Towards Causal Interpretability in Parkinson's Voice Detection

Aniruth Ananthanarayanan[1], Sudeep Senivarapu[1], Anishsairam Murari[1], Dr. Jianguo Liu[2]
[1]Texas Academy of Mathematics & Science
[2]Department of Mathematics, University of North Texas

BACKGROUND

Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects more than 10 million people worldwide. Existing non-invasive detection approaches, such as imaging, wearables, EEG, and molecular assays, lack uncertainty quantification and rarely establish causal links between biomarkers and predictions. Voice impairments, particularly hypokinetic dysarthria, can occur years before motor symptoms and represent a low-cost, accessible pathway for early detection. This study develops deep learning models for PD detection from voice data, emphasizing interpretability and uncertainty awareness to improve clinical applicability.

METHODS

We used the UCI Parkinson's Telemonitoring dataset with sustained vowel recordings from 42 subjects (37 with Parkinson's disease, 5 healthy). Each voice sample contained 22 acoustic features such as jitter, shimmer, NHR, and other MDVP measures linked to dysarthria.

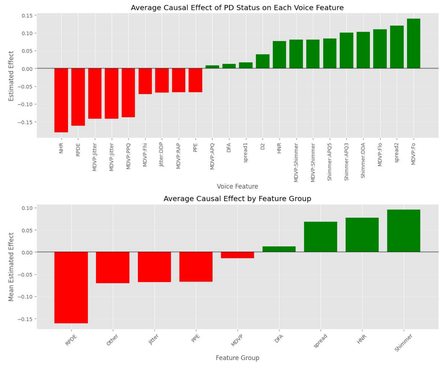
The dataset showed a strong class imbalance (24:76). To address this, we used augmentation techniques including Gaussian noise, feature perturbation, interpolation, scaling, shifting, and random masking, improving the ratio to 46:54 and increasing diversity.

Four models were trained: a Vanilla CNN with two convolutional layers and dropout; an MC-Dropout CNN with stochastic inference for uncertainty estimation; a Few-Shot Learner using prototype-based embeddings; and an Ensemble combining CNN and MC-Dropout outputs.

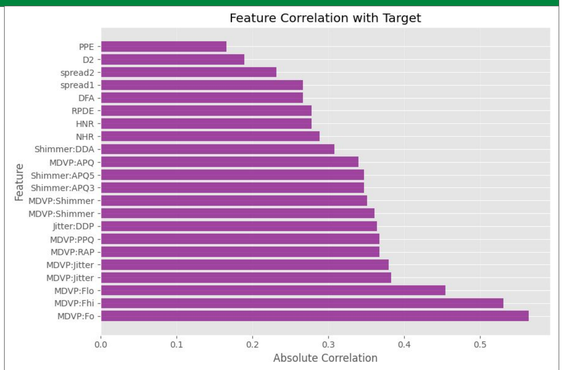
All models used Adam (lr = 1e-3), binary cross-entropy loss, batch size 32, up to 50 epochs, and early stopping. Data were split 72% training, 8% validation, and 20% testing.

For interpretability, causal inference with CausalForestDML (EconML) estimated each feature's Conditional Average Treatment Effect (CATE) on Parkinson's probability, identifying biomarkers like NHR, RPDE, and jitter as key causal drivers.

RESULTS



We used the CausalForestDML estimator to assess each voice feature's causal effect on Parkinson's probability. Noise-to-Harmonics Ratio (NHR), Recurrence Period Density Entropy (RPDE), and MDVP jitter showed the strongest effects, indicating that the models captured biomarkers with true mechanistic links to dysarthria, such as irregular vocal fold vibration and unstable phonatory control.



Feature correlation analysis of the Vanilla CNN showed that NHR, RPDE, and jitter had the strongest correlations with model predictions. The agreement between high correlation and large causal effects supports the physiological relevance of these features as vocal indicators of Parkinson's disease, suggesting the model reflects real clinical mechanisms rather than dataset artifacts.

Model	Accuracy	Precision	Rexall	F1 Score	AUC-ROC
Vanilla CNN	0.9744	1.0000	0.9655	0.9825	0.9897
MC-Dropout CNN	0.8974	0.9630	0.8966	0.9286	0.9655
Ensemble Model	0.9231	0.9643	0.9310	0.9474	0.9862

Performance evaluation demonstrated that the Vanilla CNN achieved the highest accuracy (97.4%) and AUC (0.9897), outperforming traditional shallow classifiers reported in prior literature. The MC-Dropout CNN had slightly reduced accuracy (89.7%) but provided well-calibrated estimates of predictive uncertainty, which is critical for clinical settings where overconfidence must be avoided. The Ensemble model combined the strengths of both approaches, yielding stable performance (92.3% accuracy, AUC = 0.9862). Together, these results show that deep learning models can achieve both high predictive power and clinically valuable uncertainty awareness.

DISCUSSION

This study demonstrates the feasibility of voice-based Parkinson's disease (PD) screening using deep learning models that combine high accuracy with interpretability and uncertainty estimation. A Vanilla CNN achieved 97.4% accuracy (AUC-ROC: 0.9897), outperforming traditional approaches, while an MC-Dropout CNN reached 89.7% accuracy (AUC-ROC: 0.9655) and enabled estimation of epistemic uncertainty for clinical reliability.

Causal inference with the CausalForestDML estimator identified jitter, Noise-to-Harmonics Ratio (NHR), and Recurrence Period Density Entropy (RPDE) as features with the strongest causal influence on PD probability, aligning with known vocal biomarkers.

A few-shot learner performed well in low-data settings, exceeding 66% accuracy and F1 > 0.7 with only three labeled examples per class, showing potential for deployment in data-scarce conditions.

Limitations include the small dataset, possible augmentation bias, and lack of aleatoric uncertainty modeling. Future work should use larger, longitudinal datasets with demographic controls to improve generalizability. Voice-based deep learning models, integrated into telehealth systems, could enable scalable, non-invasive early PD screening and support multimodal biomarker research for personalized neurological care.

LITERATURE REVIEW

Parkinson's disease (PD) is a neurodegenerative disorder affecting more than 10 million people worldwide [1], [2]. Diagnosis is typically based on motor symptoms that appear only after substantial neuronal loss [3]. Recent clinical work emphasizes a prodromal period during which subtle non-motor and speech-related changes emerge, offering opportunities for earlier detection [4], [5].

A wide range of non-invasive approaches have been investigated to identify PD earlier. Wearable sensors and smartphone platforms have demonstrated promise for continuous gait and motor tracking [6]. Broader efforts in digital biomarker development highlight the potential of multimodal tools for precision monitoring [7]. EEG-based models applying time-frequency representations with CNN architectures have also achieved encouraging performance [8]. On the molecular side, exploratory blood-based assays, including RNA profiling and AI-enhanced diagnostic tests, suggest that circulating biomarkers may one day enable pre-symptomatic PD detection [9]. While these technologies expand the diagnostic landscape, most rely on handcrafted features or lack interpretability and uncertainty quantification, limiting their clinical utility.

Voice represents a particularly promising biomarker pathway. Dysphonia is highly prevalent among patients with movement disorders and may precede classic motor findings [11]. Acoustic measures such as jitter, shimmer, and NHR have been validated as sensitive indicators of PD-related speech dysfunction. Early telemonitoring studies established that these features could support disease classification and monitoring, with shallow machine-learning models reporting strong accuracies. However, most of these approaches are limited by small datasets, reliance on correlation-based reasoning, and the inability to provide clinically interpretable confidence estimates.

Advances in machine learning have introduced tools to address these gaps. Monte Carlo Dropout offers a practical Bayesian approximation that allows deep networks to generate well-calibrated uncertainty estimates [10]. Few-shot learning frameworks, particularly prototypical embeddings, support classification in low-data settings typical of medical datasets. Meanwhile, causal inference approaches—such as causal forests and double machine learning—shift feature attribution beyond simple correlation, quantifying the causal effect of individual biomarkers on model predictions.

Taken together, the literature establishes the value of voice-based biomarkers for PD while highlighting key gaps in interpretability, uncertainty, and generalizability. These gaps motivate the present work, which integrates CNNs, MC-Dropout, few-shot learning, and causal inference into a unified framework for voice-based PD detection [1]-[11].

KEY REFERENCES

1. Parkinson's Foundation, "Statistics," 2024. Accessed 2025-04-22.

2. World Health Organization, "Parkinson disease fact sheet." https://www.who.int/news-room/fact-sheets/detail/parkinson-disease, 2019. Accessed 2025-04-22.

3. National Institute of Neurological Disorders and Stroke (NINDS), "Parkinson's disease: Challenges, progress, and promise," NINDS News, 2023.

4. C. Simonet, A. Schrag, A. J. Lees, and A. J. Noyce, "The motor prodromes of parkinson's disease: From bedside observation to large-scale application," Journal of Neurology, vol. 268, pp. 2099-2108, 2019.

5. Ba and Martin, "Missed insights for earlier management of parkinson's disease and dopaminergic imaging," Geriatrics, vol. 9, no. 5, p. 126, 2023.

6. T. Kalpana, R. Thamilselvan, K. Chitra, and T. Thenmalar, "Using a smartwatch and smartphone to assess early parkinson's disease in the watch-pd study," npj Parkinson's Disease, 2023.

7. J. Xu and et al., "Digital biomarkers for precision diagnosis and monitoring in parkinson's disease," npj Digital Medicine, 2024.

8. Siuly and Zhu, "An efficient parkinson's disease detection framework: Leveraging time-frequency representation and alexnet cnn on eeg," Computers in Biology and Medicine, vol. 162, p. 106602, 2024.

9. The Guardian, "Ai-enhanced blood test may detect parkinson's years before onset," 2024. Accessed 2025-04-22.

10. Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," Proceedings of the 33rd International Conference on Machine Learning, pp. 1050–1059, 2016.

11. M. Moya-Mendez, L. L. Madden, I. U. Haq, C. J. McLouth, and M. S. Siddiqui, "Analysis of the prevalence and onset of dysphonia and dysphagia symptoms in movement disorders at an academic medical center," Journal of Clinical Neuroscience, vol. 64, pp. 111-115, 2019.