NeuroPlay: A Multi-Modal Game-Based Diagnostic Tool for Early Detection and Staging for Parkinson's Disease

Ayesha Faruki*
Beaver Works Summer Institute
Massachusetts Institute of Technology
ayesha@afaruki.com

Ananya Chitnis*
Beaver Works Summer Institute
Massachusetts Institute of Technology
ananya.chitnis@gmail.com

Yeowon Yoon*
Beaver Works Summer Institute
Massachusetts Institute of Technology
yeowonyoon0109@gmail.com

Abstract— **Parkinson's disease (PD) is a progressive neurological disorder that often goes undiagnosed in its early stages due to subtle and overlapping symptoms. Early intervention can significantly improve quality of life, yet accessible diagnostic tools remain limited. In this study, we present NeuroPlay, an interactive game-based screening tool designed to detect Parkinson's disease and estimate its stage using two non-invasive modalities: tremor-based drawing analysis and voice pattern recognition. Using a curated dataset of Parkinson's handwriting and vocal recordings, we trained separate machine learning classifiers to (1) determine the presence of PD based on motor tremor patterns, and (2) predict disease stage based on speech features. Our highest-scoring models for the tremor drawing test and speech test were Random Forest and Neural Network with distinct feature engineering and preprocessing methods, with 94% and 67% accuracy, respectively. Our approach highlights the potential of gamified diagnostics in low-cost, scalable early detection, and lays the groundwork for future integration of cognitive memory assessments and mobile deployment.**

***Keywords*— Machine learning, neurodegenerative disorders, Parkinson's disease, signal processing**

I. Introduction

Neurodegenerative diseases, including Parkinson's, Alzheimer's, and ALS, affect millions worldwide and impose an increasing challenge to patients, caregivers, and healthcare systems. Among these, Parkinson's Disease is the second most prevalent neurodegenerative disorder, affecting over 8.5 million people worldwide in 2019 and contributing to approximately 329,000 deaths—a number that has doubled since 2000 [1]. Parkinson's is particularly characterized by a progressive decline in motor function, caused by the deterioration of neurons that produce dopamine [2]. Currently, there is a lack of early diagnosis methods for Parkinson's, as well as a definitive diagnosis method, such as a blood test, which means patients are often only diagnosed when they meet the clinically required symptoms [3]. Symptom-based diagnoses can be challenging because many neurodegenerative diseases share similar symptoms, leading to frequent misdiagnoses [4]. Therefore, many Parkinson's patients receive late treatment and remain undiagnosed during critical early stages, inhibiting preventive treatment that may delay the onset of later symptoms.

Recent advances in machine learning (ML) and digital health technologies present new opportunities for accessible, scalable, and early-stage Parkinson's detection. Prior research has demonstrated that Parkinson's affects multiple domains, including speech, cognition, and movement, which can all be detected with lightweight, non-invasive techniques [5]. One of the primary symptoms of Parkinson's includes tremors, which are often characterized by the shaking of the hands, feet, or jaw, occurring in a rhythmic back-and-forth movement. Individuals with Parkinson's often also exhibit speech changes, which can include speaking quietly, monotonously, or hesitantly, paired with slurred speech. Some people with later stages of Parkinson's develop cognitive or memory problems, like Parkinson's disease dementia, which may affect visuospatial skills, attention, language, and reasoning [5].

Prior studies have demonstrated that ML models trained on spiral drawing data can detect and monitor abnormalities associated with Parkinson's, while voice-based models have identified acoustic biomarkers, such as jitters, shimmers, and the harmonics-to-noise ratio, for early-stage detection [6]. However, most of these models are single modality and require controlled environments, limiting their use in real-world settings [7].

We propose a multimodal, game-based machine learning framework for accessible at-home screening and monitoring of Parkinson's disease. Our system leverages two inputs: (1) a spiral drawing test to capture tremor patterns and (2) a speech task to identify vocal biomarkers mentioned previously. By integrating motor and vocal data, we aim to build an accurate classifier for both binary diagnosis and disease staging. The tool is designed to be low-cost, non-invasive, and engaging, supporting early detection and continuous monitoring. We also discuss future expansions in diagnosing other neurodegenerative diseases and deploying the mobile app to increase accessibility, since games have been more effective in diagnostic tests related to brain changes, as patients are more motivated to participate in digital game diagnostic programs than more traditional pen-and-paper tests.

II. Methods

* Authors contributed equally to this work and are co-first authors

979-8-3315-5937-3/25/$31.00 ©2025 IEEE

The development of this project was carried out in three stages: developing an ML model for the spiral drawing test, developing an ML model for the speech task, and combining the two models for multi-modal prediction in the development of a mobile application.

A. Drawing Task

The *Parkinson Disease Spiral Drawings Using Digitized Graphics Tablet* dataset from the University of California, Irvine's machine learning repository was utilized in the training and testing of a model suitable for identifying Parkinson's in a drawing test [8]. The dataset consisted of data from three different spiral test varieties—the Static Spiral Test, Dynamic Spiral Test, and Stability Test—from 62 people with Parkinson's and 15 healthy individuals. Drawing test data was available for each participant, providing the X, Y, and Z coordinates, pressure, and grip angle for each pen stroke.

1) *Feature Engineering*: To accommodate features that could be realistically recreated in an accessible, simple-to-use mobile application, Z coordinates and grip angle were dropped from the dataset. To extract discriminative motor features from the spiral drawing data and to be able to apply them to other shapes, such as a square, circle, or triangle, we implemented a feature engineering pipeline designed to quantify tremor characteristics, kinematic profiles, and pressure control from raw stylus recordings. Raw test data kept in each participant's .txt file was concatenated into a unified dataframe, and data was scaled using the scikit-learn library's StandardScaler feature.

Tremor-specific energy was computed using a bandpass filter targeting the 3–8 Hz frequency band, which is just outside the typical Parkinson's rest tremor, which is 4–7 Hz [9]. The filtered signal's squared energy was summed to quantify tremor intensity per channel (X, Y, and pressure). This method accounts for intra-test variability and emphasizes subtle motor fluctuations.

For each type of test—the following were computed: statistical descriptors, such as mean, standard deviation, and range for X, Y, and Pressure; tremor energy as previously described, computed per signal dimension; instantaneous speeds derived from frame-to-frame Euclidean distances normalized by time intervals; mean and standard deviation of speed to capture movement consistency; and total drawing distance and test duration to assess bradykinesia and motor fatigue.

2) *Model Training & Testing:* Following feature engineering, three model architectures were selected to assist in classification of the data into classes of "Parkinson's" and "No Parkinson's:" Logistic Regression, to serve as a baseline model of comparison, and XGBoost and Random Forest, for their known competency in classification tasks with tabular data. Stratified K-fold cross-validation, with five folds, was used for training each model. This means that each split in the data preserves class proportions, which is especially crucial when one class is significantly underrepresented compared to the other, as is the case for this dataset. The Random Forest model was trained for 100 estimators.

B. Speech Task

In designing our model, we aimed not just to detect the presence of Parkinson's disease, but to estimate stages of severity, which are critical for early intervention and personalized care. Initially, we were unsure how to approach this until we discovered the *Oxford Parkinson's Telemonitoring Voice Dataset* (OPTD), which contains sustained phonation samples and corresponding UPDRS (Unified Parkinson's Disease Rating Scale) scores—a clinical scale used to assess motor function and progression in Parkinson's patients [10]. A 2010 study by Tsanas et al. pioneered the use of this dataset by applying regression models to accurately predict continuous UPDRS values, moving beyond binary classification to provide a stage-like severity estimate [11]. Most recently, in 2019, researchers used machine learning classifiers and regression to achieve multi-class stage prediction from voice features alone, demonstrating the potential for digital tools to differentiate between mild, moderate, and severe Parkinson's cases [11]. These studies validated the feasibility of our approach and informed our decision to integrate stage-sensitive modeling into the model's system.

1) *Feature Engineering*: We began by loading the complete 5,875 × 21 phonation table and verifying its integrity, noting no missing values and only minor outliers that we kept to preserve natural variability.

To prune redundancy, we computed pairwise Pearson correlations among the 16 voice measures and dropped one member of any pair with $\rho > 0.90$. For example, percent-Jitter ($\rho \approx 0.94$ with absolute Jitter) and its RAP/DDP variants were removed, retaining only the more interpretable absolute Jitter. Similarly, Shimmer(dB) proved merely a log-scaled copy of Shimmer; we removed it along with its APQ5 and DDA counterparts, retaining the two amplitude descriptors APQ11 and APQ3. Next, we eliminated deterministic transforms by discarding features that were exact multiples of others (e.g., DDP = 3×RAP, DDA = 3×APQ3) [13]. Raw and potential identifiers (subject# and test_time) were also removed.

For noise metrics, between inversely related pairs, we selected the cleaner distribution with stronger clinical relevance: harmonic-to-noise ratio (HNR) was kept over its skew-heavy inverse (NHR). Age was tested in the baseline model but ultimately excluded from our regression pipeline to prevent participant identification given our small sample size; we relied on age-stratified splits to guard against overfitting and omitted age entirely in the classification model.

Our final feature set comprised two jitter metrics, two shimmer descriptors, HNR, three nonlinear-dynamics scores

(RPDE, DFA, PPE), and sex. All features were standardized with scikit-learn's StandardScaler, then split 70:15:15 into train, validation, and test folds.

Finally, for stage prediction, we experimented with two label-splitting schemes: a four-way split (0–20, 20–30, 30–40, ≥40) to achieve more even class distributions and a three-way split (0–25, 25–33, ≥33) to reflect clinically meaningful tiers of very mild, mild, and moderate-plus severity given the absence of true severe cases in the dataset [14].

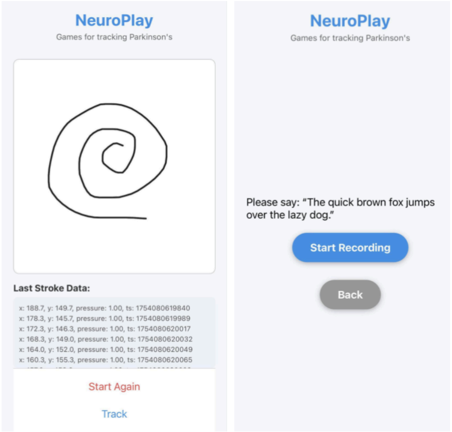
2) *Model Training & Testing*: Following initial feature engineering, we evaluated a comprehensive suite of regression algorithms, including Ordinary Least Squares, Support Vector Regression, k-Nearest Neighbors, Gradient Boosting, and XGBoost. They were evaluated using five-fold cross-validation and optimizing for root mean squared error (RMSE) to ensure clinically interpretable error bounds in UPDRS units. Random Forest and a custom Keras-based neural network emerged as the top performers, achieving the lowest validation RMSE and thus meriting deeper optimization.

3) *Random Forest Optimization:* To refine the Random Forest, we first conducted a GridSearchCV over n_estimators (100–500), max_depth (10–30), min_samples_leaf (1–5), and max_features ('auto', 'sqrt', 'log2'), optimizing for the lowest RMSE. We then applied HalvingGridSearchCV to reduce computational power. The final configuration of 439 trees, max_depth = 22, min_samples_leaf = 2, max_features = 'sqrt' yielded an RMSE of 8.3 on validation folds and 8.4 on the test set.

4) *Neural Network Optimization:* An initial randomized search over layers (4 layers from 64–256), lr (1e-4–1e-2), batch (16–64), epochs up to 300 was followed by focused grid searches and manual tuning. The resulting four-layer model (128→128→64→32 units) with batch normalization and 20% dropout trained up to 300 epochs at batch size 16 yielded an RMSE of approximately 7.3 in CV and 7.7 on the test set, suggesting slight overfitting with minimal impact.

To better reflect clinical practice, we discretized UPDRS scores into three and four severity categories and retrained the neural network as a multiclass classifier. Performance was evaluated via precision, recall, and $F1_1$-scores for each severity stage, confirming that the network maintained discriminative power across clinically relevant bins.

C. App Prototyping



x: 188.7, y: 149.7, pressure: 1.00, ts: 1754080619840
x: 178.3, y: 145.7, pressure: 1.00, ts: 1754080619989
x: 172.3, y: 146.3, pressure: 1.00, ts: 1754080620017
x: 168.3, y: 149.0, pressure: 1.00, ts: 1754080620032
x: 164.0, y: 152.0, pressure: 1.00, ts: 1754080620049
x: 160.3, y: 155.3, pressure: 1.00, ts: 1754080620065

Please say: "The quick brown fox jumps over the lazy dog."

Fig. 1 Screenshot of NeuroPlay's interface

A prototypical mobile application was developed utilizing the React Native framework (so that the application would be capable of running on both Android and iOS devices), with a Flask API backend that contained the pickled models. The drawing interface allowed for a gamified dynamic spiral test—with an Archimedean spiral flashing on the screen for two seconds, followed by an interface where the user could draw themselves, while recording keystrokes. The speech interface allowed the user to record themselves saying a phrase (i.e., a tongue twister) and then evaluate their predicted Parkinson's UPDRS. The backend of the application contained the Python script to extract features from raw data so that it could be handled by the models.

III. Results

A. Tremor Drawing Task

The performance of the three machine learning models—Logistic Regression, XGBoost, and Random Forest—was evaluated on the drawing task dataset using standard classification metrics, including precision, recall, F1 score, accuracy, and AUC (Area Under the ROC Curve). Table 1 summarizes the results for each model.

TABLE I
Tremor Drawing Task Model Evaluation Metrics

Metric	**Tremor Drawing Test Model**		
	Random Forest	**XGBoost**	**Logistic Regression**
Precision	0.93	0.91	0.85
Recall	0.95	0.93	0.89
F1 Score	0.94	0.92	0.86
Accuracy	0.94	0.93	0.87
AUC	0.99	0.99	0.92

Among the evaluated models, Random Forest demonstrated the highest overall performance, achieving the best precision, recall, F1 score, and accuracy. XGBoost closely followed, outperforming Logistic Regression across all metrics. Despite lower metrics, the Logistic Regression model still showed fairly competitive results.

The ROC curves in Figure 2 illustrate the near-perfect performance with AUC values of 0.99 for XGBoost and Random Forest models, highlighting their capacity to distinguish between Parkinson's patients and healthy controls.

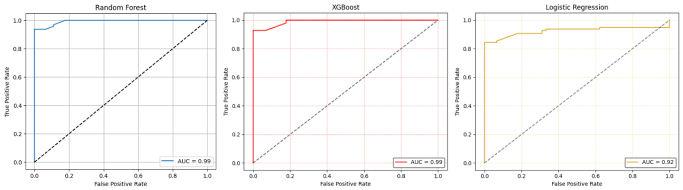


Fig. 2 ROC AUC curves for each drawing test model

Further insight into the models' classification behavior can be gleaned from the confusion matrices in Figure 3, which aggregate the results from 5-fold cross-validation.

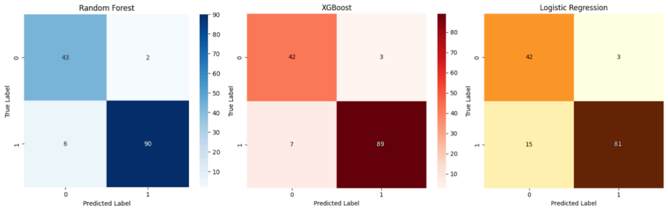


Fig. 3 Confusion matrices for each drawing test model

The Random Forest model slightly outperformed the others in terms of true positives and negatives, correctly classifying 43 out of 45 healthy individuals and 90 out of 96 Parkinson's patients. It reduced both false positives and false negatives compared to the other models. These confusion matrices reinforce that ensemble methods like Random Forest and XGBoost not only achieve higher metrics but also maintain consistent accuracy across different validation folds, making them well-suited for early-stage Parkinson's detection from drawing tasks.

B. Speech Task

Table 2 compares Random Forest and our neural network on the voice-based UPDRS regression task, reporting R², RMSE, MAE, and explained variance.

TABLE 2
Voice Regression Task Model Evaluation Metrics

Metric	Voice Regression Test Model	
	Random Forest	**Neural Network**
R^2 Score	0.358	0.463
RMSE	8.394	7.677
MAE	6.477	5.836
Explained Variance	0.358	0.466

The neural network outperformed Random Forest on the continuous UPDRS prediction, delivering an R² of 0.463 versus 0.358, alongside lower RMSE (7.677 vs. 8.394) and MAE (5.836 vs. 6.477). Its higher explained-variance (0.466) confirms that the deeper, ReLU-driven architecture captures more of the subtle voice-disorder signal than the tree ensemble.

Table 3 presents the voice-based UPDRS stage classification metrics—precision, recall, F1-score, accuracy, macro-averaged F1, and ROC AUC—for both the 4-way and 3-way severity splits.

TABLE 3
Voice Classification Model Evaluation Metrics

Metric	Voice Classification Test Model	
	4-Way Split	**3-Way Split**
Precision	0.62	0.67
Recall	0.62	0.67
F1-score	0.63	0.67
Accuracy	0.62	0.67
Macro Avg F1	0.62	0.67
ROC AUC (Macro)	0.8478	0.8579

Binning UPDRS into three or four stages yielded modest but meaningful stage-prediction accuracy. The 3-way split edged out the 4-way setup, achieving 0.67 accuracy, precision, recall, F1, and macro-AUC of 0.858. This indicates clearer separation when utilizing smaller separation bins that retains its clinical relevance. This suggests the network's strongest clinical utility lies in broad severity categories rather than fine-grained stages.

IV. Discussion and Conclusion

NeuroPlay represents a promising step toward how we screen for neurodegenerative disorders—leveraging casual gameplay to extract meaningful behavioral and cognitive biomarkers. While our initial prototype uses tremor-sensitive drawing and voice data as diagnostic input, this work is only the beginning. Our best models for the tremor drawing test and voice test are 94% and 67%, respectively (Random Forest and 3-way split Neural Network), showing promise in detecting Parkinson's and a stepping stone for identifying UPDRS scores. With further data acquisition and model fine-tuning, we are confident that this multi-modal pipeline tool and mobile app can serve as a robust system for Parkinson's prognosis.

A. Limitations and Future Direction

Future directions include expanding the scope of our model, NeuroPlay, to screen for a broader range of neurodegenerative conditions beyond Parkinson's, maximizing accessibility by deploying this as a mobile app, and refining our models through clinical validation with real patient data by partnering with institutions like UT Southwestern, one of the nation's leading medical centers for Parkinson's disease care. By embedding screening into engaging game experiences, especially on mobile devices, we aim to increase diagnostic accuracy, accessibility, and user participation, making early detection more effective and patient-centered.

1) *Broader Range*: As our model NeuroPlay evolves, one critical next step is expanding its capabilities beyond Parkinson's to screen for a broader range of neurological disorders, such as Alzheimer's, ALS, Huntington's disease, etc.. Many of these disorders present with overlapping symptoms, like slowed reaction time, memory impairment, or speech changes (to name a few), leading to frequent misdiagnoses or delayed intervention [4]. By expanding on our model's current ability to predict both if a patient has Parkinson's disease and, if they do, what stage they are in, we hope to differentiate between these conditions more accurately by creating additional "mini-games" that each test a symptom, the scores of which allow a diagnosis. Through this, we hope our future developments offer patients a more nuanced and accessible screening.

2) *Mobile Launch:* To maximize accessibility and long-term usability, we aim to develop and deploy our model, NeuroPlay, as a mobile app. Gamified diagnostic tools on smartphones and tablets increase user motivation and engagement and reduce the stress associated with formal cognitive tests. While digital health tools can pose usability challenges for geriatric populations, this group also stands to benefit significantly from early, accurate detection of neurological conditions, such as through our model, which will offer an at-home, accurate diagnosis with stage-detection to maximize the specificity and effectiveness of the care they will receive. By incorporating senior-friendly design—such as larger interfaces, simplified instructions, and voice guidance—NeuroPlay can help bridge the accessibility gap and deliver scalable, engaging screening in all demographics regardless of age.

3) *The University of Texas Southwestern*: A large limitation to the development of NeuroPlay was the lack of available datasets that could be used to train our data. Utilizing more data would enhance generalizability, and validating NeuroPlay in real-world clinical settings is essential. We plan to partner with UT Southwestern (UTSW), specifically the Peter O'Donnell Jr. Brain Institute, to test our model with actual patients across varying stages of neurological diseases. This collaboration will help us evaluate the tool's diagnostic sensitivity and specificity in clinical settings and refine our models based on diverse, real-life data. Validating with UTSW will be an important milestone toward establishing that our model is clinically reliable and practical so that it can be medically integrated as a screening tool.

Acknowledgments

A.F., A.C., and Y.Y. contributed equally to this work and are co-first authors.

We would like to thank our instructor, TAs, and program directors at MIT's Beaver Works Summer Institute for their support in this project. BWSI is transformational!

References

[1] Z. Ou et al., "Global Trends in the Incidence, Prevalence, and Years Lived with Disability of Parkinson's Disease in 204 Countries/Territories from 1990 to 2019," Frontiers in Public Health, vol. 9, Dec. 2021, doi: https://doi.org/10.3389/fpubh.2021.776847.

[2] D. G. Gadhave et al., "Neurodegenerative Disorders: Mechanisms of Degeneration and Therapeutic Approaches with Their Clinical Relevance," Ageing Research Reviews, vol. 99, pp. 102357–102357, Jun. 2024, doi: https://doi.org/10.1016/j.arr.2024.102357.

[3] M. Ugrumov, "Development of early diagnosis of Parkinson's disease: Illusion or reality?," CNS Neuroscience & Therapeutics, vol. 26, no. 10, pp. 997–1009, Jun. 2020, doi: https://doi.org/10.1111/cns.13429.

[4] Penn Medicine, "Challenges Connecting to Neurodegenerative Disease Care," Penn Medicine News, 2023. [Online]. Available: https://www.pennmedicine.org/news/challenges-connecting-to-neurodegenerative-disease-care.

[5] National Institute of Neurological Disorders and Stroke, "Parkinson's disease," *National Institute of Neurological Disorders and Stroke*, 2025. https://www.ninds.nih.gov/health-information/disorders/parkinsons-disease

[6] F. Sabermahani, M. Almasi-Dooghaee, and A. Sheikhtaheri, "Development and evaluation of serious games for diagnosis and cognitive improvement of patients with mild cognitive impairment: A study protocol," Informatics in Medicine Unlocked, vol. 32, p. 101039, Jan. 2022, doi: https://doi.org/10.1016/j.imu.2022.101039.

[7] "Image name," Pennmedicine.org, 2025. https://www.pennmedicine.org/news/challenges-connecting-to-neurodegenerative-disease-care

[8] "UCI Machine Learning Repository," archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/395/parkinson+disease+spiral+drawings+using+digitized+graphics+tablet

[9] A. Gironell, B. Pascual-Sedano, I. Aracil, J. Marín-Lahoz, J. Pagonabarraga, and J. Kulisevsky, "Tremor Types in Parkinson Disease: A Descriptive Study Using a New Classification," *Parkinson's Disease*, vol. 2018, pp. 1–5, Sep. 2018, doi: https://doi.org/10.1155/2018/4327597.

[10] A. Tsanas and M. Little. "Parkinsons Telemonitoring," UCI Machine Learning Repository, 2009. [Online]. Available: https://doi.org/10.24432/C5ZS3N.

[11] A. 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, Apr. 2010, doi: https://doi.org/10.1109/tbme.2009.2036000.

[12] Ramezani H, Khaki H, Erzin E, Akan OB. Speech features for telemonitoring of Parkinson's disease symptoms. Annu Int Conf IEEE Eng Med Biol Soc. 2017 Jul;2017:3801-3805. doi: https://doi.org/10.1109/EMBC.2017.8037685. PMID

[13] Boersma, P., & Weenink, D., Praat: *doing phonetics by computer* [Computerprogram]. Section "Perturbation measures." Available: https://www.fon.hum.uva.nl/praat/manual/Perturbation_measures.htm.

[14] P. Martínez-Martín et al., "Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale," *Parkinsonism & Related Disorders*, vol. 21, no. 1, pp. 50–54, Jan. 2015. Doi: https://doi.org/10.1016/j.parkreldis.2014.08