Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 25 July 2026
Revised 24 Aug 2026
Accepted 12 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (20s): 978–986

KEYWORDS:
Parkinson's disease; voice biomarkers; spiral drawing; wave drawing; gradient boosting; sensitivity and specificity; machine learning.

Comparative Machine Learning Analysis of Acoustic and Drawing-Based Motor Biomarkers for Parkinson's Disease Screening

Dr. Ashok R S*

Department of Electronics and Communication Engineering, Visvesvaraya Technological University, Belagavi, VTU Centre for Postgraduate Studies Mysuru, Karnataka, India. E-Mail: ashokchnagar@gmail.com

ABSTRACT: Voice and hand-drawing tests are low-cost, non-invasive behavioral probes of Parkinson's disease, yet their relative reliability is rarely examined with attention to test-set size and class balance. This study compares acoustic and drawing-based motor biomarkers using the evaluations recorded for a portable screening prototype: 200 voice recordings described by 22 acoustic measures, 200 spiral drawings, and 200 wave drawings, each analyzed with 8 to 10 classifiers. Gradient-boosted tree models performed best: XGBoost for voice (90.0% accuracy), and CatBoost for spiral (87.5%) and wave drawings (89.5%). Confusion matrices recovered from the original evaluation show that the voice test set contained 30 recordings (24 patients, 6 healthy) and the spiral test set only 8 drawings (6 patients, 2 healthy). Against majority-class baselines of 80% and 75%, the best models were not significantly better than always predicting Parkinson's disease (exact binomial $p = 0.12$ and 0.37), and 95% Wilson intervals were wide (voice 74.4–96.5%; spiral 52.9–97.8%). Class-specific analysis revealed different error profiles: the best voice model had a sensitivity of 95.8% but a specificity of 66.7%, whereas every spiral model avoided false alarms but missed patients, and linear and kernel models collapsed to predicting a single class. Feature-importance and voice–drawing fusion analyses could not be performed because raw data and participant links were unavailable, and a protocol for them is specified. Neither modality can yet be declared more reliable; about 140 test cases per modality are needed to estimate a 90% accuracy to within ±5 percentage points.

1. INTRODUCTION

Parkinson's disease is a progressive neurodegenerative disorder whose cardinal features are bradykinesia, rigidity, rest tremor, and postural instability [1]. Diagnosis is clinical, relies on the examiner's expertise, and usually follows the appearance of overt motor signs [2]. Behavioral tests that can be administered cheaply and repeatedly, such as speech recording and hand drawing, are therefore attractive for early screening and remote monitoring.

Speech is affected early in Parkinson's disease. Reduced loudness, breathiness, monotony, and irregular phonation can be quantified with perturbation and noise measures such as jitter, shimmer, and the harmonics-to-noise ratio [3, 4]. Public voice datasets and mobile recordings have supported many classifiers [5–8]. Drawing tasks probe a different system. Spirals and waves expose tremor, micrographia, hesitation, and irregular velocity, and have been analyzed with kinematic indices, handcrafted image features, and deep networks [9–13]. Voice and drawing therefore reflect distinct motor subsystems (laryngeal and respiratory control versus fine limb control), and it is plausible that each carries information the other lacks.

Comparisons between these modalities are, however, hard to interpret. Studies report single accuracy figures, often from small held-out sets, without class-specific performance, confidence intervals, or comparison with the trivial strategy of always predicting the majority class. Records from the same person that fall on both sides of a split further inflate accuracy [14]. In earlier work we developed a portable prototype with separate voice and handwriting classifiers [15, 16]. The present paper does not add new data; it re-examines the recorded evaluations of the voice, spiral, and wave tasks in order to establish what they do and do not show.

The objectives are (i) to compare machine-learning algorithms for the voice, spiral, and wave data; (ii) to describe the acoustic feature families available and specify how the strongest features can be identified; (iii) to examine whether voice and drawing models make different errors that could complement one another; (iv) to compare the spiral and wave tasks; and (v) to assess how an interpretable behavioral screening model can be obtained. A central point of the analysis is that every accuracy is interpreted against the size and class composition of the test set on which it was measured.

2. MATERIALS AND METHODS

2.1 Study Material

Three datasets from the prototype's development were analyzed (Table 1 and Figure 1). The voice dataset contains 200 recordings (151 from patients and 49 from healthy controls) described by acoustic measures. The spiral and wave datasets each contain 200 drawings, reported as equally divided between healthy and Parkinson's-indicative samples, with a 75:25 training to test split. Numbers of unique participants, ages, sex, disease stage, and medication status are not documented in the source material, and no identifier links records across the three tasks.

Table 1. Datasets analyzed

Task	Samples	Class balance (PD / healthy)	Input	Classifiers compared	Reported split
Voice	200 recordings	151 / 49	22 acoustic measures	8	Test set of 30 (inferred)
Spiral	200 drawings	100 / 100	Grayscale images	10	75:25 stated
Wave	200 drawings	100 / 100	Grayscale images	10	75:25 stated

PD, Parkinson's disease. The test size for voice is inferred from the granularity of the reported accuracies and confirmed by the confusion matrices (Section 3.2).

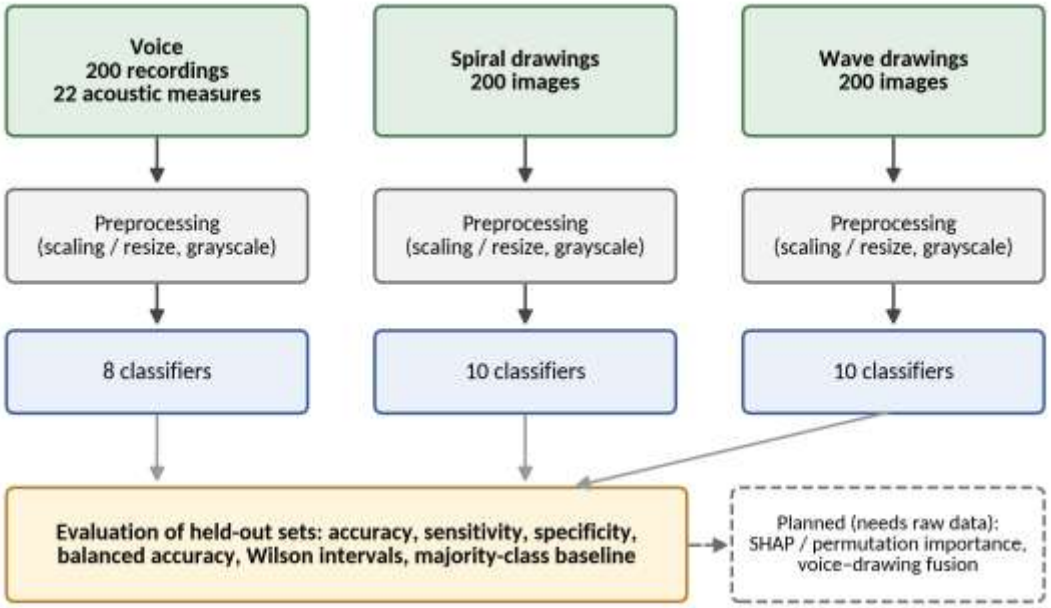


Figure 1. Analysis workflow. Solid boxes were carried out; the dashed box lists analyses that require raw data not available in the source material.

2.2 Sample Collection

Speech is captured by a microphone connected to a Raspberry Pi, and participants draw the spiral and wave patterns on paper, which are photographed with a camera and uploaded to the analysis portal. Before classification, drawings are resized, converted to grayscale, and normalized to a uniform format. The descriptors that the classical (non-deep) classifiers received from these images are not specified in the source, which limits interpretation of the drawing models (Section 3.5).

2.3 Feature Description

The voice measures follow the dysphonia feature set of Little *et al.* [4] and fall into five families (Table 2). The source text counts 24 voice features, of which 22 are named and described; the analysis therefore refers to 22 measures. No cepstral coefficients are present in the described training features.

Table 2. Acoustic feature families used for the voice task

Feature family	Measures	What it captures	Number
Fundamental frequency	Fo, Fhi, Flo	Pitch level and range	3
Frequency perturbation	Jitter (%), absolute jitter, RAP, PPQ, DDP	Cycle-to-cycle pitch variation	5
Amplitude perturbation	Shimmer, shimmer (dB), APQ3, APQ5, APQ, DDA	Cycle-to-cycle loudness variation	6
Noise and tonality	NHR, HNR	Breathiness, noise	2
Nonlinear dynamics	RPDE, DFA, spread1, spread2, D2, PPE	Signal complexity, pitch-period entropy	6

RAP, relative average perturbation; PPQ, five-point period perturbation quotient; DDP, average absolute difference of differences between periods; APQ, amplitude perturbation quotient; DDA, average absolute differences between amplitudes; NHR, noise-to-harmonics ratio; HNR, harmonics-to-noise ratio; RPDE, recurrence period density entropy; DFA, detrended fluctuation analysis; PPE, pitch period entropy.

2.4 Classification Models

Eight classifiers were compared for voice (k-nearest neighbors, support vector machine, naive Bayes, decision tree, logistic regression, random forest, linear discriminant analysis, and XGBoost) and ten for each drawing task (the same families plus gradient boosting, LightGBM, and CatBoost, without discriminant analysis). Boosting builds an additive ensemble of weak trees [17, 18]; XGBoost adds regularized tree construction [19], and CatBoost uses ordered boosting [20]. Models were scored on a held-out partition, and the best model per task was selected by accuracy.

2.5 Statistical Analysis

Confusion matrices shown in the source evaluation were read for the voice task (all eight classifiers) and the spiral task (nine of ten classifiers), from which sensitivity (recall for patients), specificity (recall for healthy controls), and balanced accuracy [21] were computed. For the wave task no per-class counts could be recovered. Uncertainty in proportions is given as 95% Wilson intervals [22]. Each best model was compared with the majority-class baseline using a one-sided exact binomial test, and the voice and spiral accuracies were compared with Fisher's exact test. Receiver operating characteristic curves, log-loss recomputation, and feature-importance analysis (permutation importance [23] or SHAP [24]) require predicted probabilities and raw data that were not available, and are specified as future analyses in Section 3.6. Metrics such as precision and recall are interpreted with the caution that class-level measures depend on the number of cases in each class [25].

3. RESULTS AND DISCUSSION

3.1 Classifier Comparison Across the Three Tasks

Table 3 shows held-out accuracy for every classifier. Tree-based ensembles led on all three tasks: XGBoost on voice (0.900), and CatBoost on spiral (0.875) and wave drawings (0.895). Single decision trees, random forests, and gradient boosting were consistent on the drawings (0.750 on both tasks), whereas logistic regression and the support vector machine were at or below

0.350 on both. Because CatBoost was not run for voice and XGBoost was weaker than CatBoost on the drawings, the data do not show which boosting implementation is generally preferable; they show only that boosted or bagged trees tolerated these small, heterogeneous feature sets better than linear and kernel methods.

Table 3. Held-out accuracy of all classifiers by task

Classifier	Voice	Spiral	Wave
Logistic regression	0.767	0.250	0.350
Support vector machine	0.867	0.250	0.250
k-nearest neighbors	0.767	0.500	0.400
Naive Bayes	0.567	0.500	0.600
Linear discriminant analysis	0.833	–	–
Decision tree	0.733	0.750	0.750
Random forest	0.867	0.750	0.750
Gradient boosting	–	0.750	0.750
XGBoost	**0.900**	0.625	0.625
LightGBM	–	0.625	0.625
CatBoost	–	**0.875**	**0.895**

Bold, best classifier per task; –, not evaluated for that task.

3.2 Class-Specific Performance and Error Analysis

The recovered confusion matrices (Tables 4 and 5, Figures 2 and 3) change the reading of the accuracies. The voice test set contained 30 recordings, 24 from patients and 6 from healthy controls, and the spiral test set contained only 8 drawings, 6 from patients and 2 from healthy controls. Neither partition is balanced, although the spiral data were described as equally divided.

Table 4. Voice task: confusion-matrix counts and class-specific metrics (30 test recordings)

Classifier	TN	FP	FN	TP	Accuracy	Sensitivity	Specificity	Balanced acc.
XGBoost	4	2	1	23	0.900	0.958	0.667	0.813
Random forest	4	2	2	22	0.867	0.917	0.667	0.792
Support vector machine	4	2	2	22	0.867	0.917	0.667	0.792
Linear discriminant analysis	4	2	3	21	0.833	0.875	0.667	0.771
k-nearest neighbors	4	2	5	19	0.767	0.792	0.667	0.729
Logistic regression	4	2	5	19	0.767	0.792	0.667	0.729
Decision tree	6	0	8	16	0.733	0.667	1.000	0.833
Naive Bayes	6	0	13	11	0.567	0.458	1.000	0.729

TN, healthy correctly identified; FP, healthy flagged as patient; FN, patient missed; TP, patient correctly identified. Sensitivity = TP/(TP+FN); specificity = TN/(TN+FP).

Natural Resources for Human Health
eISSN: 2583-1194

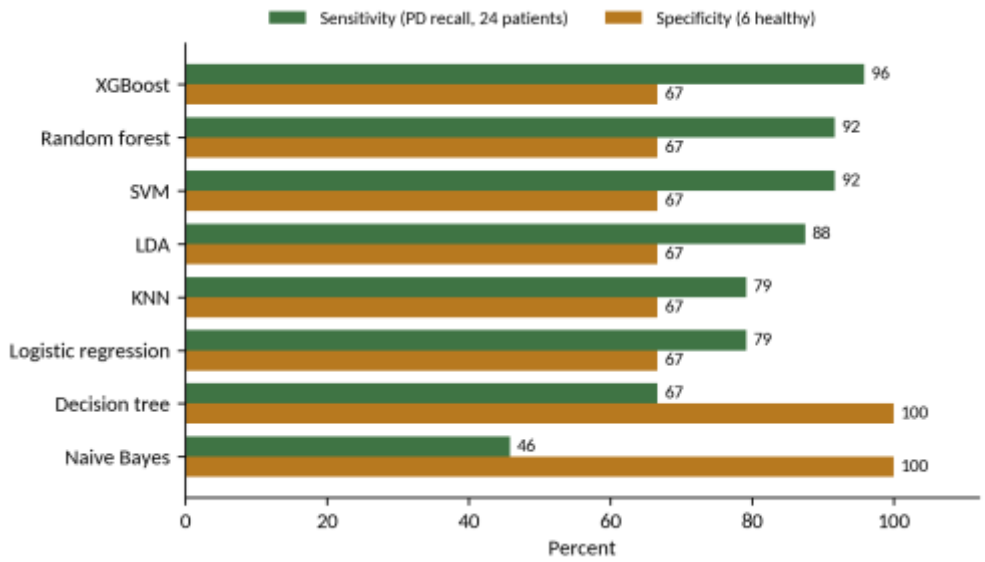


Figure 2. Sensitivity and specificity of the eight voice classifiers on the 30-recording test set.

For voice, the six linear, kernel, instance-based, and boosted models each misclassified two of the six healthy recordings (specificity 0.667) and differed only in how many patients they missed. XGBoost missed one of 24 patients (sensitivity 0.958). The two models with perfect specificity, the decision tree and naive Bayes, achieved it by under-detecting patients (sensitivity 0.667 and 0.458). By balanced accuracy the decision tree (0.833) matched or exceeded XGBoost (0.813), which shows that the ranking by plain accuracy is partly an artifact of the 4:1 class imbalance: with only six healthy cases, a single recording moves specificity by 16.7 percentage points.

Table 5. Spiral task: confusion-matrix counts and class-specific metrics (8 test drawings)

Classifier	TN	FP	FN	TP	Accuracy	Sensitivity	Specificity	Balanced acc.
CatBoost	2	0	1	5	0.875	0.833	1.000	0.917
Decision tree	2	0	2	4	0.750	0.667	1.000	0.833
Random forest	2	0	2	4	0.750	0.667	1.000	0.833
Gradient boosting	2	0	2	4	0.750	0.667	1.000	0.833
XGBoost	2	0	3	3	0.625	0.500	1.000	0.750
Naive Bayes	2	0	4	2	0.500	0.333	1.000	0.667
k-nearest neighbors	2	0	4	2	0.500	0.333	1.000	0.667
Logistic regression	2	0	6	0	0.250	0.000	1.000	0.500
Support vector machine	2	0	6	0	0.250	0.000	1.000	0.500

Abbreviations as in Table 4. LightGBM (accuracy 0.625) is omitted because its matrix was not examined.

The spiral models show a different profile. Every classifier labeled both healthy drawings correctly (specificity 1.000), and all errors were missed patients. Logistic regression and the support vector machine predicted "healthy" for all eight drawings, so their accuracy of 0.250 is exactly the share of healthy cases in the test set and reflects a degenerate constant predictor rather than a weak signal. CatBoost missed one of six patients (sensitivity 0.833). In screening terms, the voice model errs toward over-calling disease and the spiral models toward under-calling it, a contrast examined in Section 3.5.

Natural Resources for Human Health
eISSN: 2583-1194

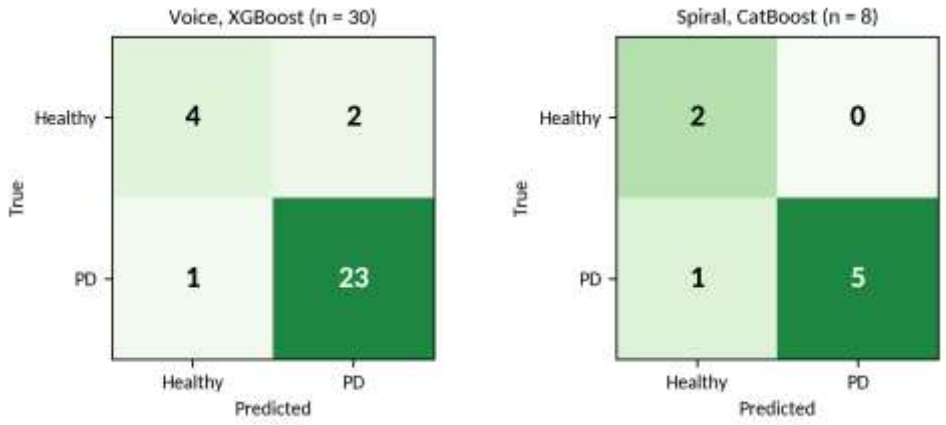


Figure 3. Confusion matrices of the best voice model (XGBoost) and the best spiral model (CatBoost).

3.3 Uncertainty and Majority-Class Baselines

Table 6 and Figure 4 place the best accuracies in context. With 30 and 8 test cases, the 95% Wilson intervals span about 22 and 45 percentage points, and the point estimates sit close to the accuracy of always predicting "Parkinson's disease" (80% for voice, 75% for spiral). One-sided exact binomial tests do not reject equality with those baselines ($p = 0.12$ and 0.37). This does not show that the models have no signal; it shows that these test sets cannot establish that they have. The voice and spiral accuracies are statistically indistinguishable (Fisher's exact $p = 1.0$).

Table 6. Best-model accuracy, uncertainty, and baseline comparison

Task	Accuracy	Test cases (PD / healthy)	95% Wilson interval	Majority baseline	Exact p vs baseline
Voice (XGBoost)	90.0%	30 (24 / 6)	74.4% – 96.5%	80.0%	0.12
Spiral (CatBoost)	87.5%	8 (6 / 2)	52.9% – 97.8%	75.0%	0.37
Wave (CatBoost)	89.5%	Not recoverable	78.0% – 95.3% if n = 50	Unknown	–

One-sided exact binomial test of the best model against the majority-class proportion. The wave interval assumes the stated 25% test partition of 200 drawings (n = 50); with 8 test drawings it would be as wide as the spiral interval.

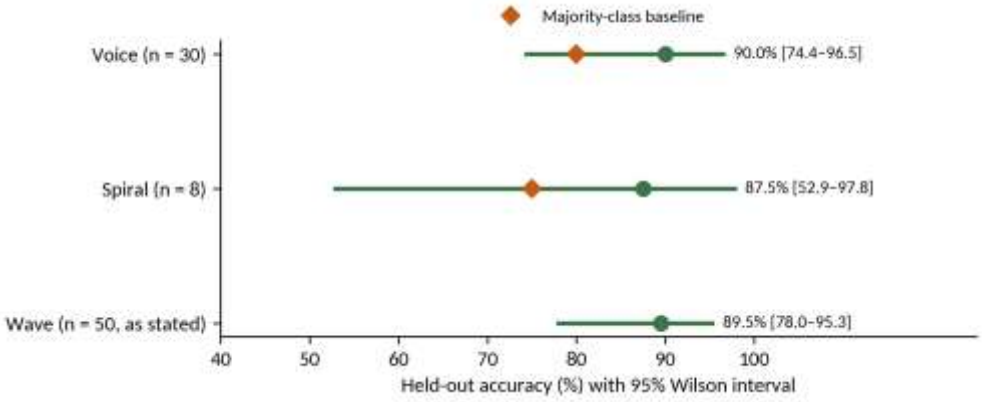


Figure 4. Best-model accuracy with 95% Wilson intervals; diamonds mark the majority-class baseline on the same test set.

The wave result needs particular care. An accuracy of 89.5% cannot arise from a test set of 8 or of 50 drawings, because neither yields that proportion, and the wave logs repeat several values from the spiral log. The wave test size and class composition therefore have to be re-established from the original code before this figure can be compared with the others.

3.4 Spiral Versus Wave Drawings

Both drawing tasks were best served by CatBoost, with accuracies of 87.5% and 89.5%. The 2-point difference is less than one test image at the spiral test size and is not interpretable. The classifier rankings were similar: decision tree, random forest, and gradient boosting scored 0.750 on both tasks, XGBoost and LightGBM scored 0.625 on both, and logistic regression and the support vector machine failed on both. The agreement suggests that the two tasks present comparable difficulty to these models, but with unresolved test sizes it cannot support the view that one task is more reliable. Conceptually the tasks differ: the spiral emphasizes amplitude decay, hesitation, and tremor over a continuously changing curvature, whereas the wave emphasizes rhythm, amplitude regularity, and speed variability during repetitive motion. Whether they carry distinct information can be tested only on drawings from the same participants.

3.5 Voice Versus Drawing: Complementarity and Interpretability

Voice and spiral models made different kinds of errors: the best voice model had high sensitivity and modest specificity, while the spiral models had perfect specificity and imperfect sensitivity. If such profiles persisted in larger aligned data, a combination rule could be designed to exploit them, for example by requiring agreement for a positive screen or by flagging disagreement for clinical review. This is a hypothesis, not a finding: it rests on six healthy voice recordings and two healthy drawings, and the voice and drawing records are not from linked participants, so complementarity cannot be measured here.

On interpretability, the voice features have clear physiological meaning (Table 2), so an interpretable model can be obtained by pairing boosted trees with SHAP or permutation importance, and it is notable that a single decision tree reached the best balanced accuracy on voice. That result favors transparent models as the first choice for a screening tool, provided they are validated on larger test sets. For drawings, interpretability is limited because the image-derived descriptors are not specified; documenting them, or replacing them with kinematic indices such as speed and pen pressure [9], would allow importance analysis. Table 7 lists the analyses that remain, with the data each requires.

Table 7. Analyses still required to complete the comparison

Analysis	Procedure	Data required
Reproduce baselines	Refit XGBoost and CatBoost with fixed seeds on identical splits	Raw features, images, split indices
Leakage-resistant validation	Repeated stratified splits grouped by participant [14]	Participant identifiers
Feature importance	Permutation importance and SHAP per task	Raw features and fitted models
ROC and calibration	ROC-AUC, reliability curves, Brier score	Predicted probabilities
Combined model	Voice, drawing, and joint models on identical folds	Participants with both tasks
Error review	Inspect misclassified recordings and drawings	Sample-level predictions

3.6 Limitations and Recommendations

The principal limitations are as follows. First, the test sets are very small, and the wave test set is undocumented. Second, the number of unique participants is unknown, so recordings or drawings from one person may appear in both training and test partitions, which would inflate accuracy [14]. Third, the class composition of the test partitions differs from that of the datasets, which suggests that splits were not stratified. Fourth, no feature-importance, ROC, or calibration results exist, and the voice–drawing comparison is indirect. To obtain an estimate of 90% accuracy with a 95% half-width of 5 percentage points, about 140 test cases are needed (about 170 at 87.5%), so the current 200-sample datasets should be evaluated by repeated participant-grouped cross-validation rather than by a single small hold-out. We also recommend reporting sensitivity and specificity together

with the majority-class baseline for every task, and recording age, sex, medication state, and disease stage so that results can be stratified. These findings relate to screening support only and do not constitute diagnosis.

4. CONCLUSION

This comparative analysis of acoustic and drawing-based biomarkers found that boosted tree models performed best on voice (XGBoost, 90.0%), spiral (CatBoost, 87.5%), and wave drawings (CatBoost, 89.5%). Recovering the confusion matrices showed that the voice and spiral accuracies came from 30 and 8 test cases, that they were not significantly better than majority-class prediction, and that the modalities had different error profiles: the voice model over-called disease, whereas the spiral models missed patients. The accuracies do not allow either modality to be called more reliable, and no difference between spiral and wave tasks can be demonstrated. Feature-level interpretation and the combined voice–drawing model remain to be done with raw, participant-linked data, for which the required analyses were specified. With adequately sized, participant-grouped evaluation, the same prototype could provide a trustworthy comparison of these low-cost biomarkers.

ACKNOWLEDGEMENTS

The authors thank the Department of Electronics and Communication Engineering, Visvesvaraya Technological University, Belagavi, VTU Centre for Postgraduate Studies Mysuru, Karnataka, India, for access to the laboratory facilities used to develop the prototype.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Ethical approval and informed-consent procedures for the voice and drawing data are not documented in the source report.

DATA AVAILABILITY STATEMENT

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Kalia, L.V., Lang, A.E., 2015. Parkinson's disease. *The Lancet* 386(9996), 896–912. https://doi.org/10.1016/S0140-6736(14)61393-3

[2] Postuma, R.B., et al., 2015. MDS clinical diagnostic criteria for Parkinson's disease. *Movement Disorders* 30(12), 1591–1601. https://doi.org/10.1002/mds.26424

[3] Rusz, J., et al., 2011. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *The Journal of the Acoustical Society of America* 129(1), 350–367. https://doi.org/10.1121/1.3514381

[4] 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

[5] Sakar, B.E., et al., 2013. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics* 17(4), 828–834. https://doi.org/10.1109/JBHI.2013.2245674

[6] Abedinzadeh Torghabeh, F., Hosseini, S.A., Ahmadi Moghadam, E., 2023. Enhancing Parkinson's disease severity assessment through voice-based wavelet scattering, optimized model selection, and weighted majority voting. *Medicine in Novel Technology and Devices* 20, 100266. https://doi.org/10.1016/j.medntd.2023.100266

[7] Iyer, A., et al., 2023. A machine learning method to process voice samples for identification of Parkinson's disease. *Scientific Reports* 13, 20615. https://doi.org/10.1038/s41598-023-47568-w

[8] Amato, F., Borzì, L., Olmo, G., Orozco-Arroyave, J.R., 2021. An algorithm for Parkinson's disease speech classification based on isolated words analysis. *Health Information Science and Systems* 9, 32. https://doi.org/10.1007/s13755-021-00162-8

[9] Zham, P., et al., 2017. Distinguishing different stages of Parkinson's disease using composite index of speed and pen-pressure of sketching a spiral. *Frontiers in Neurology* 8, 435. https://doi.org/10.3389/fneur.2017.00435

[10] Pereira, C.R., et al., 2016. A new computer vision-based approach to aid the diagnosis of Parkinson's disease. *Computer Methods and Programs in Biomedicine* 136, 79–88. https://doi.org/10.1016/j.cmpb.2016.08.005

[11] Chakraborty, S., Aich, S., Sim, J.-S., Han, E., Park, J., Kim, H.-C., 2020. Parkinson's disease detection from spiral and wave drawings using convolutional neural networks: a multistage classifier approach. In: 2020 22nd International Conference on Advanced Communication Technology (ICACT), pp. 298–303. https://doi.org/10.23919/ICACT48636.2020.9061497

[12] Huang, Y., Chaturvedi, K., Nayan, A.-A., Hesamian, M.H., Braytee, A., Prasad, M., 2024. Early Parkinson's disease diagnosis through hand-drawn spiral and wave analysis using deep learning techniques. *Information* 15(4), 220. https://doi.org/10.3390/info15040220

[13] SasiRekha, S., Shankar, R., Duraisamy, S., 2025. Parkinson's disease prediction with spiral drawings and wave frequency using deep conformal neural networks. *Journal of Neonatal Surgery* 14(4S), 11–22. https://doi.org/10.52783/jns.v14.1725

[14] Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C., Kording, K.P., 2017. The need to approximate the use-case in clinical machine learning. *GigaScience* 6(5), gix019. https://doi.org/10.1093/gigascience/gix019

[15] Ashok, R.S., Anil Kumar, D., 2025. Machine learning based early detection of Parkinson disease using handwriting and vocal features. *Research on Engineering Structures and Materials* 12(1), 1–27. https://doi.org/10.17515/resm2025-835ml0422rs

[16] Ashok, R.S., Anil Kumar, D., 2025. Multimodal machine learning framework for early detection of Parkinson disease using handwriting, vocal features, ECG and gait analysis. *Journal of Information Systems Engineering and Management* 9(4).

[17] Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[18] Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

[19] Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785

[20] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. *arXiv* https://doi.org/10.48550/arXiv.1706.09516

[21] Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition, pp. 3121–3124. https://doi.org/10.1109/ICPR.2010.764

[22] Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953

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

[24] Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. *arXiv* https://doi.org/10.48550/arXiv.1705.07874

[25] Powers, D.M.W., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* https://doi.org/10.48550/arXiv.2010.16061