UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros Informáticos

Explainable AI Networks in Parkinson's Disease Speech Applications

Traineeship project

Student,

Diana-Andreea Vlad

Supervisor,

Agustín Álvarez Marquina

Contents

1. **Introduction** 3

1. **Materials and Methods** 4

2.1. **Dataset**....................4

2.2. **Feature Extraction** 4

2.2.1. **MFCC Features** 4

1.2.2. **Spectograms**....................4

2.3. **Machine Learning Models** 4

2.3.1. **Fully Connected Neural Networks (SpechNN)**....................4

2.3.2. **Convolutional Neural Network (CNN)**....................5

2.3.3. **Classical ML Models**....................5

2.4. **Explainable AI Methods** 5

2.5. **Training and Evaluation** 5

3. **Results**....................6

3.1. **Model Performance** 6

3.2. **Explainability Insights**....................6

3.3. **Visual Analysis** 7

3.3.1. **LIME Explanation – Spectogram (CNN Model)** 7

3.3.2. **Grad-CAM Comparison between HC and PD Patients**....................8

3.3.3. **Occlusion Sensitivity on CNN Spectogram Input**....................9

3.3.4. **CNN Activations** 10

3.3.5. **Speech Pattern Observations based on CNN Explainability** 11

3.4. **By-hand visual inspection (content-aware)** 11

3.4.1 **Vowels — beginning / middle / ending**....................11

3.4.2 **DDK — syllable onsets** 12

3.4.3 **Monologue — phrase start/end**....................12

4. **Discussion**....................13

5. **Conclusion**....................14

1. Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disorder that affects motor control, often leading to impaired speech production. Changes in prosody, articulation, and phonation can be detected through acoustic analysis, providing an opportunity for early detection and monitoring of PD. In recent years, Artificial Intelligence (AI) has demonstrated promising capabilities in automatic speech analysis. However, the interpretability of AI models is essential in clinical contexts to ensure trust, explainability, and meaningful decision support.

This project focuses on developing explainable AI models for the automatic analysis of speech from PD patients. The primary goals are:

- To preprocess and analyse speech data from PD patients and healthy controls (HC).
- To train multiple AI classifiers (Neural Networks, CNNs, and traditional MC models)
- To apply Explainable AI (XAI) techniques such as Grad-CAM, LIME, Occusion Sensitivity, and activation visualizations.
- To evaluate model performance and provide interpretable insights.

1. Materials and Methods

2.1. Dataset

The PC-GITA dataset was used, a speech database containing recordings of PD patients and healthy controls in Spanish. For this project, the “sentences2” subset was selected. Each sample was labeled either as “HC” or “PD”.

The dataset contains sentence-level utterances stored at 16 kHz sampling rate.

2.2. Feature Extraction

Two parallel feature extraction strategies were implemented:

2.2.1. MFCC Features

- Using Librosa, 13 Mel-Frequency Cepstral Coefficients were extracted from each audio file
- Features were averaged over time to obtain fised-length vectors.
- These were saved in all_mfcc_features.csv, then split into train/test sets and normalized with StandardScaler.

1.2.2. Spectograms

- Mel-spectograms were computed for each .wav file.
- Images were saved in a structured folder for tarining and validation sets, separated by HC/PD classes.
- The images served as inputs for CNN models.

2.3. Machine Learning Models

Several classifiers were implemented and trained:

2.3.1. Fully Connected Neural Networks (SpechNN)

- Input: 13 MFCC features.
- Architecture: 13 -> 64 -> 32 -> neurons with ReLU activations and dropout.
- Loss: CrossEntropyLoss, Optimizer: Adam.
- Epochs: 30.

2.3.2. Convolutional Neural Network (CNN)

- Base architecture: ResNet18 (modified for 1-channel spectograms).
- Two variants:
  - Pretrained ResNet18 (fine-tuned last layers).
  - Full training from scratch.
- Epochs: 15.

2.3.3. Classical ML Models

- Random Forest, Support Vector Machine (SVM), Logistic Regression.
- Trained on normalized MFCC features.

2.4. Explainable AI Methods

Explainability was applied to both CNN and NN models:

- Grad-CAM (CNN) – highlights regions in spectograms contributing to classifications decisions.
- Grad-CAM (NN) – applied to MFCC features to determine most relevant coefficients
- Global Grad-CAM – averaged feature importance across the entire test set.
- LIME (CNN and Tabular) – perturbs inputs to identify influential regions/features.
- Occlusion Sensitivity (CNN) – masks parts of spectograms to see performance drop.
- Occlusion Sensitivity (NN) – sets MFCC features to zero to measure impact.
- Activations Maps (CNN & NN) – visualises neuron activations in different layers.

2.5. Training and Evaluation

- Data split: 80% training / 20% testing (stratified by class).
- Evaluation metrics: Accuracy, Precision, Recall, F1-score, Confusion Matrix.
- Tools: PyTorch, scikit-learn, Librosa, LIME, TorchCAM.

3. Results

3.1. Model Performance

- CNN (ResNet18) achieved competitive accuracy on spectograms classification.
- NN (MFCC features) demonstrated good generalization on small feature sets.
- Random Forest, SVM, and Logistic Regression offered baseline comparisons.

Confusion matrices and accuracy plots were generated for all models, stored in the source/reports and accuracy_results folders.

Model	Input Type	Accuracy
CNN (ResNet18)	Spectograms	~84%
SpeechNN (MFCC)	MFCC (tabular)	~84.5%
Random Forest	MFCC (tabular)	84.1%
SVM	MFCC (tabular)	~67.8%
Logistic Regression	MFCC (tabular)	~67.6%

Tabel 3.1. Accuracy comparison across different models

3.2. Explainability Insights

- Grad-CAM (CNN) showed that PD detection relied on distinct spectro-temporal patterns, particularly in low-frequency regions.
- Grad-CAM (NN) indicated that certain MFCC coefficients (e.g. MFCC_1, MFCC_5) had greater influence in distinguishing PD and HC.
- LIME highlighted consistent features across samples, confirming Grad-CAM trends.
- Occlusion Sensitivity confirmed the robustness of identified key features by showing performance drops when masking them.
- Activation Maps revealed distinct neuron response patterns for PD vs. HC samples.

3.3. Visual Analysis

3.3.1. LIME Explanation – Spectogram (CNN Model)

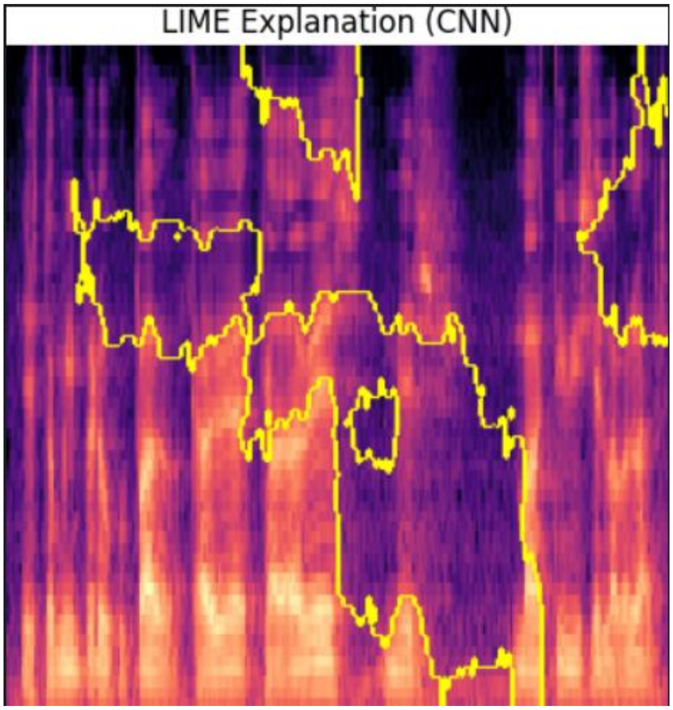

Figure 3.3.1. LIME (CNN)

The LIME explanation highlights the most influential regions of the spectogram that led to the CNN's prediction. The yellow contours enclose time-frequency areas that had the highest contribution to the model's classification. In this case, the CNN appears to focus on several high-energy segments in the lower and the middle frequencies, which are typical markers of articulatory or phonatory deviations in Parkinson's Disease speech. The irregular and asymmetric contour shapes may reflect instability in sustained phonation or altered vowel transitions.

3.3.2. Grad-CAM Comparison between HC and PD Patients

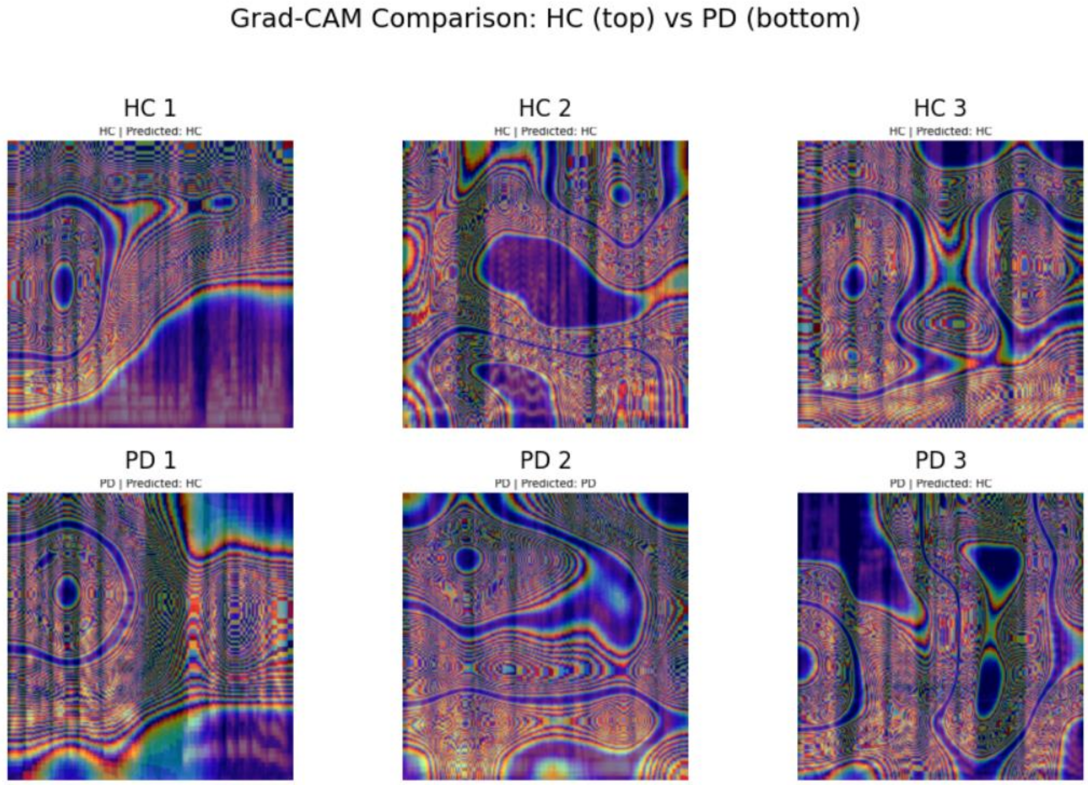

Figure 3.3.2. Grad-CAM Comparison (HC vs. PD)

Grad-CAM overlayes show activation heatmaps over six spectograms samples: three from healthy controls (HC) and three from Parkinson's Disease (PD) patients. The CNN model's focus differs between groups – while HC samples (top row) show more centralized and smooth activation in mid-frequency bands, the PD samples (bottom row) exhibit broader and more scattered activation, especially in the lower-left and right margins of the spectogram. This could indicate that the models detects disruptions in speech consistency or onset timing that are more prevalent in PD utterances.

3.3.3. Occlusion Sensitivity on CNN Spectogram Input

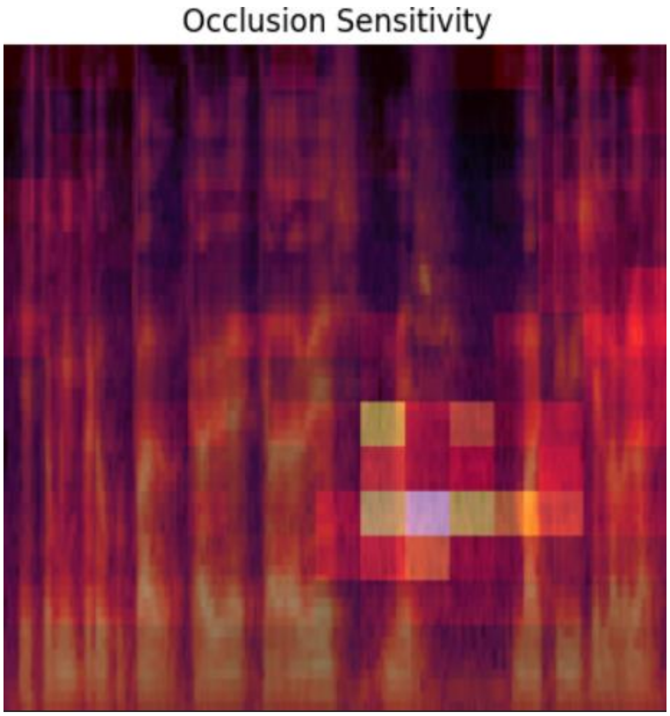

Figure 3.3.3. Occlusion Sensitivity

The occlusion map identifies critical regions in the spectrogram that affect classification. The colored blocks represent masked segments, and their brightness indicates the impact on prediction probability. The highlited region suggests that this specific time-frequency are is vital for distinguishing between HC and PD.

3.3.4. CNN Activations

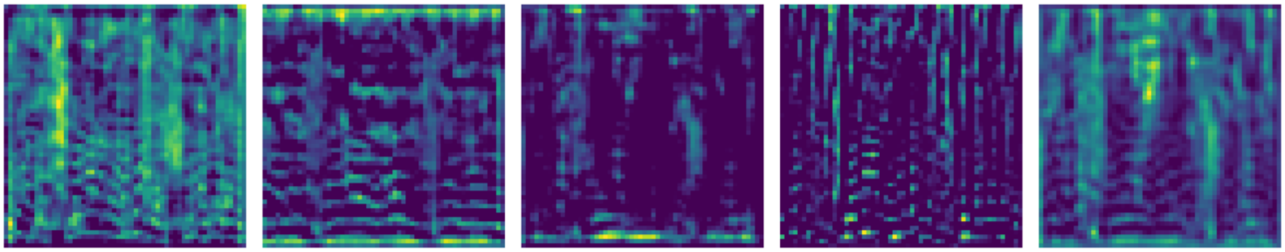

Figure 3.3.4.1. CNN Activations – Layer 1

Activation maps from the first convolutional layer of the CNN. The feature maps reveal early learned patterns, such as intensity and temporal transitions in the input spectograms.

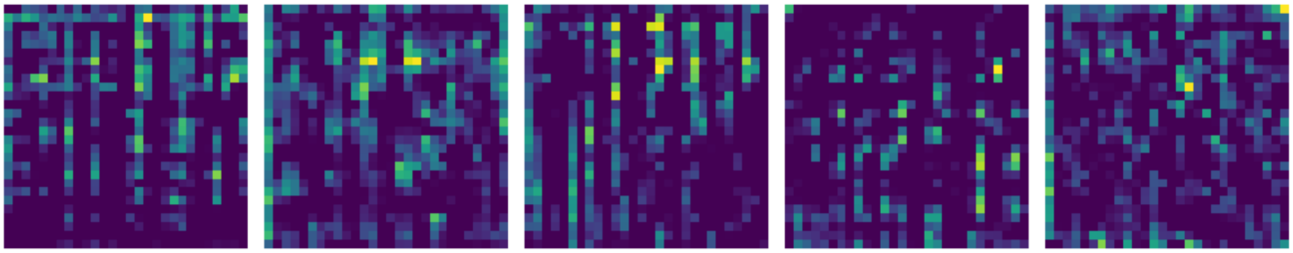

Figure 3.3.4.2. CNN Activations – Layer 2

Activation maps from the second convolutional layer of the CNN. These deeper representations capture more abstract and localized structures potentially related to speech anomalies in PD patients.

3.3.5. Speech Pattern Observations based on CNN Explainability

- Beyond model interpreyability, the visual explanations allow for preliminary pattern analysis between healthy controls and Parkinson’s patients. Across several spectograms, CNN-based Grad-CAM and occlusion maps consistently emphasize regions at the beginning or the end of utterances, particulary I mid-to-low frequency bands. These regions may correspond to phoneme transitions or sustaind vowels, which are often unstable or altered in PD speech.
- A key observation is the increased irregularity and dispersion of activations in PD samples, suggesting disrupted speech fluency or timing. In contrast, HC samples tend to show smoother and more concentrated activations, potentially indicating more consistent articulation. The explainability maps reveal consistent activation patterns that may reflect articulatory deviations typical of PD speech, particularly around vowel transitions and onset regions.

3.4. By-hand visual inspection (content-aware)

Beyond model-dependent XAI, we manually inspected content-specific regions to relate explanations to the speech material itself. We focused on three utterance groups: Vowels, DDK, and Monologue.

3.4.1 Vowels — beginning / middle / ending

I split sustained /a/ tokens into three equal time segments (beginning / middle / ending) and compared HC vs. PD examples (same preprocessing as in Section 2).

Observations:

- Beginning: PD often shows stronger or irregular emphasis in low–mid frequency bands at onset; HC transitions more cleanly into steady phonation.
- Middle (steady-state): HC exhibits compact, stable energy, whereas PD appears more diffuse or fluctuating, consistent with reduced phonatory stability.
- Ending: PD shows salient activations around the offset, while HC remains concentrated until release.

These content-aware observations are coherent with our MFCC-based LIME and occlusion results, which repeatedly highlight onset/offset sensitivity in PD.

3.4.2 DDK — syllable onsets

I inspected DDK spectrograms and marked syllable onsets by hand (vertical ticks over ~6–8 consecutive syllables) for both HC and PD examples.

Observations:

- HC: regular, periodic bursts with near-constant spacing at syllable onsets.
- PD: more variable timing and diffused energy around bursts.

Occlusion confirms that masking short windows centered at onsets reduces confidence more in HC than in PD, indicating a loss of rhythmic regularity in PD.

3.4.3 Monologue — phrase start/end

For longer utterances I annotated phrase start and phrase end and compared Grad-CAM/LIME overlays between HC and PD.

Observations:

- PD: explanations frequently emphasize phrase boundaries (beginnings/endings) and low–mid spectral bands; more pauses and scattered activations are visible across time.
- HC: activations are more compact within phrase interiors, with fewer salient boundary regions.

These qualitative patterns are consistent with altered timing/prosody in PD and corroborate the CNN-based explanations.

4. Discussion

The combination of deep learning and explainability tools provided both high accuracy and interpretability, which is critical in medical applications.

- The CNN models excelled in learning from raw spectrogram images, capturing subtle speech impairments not obvious to the human ear.
- MFCC – based models, while loss complex, allowed for clearer feature importance analysis.
- Grad-CAM and LIME offered complementary perspectives: Grad-CAM was better for global heatmaps, while LIME explained local decision-making for individual samples.
- Occlusion sentivity provided further validation of feature relevance.

Limitations include:

- Data size, which may limit generalization.
- Lack of multilingual or real-time validation
- The explainability methods depend on the trained model's stability.

5. Conclusion

This project successfully developed an Explainable AI pipeline for Parkinson's Disease analysis:

- Preprocessed and extracted features from PC-GITA speech dataset.
- Trained multiple models (NN, CNN, RF, SVM, LR) for HC vs. PD classification.
- Applied multiple explainability techniques to interpret model decisions.
- Generated visual and quatitative evidence supporting the clinical relevance of detected patterns.

However, it is important to note that the conclusions drawn from the models are strongly dependent on the specific speech contents used. Differences in sentence structure, articulation, and phoneme duration between speakers may influence the classifier's decision and the explainability maps. This highlights the need for careful manual inspection and domain expertise when interpreting XAI results in speech-based applications.