Exploring Techniques for the Classification of Early On-Set Parkinson's Disease

Sudeepthi Rebbalapalli
Indiana University Bloomington
surebbal@iu.edu

December 20, 2024

Abstract

Parkinson's disease (PD) poses significant challenges for early detection, particularly in early-onset PD, which often presents with subtle symptoms. This project implements an end-to-end deep learning model that leverages convolutional neural networks (CNNs) with Mel spectrograms to extract features and classify early-onset PD from speech data, utilizing established research methodologies. We employ an Italian dataset sourced from the Internet and incorporate modified input preprocessing techniques to enhance the model's performance. In addition, we explore NeuroSpeech, a software tool designed to assist medical professionals in evaluating biomarkers of neurodegenerative diseases through speech analysis, as well as the Librosa library for effective audio processing.

Although our findings are promising, they also reveal the limitations of deep learning in this context. Specifically, cross-language classification is ineffective without appropriately balanced training data for each language, underscoring the need for language-specific models or alternative approaches to address multi-language classification challenges. Furthermore, the complexity of deep learning models raises interpretability concerns, which may affect trust among doctors and patients in diagnostic applications. Addressing these issues with better, more diverse data, improved model interpretability, and alternative tools can lead to reliable and accessible diagnostic solutions across languages.

1 Introduction

1.1 Parkinson's Disease and the Role of Speech Analysis

Parkinson's disease (PD) is a progressive neurodegenerative disorder that significantly affects motor and non-motor functions. Early-onset PD, characterized by symptoms appearing before the age of 50, presents unique challenges due to its subtle and often overlooked symptoms. Detecting these early signs is critical for timely intervention and better patient outcomes. Among various non-invasive biomarkers, speech has emerged as a promising indicator of PD. Approximately 90% of people with PD develop dysarthria during the disease's progression and changes in vocal articulation, prosody, and other acoustic features are often early signs of PD [Tjaden, 2008], making speech analysis a valuable tool for diagnosis.

1.2 Deep Learning in Parkinson's Speech Detection

End-to-end deep learning approaches, particularly convolutional neural networks (CNNs), have shown significant potential for feature extraction and classification tasks in various medical domains.

Unlike traditional machine learning methods that rely on handcrafted features, deep learning models learn representations directly from data, allowing the discovery of complex patterns. For instance, CNNs have been widely adopted in medical image analysis [Lee et al., 2017] due to their ability to outperform humans in certain visual recognition tasks. In this project, we implement a CNN-based model for early-onset Parkinson's disease detection, leveraging Mel-spectrograms as input. These spectrograms, which represent the time-frequency characteristics of audio signals, can be treated as images, enabling the CNN to analyze subtle vocal changes associated with Parkinson's disease.

1.3 Research Objectives and Scope

This project builds on established methodologies proposed by [Quan et al., 2022] and applies them to an Italian dataset sourced online [Dimauro and Girardi, 2019]. The primary goal is to evaluate the effectiveness of an end-to-end deep learning model in detecting early-onset PD. To optimize the model's performance, we incorporate modified preprocessing techniques. Additionally, this study explores NeuroSpeech [Orozco-Arroyave et al., 2018], software tool designed to assist in the diagnostics of neurodegenerative diseases, and the Librosa library for robust audio processing.

2 Previous Work

2.1 End-to-End Deep Learning Approach for Parkinson's Disease Detection from Speech Signals

Parkinson's disease (PD) significantly impacts speech, with over 90% of individuals experiencing a condition called hypokinetic dysarthria, characterized by diminished vocal range, monotony, and difficulties with articulation. [Quan et al., 2022] proposed an end-to-end deep learning model specifically designed to analyze speech data for detecting PD. This framework eliminates the need for manual feature selection by automatically learning key spatial and temporal patterns within the audio signals.

2.1.1 Model Structure

The proposed approach begins by converting audio recordings into log Mel-spectrograms, which are then analyzed using a dual-stage convolutional architecture:

1. **Time-Distributed 2D-CNNs:** This network processes overlapping segments of the Mel-spectrogram to identify local spatial features, such as frequency dynamics. The architecture includes convolutional filters to extract features, batch normalization to stabilize learning, average pooling to reduce dimensionality, and dropout layers to prevent overfitting.

2. **1D-CNNs for Temporal Analysis:** The outputs from the 2D-CNN are passed through 1D-CNNs, which focus on capturing temporal dependencies and sequential features, such as variations in pitch or jitter over time. The extracted features are then processed through fully connected layers to output the final classification of PD versus non-PD speech.

2.1.2 Model Evaluation and Key Insights

Metrics such as accuracy, F1-score, sensitivity, specificity, and Matthews correlation coefficient (MCC) were utilized to evaluate the model. The model was assessed on two speech datasets: a Chinese dataset (collected by the authors), where it achieved accuracies of 81.6% for vowel sounds and

75.3% for short sentences, and a Spanish dataset [Orozco-Arroyave et al., 2014], where it performed notably better, achieving 92% accuracy on complex sentence samples. An important finding in the research was that low-frequency features within the Mel-spectrograms were particularly significant for PD detection, as these frequencies often reveal irregularities associated with vocal tremors and other speech impairments caused by the disease.

2.1.3 Adaptation and Relevance to This Work

This approach forms the foundation of our implementation. While the original work focused on Chinese and Spanish datasets, we applied this model to an Italian speech dataset containing samples ranging from 5 seconds to 3.5 minutes in duration. Through optimized preprocessing, including segmenting spectrograms to a frame length of 2,000 frames, we achieved a validation accuracy of over 98%. This performance demonstrates the adaptability of the model to diverse datasets and its capacity to generalize effectively to PD-affected speech.

2.2 NeuroSpeech: An open-source software for Parkinson's speech analysis

Here's a refined and more readable version of your text:

A significant number of Parkinson's disease (PD) patients develop hypokinetic dysarthria, which impairs various aspects of speech production and significantly affects their communication abilities.

In this context, NeuroSpeech [Orozco-Arroyave et al., 2018] emerges as a novel software solution specifically designed to analyze dysarthric speech in individuals with Parkinson's disease. The software provides a user-friendly interface that enables users to input patient information, initiate recordings, and select specific speech tasks for analysis. Additionally, NeuroSpeech allows for comparisons with benchmark measures from healthy control groups, offering valuable insights for result interpretation.

The authors explain that NeuroSpeech evaluates key speech dimensions—phonation, articulation, prosody, and intelligibility—using a detailed set of attributes:

- **Phonation**: Phonation refers to the ability of the vocal folds to vibrate and produce sound. NeuroSpeech assesses attributes such as fundamental frequency (F0), jitter (frequency variation), shimmer (amplitude variation), and harmonics-to-noise ratio (HNR). These measures provide insights into vocal fold function and stability, which are essential for understanding speech production efficiency in PD patients.

- **Articulation**: Articulation involves the precise movements of speech organs required to produce distinct sounds. NeuroSpeech analyzes articulation through metrics such as formant frequencies (F1, F2, F3), articulation rate (syllables per second), and speech rate. These attributes are critical for detecting slurring or distortions associated with dysarthria in PD.

- **Prosody:** Prosody encompasses the rhythm and melody of speech, including variations in pitch and loudness. NeuroSpeech measures prosodic features like pitch variability (standard deviation of pitch), segment duration, and intensity variation. These attributes are crucial for assessing the naturalness and emotional expressiveness of speech.

- **Intelligibility:** Intelligibility refers to the clarity and comprehensibility of speech. NeuroSpeech evaluates this dimension using metrics such as the percentage of words understood, articulation index, and phoneme accuracy. These measures help quantify speech clarity in PD patients.

NeuroSpeech is implemented with a graphical interface in C++, integrated with Python scripts. This design allows users to extend its functionality by training models for other neurodegenerative diseases or on custom reference datasets. By offering comprehensive analyses of multiple speech dimensions, it enhances clinicians' and researchers' ability to understand and address speech impairments.

3 Experiments

3.1 Data Collection

The Italian dataset [Dimauro and Girardi, 2019] used for training includes recordings from 15 young healthy controls (HC), 22 elderly HC, and 28 Parkinson's disease (PD) patients, comprising a total of 394 HC samples and 437 PD samples. The recordings range in duration from 5 seconds to 3.5 minutes. These samples include readings of the phonemically balanced text "Il ramarro della zia", syllable repetitions (e.g., 'pa' and 'ta'), phonemically balanced phrases, and sustained phonations of the vowels 'a,' 'e,' 'i,' 'o,' and 'u.'

For cross-linguistic testing, a Telugu dataset was collected from eight participants (four male and four female, aged 23–25), producing 24 recordings (four per person) under varying noise conditions.

Additionally, the English dataset [Jaeger et al., 2019] used was recorded at King's College London (KCL) Hospital. Participants conducted phone-based recordings, reading the text "The North Wind and the Sun" and engaging in spontaneous dialogue. Each recording is annotated with the subject's health status (HC or PD), Hoehn & Yahr stage, and Unified Parkinson's Disease Rating Scale (UPDRS) scores.

3.2 Librosa

Librosa is a Python library designed for analyzing and processing audio data, commonly utilized in speech processing, music information retrieval, and audio machine learning projects. Its key functionalities include:

- **Feature Extraction:** Librosa can extract features such as Mel-frequency cepstral coefficients (MFCCs), chroma, and spectral features, which are essential for both speech and music analysis.

- **Preprocessing:** The library converts raw audio signals into numerical features suitable for machine learning models, facilitating easier data handling and analysis.

- **Signal Processing:** Librosa offers tools for analyzing various audio properties, including tempo, beat, and rhythm, as well as capabilities to manipulate audio signals effectively.

- **Visualization:** Users can create waveforms, spectrograms, and feature plots, providing valuable insights into audio data and improving interpretability.

- **Audio Reconstruction**: The library allows for the conversion of spectrograms or MFCCs back into waveforms, which is useful for debugging or testing audio processing workflows.

3.3 Data Pre-processing

In this phase, audio recordings are processed using Librosa and Python functions to ensure uniformity and optimize the dataset for effective model training. The key steps include:

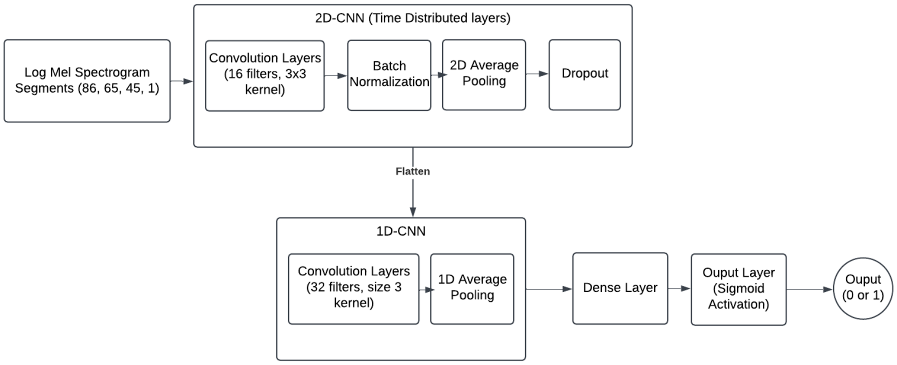
Log Mel Spectrogram Segments (86, 65, 45, 1)
2D-CNN (Time Distributed layers)
Convolution Layers (16 filters, 3x3 kernel)
Batch Normalization
2D Average Pooling
Dropout
Flatten
1D-CNN
Convolution Layers (32 filters, size 3 kernel)
1D Average Pooling
Dense Layer
Ouput Layer (Sigmoid Activation)
Ouput (0 or 1)

Figure 1: Proposed Model

1. **Loading Recordings:** The audio recordings are first loaded using Librosa, and log Mel spectrograms are generated to capture the frequency content of the recordings.

2. **Uniformity in Length:** To ensure consistency across samples, spectrograms are either padded or truncated to a fixed length of 2000 frames. This length was determined through experimentation to optimize performance.

3. **Segment Creation:** A sliding window approach is applied to each spectrogram, generating 86 overlapping segments of 45 frames each. This segmentation allows the model to focus on smaller, contextually rich portions of the audio data.

4. **Dataset Preparation:** The segments extracted from the Italian recordings are then organized into training and testing datasets.

3.4 Neural Network and Training

The proposed model (Figure 1) adopts a hybrid architecture combining 2D and 1D convolutional neural networks (CNNs) to capture spatial and temporal features from log Mel spectrograms. It consists of a 2D CNN for feature extraction and a 1D CNN to extract temporal variations, followed by a dense layer for classification. The input comprises spectrogram segments with dimensions $(86, 65, 45, 1)$, representing segments, frequency bands, time windows, and channels.

Feature extraction begins with the 2D CNN, which processes the spectrograms through a series of carefully designed layers. It starts with a TimeDistributed Conv2D layer employing 16 filters and a kernel size of $(3, 3)$ to convolve the input. This is followed by batch normalization to stabilize the learning process, average pooling to reduce dimensionality while preserving essential information, and dropout layers to mitigate overfitting. Together, these layers extract meaningful spatial features from the spectrograms, preparing the data for classification.

Once feature extraction is complete, the output is flattened and passed to the 1D CNN, which is responsible for learning the temporal variations in the extracted features. This stage includes a Conv1D layer with 32 filters and a kernel size of 3. Average pooling reduces feature dimensionality, and fully connected (dense) layers map these features to the output space. The final layer uses a sigmoid activation function to produce a binary classification, distinguishing between the two target classes (0-HC, 1-PD).

The Adam optimizer, with a learning rate of 0.0001, is used to minimize binary cross-entropy loss, making it well-suited for binary classification tasks. An EarlyStopping callback monitors the

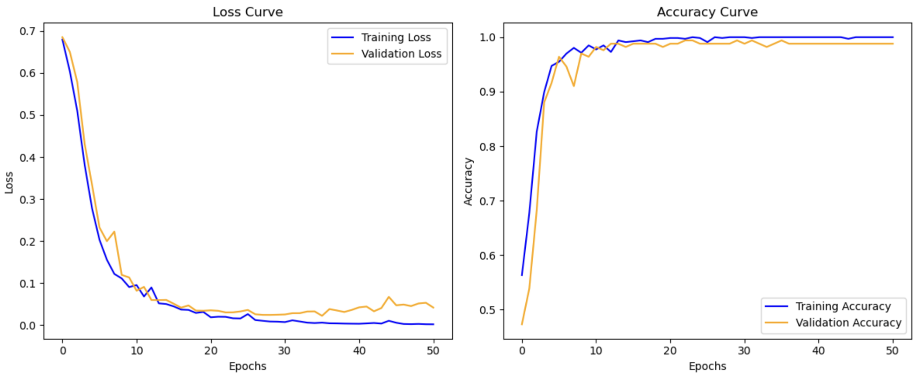

Figure 2: Loss and Accuracy Curves

validation loss and halts training if no improvement is observed for 15 consecutive epochs, thereby preventing overfitting. The dataset is divided into training (80%) and testing (20%) subsets, and training is conducted for a maximum of 200 epochs with a batch size of 16. During training, validation performance is monitored to refine hyperparameters and assess the model's generalization capabilities.

To ensure balanced learning, class distribution is carefully analyzed across the training and testing sets. Learning curves are plotted to visualize trends in loss and accuracy over epochs, offering insights into the model's convergence and performance.

After training, the model is evaluated on unseen audio segments from Italian, English and Telugu datasets. Predictions generate class probabilities for each segment, aiding in the diagnosis of early-onset Parkinson's disease by identifying patterns in audio features.

4 Evaluation

- **Performance Metrics:** The model achieved a validation accuracy of 98.8% during training on Italian audio data, demonstrating a strong ability to classify recordings accurately within this dataset. The extensive and well-structured dataset likely contributed to this performance, which is better than that reported in the original paper, providing the model with high-quality training data.

- **Learning Curves:** The loss curve (Figure 2) shows a steady decline in both training and validation loss, with minimal signs of overfitting, suggesting effective and balanced learning throughout the training process. The accuracy curve (Figure 2) exhibits rapid improvement during the initial 10 epochs, with training accuracy stabilizing around 100% and validation accuracy closely following. This indicates the model's strong generalization capability when tested on Italian data.

- **Optimal Frame Length:** Through experimentation, the optimal frame length for the sliding window was determined to be 2,000 frames. Using shorter segments not only improved accuracy but also reduced training time, significantly enhancing the model's overall efficiency.

- **Training Plateau:** The model's accuracy and loss stabilized after approximately 20 epochs, indicating that it effectively learned the underlying patterns in the Italian audio data.

- **Cross-Language Classification:** The classifier trained on Italian data was tested on Telugu recordings collected by the team and English recordings sourced online. However, the model's performance dropped drastically, achieving only 5%–20% accuracy on these datasets in different runs. This substantial decline underscores its limitations in adapting to different language patterns.

- **Error Analysis:** The model's poor performance on non-Italian recordings suggests it may have learned language-specific patterns, such as phoneme distribution, tone, or intonation, which are unique to Italian. Significant differences in prosodic features—such as stress, pitch, and rhythm—between Italian, Telugu, and English likely contribute to its inability to generalize effectively across languages.

The loss curve demonstrates effective learning with minimal overfitting, as both training and validation losses decline steadily. Additionally, the close alignment between validation accuracy and training accuracy underscores the model's generalization capabilities within the Italian dataset. Testing was conducted using Italian recordings from phone calls and other sources, where the model performed well. However, additional recordings of individuals with Parkinson's disease are needed to fully assess the model's effectiveness with Italian data. Furthermore, its inability to classify non-Italian recordings highlights the need for further development to adapt to diverse linguistic patterns.

5 Discussion

The model's performance reveals both its potential and limitations, particularly regarding language-specific patterns and the challenges of cross-language generalization. It seems to learn patterns unique to the Italian dataset, such as phoneme distribution, tone, and intonation, which effectively classify recordings within the same language. However, this approach presents significant challenges when the model encounters data from other languages. For example, Telugu and English exhibit distinct prosodic features—such as stress, pitch, and rhythm—that differ substantially from those in Italian. Because the model was trained exclusively on Italian recordings, it lacks the flexibility to adapt to these variations, leading to a significant drop in accuracy (5%–20%) when tested on Telugu and English data.

To address these limitations, future efforts can focus on multi-language training by expanding the dataset to include recordings from various languages. This expansion could help the model identify universal patterns, such as pitch variability or speech irregularities, that are less dependent on language. Such an approach may reduce overfitting to language-specific features and improve cross-language generalization. Additionally, exploring classical audio features—such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral flux, or zero-crossing rate—could minimize language dependency while preserving essential audio characteristics relevant to Parkinson's classification.

Several challenges also arise from using a deep learning architecture. The model's "black box" nature limits our understanding of its decision-making process, which poses a critical challenge in medical applications like Parkinson's classification, where interpretability is essential for ensuring trust due to the substantial consequences of healthcare decisions.

Improving model interpretability is crucial; rather than relying solely on deep learning, more interpretable machine learning models like regression or decision trees may be better suited for medical diagnostics. Furthermore, considering alternative non-predictive approaches, such as those demonstrated in the paper NeuroSpeech, which analyze specific speech features without relying on predictive algorithms, could provide reliable insights into speech impairments while circumventing the challenges of generalization across languages.

In summary, while the model demonstrates excellent performance on the Italian dataset, its inability to generalize across languages highlights the need to address language dependency. By incorporating multi-language training, predefined audio features, and interpretability measures, the model's robustness and applicability for medical diagnostics in diverse linguistic contexts can be significantly enhanced.

5.1 Conclusion

While the deep learning model achieves impressive accuracy on the Italian dataset, its significant decline in performance on Telugu and English recordings reveals critical limitations related to language dependency. These findings emphasize the need to expand the training dataset to include multiple languages, which would enable the model to learn universal speech patterns and enhance its adaptability. Additionally, integrating classical audio features could help reduce the model's reliance on language-specific characteristics, thereby improving its generalization.

Moreover, enhancing the model's interpretability is essential for its application in medical diagnostics, where understanding decision-making processes is vital for ensuring trust and reliability. Exploring alternative predictive models that are more interpretable, as well as non-predictive approaches, could provide valuable insights into speech impairments without the complexities associated with predictive modeling. Addressing these challenges will pave the way for more robust, scalable, and interpretable solutions for the early diagnosis of Parkinson's disease across diverse linguistic contexts.

References

Giovanni Dimauro and Francesco Girardi. Italian parkinson's voice and speech, 2019. URL https://dx.doi.org/10.21227/aw6b-tg17.

Hagen Jaeger, Dhaval Trivedi, and Michael Stadtschnitzer. Mobile device voice recordings at king's college london (mdvr-kcl) from both early and advanced parkinson's disease patients and healthy controls. https://doi.org/10.5281/zenodo.2867216, 2019. [Data set].

June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: general overview. *Korean journal of radiology*, 18(4):570–584, 2017.

Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia Gonzalez-Rátiva, and Elmar Nöth. New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *Lrec*, pages 342–347, 2014.

Juan Rafael Orozco-Arroyave, Juan Camilo Vásquez-Correa, Jesús Francisco Vargas-Bonilla, Raman Arora, Najim Dehak, Phani S Nidadavolu, Heidi Christensen, Frank Rudzicz, Maria Yancheva, H Chinaei, et al. Neurospeech: An open-source software for parkinson's speech analysis. *Digital Signal Processing*, 77:207–221, 2018.

Changqin Quan, Kang Ren, Zhiwei Luo, Zhonglue Chen, and Yun Ling. End-to-end deep learning approach for parkinson's disease detection from speech signals. *Biocybernetics and Biomedical Engineering*, 42(2):556–574, 2022.

K. Tjaden. Speech and swallowing in parkinson's disease. *Topics in Geriatric Rehabilitation*, 24(2): 115–126, 2008. doi: 10.1097/01.TGR.0000318899.87690.44.