Towards Causal Interpretability in Deep Learning for Parkinson's Detection from Voice Data

Aniruth Ananthanarayanan[1*†], Sudeep Senivarapu[1*], Anishsairam Murari[1*]

1 Texas Academy of Mathematics and Science, University of North Texas, Denton TX

† Correspondence: aniruth2207@gmail.com

Abstract

This research introduces a comprehensive framework for Parkinson's Disease (PD) detection using voice recording data. We implemented and evaluated multiple deep learning models, including a baseline Convolutional Neural Network (CNN), an uncertainty-aware Monte Carlo-Dropout CNN (MCD-CNN), as well as a few-shot learning approach to address dataset size limitations. Our models achieved an accuracy over 90% in classifying PD patients using vocal biomarkers, with the ensemble model demonstrating the highest performance. We employed data augmentation techniques to address class imbalance and enhance generalization. Causal feature analysis revealed that the Noise-to-Harmonics Ratio (NHR), Recurrence Period Density Entropy (RPDE), and MDVP jitter parameters were among the most significant vocal biomarkers for PD detection, in order of estimated effect magnitude. Across deep learning models, features exhibiting the strongest absolute correlation with outputs consistently showed the largest estimated effect magnitudes. The few-shot learning approach showed promising results as well, even with limited training examples. This work demonstrates the use of causal feature analysis to validate the analysis of deep learning models, potentially enabling accessible and interpretable non-invasive screening tools.

Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects more than 10 million people worldwide and can impose significant clinical and economic burdens [1, 2]. Clinical diagnosis typically relies on motor symptoms such as tremor, bradykinesia, rigidity, and postural instability that emerge only after substantial dopaminergic neuron loss [3, 4]. Although modalities such as DaT-SPECT imaging, wearable movement sensors, smartphone-based digital biomarkers, EEG screening, and blood-based molecular assays offer noninvasive or minimally invasive early detection options, they depend on handcrafted features, lack uncertainty quantification, and do not establish causal links between biomarkers and predictions [5–9]. However, hypokinetic dysarthria often manifests up to five years before motor signs, positioning voice recordings as a low-cost and widely accessible screening avenue [4].

This research introduces a comprehensive framework for PD detection from sustained vowel recordings by implementing and evaluating three deep learning models: a baseline Convolutional Neural Network (CNN), an uncertainty-aware Monte Carlo-Dropout CNN (MCD-CNN), and a few-shot learning variant to mitigate the limited amount of data [10, 11]. Our ensemble surpasses 90% accuracy in distinguishing

*Authors contributed equally to this work

PD from healthy voices, leveraging data augmentation to correct class imbalance and improve generalization. Critically, we apply causal feature analysis to quantify the effect sizes of vocal biomarkers, identifying Noise-to-Harmonics Ratio (NHR), Recurrence Period Density Entropy (RPDE), and MDVP jitter as the most influential predictors [12]. Features exhibiting the highest absolute correlations also demonstrate the largest estimated causal effects across models, confirming their mechanistic relevance.

Shallow classifiers using handcrafted acoustic features (jitter, shimmer, MFCCs) have achieved up to 95% accuracy in voice-based PD detection but remain correlational and opaque [13]. Wearable inertial sensors and gait-analysis platforms reliably quantify gait abnormalities for prodromal motor anomalies [6, 14], while smartphone and smartwatch digital biomarker systems facilitate continuous remote monitoring of both motor and non-motor symptoms [7]. EEG-based screening demonstrates high specificity in early PD neural signatures [8], and blood-based tRNA fragment assays yield AUCs (area under the receiver operating characteristic curve, a measure of classification performance where 1.0 indicates perfect accuracy) around 0.86 for molecular detection [9]. None of these approaches, however, couple deep learning performance with rigorous uncertainty quantification and causal validation of feature importance, which are gaps the new framework addresses.

Materials and Methods

Dataset.

This study utilizes the Parkinson's Telemonitoring dataset from the UCI Machine Learning Repository [15, 16]. The dataset contains voice measurements from 42 individuals, 5 of whom are healthy and 37 diagnosed with Parkinson's disease. These voice recordings are characterized by a range of vocal features such as fundamental frequency (jitter), amplitude variation (shimmer), noise-to-harmonics ratio (NHR), and other signal processing metrics commonly known to be indicative of Parkinsonian speech impairments. Each set of signal processing metrics for each voice sample is associated with a label that indicates the individual's Parkinson's diagnosis, where 0 represents healthy and 1 represents a patient diagnosed with Parkinson's. To enhance the quality and diversity of the training data, a range of data augmentation techniques was applied with the goal of balancing class distributions and expanding the dataset. These techniques included adding Gaussian noise to simulate measurement variability, perturbing a subset of features to mimic natural fluctuations, interpolating between samples of the same class to generate intermediate examples, applying global scaling and shifting to introduce broader variability, and randomly masking features to simulate missing or occluded data. Together, these strategies produced a more balanced and representative dataset that can help improve model generalization and reduce bias toward overrepresented classes.

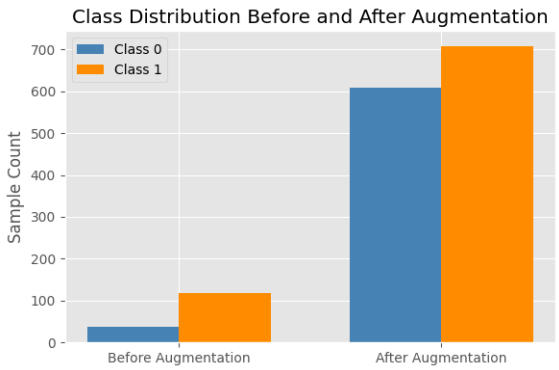


Figure 1. Visualization of class imbalance in UCI ML Repository Parkinson's Telemonitoring dataset. We performed data augmentation to address class imbalance and limited dataset size. However, we did not even out the classes completely as it would have limited the maximum dataset size without diluting the density of the unaugmented samples. Our initial class ratio was 24:76, and after augmentation, it improved to 46:54.

Model Architectures.

We designed three primary deep-learning architectures for voice-based PD classification:

1. **Vanilla CNN.** A one-dimensional convolutional neural network that treats each feature as a signal. The input (D = 22 features) is reshaped to (1 x D) and passed through two convolutional layers (4 and 8 filters respectively, kernel size = 3, padding = 1), each followed by ReLU and 20% dropout layer. The flattened representation then feeds a fully connected layer (16 units + ReLU + 20% dropout) and a sigmoid output neuron for binary classification.

2. **MC-Dropout CNN.** Identical to the Vanilla CNN, but retains dropout at inference ("Monte Carlo Dropout") to capture model uncertainty. At test time, we perform 50 stochastic forward passes and utilize the set of predicted classifications to yield a mean prediction and an estimate of epistemic uncertainty derived from variance.

3. **Few-Shot Learner.** Builds on the base CNN by adding a prototype embedding layer (8-dim) after the convolutional blocks. For a small "support set" of k examples per class, we extract embeddings and store their labels. At query time, we compute Euclidean distances between query embeddings and support embeddings and produce a weighted average prediction. This prototypical nearest-neighbor approach mitigates data scarcity.

4. **Ensemble Model.** Averages the Vanilla CNN and MC-Dropout CNN predictions to further stabilize performance, taking "the best of both worlds."

Training.

All models were trained using the same optimization settings: binary cross-entropy loss, the Adam optimizer with an initial learning rate of 1e-3, and a ReduceLROnPlateau scheduler (factor 0.5, patience of 5). Early stopping was applied with a patience of 10 epochs based on validation loss, restoring the best weights. Training was conducted with a batch size of 32 for up to 50 epochs. Throughout training, we monitored training and validation loss, validation accuracy, and AUC-ROC at each epoch, with learning rate adjustments and early stopping helping to prevent overfitting and ensure consistent convergence across models. We split the dataset into 80% training and 20% testing. Within the training set, we reserved 10% for validation. Thus, the final data split was 72% training, 8% validation, and 20% testing. All models were trained on the training set, tuned using the validation set (e.g., for early stopping and learning rate scheduling), and evaluated only on the held-out test set.

Causal Analysis.

In order to validate that our models were learning beyond correlational feature-importance, we employed Double Machine learning via the CausalForestDML estimator from the EconML library [17]. Each vocal feature X_j was treated as a "treatment", while all other features were controls. We fit two random forest regressor learners, one of the outcome model and one of the treatment model, and then estimate the conditional average treatment effect (CATE) of pertrubing X_j on the PD probability, which was then averaged to get an estimated effect size per feature. After quantifying the causal impact for each biomarker, we also took the average of biomarker groups like MDVP to get a mean estimated effect for each group. This approach quantified the causal impact of each marker, accounting for confounding among features.

Results

The Vanilla CNN achieved the highest raw accuracy (97.4%) and AUC-ROC (0.9897), while the MC-Dropout CNN provided uncertainty-aware predictions with slightly lower accuracy (89.7%, AUC-ROC = 0.9655). The ensemble balanced these, yielding 92.3% accuracy and AUC-ROC = 0.9862.

Table 1. Model comparison. Performance comparison of Vanilla CNN, MC-Dropout CNN, and Ensemble Model on training dataset.

Model	Accuracy	Precision	Recall	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

Causal feature ranking identified Noise-to-Harmonics Ratio (NHR), Recurrence Period Density Entropy (RPDE), and MDVP jitter as the top three drivers of model predictions, consistent with their high correlation and large estimated causal effects.

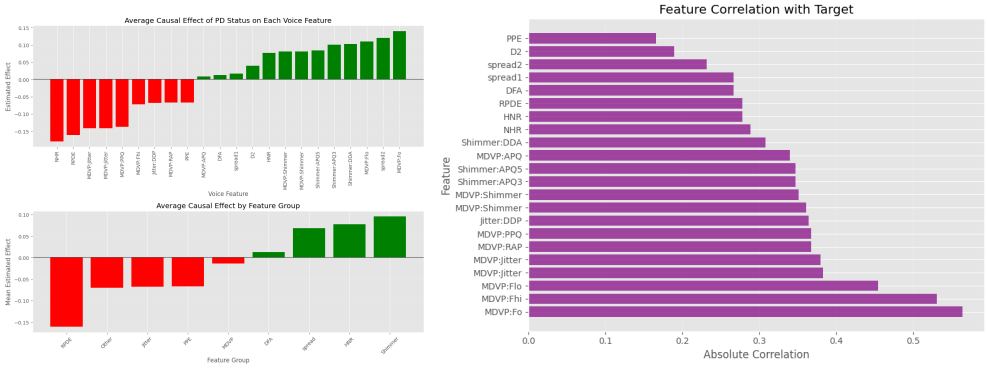


Figure 2. Causation and Correlation. Left: Visualization of estimated effect of various biomarkers and biomarker groups quantified through the CausalForestDML approach. Right: Feature correlation values for Vanilla CNN.

The few-shot learner, even with as few as three examples per class (which sums up to a total of six labeled examples), attained mean accuracy $\approx 66\%$ and F1 > 0.7 on 1,000 runs, outperforming a random forest baseline demonstrating robustness under extreme data scarcity.

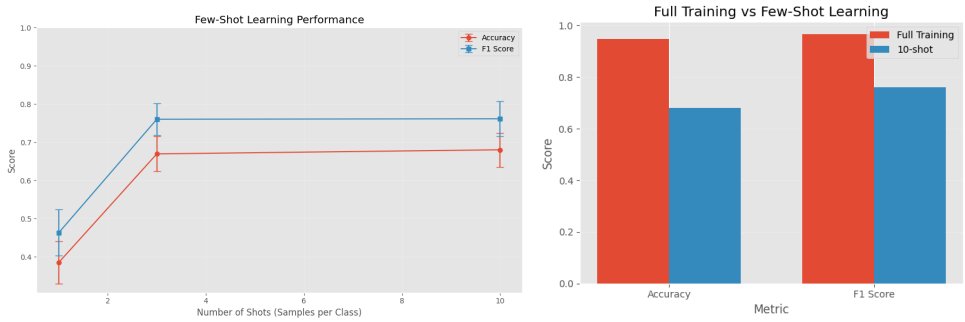


Figure 3. Few-shot learning performance and comparison to full-data training. Left: Mean accuracy (red) and F1 score (blue) $\pm$ 1 SD of the prototypical few-shot learner as a function of the number of support examples per class. Right: Bar chart contrasting the full-data CNN against the 10-shot few-shot model on test accuracy and F1 score.

Discussion

The comprehensive benchmarking of three deep learning architectures for Parkinson's disease diagnosis based on voice analysis is shown to achieve both high predictability and high interpretability. Vanilla CNN exhibited accuracy of 97.4% (AUC-ROC = 0.9897), beating conventional shallow classifiers based on hand-engineered acoustic features [13]. Although MC-Dropout CNN showed somewhat reduced raw accuracy (89.7%, AUC-ROC = 0.9655), it gave well-calibrated estimates of uncertainty, thereby resolving a key essential for clinical decision support [10].

CausalForestDML-based analysis confirmed that Noise-to-Harmonics Ratio, Recurrence Period Density Entropy, and MDVP jitter exert the largest estimated causal effects on PD probability, in line with their strong correlation with model outputs. MDVP jitter, NHR, and RPDE aren't just correlated—they likely reflect actual disease mechanisms: jitter shows unstable vocal fold movement, NHR indicates glottal closure issues, and RPDE captures irregular voice patterns [18, 19]. All three align with known pathophysiology in PD, like bradykinesia and dysphonia [20–22]. This intervention-aware approach goes beyond associational attributions, ascertaining that these vocal biomarkers do reflect underlying pathophysiology and not dataset artifacts. Agreement between correlation strength and causal effect magnitude across models can also enhance belief in mechanistic interpretations.

The few-shot learner performed about 66% accuracy and an F1 value of over 0.7 with just three examples per class, far surpassing random baseline results and highlighting its promise in scenarios with limited labeled data. Such data-efficient models could be vital if developing large-scale labeled voice corpora proves difficult.

Nevertheless, our study is limited by reliance on the UCI Telemonitoring dataset's 42 subjects [15, 16]. Although extensive data augmentation was applied, synthetic variability cannot fully substitute for diverse, real-world recordings. Future work should extend this framework to larger, multi-center cohorts and utilize longitudinal voice samples to ascertain robustness over time. Additionally, causal effect estimation assumes no unobserved confounders; the addition of demographic and recording-device covariates may provide still stronger causal inferences [17].

Looking ahead, incorporation into mobile or telehealth platforms would enable the deployment of scalable, non-invasive PD screening at large scale, with uncertainty values guiding appropriate clinical referral. Extending causal interpretability to multimodal inputs such as inertial gait sensors [6, 14], smartphone digital biomarkers [7], EEG signatures [8], or molecular assays [9] provides a more detailed diagnostic landscape and more robust early-detection hardware. Additionally, doctors mainly use motor symptoms like tremor and rigidity for diagnosis. They don't typically rely on voice features, though dysphonia is recognized. This model shows voice-based markers can detect PD even earlier and more precisely. If clinicians adopt these data-driven features, they could catch cases earlier, especially in remote settings, and improve understanding of how PD affects speech.

Data Availability

The UCI Machine Learning Repository Oxford Parkinson's Disease Detection Dataset is accessible at https://archive.ics.uci.edu/dataset/174/parkinsons [15, 16].

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*, 268:2099–2108, 2019.

5. Ba and Martin. Missed insights for earlier management of parkinson's disease and dopaminergic imaging. *Geriatrics*, 9(5):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*, 162:106602, 2024.

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

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

11. Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

12. Judea Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

13. Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, 2010.

14. Lazzaro di Biase, Alessandro Di Santo, Maria Letizia Caminiti, Alfredo De Liso, Syed Ahmar Shah, Lorenzo Ricci, and Vincenzo Di Lazzaro. Gait analysis in parkinson's disease: An overview of the most accurate markers for diagnosis and symptoms monitoring. *Sensors*, 20(12):3529, 2020.

15. Max Little. Parkinsons. UCI Machine Learning Repository, 2007. DOI: https://doi.org/10.24432/C59C74.

16. Max A. Little *, Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, and Lorraine O. Ramig. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022, 2009.

17. Maggie Hei Greg Lewis Paul Oka Miruna Oprescu Vasilis Syrgkanis Keith Battocchi, Eleanor Dillon. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML, 2019. Version 0.x.

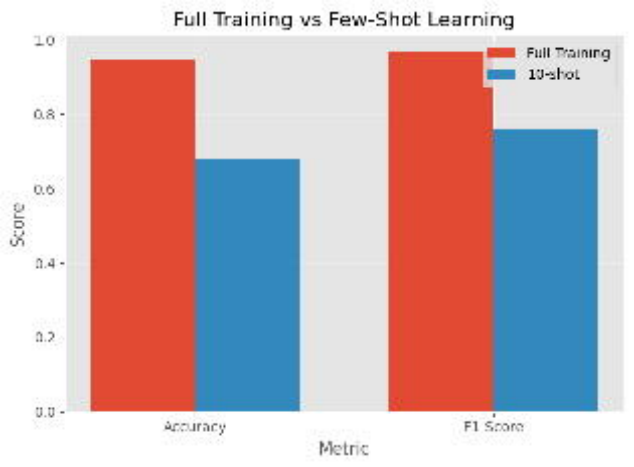
18. Shayna Zelcer, Chantal Henri, Ted L Tewfik, and Bruce Mazer. Multidimensional voice program analysis (MDVP) and the diagnosis of pediatric vocal cord dysfunction. *Ann. Allergy Asthma Immunol.*, 88(6):601–608, June 2002.

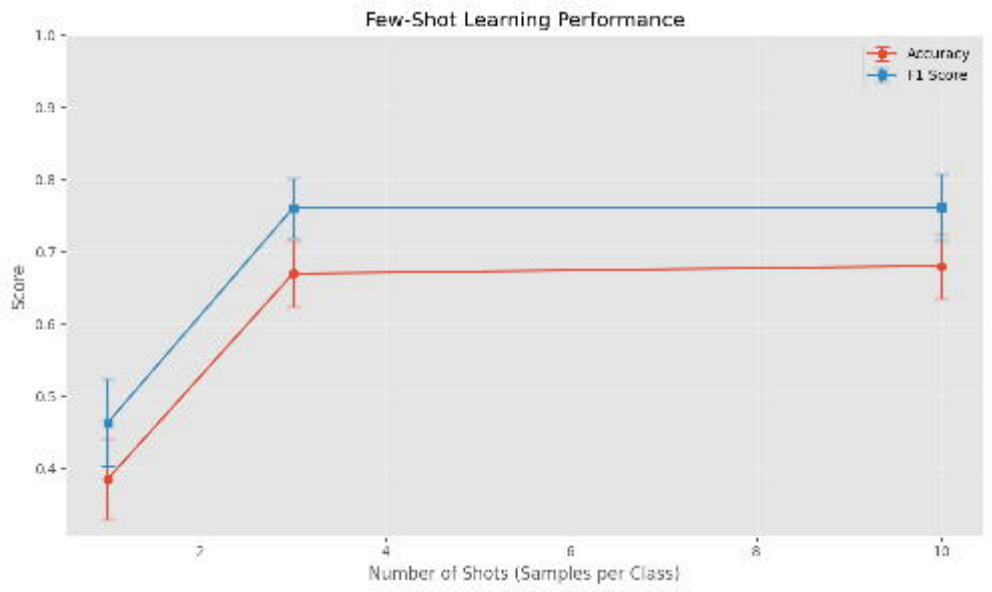
19. João Paulo Teixeira, Carla Oliveira, and Carla Lopes. Vocal acoustic analysis – jitter, shimmer and hnr parameters. *Procedia Technology*, 9:1112–1122, 2013. CENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies.

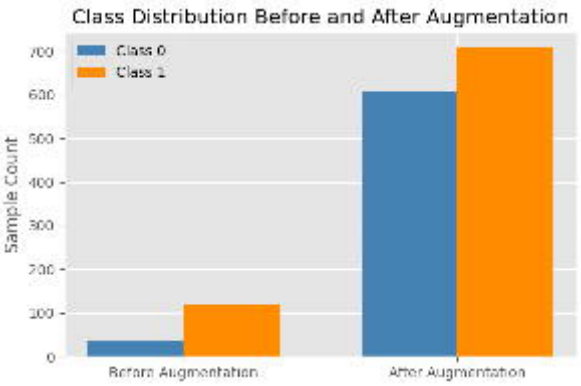
20. Fangyuan Cao, Adam P. Vogel, Puya Gharahkhani, and Miguel E. Renteria. Speech and language biomarkers for parkinson's disease prediction, early diagnosis and progression. *npj Parkinson's Disease*, 11(1):57, Mar 2025.

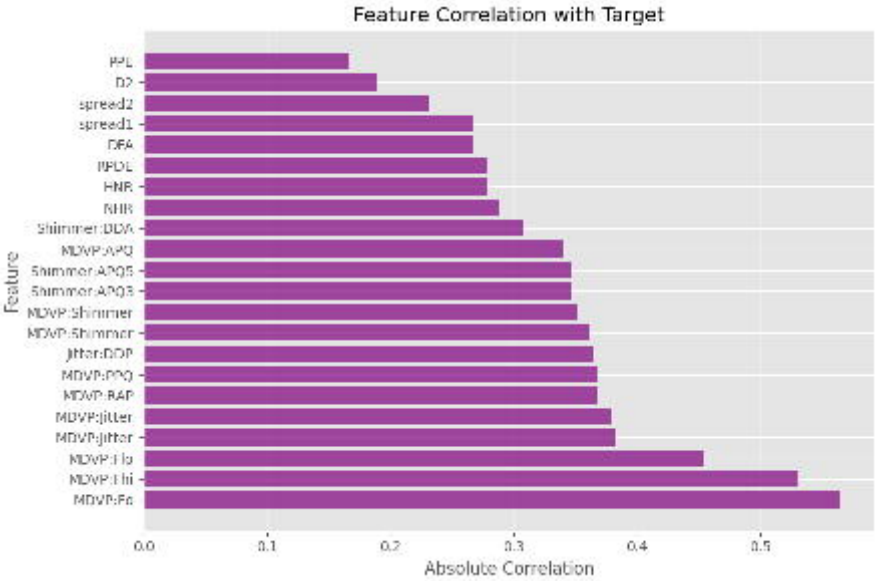
21. Andrew Ma, Kenneth K Lau, and Dominic Thyagarajan. Voice changes in parkinson's disease: What are they telling us? *J. Clin. Neurosci.*, 72:1–7, February 2020.

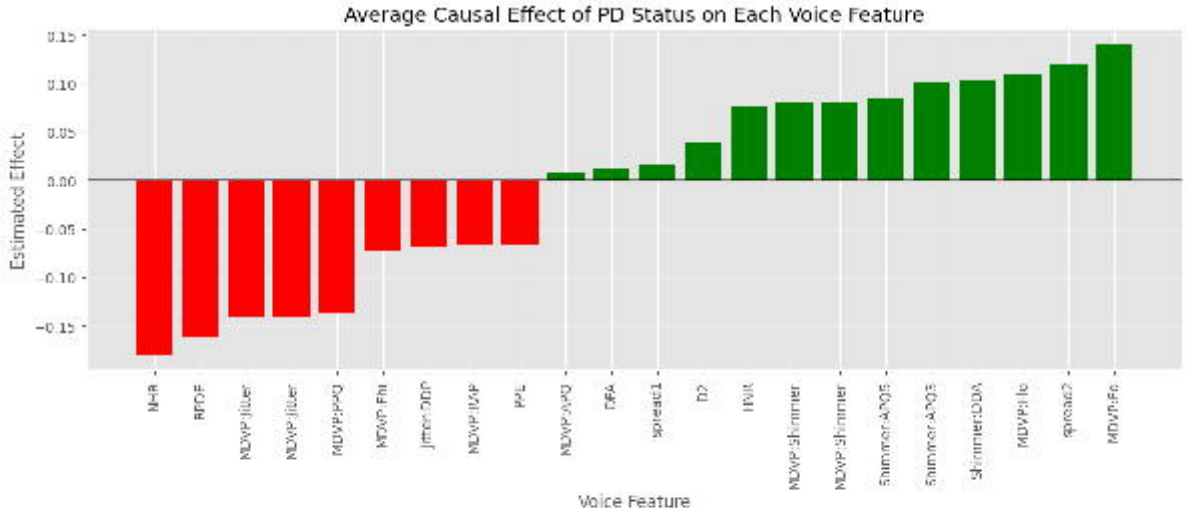
22. Mary Moya-Mendez, Lyndsay L. Madden, Ihtsham U. Haq, Christopher J. McLouth, and Mustafa S. Siddiqui. Analysis of the prevalence and onset of dysphonia and dysphagia symptoms in movement disorders at an academic medical center. *Journal of Clinical Neuroscience*, 64:111–115, 2019.

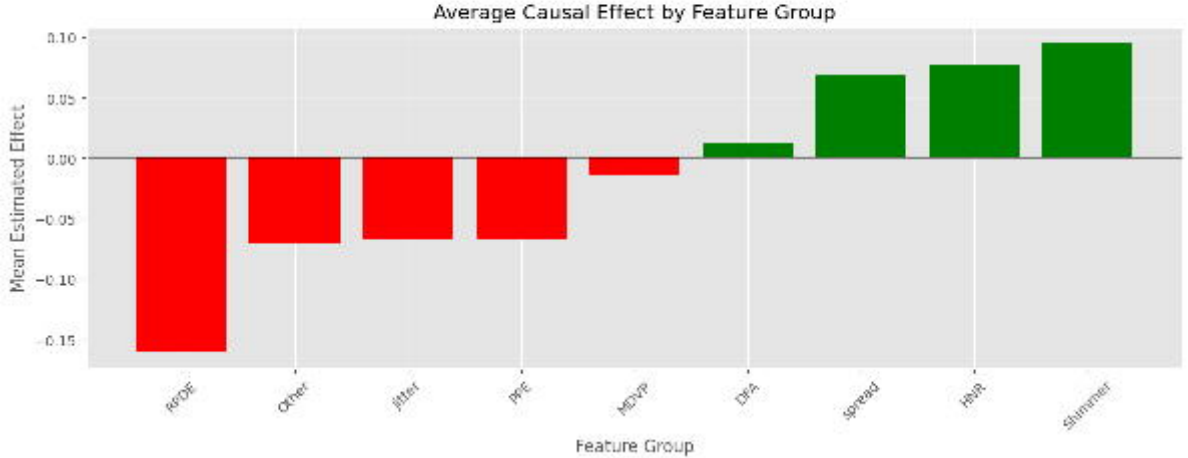

Full Training vs Few-Shot Learning
Full Training
10-shot
Score
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
F1 Score
Metric


Few-Shot Learning Performance
Accuracy
F1 Score
Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
Number of Shots (Samples per Class)


Class Distribution Before and After Augmentation
Class 0
Class 1
Sample Count
700
600
500
400
300
200
100
0
Before Augmentation
After Augmentation


Feature Correlation with Target
PPE
D2
spread2
spread1
DFA
RPDE
HNR
NHR
Shimmer:DDA
MDVP:APQ
Shimmer:APQ5
Shimmer:APQ3
MDVP:Shimmer
MDVP:Shimmer
Jitter:DDP
MDVP:PPQ
MDVP:RAP
MDVP:Jitter
MDVP:Jitter
MDVP:Flo
MDVP:Fhi
MDVP:Fo
Feature
0.0
0.1
0.2
0.3
0.4
0.5
Absolute Correlation


Average Causal Effect of PD Status on Each Voice Feature
Estimated Effect
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
NHR
RPDE
MDVP:Jitter
MDVP:Jitter
MDVP:PPQ
MDVP:Fhi
Jitter:DDP
MDVP:RAP
PPE
MDVP:APQ
DFA
spread1
D2
HNR
MDVP:Shimmer
MDVP:Shimmer
Shimmer:APQ5
Shimmer:APQ3
Shimmer:DDA
MDVP:Flo
spread2
MDVP:Fo
Voice Feature


Average Causal Effect by Feature Group
Mean Estimated Effect
0.10
0.05
0.00
−0.05
−0.10
−0.15
RPDE
Other
jitter
PPE
MDVP
DFA
spread
HNR
Shimmer
Feature Group