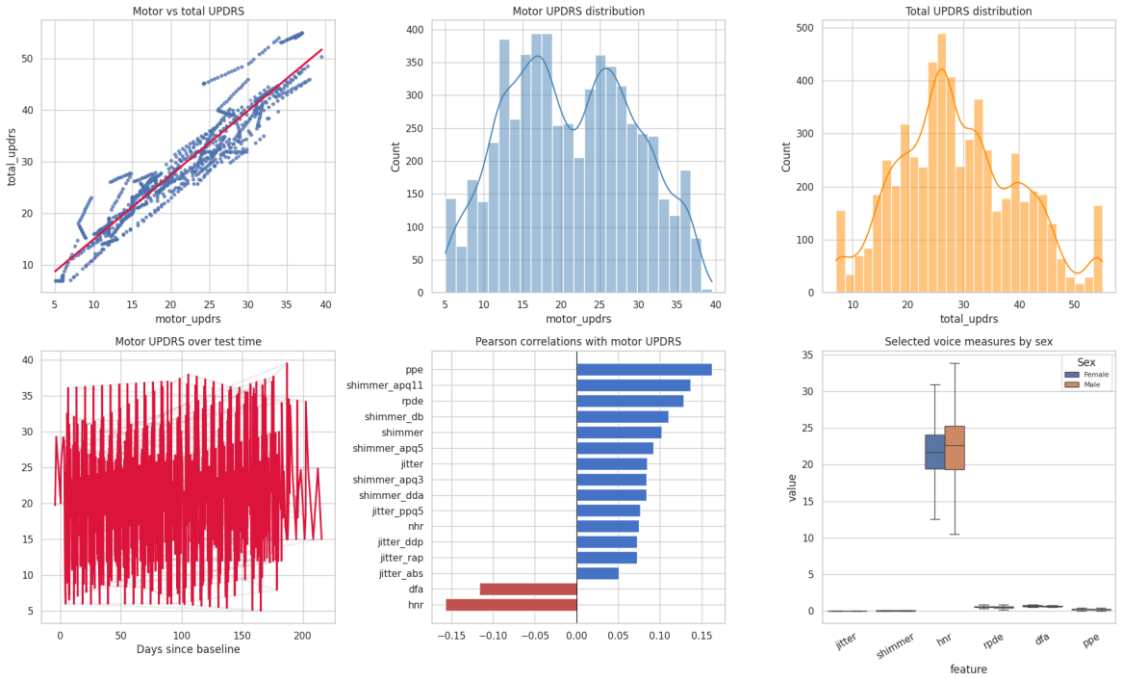
The dataset is the Oxford Parkinson's Telemonitoring dataset hosted by the UCI Machine Learning Repository. It contains 5,875 recordings from 42 people with early-stage Parkinson's disease, collected during a six-month home telemonitoring study. Each row contains subject information, time from baseline, motor and total UPDRS, and 16 biomedical voice measures (**uci189?**; **tsanas2010?**).

The original work by Tsanas et al. demonstrated that speech features could be mapped to UPDRS using linear and nonlinear regression, motivating remote and frequent symptom monitoring (**tsanas2010?**). Earlier work established the relevance of dysphonia measures such as PPE and HNR, while nonlinear recurrence and fractal measures motivated features such as RPDE and DFA (**little2007?**; **little2009?**). Importantly, the original dataset has repeated measurements per participant, and the UPDRS values associated with many recordings were linearly interpolated between clinical assessments rather than independently measured at every recording time (**barcenas2022?**). This makes the problem inherently longitudinal and also means that the target contains an interpolation component.

Research using repeated speech recordings has emphasized that treating every recording as an independent subject can distort generalization, and subject-level aggregation or leave-one-subject-out evaluation can improve the validity of predictive assessment (**sakar2013?**). More recent work has directly applied mixed-effects models to this same telemonitoring dataset. Tong et al. report that GAMM achieved MSE 6.56 and LMM MSE 7.70 when the last observation from each subject was held out, while neural mixed-effects variants performed much worse in their small-$N = 42$ setting (**tong2026?**). Their evaluation is aimed at forecasting future observations for subjects already observed, rather than predicting completely unseen patients, so its results should not be confused with leave-subject-out generalization.

3 EXPLORATORY DATA ANALYSIS

An extensive exploratory analysis reveals fundamental structural relationships within the dataset across feature distributions, correlation structures, temporal trajectories, and demographic profiles, as summarized in Figure 1.

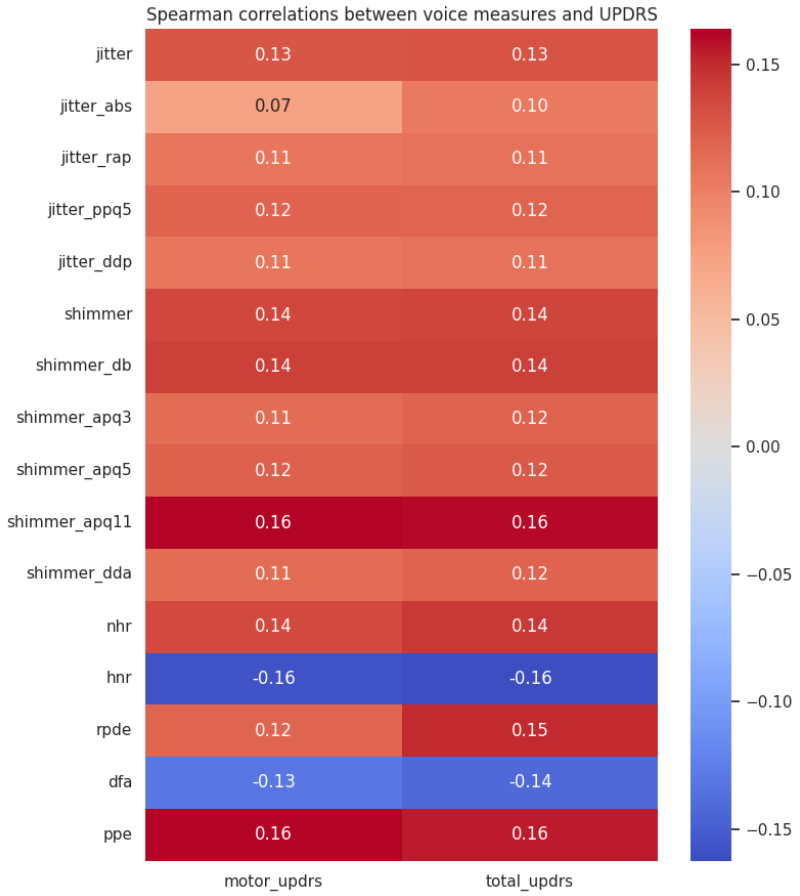


Exploratory Data Analysis of UPDRS Scores and Acoustic Voice Metrics. (Top Left) *Scatter plot showing the strong positive correlation between Motor UPDRS and Total UPDRS scores fitted with a linear trendline.* ***(Top Middle)*** *Histogram and kernel density estimation (KDE) demonstrating a bimodal distribution for Motor UPDRS scores.* ***(Top Right)*** *Distribution of Total UPDRS scores displaying a slight right-skewed pattern.* ***(Bottom Left)*** *Trajectory plot of Motor UPDRS over test time (days since baseline), highlighting individual subject variability over time.* ***(Bottom Middle)*** *Pearson correlation coefficients of voice features*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

with Motor UPDRS, showing top positive metrics (e.g., ppe, shimmer_apq11) and negative metrics (hnr, dfa). ***(Bottom Right)*** *Comparative boxplot of voice measures grouped by sex demographic (Female vs. Male).*

To evaluate monotonic dependencies across features without parametric assumptions, non-parametric Spearman rank correlations were computed between all acoustic measures and the clinical UPDRS outcomes, as illustrated in Figure 2.



Spearman Rank Correlations Between Acoustic Voice Measures and UPDRS Outcomes. *Heatmap depicting non-parametric Spearman correlation coefficients (r_s) across motor and total UPDRS scores. Acoustic measures such as ppe ($r_s = 0.16$), shimmer_apq11 ($r_s = 0.16$), hnr ($r_s = -0.16$), and dfa ($r_s = -0.13$) display consistent monotonic associations across both clinical target scores.*

4 MATHEMATICAL VIEW OF THE DATASET

Let $i = 1, \dots, 42$ index subjects and $j = 1, \dots, n_i$ index recordings. Let Y_{ij} denote either motor or total UPDRS and let $\boldsymbol{X}_{ij} \in \mathbb{R}^{16}$ denote the voice feature vector.

A naive regression assumes

$$Y_{ij} = \beta_0 + \boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{iid}}{\sim} (0, \sigma^2).$$

This is inappropriate if repeated recordings from the same participant share a persistent latent severity component. A more realistic decomposition is

$$Y_{ij} = \mu + u_i + \varepsilon_{ij},$$

where u_i represents persistent subject-specific severity. Then

$$\text{Var}(Y) = \underbrace{\text{Var}(u_i)}_{\text{between subjects}} + \underbrace{\text{Var}(\varepsilon_{ij})}_{\text{within subjects}}.$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

The intraclass correlation is

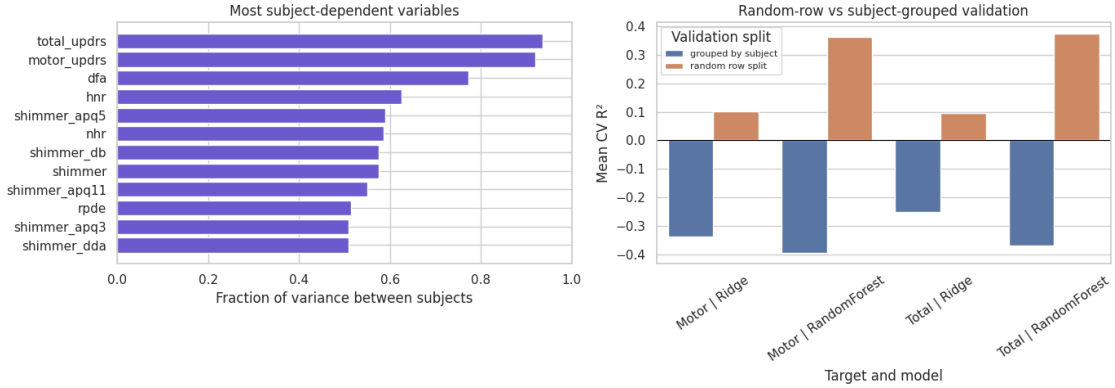
$$\mathrm{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}.$$

For this dataset, the empirical variance decomposition gives approximately

$$\mathrm{ICC}_{motor} = 0.919, \qquad \mathrm{ICC}_{total} = 0.935.$$

Therefore two random recordings from the same subject are highly correlated in their latent target level. Random row splitting puts that subject-level information into both training and test sets.

Figure 3 quantifies this subject-level variance fraction alongside the empirical impact of random-row vs. subject-grouped evaluation across different predictive algorithms.



Subject-Level Variance and Machine Learning Validation Vulnerabilities. (Left) *Fraction of variance explained across subject boundaries, revealing high subject dependency ($>$ 60% to 90% +) for clinical outcomes and core acoustic metrics.* ***(Right)*** *Cross-validation performance comparison (R^2) across model targets and algorithms. Random-row splits yield artificially inflated R^2 scores ($\approx 0.10 - 0.37$) due to severe data leakage, whereas proper subject-grouped validation results in negative mean R^2 values.*

4.1 Why random row validation is optimistic

Suppose a predictive model learns

$$\hat{Y}_{ij} = f(\boldsymbol{X}_{ij}) \approx \hat{u}_i + g(\boldsymbol{X}_{ij}).$$

If the same subject appears in both train and test, the model can exploit the participant's stable voice phenotype to infer u_i. Under subject-grouped validation, u_i is unseen, and the model must predict

$$Y_{new,j} = \mu + u_{new} + g(\boldsymbol{X}_{new,j}) + \varepsilon_{new,j}$$

without observing u_{new}. If $\mathrm{Var}(u_i)$ dominates $\mathrm{Var}(Y)$, the out-of-subject prediction problem is intrinsically difficult.

This explains why negative grouped R^2 is not evidence that voice contains no information. It indicates that voice features contain much more information about *who the participant is and their persistent baseline* than about short-term changes in severity.

5 WITHIN–BETWEEN DECOMPOSITION

A central modeling improvement is to decompose each voice feature into a subject mean and a within-subject deviation:

$$X_{ijk} = \bar{X}_{ik} + \left(X_{ijk} - \bar{X}_{ik}\right).$$

This leads to a correlated-random-effects/Mundlak-style model (**mundlak1978?**):

$$Y_{ij} = \beta_0 + \bar{\boldsymbol{X}}_i^{\top}\boldsymbol{\beta}_B + \left(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i\right)^{\top}\boldsymbol{\beta}_W + f(t_{ij}) + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij}.$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Here $\boldsymbol{\beta}_B$ measures a **between-subject association**: do people with different typical voice profiles have different typical UPDRS? In contrast, $\boldsymbol{\beta}_W$ measures a **within-subject association**: when a person's voice deviates from their own typical voice, does UPDRS deviate in the predicted direction?

These parameters answer different scientific questions and should not be conflated. A large β_B with a small β_W is exactly the structure suggested by the present data.

6 LONGITUDINAL MIXED-EFFECTS MODEL

For a practical model, define standardized selected voice features Z_{ijk} and standardized time T_{ij}. A linear mixed-effects model is

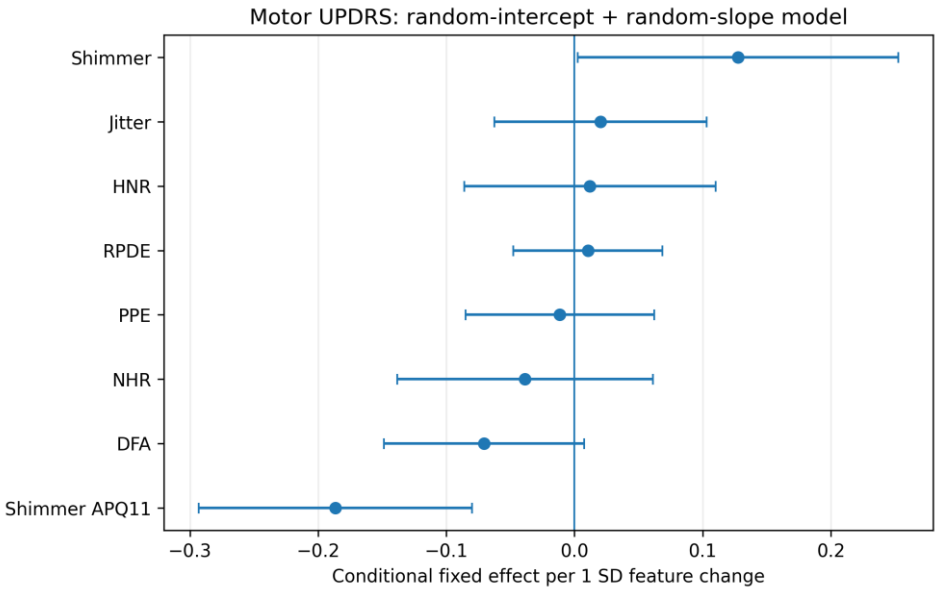
$$Y_{ij} = \beta_0 + \beta_t T_{ij} + \sum_{k=1}^{K} \beta_k Z_{ijk} + b_{0i} + b_{1i} T_{ij} + \varepsilon_{ij},$$

with

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{G} \right), \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

The random intercept b_{0i} represents baseline severity. The random slope b_{1i} represents individual progression rate. This is substantially more faithful to the observed trajectories than a single global time coefficient.

Figure 4 summarizes the conditional fixed-effect estimates obtained from fitting a random-intercept + random-slope mixed-effects model.



Mixed-Effects Conditional Fixed Effects for Motor UPDRS. *Forest plot illustrating the conditional fixed-effect estimates per 1 SD feature change in a random-intercept + random-slope model. Shimmer exhibits a slight positive effect (≈ +0.13), while Shimmer APQ11 shows a negative effect (≈ −0.18). Most remaining voice parameters feature confidence intervals crossing zero once random subject effects are accounted for.*

For motor UPDRS, fitting a selected-feature random-intercept model gave a subject variance of approximately 60.88 and residual variance of 5.44, yielding an ICC of 0.918. For total UPDRS, the corresponding subject and residual variances were approximately 106.98 and 7.63, yielding an ICC of 0.933.

Adding a random slope for time reduced the motor residual variance to approximately 1.64 and estimated a random-slope standard deviation of approximately 0.036 points/day, close to the empirical standard deviation of subject-specific slopes (0.037 points/day). The conditional fit is therefore very high for already-observed subjects, but this must not be interpreted as unseen-subject predictive performance because the random effects are estimated from the same subjects.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

6.1 Selected conditional effects

A parsimonious mixed model using jitter, shimmer, shimmer_apq11, NHR, HNR, RPDE, DFA, and PPE gives the following standardized fixed-effect estimates:

Selected fixed effects from the random-intercept mixed model. Coefficients are conditional on the other selected variables and therefore need not match marginal Spearman correlations.

Feature	Motor β	Motor p	Total β	Total p
jitter	0.196	0.009	0.205	0.022
shimmer	0.079	0.493	0.096	0.481
shimmer_apq11	-0.116	0.230	-0.147	0.201
nhr	-0.204	0.025	-0.157	0.144
hnr	0.141	0.114	0.257	0.015
rpde	-0.083	0.117	-0.041	0.508
dfa	-0.214	0.002	-0.113	0.173
ppe	0.064	0.340	0.063	0.430

The conditional signs should not be treated as causal effects. The difference between marginal correlations and conditional coefficients is a direct consequence of collinearity and the repeated-measures structure. The scientifically safer conclusion is at the signal-family level: perturbation, noise/harmonic structure, and nonlinear vocal dynamics are associated with severity, but no single feature has a stable causal interpretation in this dataset.

7 MODEL COMPARISON

Table [tab:models] summarizes the main model classes. The linear-model results were recomputed from the supplied CSV using five-fold cross-validation. Scaling was performed inside each training fold. The random-forest values are the approximate values shown in the supplied validation analysis. Hyperparameter optimization was intentionally limited for the baseline comparison; the purpose is to diagnose the structure of the problem rather than to claim a leaderboard result.

*Approximate values from the supplied validation figure; not independently re-tuned in this report.

The important result is not which conventional machine-learning algorithm wins. All ordinary models suffer the same validation collapse when subject identity is removed. This is a structural signal that model capacity alone cannot solve the problem.

7.1 Why nonlinear models do not automatically solve the problem

A random forest or RBF-SVR can represent

$$Y = f(\boldsymbol{X}) + \varepsilon$$

with substantial nonlinear flexibility. But if the dominant component is

$$Y = \mu + u_i + g(\boldsymbol{X}) + \varepsilon,$$

then increasing the complexity of f does not create information about u_i for an unseen subject. The fundamental missing quantity is the individual's baseline.

This is consistent with recent literature. Bárcenas et al. used a mixed-kernel SVR and emphasized nonlinear relationships and the importance of patient dynamics (**barcenas2022?**). More importantly for the present analysis, Tong et al. explicitly compared LMM, GAMM, GNMM and neural mixed-effects models on the same dataset and found that statistical mixed-effects models generalized much better in their last-visit-per-subject forecasting design (**tong2026?**).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

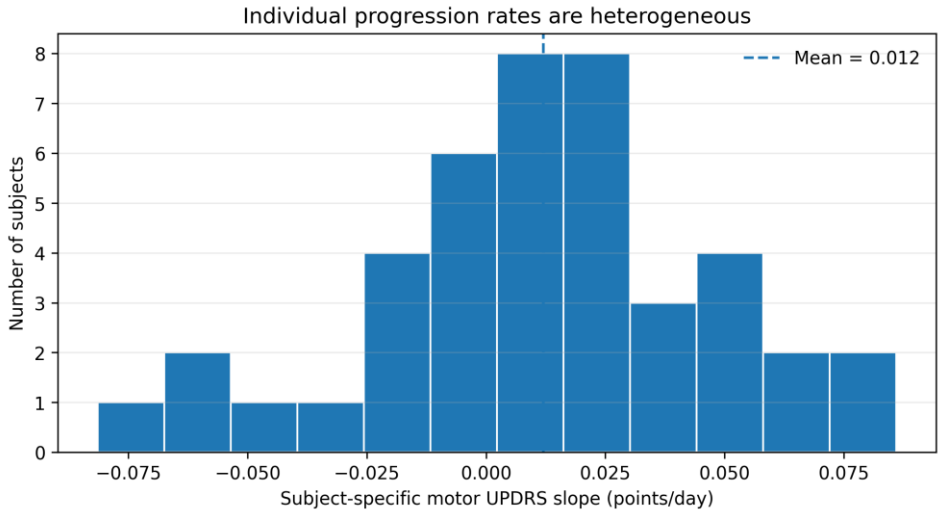
8 Generalized Additive Mixed Model

The next model to test is a generalized additive mixed model:

$$Y_{ij} = \beta_0 + s(t_{ij}) + \boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta} + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij},$$

where $s(\cdot)$ is a low-dimensional spline. This model is attractive because the time trajectory in the supplied data is not perfectly linear. It also avoids asking a neural network to learn a complicated temporal function from only 42 subjects.

Figure 5 illustrates the distribution of subject-level progression rates, displaying substantial inter-individual trajectory heterogeneity.



Heterogeneity of Individual Motor UPDRS Progression Rates. *Histogram showing the distribution of subject-specific motor UPDRS progression slopes (points per day). The population mean rate of change is* 0.012 *points/day, with individual trajectories ranging from negative slopes (stable/improving) to steep positive progression rates exceeding* +0.075 *points/day.*

Tong et al. reported that a GAMM with a smooth time effect achieved MSE 6.56 and MAE 2.00 on a test set containing the final observation from each subject, compared with MSE 7.70 and MAE 2.25 for their LMM. Their neural models had MSE values above 96 in the same setup (**tong2026?**). These results are not directly comparable to the grouped cross-validation numbers in Table [tab:models], because their task is forecasting future observations for subjects already observed. Nevertheless, the agreement is strong: once the longitudinal structure is modeled explicitly, simple statistical models become highly competitive.

9 Personalized Baseline Model for Telemonitoring

For actual telemonitoring, the most clinically meaningful target is arguably not raw UPDRS but change from a subject's own baseline. Let 0 denote a calibration period and define

$$\Delta Y_{ij} = Y_{ij} - Y_{i0}, \qquad \Delta \boldsymbol{X}_{ij} = \boldsymbol{X}_{ij} - \boldsymbol{X}_{i0}.$$

A personalized monitoring model can then be written as

$$\Delta Y_{ij} = \alpha + s(t_{ij}) + \Delta \boldsymbol{X}_{ij}^{\top}\boldsymbol{\theta} + c_i(t_{ij}) + \eta_{ij},$$

where $c_i(t)$ is a subject-specific progression component.

An even more robust version uses a rolling baseline:

$$B_{ik}(t) = \text{median}\{X_{ik}(t-h), \dots, X_{ik}(t)\},$$

then defines

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

$$Z_{ijk}(t) = X_{ijk}(t) - B_{ik}(t).$$

The prediction target can similarly be smoothed:

$$\tilde{Y}_{ij} = \text{median}\{Y_{i,j-m}, \dots, Y_{ij}\}.$$

This reduces recording-level noise and converts the model from a cross-sectional severity estimator into a change detector.

10 A PROPOSED FINAL MODEL

The recommended model for a future study is a hierarchical additive model with explicit within/between effects:

$$Y_{ij} = \beta_0 + s_1(t_{ij}) + \bar{\boldsymbol{Z}}_i^{\top}\boldsymbol{\beta}_B + \boldsymbol{Z}_{ij}^{*\top}\boldsymbol{\beta}_W + s_2(t_{ij})\boldsymbol{Z}_{ij}^{*\top}\boldsymbol{\gamma} + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij},$$

where

$$\boldsymbol{Z}_{ij}^{*} = \boldsymbol{Z}_{ij} - \bar{\boldsymbol{Z}}_i.$$

The terms have the following interpretation:

- $\boldsymbol{\beta}_B$: stable between-person voice phenotype associated with typical severity;
- $\boldsymbol{\beta}_W$: within-person voice deviations associated with symptom deviations;
- $s_1(t)$: population-level nonlinear disease trajectory;
- b_{0i}: personalized baseline severity;
- b_{1i}: personalized progression rate;
- $s_2(t)\boldsymbol{Z}^*$: time-varying voice–severity relationship.

This model can be regularized by grouping redundant features. For example, because

$$\text{jitter\_ddp} \approx 3\,\text{jitter\_rap}, \qquad \text{shimmer\_dda} \approx 3\,\text{shimmer\_apq3},$$

we can remove one feature from each pair or construct group-level latent scores. A principal-component or partial-least-squares representation can then be applied separately to jitter and shimmer families.

11 INTERPRETATION OF THE MAIN SCIENTIFIC RELATIONSHIP

The evidence supports a three-layer interpretation.

11.0.0.1 Layer 1: physiological signal.

Voice contains measurable information related to Parkinsonian impairment. Perturbation, harmonic/noise structure, and nonlinear dynamics are associated with UPDRS, consistent with prior voice-biomarker literature (**little2007?**; **little2009?**; **tsanas2010?**).

11.0.0.2 Layer 2: person-specific phenotype.

The majority of UPDRS variance is persistent across recordings of the same participant. A person's voice characteristics therefore act partly as a stable phenotype that correlates with their typical severity. This is why random row-level prediction can look substantially better than subject-level prediction.

11.0.0.3 Layer 3: longitudinal change.

The within-person correlations are small and heterogeneous. Therefore the cross-sectional mapping

$$\text{voice phenotype} \rightarrow \text{typical UPDRS}$$

should not be interpreted as the longitudinal mapping

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

voice change → UPDRS change.

The second relationship requires subject calibration, temporal smoothing, and explicit modeling of individual trajectories.

12 SEX, AGE, AND CONFOUNDING

The supplied CSV encodes sex as Boolean values. According to the UCI documentation, sex=0 denotes male and sex=1 denotes female (**uci189?**). Therefore the supplied file contains approximately 4,008 male recordings and 1,867 female recordings. This reverses the male/female labels in the supplied executive summary and should be corrected in any manuscript using this dataset.

Age is constant within each subject, so its apparent relationship with UPDRS is a between-subject relationship. It cannot explain within-person changes over time. Sex has the same limitation. These variables should therefore be modeled as subject-level covariates rather than treated as time-varying predictors.

13 LIMITATIONS

1. **Only 42 subjects.** The nominal sample size is 5,875, but the effective number of independent participants for subject-level inference is 42.
2. **Interpolated target values.** Many UPDRS values are interpolated between clinical assessments, so adjacent observations do not represent independent clinical measurements (**barcenas2022?**).
3. **Single disease cohort.** All participants have early-stage Parkinson's disease. The analysis does not establish discrimination between Parkinson's disease and healthy controls.
4. **Potential confounding.** Medication, recording environment, fatigue, microphone/device effects, and other time-varying factors may affect voice measures.
5. **Validation target matters.** Leave-subject-out prediction and future-within-subject prediction answer different questions and should be reported separately.
6. **Causality is not established.** Regression coefficients and correlations are associations, not evidence that changing a particular acoustic feature will change UPDRS.

14 CONCLUSION

The dataset supports a clear mathematical and empirical story. Voice measurements contain real but modest cross-sectional information about Parkinson's severity. The strongest signals involve perturbation, shimmer, harmonic/noise structure, and nonlinear vocal dynamics. However, approximately 92–94% of target variance is between subjects, while within-subject voice–UPDRS relationships are weak and heterogeneous.

The correct mathematical abstraction is therefore not an ordinary regression problem but a **hierarchical longitudinal measurement problem**. A suitable model should be written as

$$\mathrm{UPDRS}_{ij} = \underbrace{\mu + s_1(t_{ij})}_{\text{population trajectory}} + \underbrace{\bar{\boldsymbol{Z}}_i^{\top}\boldsymbol{\beta}_B + b_{0i}}_{\text{between-subject phenotype}} + \underbrace{(\boldsymbol{Z}_{ij} - \bar{\boldsymbol{Z}}_i)^{\top}\boldsymbol{\beta}_W}_{\text{within-subject voice deviation}} + \underbrace{b_{1i}t_{ij}}_{\text{subject-specific trajectory}} + \underbrace{\epsilon_{ij}}_{\text{measurement noise}} .$$

For a research paper, the strongest next experiment is a direct comparison of four evaluation tasks: (1) random-row prediction, (2) leave-subject-out prediction, (3) last-visit forecasting for already-calibrated subjects, and (4) baseline-change prediction. The primary model should be a regularized LMM/GAMM with random intercept and slope, while Elastic Net/PLS, random forest, SVR, and a small neural baseline serve as secondary comparisons.

The final scientific objective should be personalized monitoring: first learn a participant's stable voice phenotype, then estimate whether subsequent deviations from that phenotype are consistent with a meaningful change in motor severity. This is more defensible than claiming that a single voice recording can universally predict an absolute UPDRS score.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

15. REFERENCES

[1] Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010. doi: 10.1109/TBME.2009.2036000.

[2] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009. doi: 10.1109/TBME.2008.2005954.

[3] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, article 23, 2007. doi: 10.1186/1475-925X-6-23.

[4] A. Tsanas and M. Little, "Parkinsons Telemonitoring," UCI Machine Learning Repository, Dataset 189, 2009. doi: 10.24432/C5ZS3N.

[5] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013. doi: 10.1109/JBHI.2013.2245674.

[6] R. Bárcenas, R. Fuentes-García, and L. Naranjo, "Mixed kernel SVR addressing Parkinson's progression from voice features," *PLOS ONE*, vol. 17, no. 10, e0275721, 2022. doi: 10.1371/journal.pone.0275721.

[7] Y. Mundlak, "On the pooling of time series and cross section data," *Econometrica*, vol. 46, no. 1, pp. 69–85, 1978. doi: 10.2307/1913646.

[8] R. Tong, L. Wang, T. Wang, and W. Yan, "Modeling Parkinson's disease progression from longitudinal voice biomarkers: A comparative study of statistical and neural mixed effects models," *Computer Methods and Programs in Biomedicine Update*, vol. 9, article 100242, 2026. doi: 10.1016/j.cmpbup.2026.100242.