POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering

Master's Degree Thesis

Voice Classification in Parkinson's Disease Using Transformer Models and Error Rate Metrics

Supervisors

Prof. Gabriella OLMO

PhD. Federica AMATO

Candidate

Benedetta PERRONE

Academic Year 2023/2024

Summary

Parkinson's disease is a neurodegenerative disorder resulting from the progressive degeneration of dopaminergic neurons in the substantia nigra pars compacta. Alongside motor symptoms like bradykinesia, tremor, and rigidity, Parkinson's disease is also associated with non-motor impairments, including cognitive decline, depression, sleep disorders, and autonomic dysfunctions. One of the most prevalent non-motor symptoms is the alteration of voice and speech, affecting up to 90% of patients. The progressive decline in vocal function can lead to hypokinetic dysarthria, reducing speech intelligibility, volume, and prosody, with a significant impact on patients' quality of life.

This study has two main objectives: (1) to distinguish between healthy individuals and those with Parkinson's disease based on vocal characteristics, and (2) to assess disease severity using Word Error Rate and Character Error Rate, exploring their correlation with Unified Parkinson's Disease Rating Scale scores. The models used include the Vision Transformer and the Audio Spectrogram Transformer, trained on vocal recordings from datasets comprising both Parkinsonian patients and healthy controls.

The preprocessing pipeline included resampling to 16 kHz, volume normalization, and outlier removal. Mel-spectrograms were generated for Vision Transformer, while Audio Spectrogram Transformer directly processed the waveform. Both models were trained using 5-fold cross-validation to ensure robustness, and their performance was evaluated in terms of accuracy, precision, recall, and F1-score.

Word Error Rate and Character Error Rate metrics were calculated using OpenAI's Whisper model and compared between patients and healthy controls. Statistical analysis, including Shapiro-Wilk and Mann-Whitney U tests, revealed significant differences in Word Error Rate and Character Error Rate between patients and controls, indicating a correlation between increased vocal production errors and disease severity.

The results suggest that integrating deep learning methodologies in clinical settings could offer promising, non-invasive tools for early diagnosis and continuous monitoring of Parkinson's disease, such as through mobile voice recordings for remote, non-invasive monitoring. Additionally, explainability techniques could generate heatmaps highlighting critical areas of the mel-spectrogram, enabling the identification and restoration of unintelligible speech elements directly from spectrograms. This approach could support the development of personalized text-to-speech systems to aid communication for patients with severe vocal impairments.

A chi mi ha trasmesso emozioni

Table of Contents

List of Tables X

List of Figures XII

Acronyms XV

1 Introduction 1
- 1.1 Overview of Parkinson's Disease 1
  - 1.1.1 Pathophysiology 1
  - 1.1.2 Diagnosis 2
  - 1.1.3 Treatment 3
  - 1.1.4 Unified Parkinson's Disease Rating Scale 3
- 1.2 Impact of Parkinson's Disease on Vocal Characteristics 3
  - 1.2.1 Anatomy and Physiology of Speech Production 4
  - 1.2.2 Anatomical Alterations in Parkinson's Disease 5
  - 1.2.3 Consequences of Anatomical Alterations on Speech in Parkinson's Disease 6

2 State of the Art 7
- 2.1 Sound Automatic Analysis 7
  - 2.1.1 Deep Learning Approaches for Sound Automatic Analysis . . 7
- 2.2 Speech Analysis in Parkinson's Disease 8
  - 2.2.1 Automatic Speech Recognition with Whisper 8
  - 2.2.2 Data Augmentation in Speech Analysis 9
  - 2.2.3 State-of-the-Art Challenges 9

3 Materials and Methods 11
- 3.1 Data Acquisition Protocol 11
  - 3.1.1 Voice Recording Protocol at Molinette Hospital 11
  - 3.1.2 Description of Datasets Used (PC-GITA, Bari, Molinette) . 14
- 3.2 Word Error Rate and Character Error Rate Calculation Using Whisper 19

3.2.1 Preprocessing Pipeline for Speech Signals 19
3.2.2 Calculation and Analysis of Word Error Rate and Character Error Rate 21
3.2.3 Statistical Analysis of Word Error Rate and Character Error Rate 22
3.3 Classification Task 24
3.4 Preprocessing Pipeline 25
3.4.1 Common Preprocessing Steps 25
3.4.2 Vision Transformer Specific Preprocessing 26
3.4.3 Audio Spectrogram Transformer Specific Preprocessing . . . 28
3.5 Parallel Processing of Audio 29
3.6 Cross-Validation and Stratified Group Splitting 30
3.7 Evaluation Metrics 31
3.8 Model Configurations and Hyperparameters 32
3.8.1 Vision Transformer Model 32
3.8.2 Audio Spectrogram Transformer Model 33
3.8.3 Parallel Vision Transformer Model 34
3.8.4 Parallel Audio Spectrogram Transformer Model 35
3.9 UPDRS-based Classification 36

4 Results 38
4.1 Speech Recognition Results 38
4.1.1 PC-GITA Dataset Results 38
4.1.2 Bari Dataset Results 41
4.1.3 Molinette Dataset Results 46
4.1.4 Combined Results (Bari + Molinette) 48
4.2 Classification Results 52
4.2.1 Vision Transformer Results 53
4.2.2 Parallel Vision Transformer Results 54
4.2.3 Audio Spectrogram Transformer Results 57
4.2.4 Parallel Audio Spectrogram Transformer Results 58

5 Discussion and Conclusion 60
5.1 Word Error Rate and Character Error Rate 60
5.1.1 Dataset Differences and Their Impact on Word Error Rate and Character Error Rate 60
5.1.2 Statistical Significance of Differences Between Groups 61
5.1.3 Statistical Differences Across UPDRS Levels 62
5.1.4 Word Error Rate and Character Error Rate as Measures of Speech Impairment 65
5.2 Classification 66

5.2.1 Overview of Classification Performance 66
5.2.2 UPDRS-Based Probability Analysis 67
5.3 Limitations and Future Directions 68
5.3.1 Limitations of the Study 68
5.3.2 Future Research Directions 69

Bibliography 72

List of Tables

3.1 Demographic data and UPDRS scores for patients in Molinette and Politecnico studies . . . 13
3.2 Phonemically balanced sentences used in the PC-GITA dataset . . . 15
3.3 DDK sequences used in the PC-GITA dataset . . . 15
3.4 Words used in the PC-GITA dataset . . . 16
3.5 Number of recordings per task in the PC-GITA dataset. . . . 16
3.6 Words used in the Bari dataset, with their English translations. . . 17
3.7 Phonemically balanced text used in the Bari dataset . . . 18
3.8 Number of recordings per task in the Bari dataset. . . . 18
3.9 Proverbs used in the Molinette dataset with their English translations. 19
3.10 Total number of recordings per exercise for each patient of the Molinette dataset . . . 19
3.11 Hyperparameters used for training the ViT model. . . . 33
3.12 Hyperparameters used for training the AST model. . . . 34
3.13 Hyperparameters for training the parallel ViT model. . . . 35
3.14 Hyperparameters used for training the parallel AST model. . . . 36

4.1 Cumulative Means, Medians, and Standard Deviations for WER in the PC-GITA Dataset . . . 39
4.2 Cumulative Means, Medians, and Standard Deviations for CER in the PC-GITA Dataset . . . 40
4.3 Cumulative Means, Medians, and Standard Deviations for WER in the Bari Dataset . . . 43
4.4 Cumulative Means, Medians, and Standard Deviations for CER in the Bari Dataset . . . 43
4.5 Cumulative Means, Medians, and Standard Deviations for WER by UPDRS Levels in the Bari Dataset . . . 43
4.6 Cumulative Means, Medians, and Standard Deviations for CER by UPDRS Levels in the Bari Dataset . . . 44
4.7 Cumulative Means, Medians, and Standard Deviations for WER by UPDRS Levels in the Molinette Dataset . . . 48

4.8 Cumulative Means, Medians, and Standard Deviations for CER by UPDRS Levels in the Molinette Dataset 48
4.9 Cumulative Means, Medians, and Standard Deviations for WER by UPDRS Levels in the Combined Bari and Molinette Dataset 50
4.10 Cumulative Means, Medians, and Standard Deviations for CER by UPDRS Levels in the Combined Bari and Molinette Dataset 50
4.11 Training and Validation Metrics for the ViT Model (Averaged Across 5 Folds). 54
4.12 UPDRS-Based Analysis for the ViT Model. 55
4.13 Training and Validation Metrics for the ViT Parallel Model (Averaged Across 5 Folds). 56
4.14 UPDRS-Based Analysis for the ViT Parallel Model. 57
4.15 Training and Validation Metrics for the AST Model (Averaged Across 5 Folds). 58
4.16 UPDRS-Based Analysis for the AST Model. 58
4.17 Training and Validation Metrics for the AST Parallel Model (Averaged Across 5 Folds). 59
4.18 UPDRS-Based Analysis for the AST Parallel Model. 59

5.1 Mean WER and CER Comparison Across Datasets 61
5.2 Median WER and CER Comparison Across Datasets 61
5.3 Mann-Whitney U Test p-values for WER and CER in the PC-GITA and Bari Datasets 62
5.4 Kruskal-Wallis Test p-values for WER and CER across UPDRS Levels in the Bari, Molinette, and Combined Datasets 65
5.5 Summary of Validation Metrics for All Classifiers. 66

List of Figures

1.1 Progression of motor and non-motor symptoms in Parkinson's disease [1]. 2
1.2 Anatomical structures involved in vocalization and muscles responsible for vocal fold movement 5

3.1 Overview of the preprocessing pipeline for audio data used in the classification task. 21
3.2 Flowcharts of statistical analysis for WER and CER metrics 25
3.3 Distribution of audio lengths before and after outlier removal. . . . 26
3.4 Overview of common preprocessing steps applied to all audio recordings 27
3.5 Overview of the ViT-specific preprocessing pipeline 28
3.6 Diagram of 5-fold cross-validation with stratified group splitting . . 31

4.1 WER distribution for controls and Parkinsonians (PC-GITA) 39
4.2 CER distribution for controls and Parkinsonians (PC-GITA). 40
4.3 WER distribution for controls and Parkinsonians (Bari). 41
4.4 CER distribution for controls and Parkinsonians (Bari). 42
4.5 WER distribution by control and UPDRS levels (Bari dataset) . . . 44
4.6 CER distribution by control and UPDRS levels (Bari dataset). . . . 45
4.7 Dunn's post hoc test results for pairwise comparisons of WER across UPDRS levels in the Bari dataset. 46
4.8 Dunn's post hoc test results for pairwise comparisons of CER across UPDRS levels in the Bari dataset. 47
4.9 Boxplots showing the distribution of WER and CER for Parkinsonian patients in the Molinette dataset (n = 80). 49
4.10 WER distribution by UPDRS levels in the Molinette dataset. . . . 50
4.11 CER distribution by UPDRS levels in the Molinette dataset. 51
4.12 Dunn's post hoc test results for pairwise comparisons of WER across UPDRS levels in the Molinette dataset. 52
4.13 Dunn's post hoc test results for pairwise comparisons of CER across UPDRS levels in the Molinette dataset. 53

4.14 WER distribution by UPDRS levels in the combined Bari and Molinette dataset. 54
4.15 CER distribution by UPDRS levels in the combined Bari and Molinette dataset. 55
4.16 Dunn's post hoc test results for pairwise comparisons of WER across UPDRS levels in the combined Bari and Molinette dataset. 56
4.17 Dunn's post hoc test results for pairwise comparisons of CER across UPDRS levels in the combined Bari and Molinette dataset. 57

5.1 Mean WER and CER for Different UPDRS Levels across Bari and Molinette datasets. 63
5.2 Median WER and CER for Different UPDRS Levels across Bari and Molinette datasets. 64
5.3 UPDRS-Based Mean Probabilities for All Classifiers 68

Acronyms

PD

Parkinson's disease

DBS

Deep Brain Stimulation

SPECT

Single Photon Emission Computed Tomography

PET

Positron Emission Tomography

CSF

Cerebrospinal Fluid

NfL

Neurofilament Light Chain

UPDRS

Unified Parkinson's Disease Rating Scale

CNN

Convolutional Neural Network

AST

Audio Spectrogram Transformer

ViT

Vision Transformer

MFCC

Mel-Frequency Cepstral Coefficients

HNR

Harmonics-to-Noise Ratio

ASR

Automatic Speech Recognition

WER

Word Error Rate

CER

Character Error Rate

EBU

European Broadcasting Union

DDK

Diadochokinetic

IQR

Interquartile Range

XAI

Explainable Artificial Intelligence

TTS

Text-to-Speech

Chapter 1

Introduction

1.1 Overview of Parkinson's Disease

Parkinson's disease (PD) is a chronic, progressive neurodegenerative disorder primarily caused by the degeneration of dopaminergic neurons in the substantia nigra pars compacta, a region of the brain responsible for motor control. The resulting dopamine deficiency leads to a range of motor symptoms such as bradykinesia, resting tremor, rigidity, and postural instability. Beyond these motor symptoms, PD also manifests in non-motor areas including cognitive decline, depression, sleep disturbances, and autonomic dysfunction, often preceding motor symptoms by several years [1, 2].

The incidence of Parkinson's disease worldwide ranges from 5 to over 35 new cases per 100,000 individuals annually [1]. This rate increases consistently with age, particularly between 60 and 90 years [2]. The prevalence of the disease has been rising significantly, driven by an aging population, advancements in diagnostic techniques, and increased exposure to environmental risks, including pesticides and industrial pollutants [2]. It is estimated that approximately 0.3% of the general population is affected, with prevalence exceeding 3% in individuals over 80 years old [1].

1.1.1 Pathophysiology

The primary factor in the pathophysiology of Parkinson's disease is the progressive loss of dopaminergic neurons in the substantia nigra. However, the disease involves an interplay of additional factors such as the aggregation of misfolded α-synuclein proteins, mitochondrial dysfunction, impairment of lysosomal degradation, issues with vesicle transport, problems with synaptic transmission, and neuroinflammation [1, 2, 3]. Most cases result from the interaction between common genetic variants and environmental factors, making the etiology heterogeneous and multifactorial [3].

1.1.2 Diagnosis

The diagnosis of Parkinson's disease is primarily clinical, relying on a neurological examination, the patient's medical history, and an assessment of both motor and non-motor symptoms [1]. However, early diagnosis can be challenging due to the overlap of symptoms with other neurodegenerative disorders, such as multiple system atrophy (MSA) and progressive supranuclear palsy (PSP) [4]. Bradykinesia, rigidity, and tremor are important clinical signs of Parkinson's disease, but they are not always sufficient on their own to ensure an early and accurate diagnosis [1].

Figure 1.1 illustrates the progression of Parkinson's disease through distinct stages. The prodromal phase is characterized by non-motor symptoms such as depression, anxiety, and constipation, which precede the development of motor symptoms like bradykinesia, rigidity, and tremor.

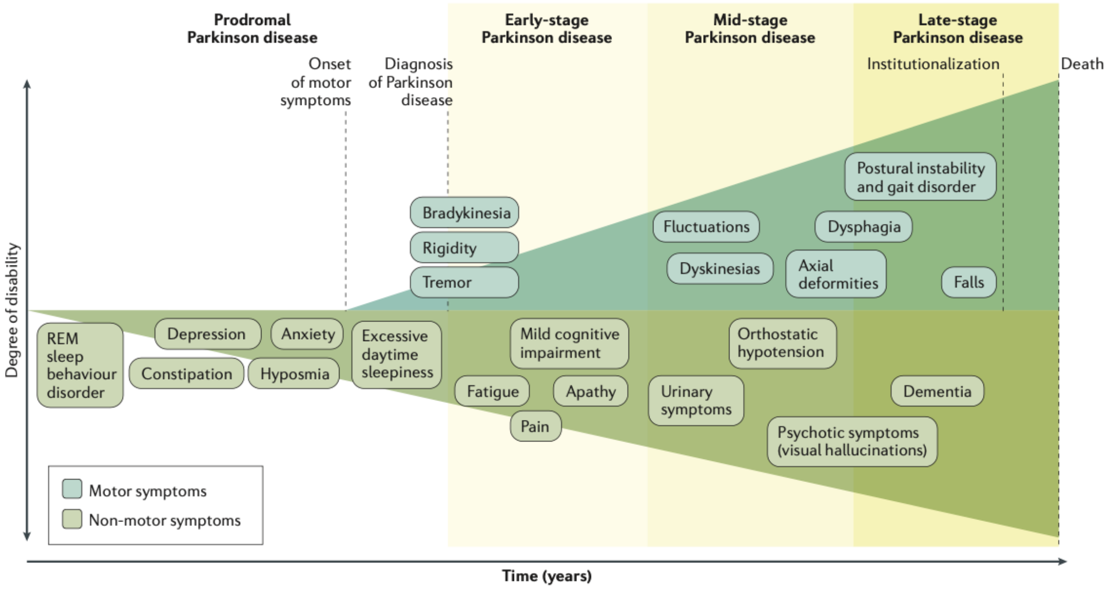

Prodromal Parkinson disease
Early-stage Parkinson disease
Mid-stage Parkinson disease
Late-stage Parkinson disease
Onset of motor symptoms
Diagnosis of Parkinson disease
Institutionalization
Death
Degree of disability
Bradykinesia
Rigidity
Tremor
Fluctuations
Dysphagia
Postural instability and gait disorder
Dyskinesias
Axial deformities
Falls
REM sleep behaviour disorder
Depression
Anxiety
Constipation
Hyposmia
Excessive daytime sleepiness
Mild cognitive impairment
Orthostatic hypotension
Fatigue
Apathy
Pain
Urinary symptoms
Dementia
Psychotic symptoms (visual hallucinations)
🟩 Motor symptoms
🟩 Non-motor symptoms
Time (years)

Figure 1.1: Progression of motor and non-motor symptoms in Parkinson's disease [1].

Advanced imaging methods, including SPECT and PET, allow for the visualization of dopaminergic deficits and metabolic alterations in the brain [5]. Furthermore, biomarkers such as α-synuclein in CSF and neurofilaments (NfL) in both blood and CSF are under investigation to improve the accuracy of Parkinson's disease diagnosis [2, 6].

Nevertheless, definitive confirmation of Parkinson's disease is only possible post-mortem through the identification of neuropathological features, such as Lewy bodies in the brain [7].

1.1.3 Treatment

There is currently no cure for PD, and treatment focuses on alleviating symptoms. Levodopa, often administered in combination with carbidopa or benserazide to improve its efficacy and reduce side effects such as nausea, remains the most commonly used therapy [5]. Over time, most patients require Levodopa, which is considered the gold standard for managing motor symptoms. However, prolonged use can lead to motor complications, including fluctuations between "on" phases, characterized by improved motor control, and "off" phases, during which symptoms reappear as the medication's effects wear off [8]. Additional treatment options include dopamine agonists, monoamine oxidase B (MAO-B) inhibitors, and catechol-O-methyltransferase (COMT) inhibitors, which help stabilize dopamine levels and manage the disease [5].

Deep brain stimulation (DBS) is a surgical treatment option for patients whose symptoms cannot be effectively managed with medication. This procedure involves placing electrodes in targeted areas of the brain to regulate abnormal neuronal activity, offering more consistent symptom relief [5, 7].

1.1.4 Unified Parkinson's Disease Rating Scale

The Unified Parkinson's Disease Rating Scale (UPDRS) is a comprehensive tool used to assess the severity and progression of Parkinson's disease. It evaluates multiple aspects of the disease, including motor function, activities of daily living, and treatment-related complications. The scale, widely used both in clinical practice and research, is divided into four parts: Mentation, Behavior, and Mood; Activities of Daily Living; Motor Examination; and Complications of Therapy. Each item on the UPDRS is rated on a scale from 0 to 4, with higher scores indicating greater impairment [9].

1.2 Impact of Parkinson's Disease on Vocal Characteristics

Voice disorders are among the most common non-motor symptoms in PD, affecting approximately 70-90% of patients, with manifestations often emerging in the early stages of the disease [10, 11]. These disorders primarily manifest through changes in *phonation*, *articulation*, and *prosody* (Section 1.2.1). The progressive impairment of the motor mechanisms involved in vocal production makes these symptoms particularly debilitating, reducing speech intelligibility and communicative effectiveness [10].

1.2.1 Anatomy and Physiology of Speech Production

Vocalization is a process that requires the integration of multiple anatomical and physiological components. The main structures involved include the lungs, larynx, and vocal cords, which work together to produce sound [10]. As shown in Figure 1.2, Panel A, the lungs generate the airflow, the larynx modulates this airflow through the vibration of the vocal cords, and the articulatory system (including the tongue, lips, and palate) shapes the sound to produce comprehensible speech [10]. The vocal cords are controlled by intrinsic muscles of the larynx, as shown in Panel B of Figure 1.2. Abduction of the vocal folds is facilitated by the posterior cricoarytenoid muscle, while adduction is primarily controlled by the thyroarytenoid, lateral cricoarytenoid, and interarytenoid muscles [12, 11].

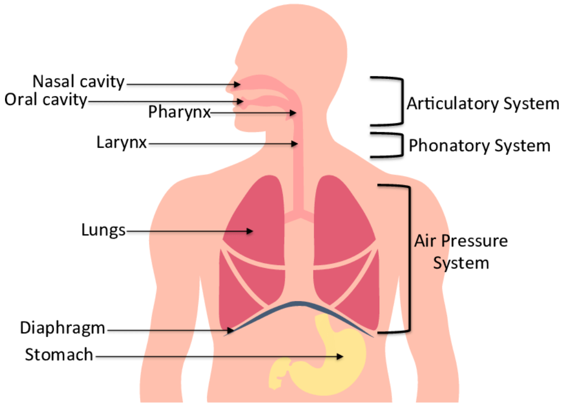
The processes involved in vocalization can be divided into three main components:

- **Phonation**: The process of sound production, which depends on the controlled oscillation of the vocal cords within the larynx, driven by the airflow exhaled from the lungs. The ability to modulate the tension of the vocal cords determines the pitch (fundamental frequency) of the voice, while subglottic pressure and the configuration of the resonant cavities contribute to the volume and quality of the voice [10, 11].

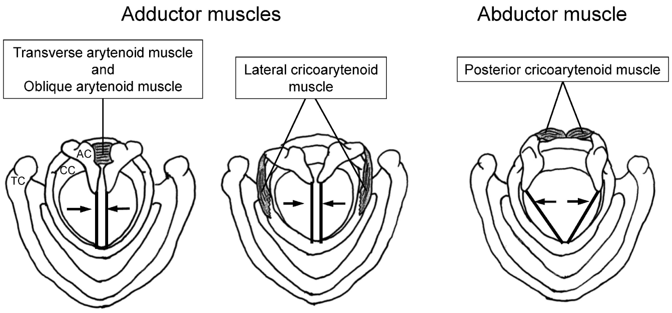
- **Articulation**: The modification of the sound produced during phonation into distinct phonemes. This process involves coordinated movements of the tongue, lips, jaw, and soft palate, which together produce the vowels and consonants essential for intelligible speech. Proper articulation ensures clarity in verbal communication, allowing listeners to differentiate between sounds and accurately interpret words [14].

- **Prosody**: The rhythm, intonation, and stress patterns in speech, encompassing variations in pitch, loudness, and duration that convey emotional tone, emphasis, and sentence structure. Prosody makes speech expressive and engaging, helping listeners interpret meaning beyond the literal words spoken [15].

In a healthy individual, these components work in harmony, allowing for a wide range of vocal tones, volumes, and expressive qualities [10, 11, 15].



(A) Vocal Apparatus


Adductor muscles
Abductor muscle
Transverse arytenoid muscle and Oblique arytenoid muscle
Lateral cricoarytenoid muscle
Posterior cricoarytenoid muscle
AC
CC
TC

(B) Muscles of the Larynx

Figure 1.2: (A) Anatomical structures involved in vocalization, including the air pressure, phonatory, and articulatory systems [13]. (B) Muscles responsible for vocal fold movement [12].

1.2.2 Anatomical Alterations in Parkinson's Disease

In PD, anatomical and physiological alterations occur that disrupt normal vocalization. Muscle rigidity and bradykinesia negatively affect the ability to control the tension of the vocal cords, causing difficulty in maintaining a consistent pitch, resulting in variations in fundamental frequency and a less stable vocal quality. Laryngeal joints may become stiff, limiting the movement of the vocal cords and causing further problems in speech production [14]. Additionally, the impairment of the basal ganglia leads to dysfunction in the smooth execution of speech movements, contributing to the characteristic hypokinetic dysarthria observed in PD patients [15].

1.2.3 Consequences of Anatomical Alterations on Speech in Parkinson's Disease

The anatomical and functional alterations described above result in several speech anomalies in patients with PD, including issues in *phonation*, *articulation*, and *prosody*.

Phonation: Due to the reduced ability to adequately control the vocal cords, Parkinson's patients develop hypophonia, characterized by reduced vocal volume, and dysphonia, manifesting as a hoarse, weak, and often unstable voice [11]. The reduction in fundamental frequency (F0) and decreased pitch variability (monopitch) are common symptoms, reflecting the inability to properly modulate the tension of the vocal cords [15]. Additionally, parameters such as jitter (frequency variation) and shimmer (amplitude variation) are significantly increased, indicating ineffective vocal control [14].

Articulation: Rigidity and bradykinesia affecting the articulatory muscles lead to reduced amplitude and speed of articulatory movements. This results in less precise pronunciation of words, particularly for phonemes requiring precise closure, such as plosives (e.g., /p/, /t/, /k/) [10]. Spirantization, where plosives are produced as fricatives (e.g., /f/, /v/, /z/), is a common phenomenon in Parkinson's patients and further compromises speech intelligibility [15].

Prosody: Difficulty in modulating pitch and vocal intensity leads to monotonous speech, with a lack of variation in pitch (monopitch) and intensity (monoloudness) [11]. These changes reduce the patient's ability to express emotions through speech, making communication less effective and harder to understand for listeners [14].

The combination of alterations in *phonation*, *articulation*, and *prosody* leads to hypokinetic dysarthria, a speech disorder characterized by reduced verbal fluency and an increased frequency of pauses during conversation. This disorder reflects both motor difficulties and cognitive impairments typical of PD. Studies have shown that PD patients may produce approximately 30% fewer words than healthy controls in verbal fluency tests, indicating a negative impact on language production [14, 15, 16].

Speech alterations in Parkinson's patients not only impair the ability to communicate effectively but also have a profound impact on quality of life. Difficulty in producing clear and intelligible speech can lead to social isolation, frustration, and depression [15]. Additionally, reduced facial expressivity (hypomimia) and difficulties in recognizing and interpreting emotions in others further complicate daily communication, exacerbating feelings of isolation and reducing the overall well-being of the patient [14].

Chapter 2

State of the Art

2.1 Sound Automatic Analysis

Sound automatic analysis involves a set of techniques aimed at identifying and distinguishing between different types of acoustic signals. These techniques find application in several fields, including emotion recognition, disease detection through speech, and the classification of environmental sounds, such as urban or natural noise recognition [17, 18].

2.1.1 Deep Learning Approaches for Sound Automatic Analysis

Deep learning models have revolutionized sound classification due to their ability to learn complex representations directly from the data, without the need for manual feature extraction.

Convolutional neural networks (CNNs) are widely used in sound classification, thanks to their ability to analyze spectral representations such as log-Mel-Frequency Cepstral Coefficients (log-MFCCs) [19, 20]. A MFCC represents the audio signal on the Mel scale, designed to reflect human perception of frequencies, making it ideal for tasks like speech recognition or sound classification [20].

Transformers, such as the Vision Transformer (ViT) and the Audio Spectrogram Transformer (AST), use self-attention mechanisms to analyze sequences of audio data. The ViT, initially developed for images, has been successfully adapted for sound analysis, demonstrating excellent performance in audio classification benchmarks [20, 21]. The AST, specifically designed for audio signal classification, has outperformed CNNs in various classification tasks [20].

Another widely used approach in sound classification is the use of embedding techniques. Embeddings are compact representations of data in vector spaces, allowing machine learning models to capture relevant information from acoustic signals. Embeddings such as MFCCs and those derived from pre-trained neural models like VGGish and YAMNet are widely used to represent the spectral features of speech signals [17].

End-to-end models, on the other hand, have revolutionized traditional speech recognition approaches by eliminating the need for separate pipelines for feature extraction and classification [20].

2.2 Speech Analysis in Parkinson's Disease

CNNs have been applied to analyze spectral representations like log-Mel spectrograms in PD, demonstrating accuracy above 85% in detecting vocal anomalies [19].

Acoustic features like jitter, shimmer, and the Harmonics-to-Noise Ratio (HNR) are commonly analyzed to evaluate the stability of vocal cord vibrations in patients with PD. Changes in these parameters are indicative of typical impairments associated with the disease, such as vocal instability and reduced sound quality [22, 23]. Moreover, MFCCs are often employed to capture the spectral characteristics of speech. These features are particularly useful for processing signals affected by noise, making them effective for analyzing the vocal output of PD patients [24].

Regarding embedding techniques, i-vectors and x-vectors have proven particularly useful. i-vectors provide a compact representation of vocal features, but x-vectors, derived from deep neural networks, have shown superior performance in classifying vocal anomalies related to PD. X-vectors capture better phonatory and articulatory variations, ensuring greater accuracy in pathological speech recognition compared to i-vectors [19].

Another relevant approach is the use of Transformer models, such as the ViT, which has been used to classify sustained vowels of PD patients, showing promising results with 78% accuracy [21]. The AST, though effective in other sound classification applications, has not yet been specifically applied to PD or other neurodegenerative diseases [20].

2.2.1 Automatic Speech Recognition with Whisper

Automatic speech recognition (ASR) refers to a technology capable of transcribing spoken language into text, facilitating the analysis of vocal signals. In the context of PD, ASR systems have been employed to identify vocal abnormalities, contributing to both the diagnosis and ongoing monitoring of the disease. Whisper, an advanced ASR model developed by OpenAI, has proven effective for this purpose.

Whisper is based on a Transformer architecture and has been trained on over 680,000 hours of multilingual and multitask supervised data, allowing it to perform robust speech recognition across a variety of conditions, including noisy environments and diverse accents [25]. This makes it particularly well-suited for analyzing pathological speech, which is often more variable than typical speech.

Whisper has been successfully applied in studies analyzing speech disorders, such as dysarthria, which is common in PD. For instance, Whisper has outperformed other ASR models in recognizing dysarthric speech, achieving high levels of Word Recognition Accuracy (WRA) even with limited data [26].

Whisper's versatility makes it suitable for real-time applications, including deployment on edge devices such as Raspberry Pi. This allows it to provide immediate transcription for individuals with speech impairments, even in environments where access to cloud-based resources is limited [27]. Additionally, Whisper's robust performance in multilingual ASR tasks enables it to process speech across various languages and accents, making it particularly useful when working with diverse patient populations [28].

However, Whisper is not very reliable when applied to highly disordered speech, as in the advanced cases of PD. The variability of speech patterns and the presence of background noise can sometimes compromise the accuracy of the model. In addition, the computational requirements of the larger variants of the Whisper model may limit its applicability in resource-limited environments [29].

2.2.2 Data Augmentation in Speech Analysis

Data augmentation techniques are very useful when working with small datasets, such as those involving PD patients. Methods including the addition of background noise, time stretching, and pitch shifting are commonly employed to artificially expand the dataset size, enhancing the model's ability to generalize [30, 23]. These approaches not only mitigate the risk of overfitting but also improve the model's robustness, enabling it to perform better on unseen data [23].

2.2.3 State-of-the-Art Challenges

Despite advances in sound classification and speech analysis, many open challenges remain in the context of neurodegenerative diseases. Diagnosing diseases like PD through vocal analysis is complex due to the overlap of vocal symptoms with other conditions, such as aging or other neurodegenerative diseases [22]. Vocal problems such as dysarthria, typical of PD, are also common in diseases like amyotrophic lateral sclerosis (ALS) and Alzheimer's, further complicating the diagnosis [24].

Additionally, variability in recording conditions, including differences in microphones and environments, along with individual factors such as age, gender, and disease severity, complicates the development of accurate models [22]. Addressing these challenges requires further research into the application of deep learning methods, with a particular focus on enhancing model generalization and robustness [22].

Chapter 3

Materials and Methods

3.1 Data Acquisition Protocol

This study utilized three datasets: the Molinette dataset, the PC-GITA dataset, and the Bari dataset, each containing speech recordings from PD patients, with the latter two also including healthy controls. Detailed descriptions of the datasets, including participant demographics, recording protocols, and task types, are provided in Sections 3.1.1 and 3.1.2.

- **Molinette Dataset**: Recordings from 14 PD patients performing sustained vowels, proverbs, and monologues. Data include the ON and OFF phases for six patients (Section 3.1.1).
- **PC-GITA Dataset**: Recordings from 50 PD patients and 50 controls, including tasks such as vowels, DDK, sentences, and monologues (Section 3.1.2).
- **Bari Dataset**: Recordings from 28 PD patients and 22 controls performing vowels, syllables, and phonemically balanced text reading (Section 3.1.2).

The datasets were used for deep learning classification (sustained vowels), WER and CER analysis (sentences, proverbs, and text), and statistical analysis (UPDRS scores). Data preprocessing and feature extraction are described in Section 3.2.1.

3.1.1 Voice Recording Protocol at Molinette Hospital

The voice recording protocol at Molinette Hospital was designed to ensure consistency in data collection across all subjects. Recordings were performed using a smartphone (Samsung Galaxy S5 Mini) equipped with a voice recording application (Registratore Vocale, Version 3.28 [31]), configured to record at a sampling rate of

44.1 kHz in 16-bit PCM (WAV format). The smartphone was placed approximately 20-30 cm from the speaker's mouth and held by hand during the recording. All sessions took place in the *Neurology Department* of Molinette Hospital.

The study involved two groups of patients: six patients recorded at Molinette Hospital, consisting of four men and two women aged between 57 and 66 years, and eight additional patients recorded at the Politecnico di Torino, under the same recording conditions and using the same protocol as at Molinette. The second group consisted of five men and three women, aged between 58 and 77 years. All patients at Molinette were recorded in both OFF and ON phases, while patients at the Politecnico were recorded exclusively in the ON phase after their standard daily medication dose.

As part of the pre-operative evaluation for DBS surgery, Molinette patients underwent a detailed clinical assessment, which included the administration of the UPDRS questionnaire. This study focuses on subitem 3.1 of the UPDRS, which evaluates the quality of speech. For Molinette patients, UPDRS scores were collected during both the OFF and ON medication phases, while for Politecnico patients, these scores were recorded only in the ON phase.

The audio recordings were carried out under controlled conditions to ensure consistency across sessions. Each participant was seated in a comfortable position, with their back supported, arms resting, and feet flat on the ground. They were asked to maintain an upright posture to avoid any restriction of the diaphragm, which could affect their vocal performance. Recordings were conducted in a quiet room to minimize background noise, and the speaker was positioned centrally to reduce sound reflections that might compromise the quality of the recordings.

Before each session, the recording device was prepared by setting up the application and positioning the smartphone at an appropriate distance from the participant. Participants were given a trial run to familiarize themselves with the task and ensure they understood the instructions clearly. To avoid unnecessary sounds during the recordings, non-verbal cues, such as hand signals, were used to indicate the start of each recording.

Recorded Exercises:
Each patient completed the following exercises:

- **Vocalization**: Patients sustained the vowel sound /a/ for as long as possible (labeled A1, A2 for each phase for Molinette patients).
- **Proverbs**: Patients recited two proverbs, with each proverb recorded twice during both the OFF and ON phases for Molinette patients:
  - Proverb 1: "Meglio soli che male accompagnati" (labeled P1a, P1b for each phase).

Table 3.1: Demographic data and UPDRS scores for patients in the Molinette and Politecnico di Torino studies, including gender, age, and UPDRS scores in the OFF and ON phases

Patients Recorded at Molinette Hospital

ID	Gender	Age	UPDRS (ON)	UPDRS (OFF)
1	M	66	0	1
2	M	61	2	2
3	M	60	2	2
4	F	58	0	1
5	F	60	0	0
6	M	57	0	1

Patients Recorded at Politecnico di Torino (ON Phase Only)

ID	Gender	Age	UPDRS (ON)
OP01	M	73	2
OP02	M	76	2
OP03	F	62	1
OP04	M	77	2
OP05	M	71	0
OP06	M	58	1
OP07	F	74	2
OP08	F	67	1

  - Proverb 2: "A caval donato non si guarda in bocca" (labeled P2a, P2b for each phase).

- **Monologue**: Patients delivered a short, spontaneous monologue lasting 15-30 seconds, recorded once for each phase for Molinette patients (labeled D for each phase).

The proverbs *"Meglio soli che male accompagnati"* and *"A caval donato non si guarda in bocca"* were selected for their simplicity and phonetic properties. Their structure includes plosive consonants (/p/, /b/, /k/, /t/), which are produced through a complete closure of the vocal tract followed by a release of air. These sounds are particularly sensitive to articulatory difficulties typical of Parkinsonian speech, such as incomplete closures or transformations into fricatives (spirantization), as highlighted in Section 1.2.2 [15].

Moreover, the choice of proverbs ensures that participants do not need to rely on reading skills, making the task accessible even to those with literacy difficulties or reading impairments. This approach minimizes variability and ensures that all patients, regardless of their reading ability, can complete the task consistently [11].

For the Molinette group, all exercises, with the exception of the monologue, were recorded twice during each phase, resulting in two recordings per phase for each exercise. The Politecnico group followed a similar protocol but was recorded only during the ON phase. This produced an additional 16 recordings of sustained vowels, 16 recordings of each proverb, and 8 recordings of the monologue. Each audio file was labeled systematically to include the patient ID, the exercise type, and the phase of the recording. For example, the label `P1_OFF_A1` represents the first sustained vowel recording for Patient 1 during the OFF phase.

3.1.2 Description of Datasets Used (PC-GITA, Bari, Molinette)

PC-GITA Dataset

The PC-GITA dataset contains recordings from 50 PD patients and 50 healthy controls, with an even distribution of gender. The ages of the participants range from 33 to 77 years for PD patients (mean age: 62.2 ± 11.2 for men, 60.1 ± 7.8 for women) and from 31 to 86 years for healthy controls (mean age: 61.2 ± 11.3 for men, 60.7 ± 7.7 for women). Recordings were conducted in a noise-controlled environment at Clínica Noel in Medellín, Colombia [32].

The following tasks were performed by participants:

- Sustained phonation of all vowels: /a/, /e/, /i/, /o/, /u/.
- Repetition of DDK sequences (e.g., /pa-ta-ka/) (see Table 3.3).
- Reading of phonemically balanced sentences (see Table 3.2).
- Spontaneous monologue describing daily activities.
- Pronunciation of words (see Table 3.4).

The total number of recordings in the PC-GITA dataset is 6090 (see Table 3.5), with tasks such as sustained vowels being used for machine learning classification, and sentences, words, and DDK sequences being used for WER and CER analysis.

Recordings were made in a soundproof environment using professional-grade microphones connected to digital audio recorders. Participants were seated in a comfortable position with a microphone placed approximately 20 cm from their mouth, ensuring consistent sound quality across all recordings. Speech therapists supervised the recordings to ensure uniform execution of tasks [32].

Table 3.2: Phonemically balanced sentences used in the PC-GITA dataset, with their English translations.

Sentence (Spanish)	English Translation
Juan se rompió una pierna cuando iba en la moto.	Juan broke his leg while riding the motorcycle.
Los libros nuevos no caben en la mesa de la oficina.	The new books don't fit on the office table.
Laura sube al tren que pasa.	Laura gets on the passing train.
Luisa Rey compra el colchón duro que tanto le gusta.	Luisa Rey buys the hard mattress she likes so much.
Mi casa tiene tres cuartos.	My house has three rooms.
Omar, que vive cerca, trajo miel.	Omar, who lives nearby, brought honey.
Estoy muy preocupado, cada vez me es más difícil hablar.	I am very worried, it is becoming harder for me to speak.
Rosita Niño, que pinta bien, donó sus cuadros ayer.	Rosita Niño, who paints well, donated her paintings yesterday.
Estoy muy triste, ayer vi morir a un amigo.	I am very sad, yesterday I saw a friend die.
¿Viste las noticias? Yo vi ganar la medalla de plata en pesas. Ese muchacho tiene mucha fuerza!	Did you see the news? I saw him win the silver medal in weightlifting. That guy is very strong!

Table 3.3: DDK sequences used in the PC-GITA dataset. These sequences assess the rapid repetition of syllables to evaluate motor speech function.

DDK Sequences
ka-ka-ka
pa-pa-pa
pakata
pataka
petaka
ta-ta-ta

Bari Dataset

The Bari dataset contains recordings from 28 PD patients (19 men and 9 women) and 22 healthy elderly controls (10 men and 12 women). Participants' ages range

Table 3.4: Words used in the PC-GITA dataset, with their English translations.

Word (Spanish)	**English Translation**
apto, atleta, blusa, bodega, brazo, campana, caucho, clavo, coco, crema, drama, flecha, fruta, gato, globo, grito, llueve, ñame, pato, petaka, plato, presa, reina, trato, viaje	apt, athlete, blouse, warehouse, arm, bell, rubber, nail, coconut, cream, drama, arrow, fruit, cat, balloon, shout, it rains, yam, duck, suitcase, plate, dam, queen, deal, journey

Table 3.5: Number of recordings per task in the PC-GITA dataset.

Task	**Number of Recordings**
Sustained vowels	1472
Modulated vowels	500
Monologue	100
Read Text	117
Sentences	1002
Words	2299
DDK (syllables)	600

from 40 to 80 years for PD patients (mean age: 63.5 ± 10.3) and 60 to 77 years for controls (mean age: 68.1 ± 8.4).The Bari dataset is publicly available at [33].

The following tasks were performed by participants:

- Sustained phonation of vowels: /a/, /e/, /i/, /o/, /u/.
- Repetition of syllables (/pa/, /ta/) for DDK analysis.
- Reading of a phonemically balanced text (see Table 3.7).
- Pronunciation of specific words (see Table 3.6).

The total number of recordings in the Bari dataset is 870, including sustained vowel phonations, syllable repetitions, and readings of words and a phonemically balanced text. The sustained vowels were used for machine learning classification, while the phonemically balanced text, words, and syllables were analyzed for WER and CER.

The acquisition protocol for the Bari dataset is explained in [34]. Recordings were conducted in a quiet room at the Neurology Department of the Policlinico in Bari, Italy, under controlled conditions. Speech therapists supervised the sessions to

Table 3.6: Words used in the Bari dataset, with their English translations.

Word (Italian)	**English Translation**
pipa, buco, topo, dado, casa, gatto, filo, vaso, muro, neve, luna, rete, zero, scia, ciao, giro, sole, uomo, iuta, gnomo, glielo, pozzo, brodo, plagio, treno, classe, grigio, flotta, creta, drago, frate, spesa, stufa, scala, slitta, splende, strada, scrive, spruzzo, sgrido, sfregio, sdraio, sbrigo, prova, calendario, autobiografia, monotono, pericoloso, montagnoso, prestigioso	pipe, hole, mouse, die, house, cat, thread, vase, wall, snow, moon, net, zero, wake, hello, turn, sun, man, jute, gnome, give it to him, well, broth, plagiarism, train, class, gray, fleet, clay, dragon, friar, shopping, stove, staircase, sleigh, shines, road, writes, splash, scold, scar, deckchair, handle, proof, calendar, autobiography, monotonous, dangerous, mountainous, prestigious

ensure consistency among all participants. For Parkinson's disease patients, UPDRS scores were collected with the voice recordings, allowing for a more comprehensive analysis of the relationship between speech impairments and disease severity.

Molinette Dataset

The Molinette dataset was collected at Molinette Hospital in Turin and contains recordings from six PD patients (four men and two women), aged between 57 and 66 years.

In addition to the recordings made at Molinette Hospital, eight additional PD patients were recorded at the Politecnico di Torino, under the same recording conditions and using the same protocol as at Molinette Hospital. These eight patients were recorded during their daily ON phase (no high-dose medication was administered; rather, their standard daily dose), providing additional data for analysis. This brings the total number of patients in the dataset to 14.

The participants performed the following tasks:

- Sustained phonation of the vowel /a/.
- Reading of proverbs.
- Spontaneous monologue describing daily activities.

The acquisition protocol is described in Section 3.1.1.

The total number of recordings in the Molinette dataset, including those recorded at the Politecnico di Torino, is shown in Table 3.10. In this dataset, sustained

Table 3.7: Phonemically balanced text used in the Bari dataset, with its English translation.

Text (Italian)	**English Translation**
Il ramarro della zia. Il papà (o il babbo come dice il piccolo Dado) era sul letto. Sotto di lui, accanto al lago, sedeva Gigi, detto Ciccio, cocco della mamma e della nonna. Vicino ad un sasso c'era una rosa rosso vivo e lo sciocco, vedendola, la volle per la zia. La zia Lulù cercava zanzare per il suo ramarro, ma dato che era giugno (o luglio non so bene) non ne trovava. Trovò invece una rana che saltando dalla strada finì nel lago con un grande spruzzo. Sai che fifa, la zia! Lo schizzo bagnò il suo completo rosa che divenne giallo come un taxi. Passava di lì un signore cosmopolita di nome Sardanapalo Nabucodonosor che si innamorò della zia e la portò con sé in Afghanistan.	The aunt's lizard. Dad (or "babbo" as little Dado says) was on the bed. Below him, beside the lake, sat Gigi, nicknamed Ciccio, the apple of his mother's and grandmother's eye. Next to a rock was a bright red rose, and the fool, seeing it, wanted it for the aunt. Aunt Lulù was looking for mosquitoes for her lizard, but since it was June (or July, I'm not sure), she couldn't find any. Instead, she found a frog that, jumping from the road, landed in the lake with a big splash. What a fright for the aunt! The splash soaked her pink suit, which turned yellow like a taxi. A cosmopolitan man named Sardanapalo Nabucodonosor passed by, fell in love with the aunt, and took her with him to Afghanistan.

Table 3.8: Number of recordings per task in the Bari dataset.

Task	**Number of Recordings**
Sustained vowels	534
Syllable repetition	100
Read Text	124
Words	65
Sentences	47

vowels were used for machine learning classification, while proverbs and monologues were analyzed to calculate WER and CER.

Table 3.9: Proverbs used in the Molinette dataset with their English translations.

Proverb (Italian)	English Translation
Meglio soli che male accompagnati.	Better alone than in bad company.
A caval donato non si guarda in bocca.	Don't look a gift horse in the mouth.

Table 3.10: Total number of recordings per exercise for each patient of the Molinette dataset, including both ON and OFF phases (only ON for the eight patients recorded at Politecnico di Torino).

Exercise	Phase	N. of Recordings per Patient	Total Recordings
Vowel "a"	OFF	2	12
Vowel "a"	ON	2	28
Proverb 1	OFF	2	12
Proverb 1	ON	2	28
Proverb 2	OFF	2	12
Proverb 2	ON	2	28
Monologue	OFF	1	6
Monologue	ON	1	14

3.2 Word Error Rate and Character Error Rate Calculation Using Whisper

3.2.1 Preprocessing Pipeline for Speech Signals

The preprocessing pipeline for speech signals was designed to ensure consistency and optimize the quality of audio data across multiple datasets (Molinette, Bari, and PC-GITA). This process follows the methodology described by Favaro et al. in [35], which includes resampling, loudness normalization, and transcription extraction (Figure 3.1). The corresponding open-source code is available in the following repository.

```
$ git clone https://github.com/Neuro-Logical/speech.git
```

In the work of Favaro et al. [35], resampling was performed using SoX, and loudness normalization was applied using the EBU R128 standard through the `ffmpeg-normalize` library. This procedure ensures a uniform loudness level across the dataset. Furthermore, all spontaneous speech tasks were automatically transcribed using OpenAI Whisper [35].

Before beginning the automated processing, all recordings were manually reviewed to remove non-task-related speech. This included trimming any irrelevant speech at the start or end of the recordings to focus exclusively on the tasks at hand.

The key steps of the preprocessing pipeline are as follows:

- **Step 1: Resampling**
  The audio files from the PC-GITA and Molinette datasets were originally sampled at 44.1 kHz, while those from the Bari dataset were already sampled at 16 kHz. To ensure uniformity across all recordings, all files were resampled to 16 kHz. The resampling process was implemented using the `librosa` Python library, as demonstrated in the following code snippet:

```
# Resample the audio to the target sample rate
resampled_audio = librosa.resample(audio, orig_sr=sr, target_sr=16000)
```

- **Step 2: Loudness Normalization**
  After resampling, a loudness normalization step was performed to standardize the volume levels across all audio files. This was done to avoid any volume discrepancies that could affect ASR performance. The normalization was performed using the `ffmpeg-normalize` tool, which applies the EBU R128 loudness standard for more uniform audio levels compared to peak-based normalization methods. The following command was used in a batch script:

```
ffmpeg-normalize input_file.wav -o output_file.wav
```

- **Step 3: Speech-to-Text Conversion**
  The resampled and normalized audio files were then transcribed into text using the base version of the OpenAI Whisper model. For each dataset, the appropriate language setting (Italian for Bari and Molinette, Spanish for PC-GITA) was specified during transcription. The code used for the transcription process is shown below:

```
# Loading the Whisper model
model = whisper.load_model("base")

# Transcribing the audio files
result = model.transcribe(str(path), language=language)
```

The transcriptions were saved in `.txt` format for each corresponding audio file and used in the subsequent WER and CER calculations.

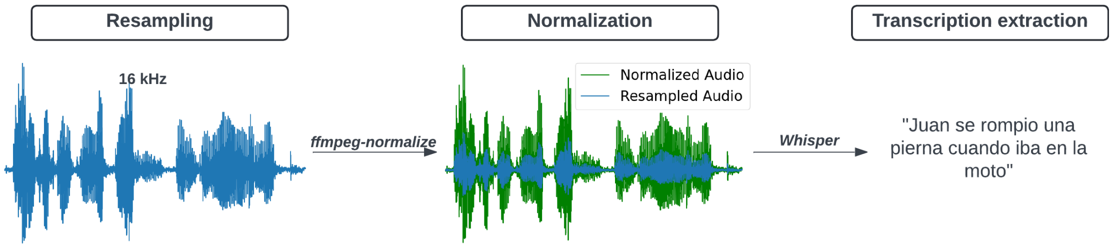


Figure 3.1: Overview of the preprocessing pipeline for audio data used in the classification task.

3.2.2 Calculation and Analysis of Word Error Rate and Character Error Rate

The WER and CER are widely used metrics for assessing the performance of ASR systems [28]. These metrics are calculated by comparing the transcription generated by the ASR system (hypothesis) with the corresponding correct transcription (reference). In this work, Whisper is used as a gold standard, and WER and CER are applied to quantify the pronunciation accuracy of Parkinsonian patients. The goal is not to evaluate Whisper's performance but to analyze the speech production errors in patients with Parkinson's disease, which are expected to increase with disease severity.

WER is calculated by determining the minimum number of word-level substitutions, deletions, and insertions required to transform the hypothesis into the reference. The formula is as follows:

$$WER = \frac{S + D + I}{N} \tag{3.1}$$

where:

- S is the number of word substitutions,
- D is the number of word deletions,
- I is the number of word insertions, and
- N is the total number of words in the reference.

CER operates on a similar principle but is applied at the character level. It is calculated as:

$$CER = \frac{S + D + I}{C} \tag{3.2}$$

where:

- S is the number of character substitutions,
- D is the number of character deletions,
- I is the number of character insertions, and
- C is the total number of characters in the reference.

The `jiwer` Python library was used to calculate both WER and CER across the different datasets. The reference transcriptions were sourced from the following studies:

- For the Bari dataset, the reference transcriptions were obtained from the work detailed in [34], which describes the tasks and sentences used.
- For the PC-GITA dataset, the reference phrases were taken from the dataset documentation in [32], where the speech assessments are explained.
- For the Molinette dataset, the reference transcriptions were based on the phrases recorded during the speech protocol, as described in Section 3.1.1.

WER and CER were calculated separately for control subjects and Parkinsonian patients across all datasets, enabling a direct comparison of pronunciation accuracy between the two groups. For the Bari dataset, the analysis was further stratified by UPDRS levels associated with speech impairment. This allowed for a closer examination of how increasing speech impairments, as reflected by higher UPDRS scores, influenced transcription accuracy.

3.2.3 Statistical Analysis of Word Error Rate and Character Error Rate

Outliers were removed from the WER and CER data using Tukey's Interquartile Range (IQR) method [36]. This approach excludes data points that fall outside 1.5 times the interquartile range from the first and third quartiles, reducing the influence of extreme values on the analysis. The IQR is calculated as follows:

$$IQR = Q_3 - Q_1 \tag{3.3}$$

where Q_1 is the first quartile (the 25th percentile) and Q_3 is the third quartile (the 75th percentile). Data points that fall below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$ were removed as outliers. This filtering ensures that subsequent statistical analyses are not disproportionately affected by outlier values.

After removing outliers, the cumulative WER and CER values were calculated separately for control and Parkinsonian subjects. After that, statistical analyses were conducted to determine whether the differences in WER and CER between the two groups were significant. These analyses were performed at a significance level ($\alpha = 0.05$) and followed the steps outlined below.

- **Shapiro-Wilk Test for Normality**: The Shapiro-Wilk test [37] was used to assess the normality of the data distributions within each group. This test is necessary to determine whether parametric or non-parametric methods are appropriate for comparing control and Parkinsonian groups.

- **Parametric vs. Non-Parametric Testing**: Depending on the Shapiro-Wilk test results, an unpaired statistical test was chosen for each metric. If both groups displayed a normal distribution ($p > 0.05$), an unpaired t-test [38] was applied to compare group means. Otherwise, if at least one group did not follow a normal distribution ($p < 0.05$), a Mann-Whitney U test [39] was used as a non-parametric alternative.

- **Hypotheses**: The null hypothesis (H_0) assumed no difference in WER or CER between controls and Parkinsonian patients, while the alternative hypothesis (H_1) indicated a significant difference.

These statistical analyses were conducted across all three datasets (PC-GITA, Bari, and Molinette). However, further analysis of WER and CER values according to UPDRS levels was only possible for the Bari and Molinette datasets, as the PC-GITA dataset does not contain UPDRS scores.

UPDRS Group Comparisons for Bari and Molinette Datasets

The Bari and Molinette datasets allowed for an additional level of analysis based on UPDRS scores, enabling an examination of potential trends in WER and CER values as a function of disease severity. The following steps were performed for each metric, divided by UPDRS levels:

- **Normality Assessment**: For each UPDRS group, the Shapiro-Wilk test was again applied to verify normality. This assessment was necessary to choose between parametric and non-parametric approaches to detect differences among multiple UPDRS levels.

- **One-Way ANOVA or Kruskal-Wallis Test**: Based on the normality results, different tests were used to analyze WER and CER values across UPDRS levels. If all groups exhibited normal distributions, a one-way ANOVA [40] was conducted to compare mean values across UPDRS levels. Conversely, if at least one group deviated from normality, the Kruskal-Wallis test [41] was employed as a non-parametric alternative.
- **Dunn's Post Hoc Test**: If the Kruskal-Wallis test indicated significant differences ($p < 0.05$), Dunn's post hoc test [42] was applied to compare specific UPDRS levels. This test enabled pairwise comparisons while accounting for the non-normal distribution of the groups.
- **Hypotheses**: For both the ANOVA and Kruskal-Wallis tests, the null hypothesis (H_0) posited no differences in WER or CER across the UPDRS levels, while the alternative hypothesis (H_1) indicated significant differences between these groups.

All analyses were performed using Python libraries. Statistical tests were carried out with SciPy, while Dunn's post hoc test was conducted using `scikit-posthocs`. These tools provided a systematic approach for evaluating differences between control, Parkinsonian, and stratified UPDRS groups. An overview of the statistical procedures is presented in Figure 3.2.

3.3 Classification Task

The aim of this classification task was to distinguish between healthy controls and Parkinsonian patients using vocal recordings from the Bari, PC-GITA and Molinette datasets. To accomplish this, two primary transformer-based models were used: the Vision Transformer (ViT) and the Audio Spectrogram Transformer (AST). Each model was implemented in two variations—a standard version and a parallel-processing version—resulting in four distinct configurations: `ViT`, `AST`, `ViT_parallel`, and `AST_parallel`. In the parallel versions, audio data was divided into segments (first and last halves) to improve model sensitivity to distinct portions of each recording (see Section 3.5 for more details).

The 5-fold cross-validation procedure, detailed in Section 3.6, was applied uniformly across all four configurations. This approach was used to ensure model robustness and to evaluate generalizability on different subsets of data. Performance was assessed using metrics such as accuracy, precision, recall, and F1-score.

Despite these shared elements, the preprocessing steps differed slightly between ViT and AST to accommodate the unique input requirements of each model. ViT utilized mel-spectrograms converted into RGB images as input (detailed in Section 3.4.2), while AST operated directly on raw audio waveforms (see Section 3.4.3).

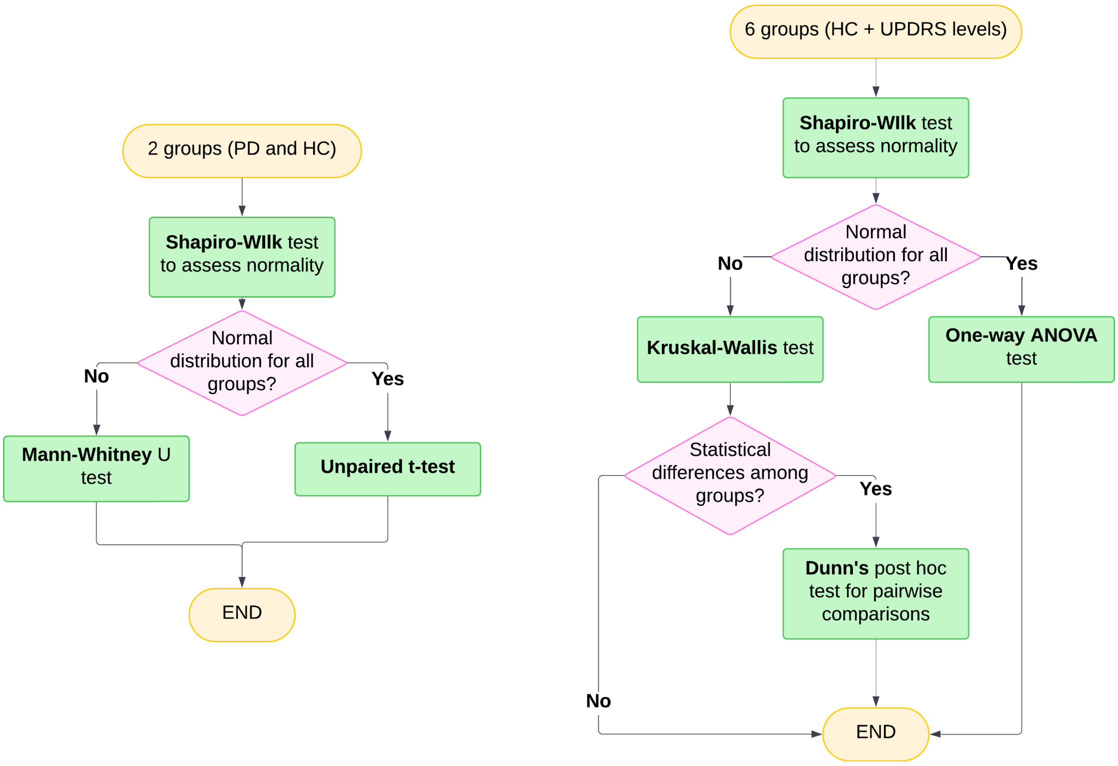

2 groups (PD and HC)
Shapiro-WIlk test to assess normality
Normal distribution for all groups?
No
Yes
Mann-Whitney U test
Unpaired t-test
END
6 groups (HC + UPDRS levels)
Shapiro-WIlk test to assess normality
Normal distribution for all groups?
No
Yes
Kruskal-Wallis test
One-way ANOVA test
Statistical differences among groups?
Yes
Dunn's post hoc test for pairwise comparisons
No
END

Figure 3.2: Flowcharts summarizing the statistical analysis steps applied to WER and CER metrics, including procedures for comparing two groups (controls and Parkinsonian patients) and six groups (control and UPDRS levels).

Detailed descriptions of these preprocessing pipelines are provided in the following sections.

3.4 Preprocessing Pipeline

The audio data underwent a series of preprocessing steps before being used for model training. These steps were designed to ensure the uniformity and quality of the audio recordings. Some preprocessing stages were common to both the ViT and AST models, while others were tailored to the specific requirements of each model.

3.4.1 Common Preprocessing Steps

The following steps were applied to all audio recordings:

- **Silence Trimming**: The leading and trailing silences in the audio files were removed using the `librosa.effects.trim` function with a 20 dB threshold. This ensured that the focus was solely on the vocal content.

- **Outlier Removal**: Audio files with unusually long or short durations were removed using the Interquartile Range (IQR) method to ensure a consistent dataset. Initially, the dataset contained 1997 files, with lengths ranging from 0.36 to 34.66 seconds. After removing outliers, 1841 audio files remained, with a final length range of 0.36 to 11.73 seconds (Figure 3.3).
- **Trimming**: Two different trimming strategies were applied:
  - **Non-parallel approach**: The central portion of each audio file was retained, with the length being equal to that of the shortest audio after outlier removal (0.36 seconds).
  - **Parallel approach**: Each audio was split into two segments—the first and the last halves—both having a length equal to half of the shortest audio (0.18 seconds).

These common preprocessing steps are summarized in Figure 3.4.

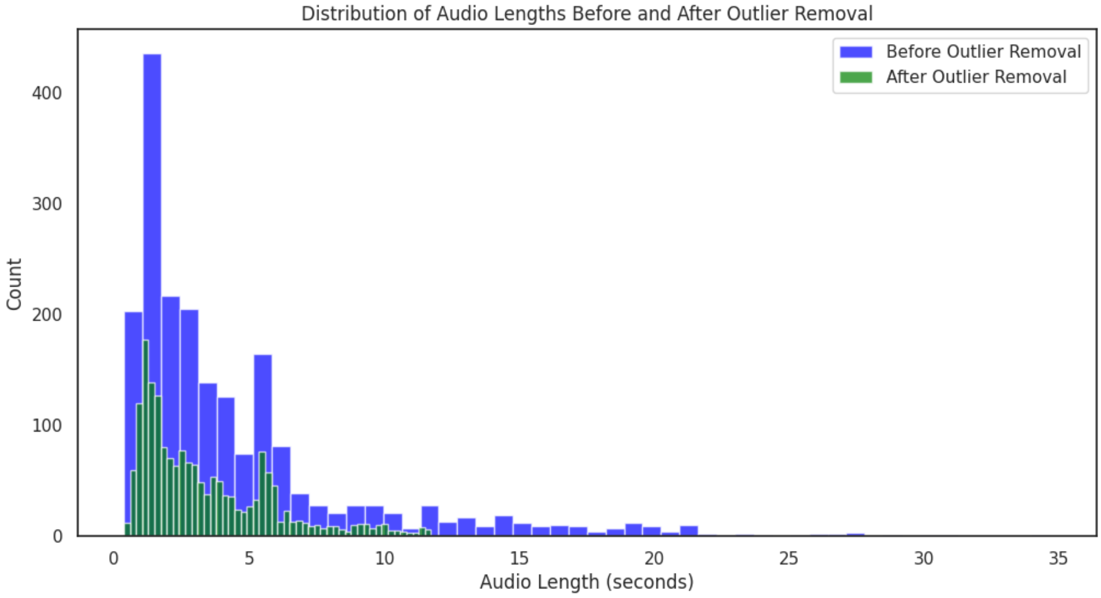


Figure 3.3: Distribution of audio lengths before and after outlier removal.

3.4.2 Vision Transformer Specific Preprocessing

For the ViT model, additional preprocessing steps were required to adapt the audio data into a format suitable for input:

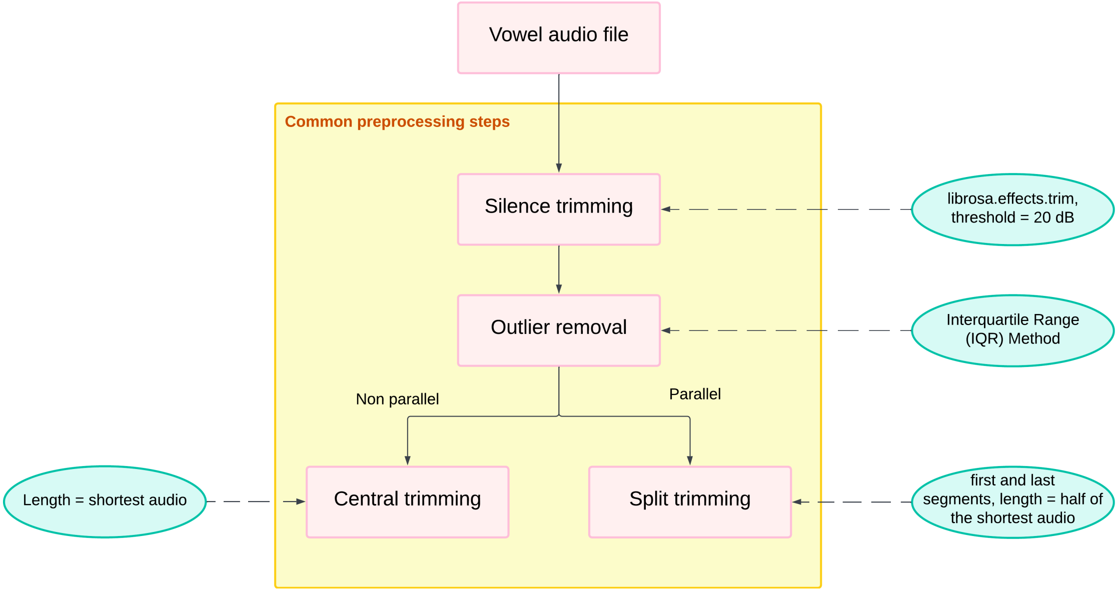

Vowel audio file
Common preprocessing steps
Silence trimming
librosa.effects.trim, threshold = 20 dB
Outlier removal
Interquartile Range (IQR) Method
Non parallel
Parallel
Length = shortest audio
Central trimming
Split trimming
first and last segments, length = half of the shortest audio

Figure 3.4: Overview of common preprocessing steps applied to all audio recordings, including silence trimming, outlier removal, and two possible trimming strategies: central trimming for the non-parallel approach and splitting into two segments for the parallel approach.

- **Mel-Spectrogram Generation**: The audio recordings were transformed into mel-spectrograms using a window length of 2048 samples and a hop length of 512 samples, which are the default parameters of the `librosa.feature.melspectrogram` function from the `librosa` library [43]. These default values are widely adopted in audio signal processing and provide a compromise between temporal and frequency resolution, suitable for capturing speech features in the time-frequency domain. The `librosa` library itself is a robust Python package for music and audio analysis, designed to support reproducibility and advanced audio processing [44].

- **RGB Conversion**: Since ViT requires a 3-channel input designed for image data, the mel-spectrogram was resized to 224 × 224 pixels and normalized to the range $[0, 1]$. To meet the model's input requirements, the spectrogram was replicated across three channels, producing an RGB image compatible with the ViT architecture.

- **Data Augmentation**: To improve generalization and reduce overfitting, the mel-spectrograms were converted into PyTorch tensors and normalized with a mean and standard deviation of 0.5. Additionally, random horizontal flips and random rotations up to 5 degrees were applied. These augmentations introduced slight variations in the data, encouraging the model to learn more robust features.

The specific preprocessing steps for ViT are illustrated in Figure 3.5.

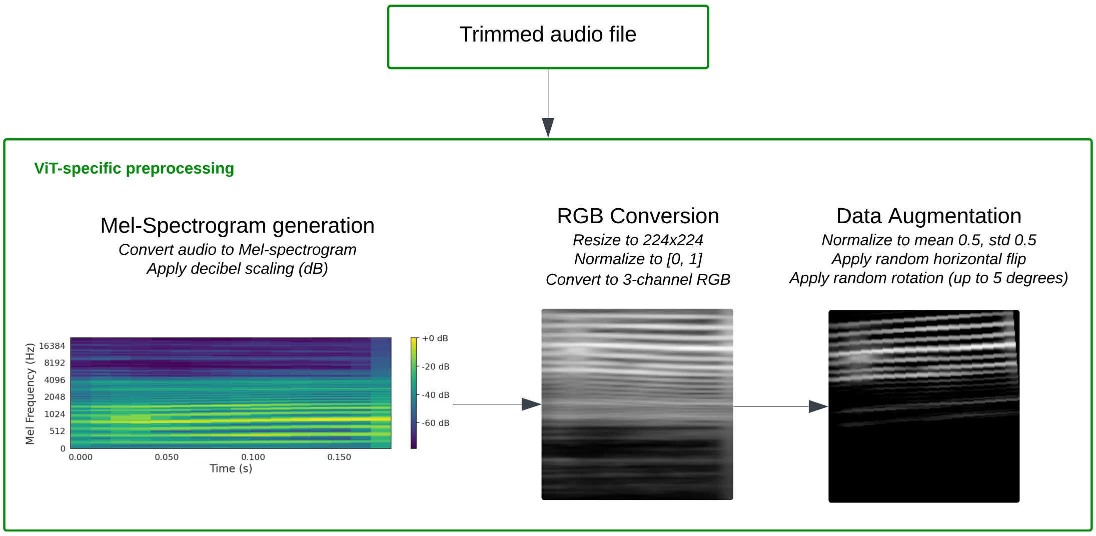

Trimmed audio file
ViT-specific preprocessing
Mel-Spectrogram generation
Convert audio to Mel-spectrogram
Apply decibel scaling (dB)
Mel Frequency (Hz)
16384
8192
4096
2048
1024
512
0
0.000
0.050
0.100
0.150
Time (s)
+0 dB
-20 dB
-40 dB
-60 dB
RGB Conversion
Resize to 224x224
Normalize to [0, 1]
Convert to 3-channel RGB
Data Augmentation
Normalize to mean 0.5, std 0.5
Apply random horizontal flip
Apply random rotation (up to 5 degrees)

Figure 3.5: Overview of the ViT-specific preprocessing pipeline, including mel-spectrogram generation, RGB conversion, and data augmentation.

3.4.3 Audio Spectrogram Transformer Specific Preprocessing

The AST model, designed to handle raw audio waveforms, followed a distinct preprocessing approach compared to ViT. The preprocessing steps for AST are managed by the `ASTFeatureExtractor` from the Hugging Face Transformers library, which prepares the raw audio for input into the model:

- **Normalization**: The raw audio waveform was normalized to a range of $[-1, 1]$ to ensure consistent amplitude levels across recordings. This step was explicitly implemented in the preprocessing pipeline to standardize the audio data.

- **Patch Embedding**: Unlike ViT, AST processes raw waveforms directly by segmenting the audio into overlapping patches. Each patch is linearly projected into embeddings that serve as input to the transformer encoder. This process is automatically handled by the `ASTFeatureExtractor` and allows AST to capture long-range temporal dependencies in the audio data without requiring any transformation into spectrograms or images [20].

The AST model works directly with raw audio inputs, removing the need to create spectrograms. This simplifies the preprocessing steps and allows the model to analyze the original waveform data without additional transformations.

3.5 Parallel Processing of Audio

In addition to the standard preprocessing pipeline, a parallel processing approach was explored. The objective was to determine whether processing separate portions of the audio could improve classification performance by capturing additional temporal features from distinct parts of the recording. In this approach, each audio file was divided into two segments, allowing the model to analyze different parts of the vocal recording independently before combining the results. This parallel processing approach aimed to enhance the model's ability to capture variations in vocal patterns, which may be useful in distinguishing between control and Parkinsonian speech.

The steps in this approach were as follows:

- **Segment Division and Model Processing**: Each audio file was divided into two segments, specifically the beginning and the end of the recording. Each segment was trimmed to ensure that its length equaled half of the shortest audio file in the dataset (0.18 seconds). For both the ViT and AST models, each half was processed independently, and the outputs (logits) of each segment were averaged to produce the final classification result. For ViT, each segment was first transformed into a mel-spectrogram, while AST processed the raw audio segments directly. The following code snippet illustrates this process:

```
    with torch.no_grad():
        for inputs_first, inputs_last, labels in val_loader:
            # Move data to device
            inputs_first, inputs_last = inputs_first.to(device), inputs_last.to(device)
            labels = labels.to(device)

            # Forward pass for each half
            outputs_first = model(inputs_first).logits  # \gls{ViT} uses mel-spectrograms; \gls{AST} uses raw audio
            outputs_last = model(inputs_last).logits

            # Combine outputs by averaging
            outputs_combined = (outputs_first + outputs_last) / 2

```

If dropout was applied, as in the case of the AST parallel model, it was added after averaging to improve generalization.

The parallel processing approach utilized the distinct information present in the first and last halves of each recording. By combining the outputs from both segments, this method aimed to enhance classification accuracy and robustness by capturing a wider range of features associated with Parkinsonian speech patterns.

3.6 Cross-Validation and Stratified Group Splitting

Given the relatively limited size of the dataset, a 5-fold cross-validation approach was implemented to robustly evaluate the model's performance. This method divides the data into five equally sized folds, iteratively using one fold (20% of the data) as the validation set and the remaining four folds (80% of the data) as the training set. This ensures that each sample is used exactly once for validation, providing a comprehensive evaluation across all data.

To maintain a balanced representation of Parkinsonian and control groups within each fold, a stratified group splitting function was employed. This function grouped the data by subject ID to ensure that all recordings from the same individual were included exclusively in either the training or validation set for a given fold. This strategy eliminates the risk of data leakage, where recordings from the same individual might appear in both training and validation sets, thereby introducing bias and inflating performance metrics.

The splitting function first divided the dataset into two groups based on class labels: controls and Parkinsonian subjects. Each group was randomly shuffled, and its samples were split into five roughly equal parts. For each fold, one part from each group was assigned to the validation set, while the remaining four parts were combined to create the training set. This approach ensured that the proportions of controls and Parkinsonian samples in each fold reflected the overall dataset distribution, maintaining class balance across both training and validation sets.

Additionally, the classifier was reinitialized at the start of each fold to prevent it from retaining any information or patterns from previous folds. This step ensured the independence of the folds and upheld the validity of the cross-validation process.

Figure 3.6 provides a visual representation of the 5-fold cross-validation process, highlighting these key considerations.

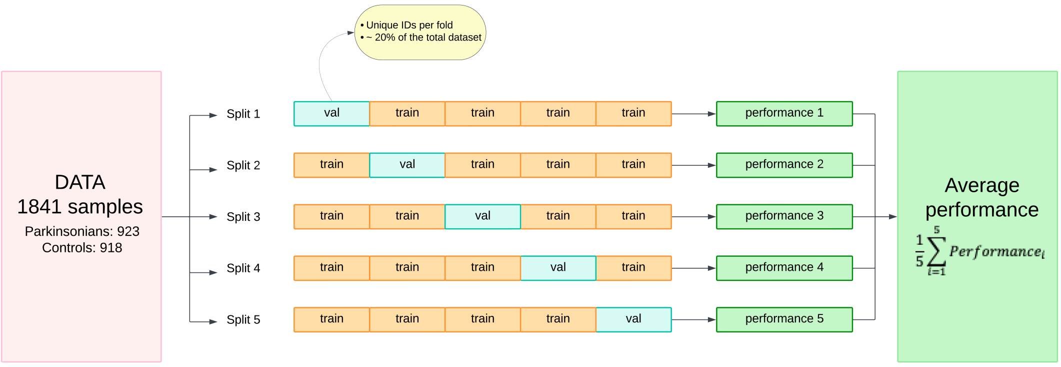


Figure 3.6: Diagram of 5-fold cross-validation with stratified group splitting to ensure balanced representation of control and Parkinsonian groups.

3.7 Evaluation Metrics

To assess the performance of the models, several evaluation metrics were calculated [45]. These metrics were computed for both the training and validation sets and averaged across the 5 cross-validation folds to ensure a robust assessment. Below are the definitions of the metrics used:

- **Accuracy**: The proportion of correctly classified samples among the total number of samples. It is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

  where TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives, respectively.

- **Precision**: The proportion of true positive predictions among all samples predicted as positive. It is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall (Sensitivity)**: The proportion of true positive samples that are correctly identified by the model. It reflects the model's ability to identify positive cases and is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure when both metrics are equally important. It is given by:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

For each fold, the metrics were computed at the end of every epoch for both the training and validation sets. To ensure a fair comparison and capture model generalizability, the final reported values represent the averages across the 5 folds.

3.8 Model Configurations and Hyperparameters

This section details the characteristics, architecture, and hyperparameters of the four model configurations used in this study: ViT, AST, ViT_parallel, and AST_parallel. Each configuration is tailored to optimize the classification task by capturing relevant features from vocal recordings of control and Parkinsonian groups.

3.8.1 Vision Transformer Model

The Vision Transformer (ViT) is a transformer-based architecture developed for image recognition tasks. In this study, the `vit-base-patch16-224` model was employed, pretrained on the ImageNet-21k dataset, which comprises 14 million images and 21,843 classes with an input resolution of 224x224 pixels [46, 47].

Initially introduced by Dosovitskiy et al. [48], the ViT model was subsequently adapted to PyTorch from JAX by Ross Wightman [47]. The architecture consists of a transformer encoder with 12 layers that process a sequence of embedded image patches, each of size 16x16 pixels. A special `[CLS]` token is prepended to the sequence for classification, while absolute position embeddings preserve spatial information.

In this implementation, the ViT model was fine-tuned for the binary classification task of distinguishing between healthy controls and Parkinsonian patients. The pretrained model, available through Hugging Face, does not include a pretrained classification head. Therefore, a custom linear classifier was added on top of the final hidden state of the `[CLS]` token. During training, only the classification head was left unfrozen to allow task-specific learning, while all other model parameters were frozen to retain the pretrained knowledge from ImageNet-21k.

The training process was configured to utilize the AdamW optimizer, which combines Adam's adaptive learning rates with decoupled weight decay to mitigate overfitting. The initial learning rate was set to 0.001, and an ExponentialLR scheduler with a decay factor of $\gamma = 0.995$ was employed to gradually reduce the learning rate during training. Early stopping was applied with a patience of 30 epochs to prevent overfitting.

The hyperparameters used for training the ViT model are summarized in Table 3.11.

Table 3.11: Hyperparameters used for training the ViT model.

Hyperparameter	**Value**
Batch Size	64
Learning Rate	0.001
Number of Epochs	100
Early Stopping Patience	30 epochs
Optimizer	AdamW (`weight decay`: 1e-5)
Scheduler	ExponentialLR (`gamma`: 0.995)

The ExponentialLR scheduler enabled a gradual adjustment of the learning rate, contributing to training stability and convergence. By leveraging these configurations, the ViT model effectively captured relevant features from the mel-spectrogram representations of vocal recordings.

3.8.2 Audio Spectrogram Transformer Model

The Audio Spectrogram Transformer (AST) is a model designed for audio classification tasks, introduced by Gong et al. [20]. It shares similarities with the (ViT) but is specifically adapted to process audio by transforming it into spectrogram images. This approach allows for detailed analysis of audio features. For this study, the `ast-finetuned-audioset-10-10-0.4593` model from Hugging Face was used, as it is pretrained on the AudioSet dataset, providing a strong starting point for audio classification tasks.

To adapt AST to classify between healthy controls and Parkinsonian patients, additional fine-tuning was performed. Only the last two transformer layers (layers 10 and 11) and the classification head were unfrozen, while the remaining layers were frozen to retain pretrained knowledge from AudioSet. This approach allowed the model to adjust to the binary classification requirements of this task without losing its pretrained audio analysis capabilities.

Table 3.12 summarizes the hyperparameters used for training. The optimizer chosen was AdamW with a learning rate scheduler, and early stopping was applied to prevent overfitting.

Fine-tuning these specific layers allowed AST to utilize its pretrained spectrogram analysis while effectively distinguishing between control and Parkinsonian groups.

Table 3.12: Hyperparameters used for training the AST model.

Hyperparameter	**Value**
Batch Size	64
Learning Rate	0.0005
Number of Epochs	100
Early Stopping Patience	30 epochs
Optimizer	AdamW (`weight decay`: 1e-5)
Scheduler	ExponentialLR (gamma: 0.995)

3.8.3 Parallel Vision Transformer Model

The parallel ViT model configuration (`ViT_parallel`) was developed to process audio recordings in two separate temporal segments: the first half and the last half of each sample. By analyzing these portions independently, the model aims to capture temporal variations that can improve its ability to differentiate between control and Parkinsonian speech patterns.

This configuration is based on the pretrained `vit-base-patch16-224` model from the ImageNet-21k dataset, as described in Section 3.8.1. For each segment, the corresponding spectrogram was fed into the model, which produced logits for both halves. The final classification result was determined by averaging the logits from the two segments, ensuring equal contribution from both parts of the audio.

To improve the model's ability to generalize, the following regularization techniques were applied:

- **Dropout**: A dropout rate of 0.2 was applied to both the last transformer block and the classifier head. This helped reduce overfitting by randomly dropping units during training, making the model more robust to variations in the training data.
- **Batch Normalization**: Batch normalization was incorporated into the classifier head to stabilize the training process and improve convergence. This technique helps mitigate internal covariate shifts by normalizing the activations.
- **Layer Normalization**: Layer normalization was applied to the output of the last transformer block to maintain well-scaled activations, improving gradient stability and aiding in the overall training process.

Fine-tuning in this setup was limited to the last transformer block and the classifier head, while all other layers were kept frozen to retain the pretrained weights. This selective approach allowed the model to adapt to the specific features of Parkinsonian speech while leveraging its pretrained knowledge from ImageNet-21k.

The hyperparameters used for training this configuration are detailed in Table 3.13. The training process utilized the AdamW optimizer with weight decay, alongside an ExponentialLR scheduler to reduce the learning rate by a factor of 0.95 per epoch. Early stopping based on validation loss was applied with a patience of 25 epochs to prevent overfitting and ensure efficient training.

Table 3.13: Hyperparameters for training the parallel ViT model.

Hyperparameter	**Value**
Batch Size	128
Learning Rate	0.0001
Number of Epochs	100
Early Stopping Patience	25 epochs
Optimizer	AdamW (`weight decay`: 1e-5)
Scheduler	ExponentialLR (gamma: 0.995)
Dropout Rate	0.2

3.8.4 Parallel Audio Spectrogram Transformer Model

The parallel version of the Audio Spectrogram Transformer (AST) was specifically designed to process audio data in two separate segments (the beginning and the end of each recording). This approach aims to capture temporal variations that could improve the classification between control and Parkinsonian groups.

The model builds upon the same AST architecture described in Section 3.8.2, leveraging its pretrained weights for audio classification. To adapt the model for the binary task, only the last two transformer layers (layers 10 and 11) and the classifier were unfrozen for fine-tuning, while all other layers were frozen to retain the pretrained knowledge from AudioSet. Each segment is processed independently through the network, with logits regularized using dropout layers set to a rate of 0.3. The final prediction is obtained by averaging the logits from both segments.

The training procedure made use of the AdamW optimizer, which combines Adam's adaptive learning rates with decoupled weight decay to enhance generalization and reduce overfitting. The learning rate was set to 10^{-4}, with a weight decay parameter of 0.05 to provide additional regularization. To ensure gradual and stable adjustments to the learning rate throughout training, an ExponentialLR scheduler with a decay factor of $\gamma = 0.995$ was implemented. Early stopping was incorporated with a patience of 10 epochs, allowing the training process to terminate early if the validation performance ceased to improve, avoiding overfitting. A summary of the hyperparameters used in this training process is provided in Table 3.14.

Table 3.14: Hyperparameters used for training the parallel AST model.

Hyperparameter	**Value**
Batch Size	16
Learning Rate	0.0001
Number of Epochs	100
Early Stopping Patience	10 epochs
Optimizer	AdamW (`weight decay`: 1e-5)
Scheduler	ExponentialLR (`gamma`: 0.995)
Dropout Rate	0.3

This configuration allowed the model to balance regularization and effective learning across cross-validation folds. The use of dropout, weight decay, and early stopping contributed to enhanced generalization performance, while the ExponentialLR scheduler ensured gradual and stable optimization throughout the training process.

3.9 UPDRS-based Classification

An additional analysis was carried out to study the relationship between UPDRS levels and the probability assigned by the model to the Parkinsonian class. These probabilities were derived from the model's logits prior to applying the binary classification threshold. The goal was to examine whether higher UPDRS levels corresponded to increased predicted probabilities, potentially reflecting a link between disease severity and the model's confidence in classification.

For each audio sample, the logits associated with the Parkinsonian class were transformed into probabilities using the softmax function. These probabilities were grouped according to UPDRS levels, enabling the calculation of average probabilities for each level, from UPDRS 0 to UPDRS 4. It is worth noting that UPDRS levels apply exclusively to Parkinsonian samples (label 1), while control samples (label 0) were evaluated separately.

For the Parkinsonian group, the analysis considered both the average predicted probabilities and the classification accuracy within each UPDRS level. Accuracy was determined by comparing predictions to true labels using a threshold of 0.5.

This analysis was restricted to datasets containing UPDRS annotations, specifically the Bari and Molinette datasets. Other datasets, such as PC-GITA, lacked these annotations, preventing a broader evaluation. This limitation underscores the importance of consistent and detailed metadata across datasets to enable a more comprehensive analysis.

The findings offer insight into how the model's predicted probabilities align with disease severity as indicated by UPDRS levels. However, the incomplete availability of UPDRS annotations limits the scope of this evaluation, highlighting the need for richer datasets in future research.

Chapter 4

Results

4.1 Speech Recognition Results

This section presents the results of the speech recognition analysis performed on the PC-GITA, Bari, and Molinette datasets. For each dataset, WER and CER values were calculated for both control subjects and Parkinsonian patients, as described in Chapter 3. Additionally, results from a combined analysis of the Bari and Molinette datasets are presented. Statistical tests were conducted to evaluate the significance of observed differences between the groups, with a significance level (α) set at 0.05.

4.1.1 PC-GITA Dataset Results

For the PC-GITA dataset, Tukey's method was applied to identify and remove outliers. Specifically, 16 outliers were excluded from the control group for WER, while 5 outliers were removed from the Parkinsonian group. Regarding CER, 47 outliers were excluded from the control group, and 30 from the Parkinsonian group. Although these outliers were not considered in the statistical analysis, they were included in the calculations of cumulative means, medians, and standard deviations.

As detailed in Tables 4.1 and 4.2, both WER and CER were found to be higher in the Parkinsonian group compared to the controls. The mean WER for Parkinsonian patients was 0.427, representing an increase of approximately 63% compared to the control group's mean of 0.262. Similarly, the mean CER for the Parkinsonian group was 0.172, more than double the mean value of 0.080 observed in the control group.

Figures 4.1 and 4.2 illustrate these differences with boxplots, showing the distribution of WER and CER for both control and Parkinsonian groups within the PC-GITA dataset. These visualizations highlight the disparity in transcription accuracy between the two groups, reflecting the impact of Parkinson's disease on speech clarity.

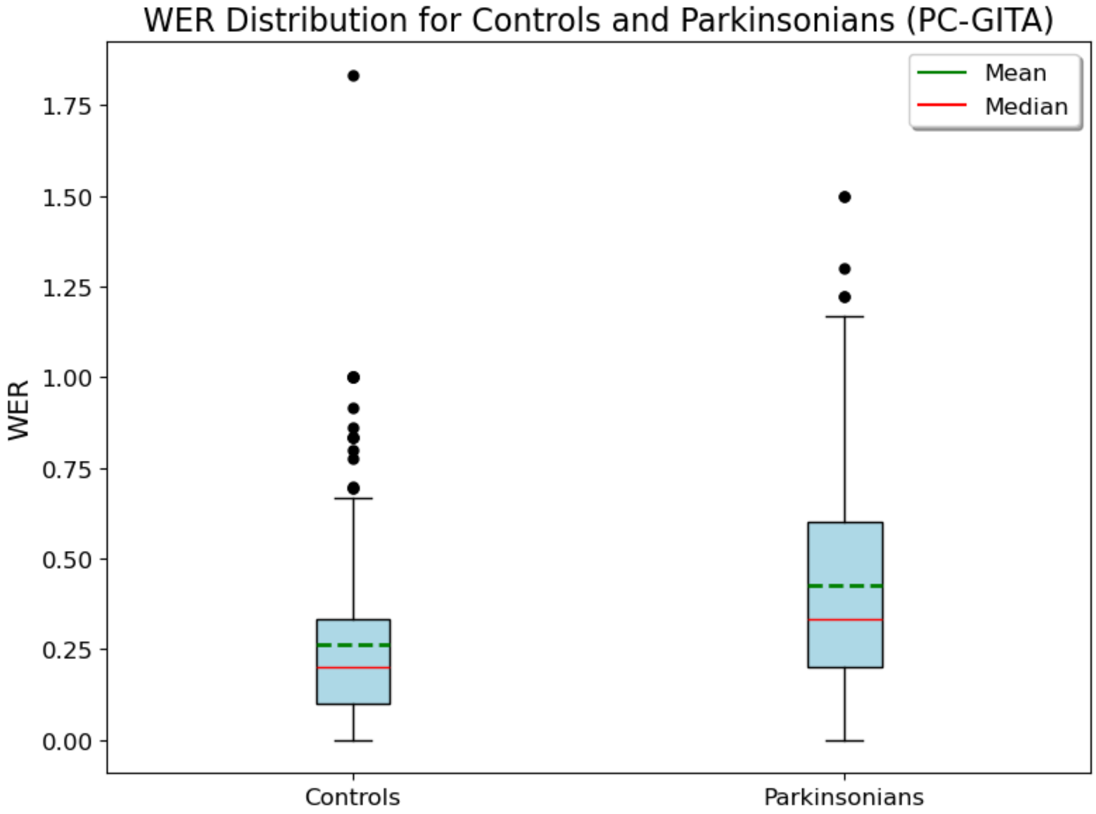


Figure 4.1: WER distribution for controls and Parkinsonians (PC-GITA). Sample sizes: Controls (n=484), Parkinsonians (n=500). Outliers were removed using Tukey's method (represented by the dots). Statistical significance was assessed using the Mann-Whitney U test (p-value=2.76e-21).

Group	WER Mean	WER Median	WER Standard Deviation
Controls	0.262	0.200	0.199
Parkinsonians	0.427	0.333	0.298

Table 4.1: Cumulative Means, Medians, and Standard Deviations for WER in the PC-GITA Dataset

As shown in Figures 4.1 and 4.2, the distribution of WER and CER reveals higher values for the Parkinsonian group compared to controls.

The Shapiro-Wilk test revealed that the distributions of both WER and CER were not normal (p-values < 0.05). Therefore, the Mann-Whitney U test was applied, confirming statistically significant differences between the control and

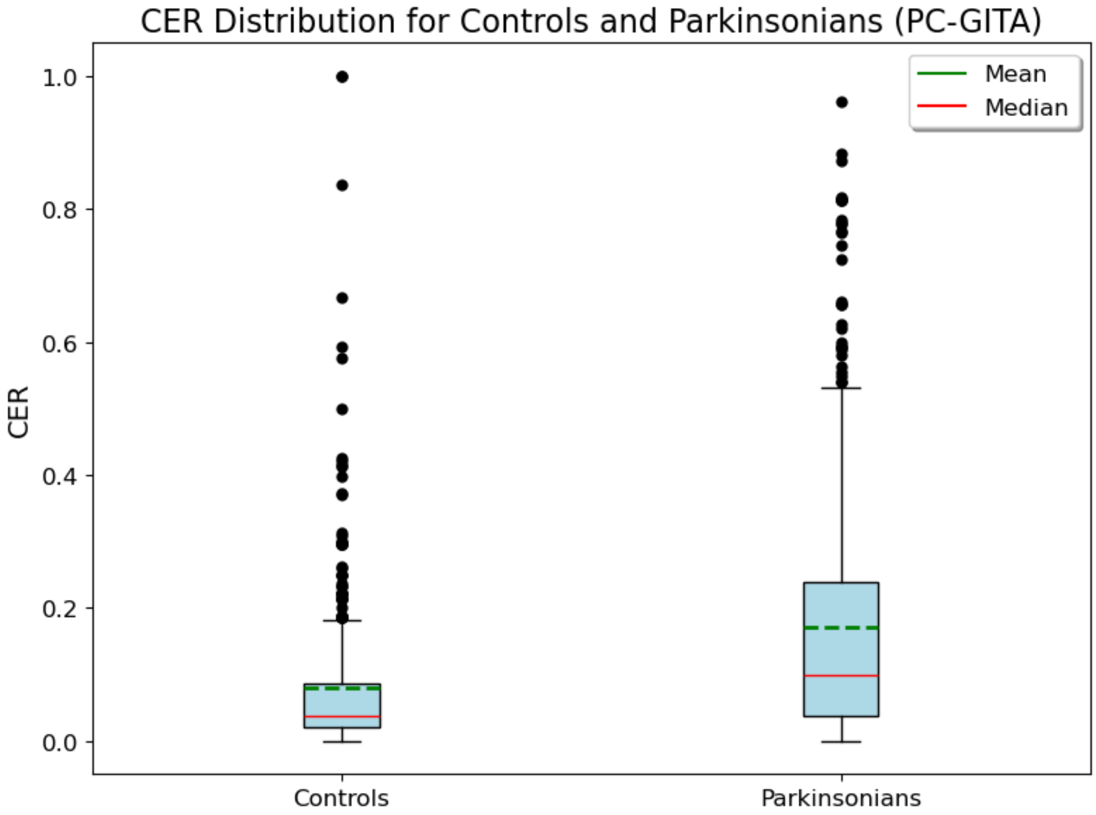


Figure 4.2: CER distribution for controls and Parkinsonians (PC-GITA). Sample sizes: Controls (n=453), Parkinsonians (n=475). Outliers were removed using Tukey's method (represented by the dots). The Mann-Whitney U test confirmed a statistically significant difference (p-value=4.36e-26).

Table 4.2: Cumulative Means, Medians, and Standard Deviations for CER in the PC-GITA Dataset

Group	**CER Mean**	**CER Median**	**CER Standard Deviation**
Controls	0.080	0.039	0.109
Parkinsonians	0.172	0.100	0.182

Parkinsonian groups for both WER (p = 2.76e-21) and CER (p = 4.36e-26).

4.1.2 Bari Dataset Results

The same analysis was conducted for the Bari dataset, where outliers were identified and removed using Tukey's method. Specifically, only one outlier was removed from the WER distribution of the Parkinsonian group, while no outliers were present in the CER distribution for either group, as shown in the boxplots in Figures 4.3 and 4.4. Similar to the PC-GITA dataset, Parkinsonian patients exhibited higher WER and CER compared to control subjects.

Figures 4.3 and 4.4 present the boxplots for WER and CER, respectively. These results show a pattern consistent with the PC-GITA dataset, where Parkinsonian patients exhibit higher error rates and greater variability compared to control subjects. This reflects the increased difficulty Parkinsonians face in maintaining accurate and consistent speech, as captured by these metrics.

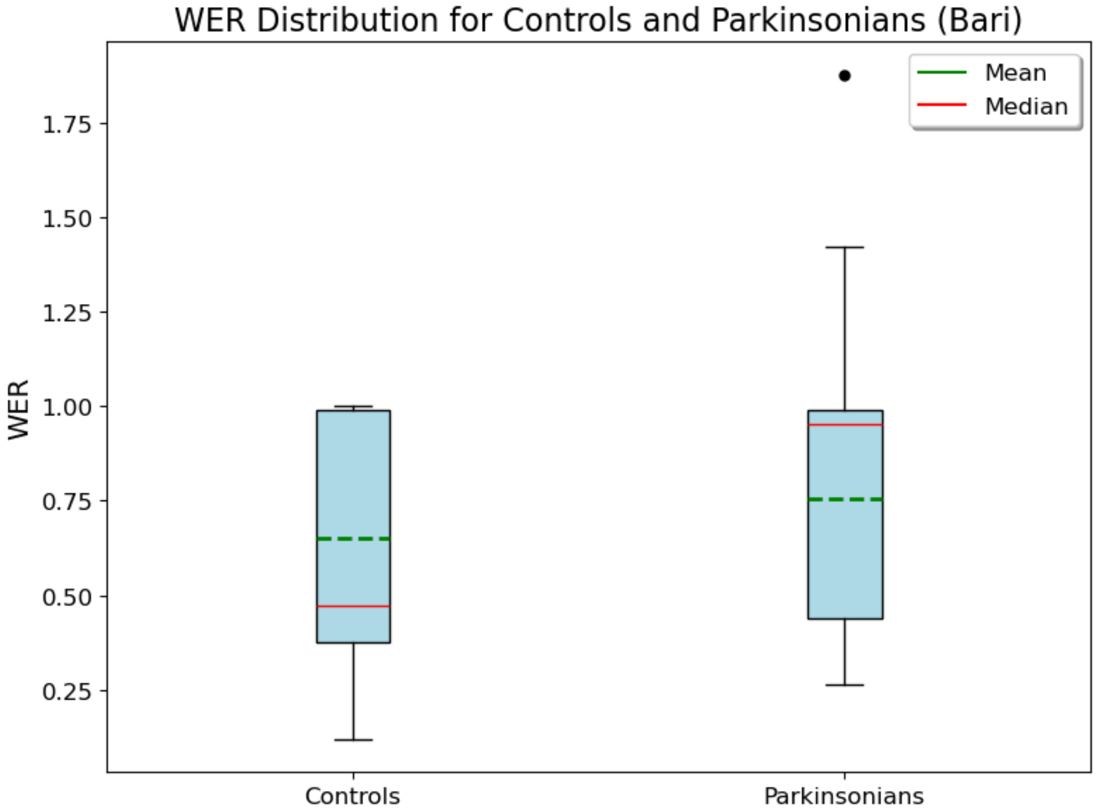


Figure 4.3: WER distribution for controls and Parkinsonians (Bari). Sample sizes: Controls (n=130), Parkinsonians (n=106). Outliers were removed using Tukey's method (represented by the dots). Statistical significance was assessed using the Mann-Whitney U test (p-value=5.27e-02).

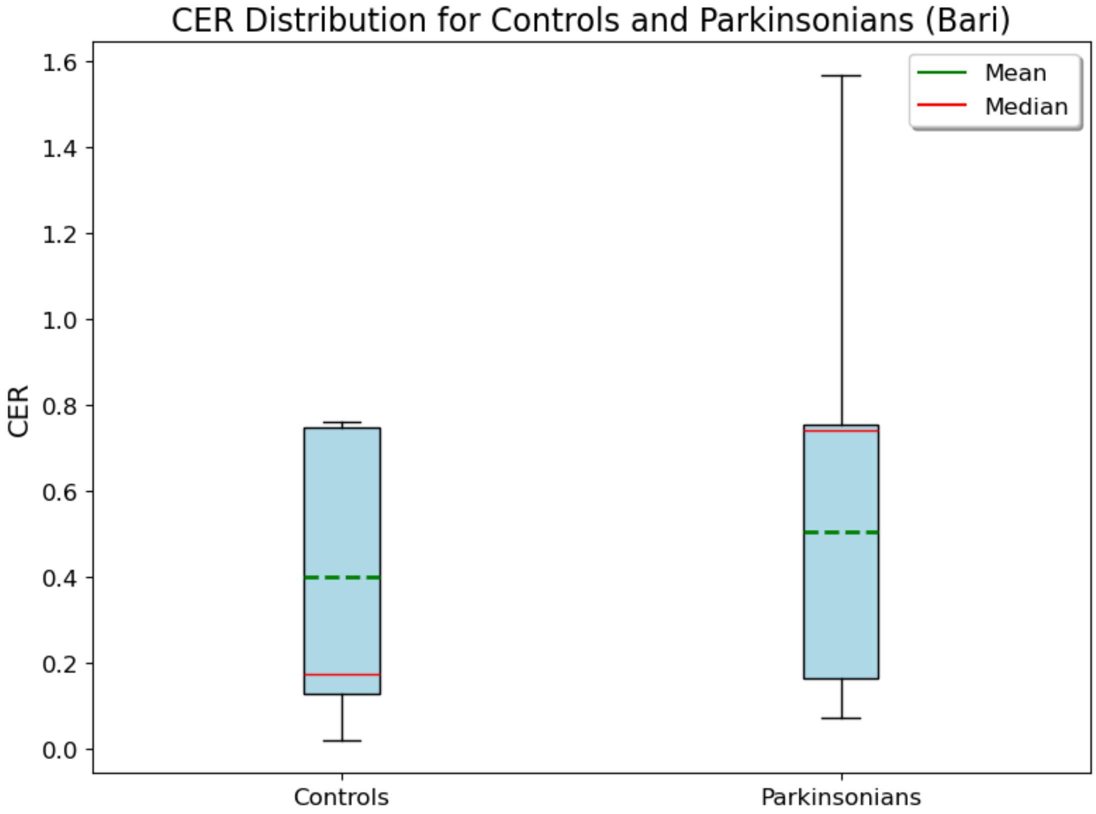


Figure 4.4: CER distribution for controls and Parkinsonians (Bari). Sample sizes: Controls (n=130), Parkinsonians (n=106). Outliers were removed using Tukey's method (represented by the dots). The Mann-Whitney U test confirmed a statistically significant difference (p-value=2.78e-03).

As shown in Tables 4.3 and 4.4, both WER and CER are higher in Parkinsonian patients compared to control subjects. The mean WER for Parkinsonians is 0.7534, approximately 16% higher than the control group's mean of 0.6498, while the median WER for Parkinsonians (0.9531) is more than twice that of controls (0.472).

Similarly, the mean CER for Parkinsonians is 0.5044, showing a 26% increase compared to the control group's mean of 0.400. The median CER for Parkinsonians (0.740) is also substantially higher than the control median (0.175).

Again, the Shapiro-Wilk test was performed to check the normality of the data distribution. Since the data did not follow a normal distribution (p-values < 0.05 for both groups), the Mann-Whitney U test was applied, confirming statistically significant differences between controls and Parkinsonians for CER ($p = 2.80e-03$). For WER, the p-value is 5.27e-02, which is slightly above the significance level

Table 4.3: Cumulative Means, Medians, and Standard Deviations for WER in the Bari Dataset

Group	WER Mean	WER Median	WER Standard Deviation
Controls	0.650	0.473	0.298
Parkinsonians	0.753	0.953	0.309

Table 4.4: Cumulative Means, Medians, and Standard Deviations for CER in the Bari Dataset

Group	CER Mean	CER Median	CER Standard Deviation
Controls	0.400	0.175	0.305
Parkinsonians	0.504	0.740	0.322

($\alpha = 0.05$), making the result borderline.

UPDRS Analysis in Bari Dataset

In the case of the Bari dataset, the availability of UPDRS scores allowed for a study of the correlation between these values and disease severity. Figures 4.5 and 4.6 show the distribution of WER and CER across control subjects and Parkinsonian patients with different UPDRS 3.1 scores.

As shown in Tables 4.5 and 4.6, both WER and CER show a clear upward trend as UPDRS levels increase. For instance, the mean WER for control subjects is 0.650, rising to 0.951 for UPDRS 4, representing a 46% increase. Similarly, the mean CER increases from 0.400 for controls to 0.682 for UPDRS 4, a rise of 70%. This pattern is consistent for the medians, where higher values are observed in more severe stages of Parkinson's disease.

Table 4.5: Cumulative Means, Medians, and Standard Deviations for WER by UPDRS Levels in the Bari Dataset

Group	WER Mean	WER Median	WER Standard Deviation
Control	0.650	0.473	0.298
UPDRS 0	0.720	0.836	0.270
UPDRS 1	0.728	0.961	0.361
UPDRS 2	0.809	0.953	0.314
UPDRS 3	0.909	0.961	0.103
UPDRS 4	0.951	0.949	0.031

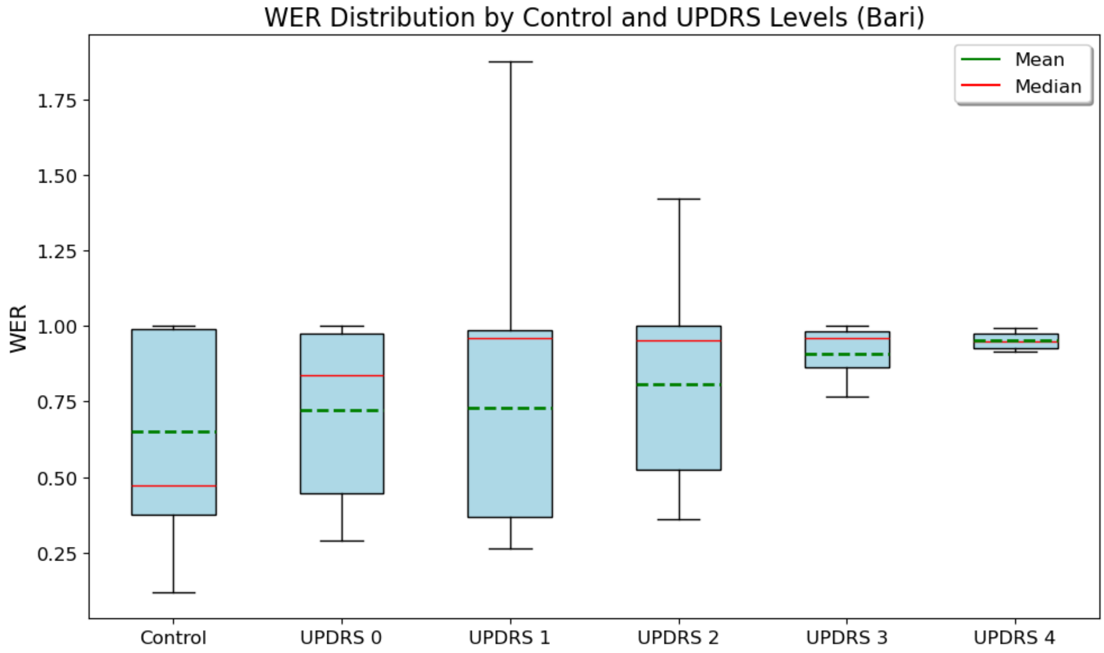


Figure 4.5: WER distribution by control and UPDRS levels (Bari dataset). Sample sizes: Control (n=130), UPDRS 0 (n=44), UPDRS 1 (n=35), UPDRS 2 (n=20), UPDRS 3 (n=3), and UPDRS 4 (n=4). Outliers were removed using Tukey's method.

Table 4.6: Cumulative Means, Medians, and Standard Deviations for CER by UPDRS Levels in the Bari Dataset

Group	**CER Mean**	**CER Median**	**CER Standard Deviation**
Control	0.400	0.175	0.305
UPDRS 0	0.462	0.536	0.292
UPDRS 1	0.486	0.740	0.334
UPDRS 2	0.571	0.743	0.376
UPDRS 3	0.666	0.762	0.142
UPDRS 4	0.682	0.691	0.104

The Shapiro-Wilk test revealed that the distributions for WER and CER did not follow a normal distribution for most of the UPDRS levels. Specifically, only the distributions for UPDRS 3 and UPDRS 4 showed normality. Given the non-normality of the other groups, the Kruskal-Wallis test was applied to evaluate significant differences across the UPDRS levels. The results revealed no statistically significant difference in WER (p = 0.226), while significant differences were observed

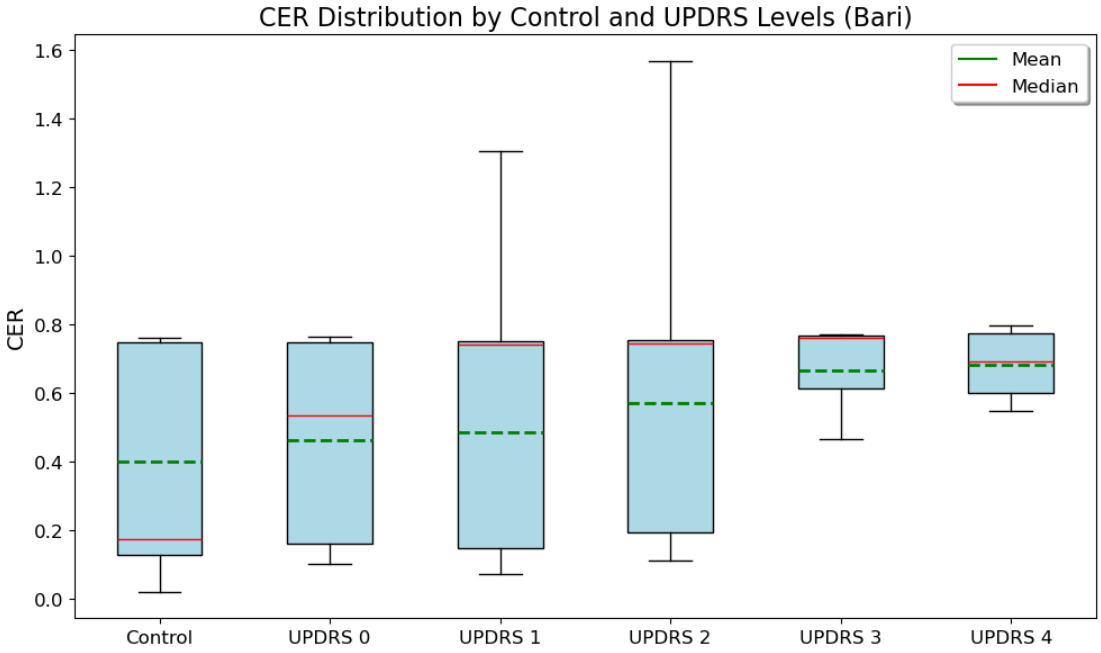


Figure 4.6: CER distribution by control and UPDRS levels (Bari dataset). Sample sizes: Control (n=130), UPDRS 0 (n=44), UPDRS 1 (n=35), UPDRS 2 (n=20), UPDRS 3 (n=3), and UPDRS 4 (n=4). Outliers were removed using Tukey's method.

for CER (p = 0.011), suggesting that CER varies significantly between UPDRS levels.

Following the Kruskal-Wallis test, Dunn's post hoc test was used to perform pairwise comparisons between the groups. For CER, significant differences were observed between Controls and UPDRS 2 (p = 0.017), Controls and UPDRS 3 (p = 0.029), and Controls and UPDRS 4 (p = 0.040). These findings suggest that higher UPDRS levels, indicating greater disease severity, are associated with increased CER values. In contrast, no significant differences were found for WER across the groups.

Figures 4.7 and 4.8 illustrate the mean and standard deviation of WER and CER for each UPDRS level. In Figure 4.8, the significant differences identified by Dunn's test are marked, highlighting the increased CER values in Parkinsonian patients with more advanced disease compared to controls.

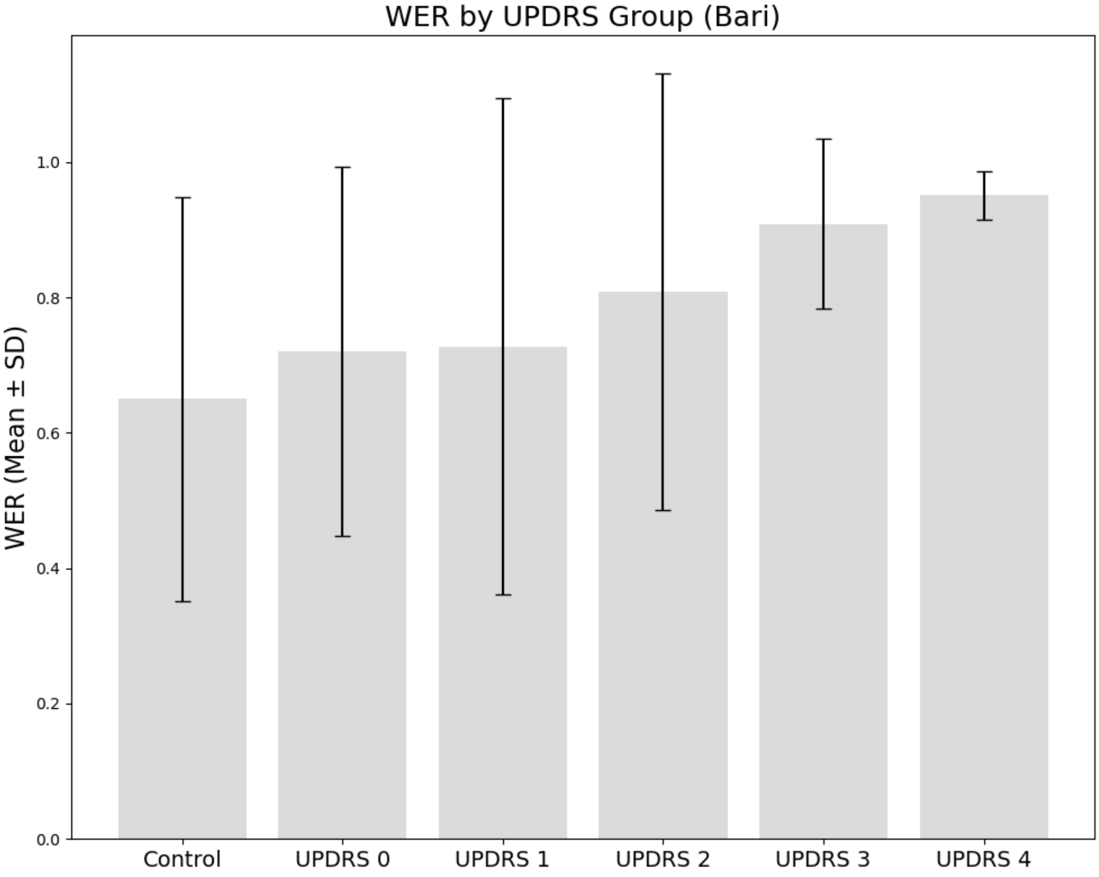


Figure 4.7: Dunn's post hoc test results for pairwise comparisons of WER across UPDRS levels in the Bari dataset. Values are measured as mean ± SD. Significance levels: *p $<$ 0.05, **p $<$ 0.01, ***p $<$ 0.001, ****p $<$ 0.0001.

4.1.3 Molinette Dataset Results

For the Molinette dataset, the analysis followed the same methodology as the other datasets, but it focused on Parkinsonian patients only, as there are no control subjects in this dataset. Parkinsonian patients were divided into 3 groups based on their UPDRS scores: UPDRS 0, UPDRS 1, and UPDRS 2.

Outliers were identified and removed using Tukey's method, resulting in the exclusion of 2 outliers for WER and 10 for CER. Figure 4.9 shows the cumulative statistics for the Parkinsonian group before outlier removal, with a mean WER of 0.647 and a mean CER of 0.243. The distribution of WER and CER across different UPDRS levels is depicted in Figures 4.10 and 4.11, where both metrics exhibit a clear increasing trend as UPDRS scores rise.

Tables 4.7 and 4.8 provide summary statistics for WER and CER by UPDRS

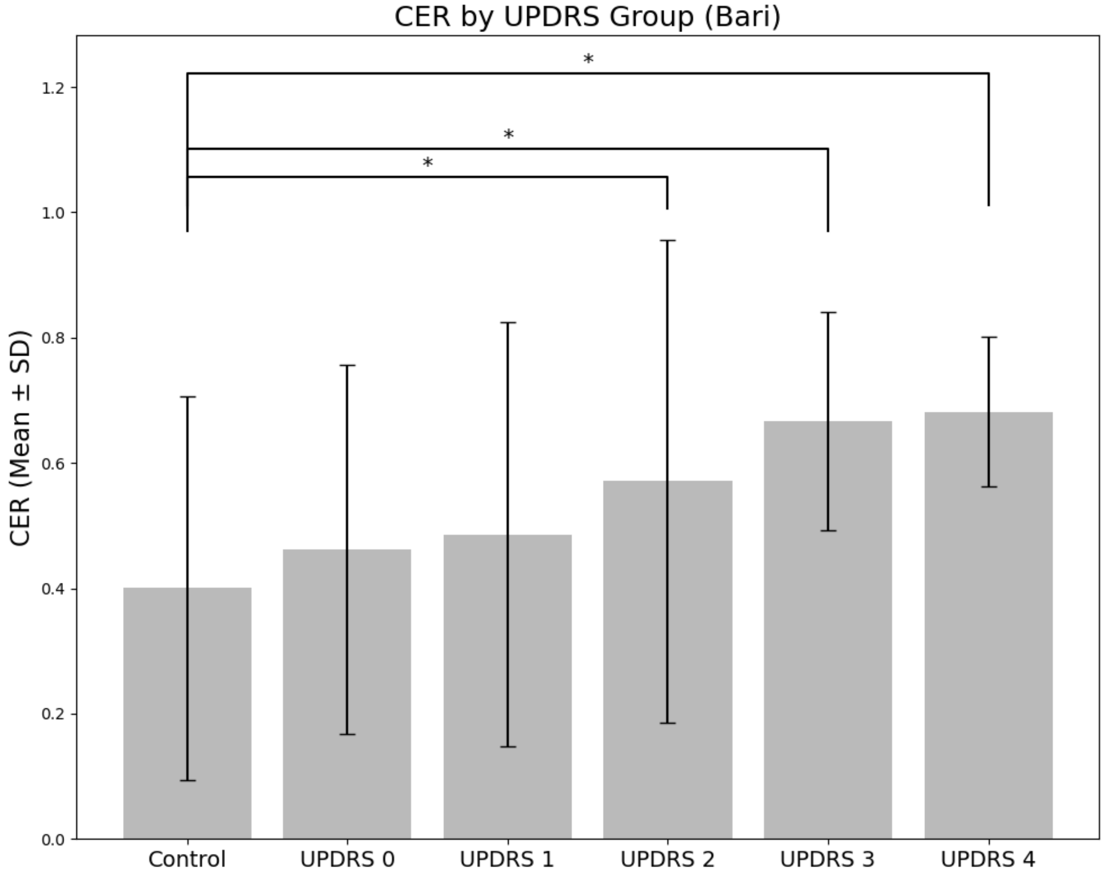


Figure 4.8: Dunn's post hoc test results for pairwise comparisons of CER across UPDRS levels in the Bari dataset. Values are measured as mean ± SD. Significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, $****p < 0.0001$.

level. The mean WER increases from 0.460 at UPDRS 0 to 0.858 at UPDRS 2, representing an 86% rise. Similarly, the mean CER increases from 0.122 at UPDRS 0 to 0.394 at UPDRS 2, more than tripling. Median values follow a similar trend, further confirming the association between higher UPDRS scores and greater speech recognition errors.

The Shapiro-Wilk test revealed that the distributions of WER and CER were not normally distributed for UPDRS 0 and UPDRS 1, with p-values below 0.05.

Due to the non-normality of the distributions, the Kruskal-Wallis test was applied to assess the significance of differences between the groups. The test revealed significant differences in both WER (p = 3.58e-04) and CER (p = 1.45e-05) across the UPDRS levels, confirming that higher UPDRS levels are associated with significantly greater speech recognition errors.

Table 4.7: Cumulative Means, Medians, and Standard Deviations for WER by UPDRS Levels in the Molinette Dataset

Group	WER Mean	WER Median	WER Std Dev
UPDRS 0	0.460	0.500	0.198
UPDRS 1	0.551	0.500	0.201
UPDRS 2	0.858	0.838	0.426

Table 4.8: Cumulative Means, Medians, and Standard Deviations for CER by UPDRS Levels in the Molinette Dataset

Group	CER Mean	CER Median	CER Std Dev
UPDRS 0	0.122	0.117	0.064
UPDRS 1	0.163	0.125	0.151
UPDRS 2	0.394	0.219	0.379

Figures 4.12 and 4.13 display the results of Dunn's post hoc test for pairwise comparisons. For both WER and CER, significant differences were found between the same pairs of UPDRS levels: UPDRS 0 and UPDRS 2 (p = 2.23e-04 for WER and p = 6.16e-05 for CER), and UPDRS 1 and UPDRS 2 (p = 4.21e-03 for WER and p = 8.36e-05 for CER). This highlights the increasing severity of speech recognition impairment as the UPDRS score rises.

4.1.4 Combined Results (Bari + Molinette)

The results presented in this section combine data from the Bari and Molinette datasets, with UPDRS levels 0, 1, and 2 grouped together for the analysis. Outlier removal using Tukey's method led to the exclusion of 2 outliers from the WER values (1 from UPDRS 1 and 1 from UPDRS 2), while no outliers were identified for CER.

As shown in Tables 4.9 and 4.10, both mean and median values for WER and CER consistently increase with higher UPDRS levels, despite the combination of data from two distinct datasets. The mean WER starts at 0.629 for UPDRS 0 and rises to 0.839 for UPDRS 2, reflecting a 34% increase. Similarly, the mean CER increases from 0.342 for UPDRS 0 to 0.462 for UPDRS 2, corresponding to a 35% increase. Median values also display a clear upward trend, underscoring the association between higher UPDRS levels and increased speech recognition errors.

Figures 4.14 and 4.15 show the boxplots for WER and CER distributions across the combined dataset. Both metrics display a clear upward trend in relation to

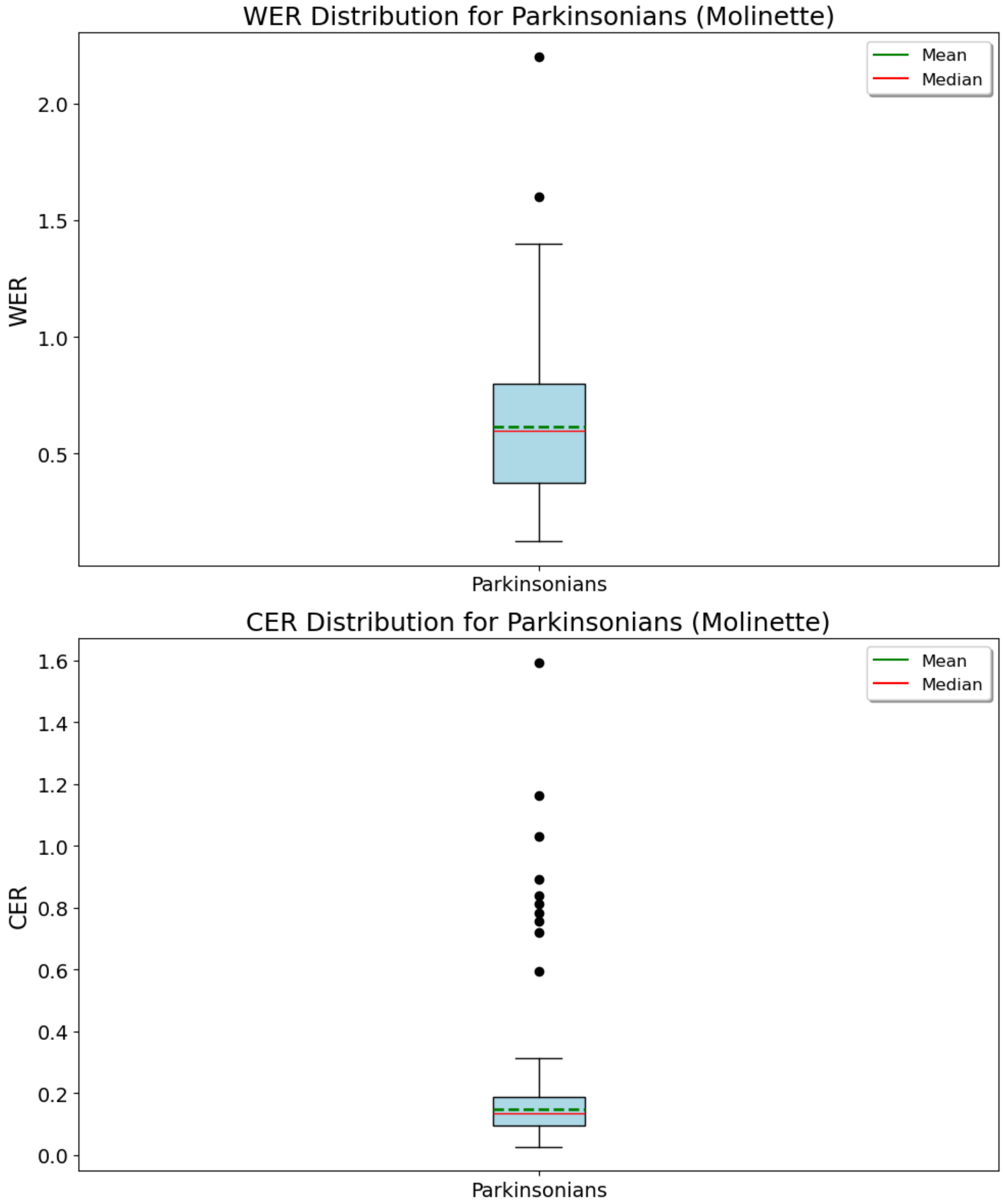


Figure 4.9: Boxplots showing the distribution of WER and CER for Parkinsonian patients in the Molinette dataset (n = 80). Values are measured as mean ± SD.

increasing UPDRS levels.

The Shapiro-Wilk test was conducted to examine the normality of the WER

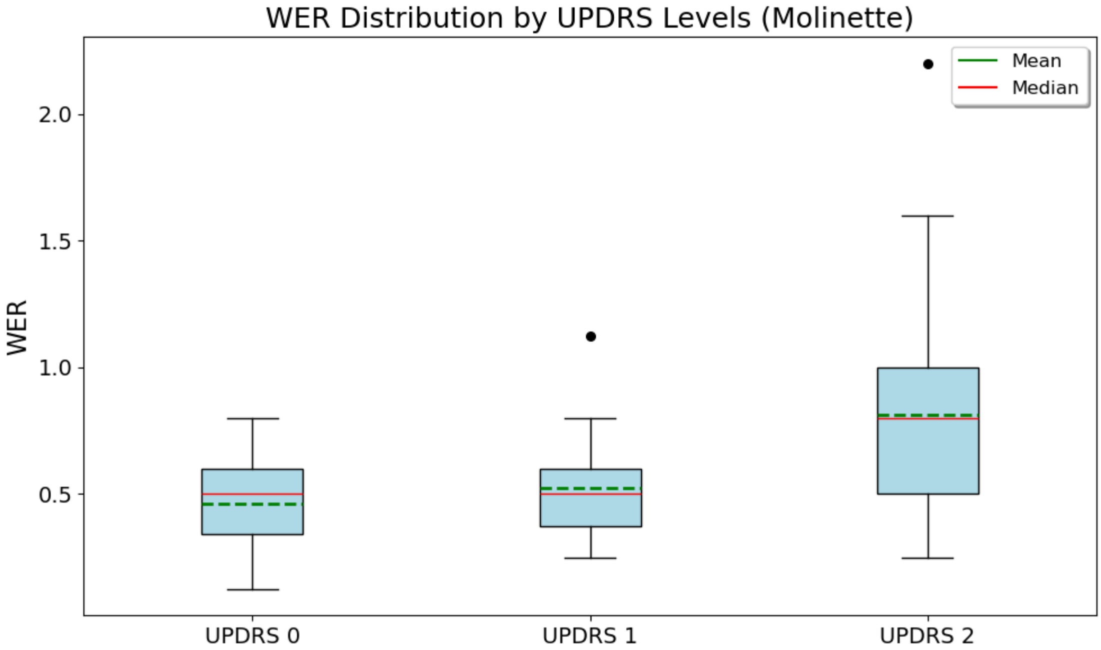


Figure 4.10: WER distribution by UPDRS levels in the Molinette dataset. Values are measured as mean ± SD. Sample sizes: UPDRS 0 (n=24), UPDRS 1 (n=24), UPDRS 2 (n=32). The mean WER for UPDRS 0, UPDRS 1, and UPDRS 2 are 0.460, 0.551, and 0.858, respectively.

Table 4.9: Cumulative Means, Medians, and Standard Deviations for WER by UPDRS Levels in the Combined Bari and Molinette Dataset

Group	WER Mean	WER Median	WER Std Dev
UPDRS 0	0.629	0.600	0.277
UPDRS 1	0.656	0.600	0.320
UPDRS 2	0.839	0.914	0.387

Table 4.10: Cumulative Means, Medians, and Standard Deviations for CER by UPDRS Levels in the Combined Bari and Molinette Dataset

Group	CER Mean	CER Median	CER Std Dev
UPDRS 0	0.342	0.167	0.290
UPDRS 1	0.354	0.171	0.319
UPDRS 2	0.462	0.266	0.388

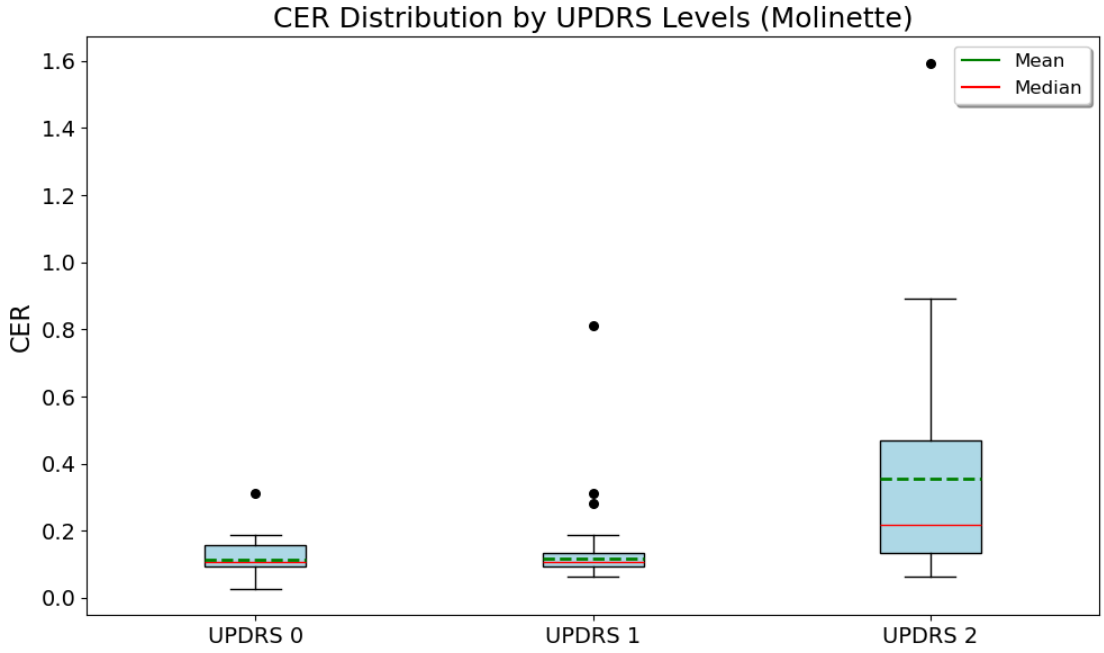


Figure 4.11: CER distribution by UPDRS levels in the Molinette dataset. Values are measured as mean ± SD. Sample sizes: UPDRS 0 (*n*=24), UPDRS 1 (*n*=24), UPDRS 2 (*n*=32). The mean CER for UPDRS 0, UPDRS 1, and UPDRS 2 are 0.122, 0.163, and 0.394, respectively.

and CER distributions across UPDRS levels. The results showed that none of the distributions followed a normal distribution, as all p-values were below 0.05. Given this, the Kruskal-Wallis test was applied to evaluate differences across UPDRS levels. The Kruskal-Wallis test identified statistically significant differences for both WER (p = 0.006) and CER (p = 0.033), indicating that the variations observed across the UPDRS levels are unlikely to be due to random chance.

Figures 4.16 and 4.17 illustrate the outcomes of Dunn's post hoc test, which was performed to assess pairwise comparisons between UPDRS groups. For WER, significant differences were found between UPDRS 0 and UPDRS 2 (p = 0.006) and between UPDRS 1 and UPDRS 2 (p = 0.004). Similarly, for CER, significant differences were observed between UPDRS 0 and UPDRS 2 (p = 0.034) and between UPDRS 1 and UPDRS 2 (p = 0.015). These findings highlight a clear trend: higher UPDRS levels, which reflect more advanced stages of Parkinson's disease, are associated with increased speech recognition errors.

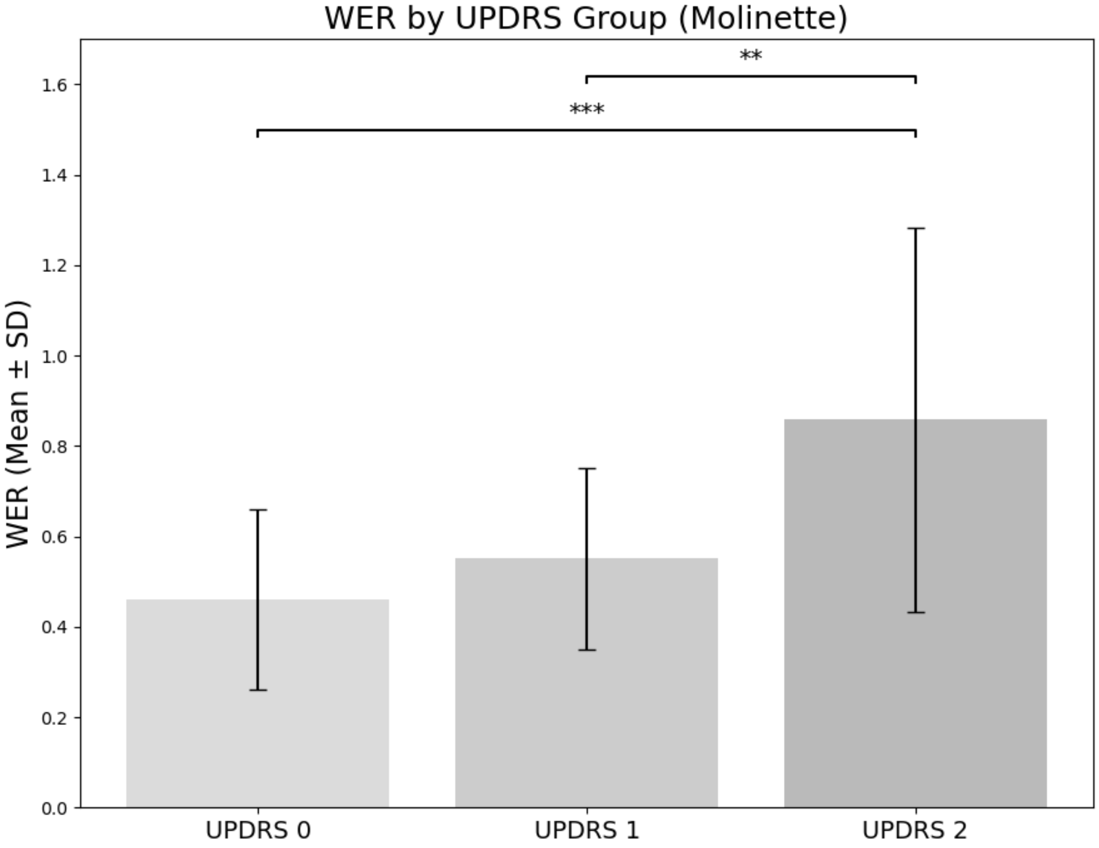


Figure 4.12: Dunn's post hoc test results for pairwise comparisons of WER across UPDRS levels in the Molinette dataset. Values are measured as mean ± SD. Significance levels: *p $<$ 0.05, **p $<$ 0.01, ***p $<$ 0.001, ****p $<$ 0.0001.

4.2 Classification Results

This section presents the performance metrics for the four architectures evaluated in this study: Vision Transformer (ViT), Audio Spectrogram Transformer (AST), Parallel Vision Transformer (ViT Parallel), and Parallel Audio Spectrogram Transformer (AST Parallel). For each model, accuracy, precision, recall, and F1-score are reported for both the training and validation sets, averaged across the 5 cross-validation folds. Additionally, UPDRS-based analyses provide insights into the relationship between predicted probabilities and disease severity.

The UPDRS-based analyses include the following sample sizes: 73 for UPDRS Level 0, 69 for UPDRS Level 1, 38 for UPDRS Level 2, 5 for UPDRS Level 3, and 19 for UPDRS Level 4. Control subjects (labeled as 0) were excluded from this analysis.

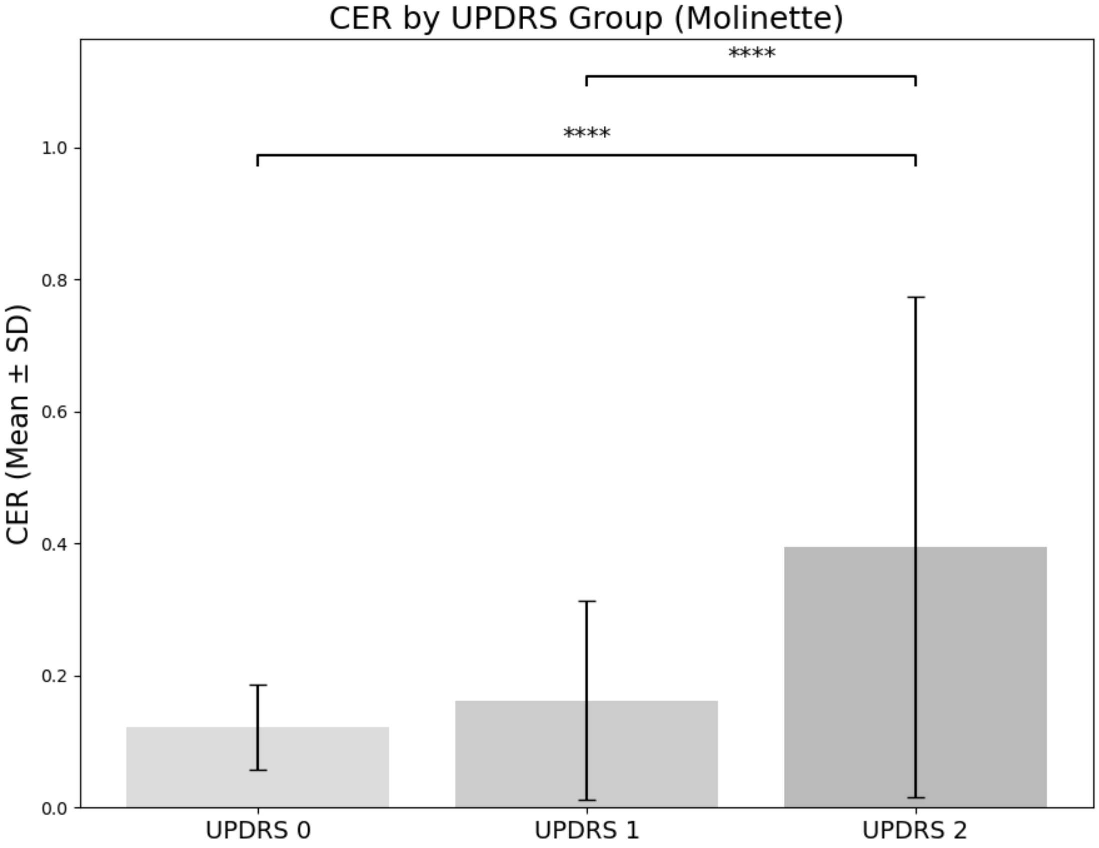


Figure 4.13: Dunn's post hoc test results for pairwise comparisons of CER across UPDRS levels in the Molinette dataset. Values are measured as mean ± SD. Significance levels: *p $<$ 0.05, **p $<$ 0.01, ***p $<$ 0.001, ****p $<$ 0.0001.

4.2.1 Vision Transformer Results

Table 4.11 summarizes the training and validation metrics for the ViT model, averaged across the 5 cross-validation folds. The model achieved an accuracy of 0.72 on the training set and 0.65 on the validation set, with precision, recall, and F1-score values showing a consistent performance across both datasets.

The UPDRS-based analysis results are shown in Table 4.12. The mean probability of belonging to the Parkinsonian class (PD) increases along the UPDRS levels, starting from 0.57 for UPDRS Level 0, 0.59 for UPDRS Level 1, and 0.63 for UPDRS Level 2. For UPDRS Level 4, the mean probability reaches 0.65.

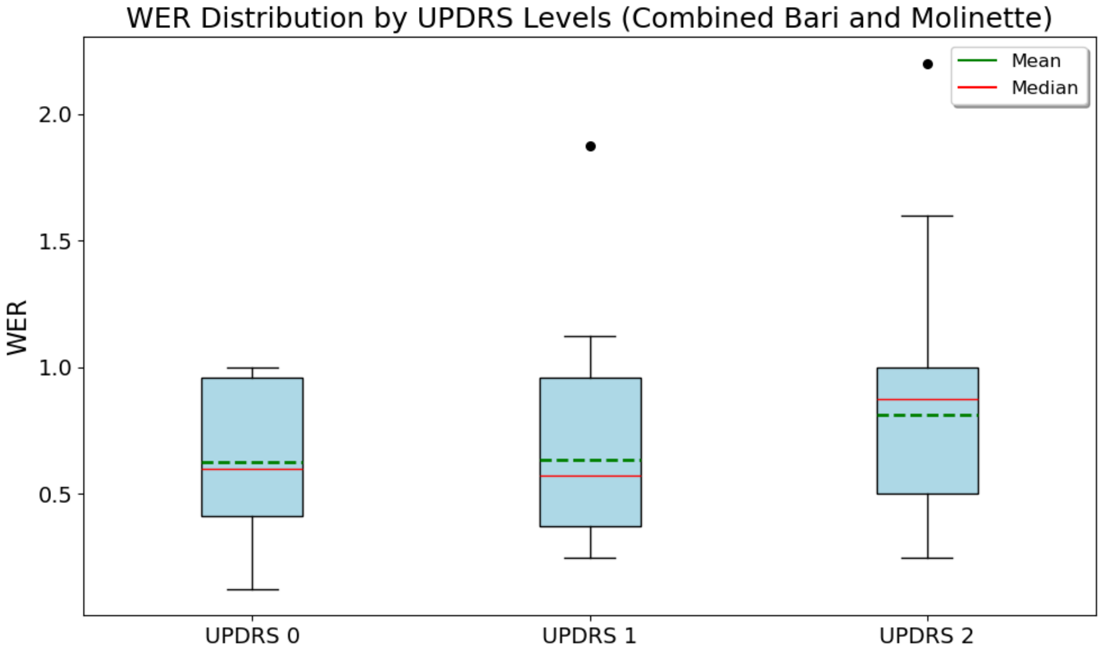


Figure 4.14: WER distribution by UPDRS levels in the combined Bari and Molinette dataset. Values are measured as mean ± SD. Sample sizes: UPDRS 0 (n=68), UPDRS 1 (n=59), UPDRS 2 (n=52). Outliers were removed using Tukey's method.

Table 4.11: Training and Validation Metrics for the ViT Model (Averaged Across 5 Folds).

Metric	Training Set	Validation Set
Accuracy	0.72	0.65
Precision	0.73	0.68
Recall	0.70	0.60
F1-Score	0.71	0.63

4.2.2 Parallel Vision Transformer Results

Table 4.13 summarizes the training and validation metrics for the ViT Parallel model, averaged across the 5 cross-validation folds. The model demonstrates consistent performance across the metrics, with an accuracy, precision, recall, and F1-score of 0.83 on the training set. On the validation set, the model achieves an accuracy of 0.67 and similar values for precision, recall, and F1-score.

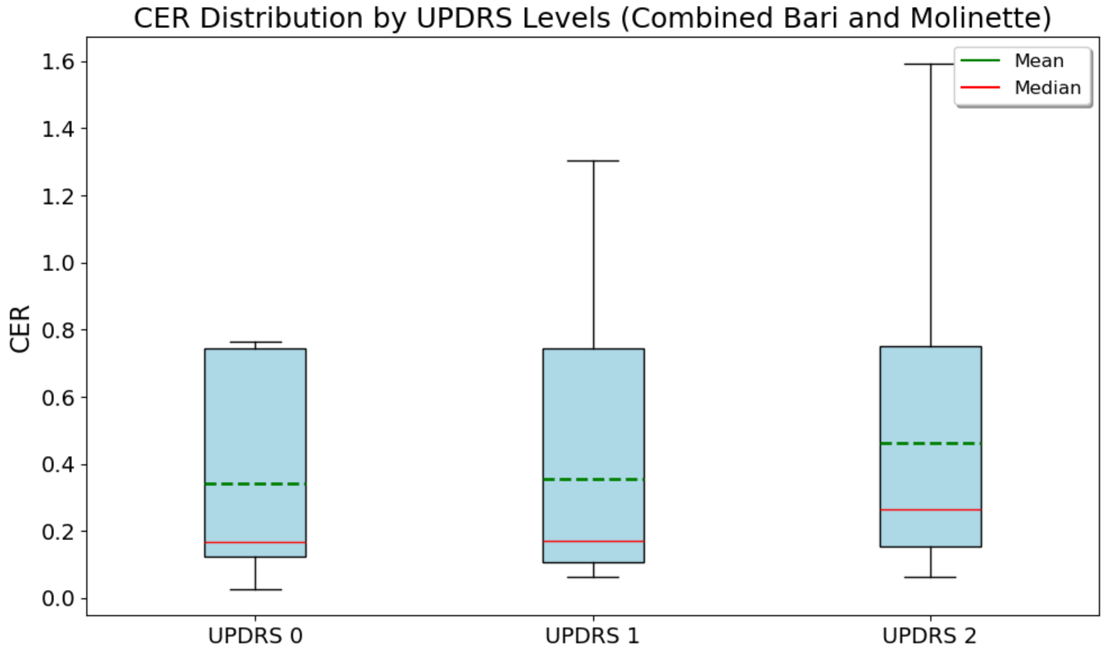


Figure 4.15: CER distribution by UPDRS levels in the combined Bari and Molinette dataset. Values are measured as mean ± SD. Sample sizes: UPDRS 0 (n=68), UPDRS 1 (n=59), UPDRS 2 (n=52). Outliers were removed using Tukey's method.

Table 4.12: UPDRS-Based Analysis for the ViT Model.

UPDRS Level	Mean Probability (PD)	Percentage Classified as PD
0	0.57	58.8%
1	0.59	63.8%
2	0.63	76.3%
3	0.83	100.0%
4	0.65	78.7%

The UPDRS-based analysis results are shown in Table 4.14. The mean probability of belonging to the Parkinsonian class shows a progressive increase from UPDRS Level 0 to UPDRS Level 4, starting at 0.77 for UPDRS Level 0 and Level 1, rising to 0.84 for UPDRS Level 2, and reaching 0.88 for UPDRS Level 4.

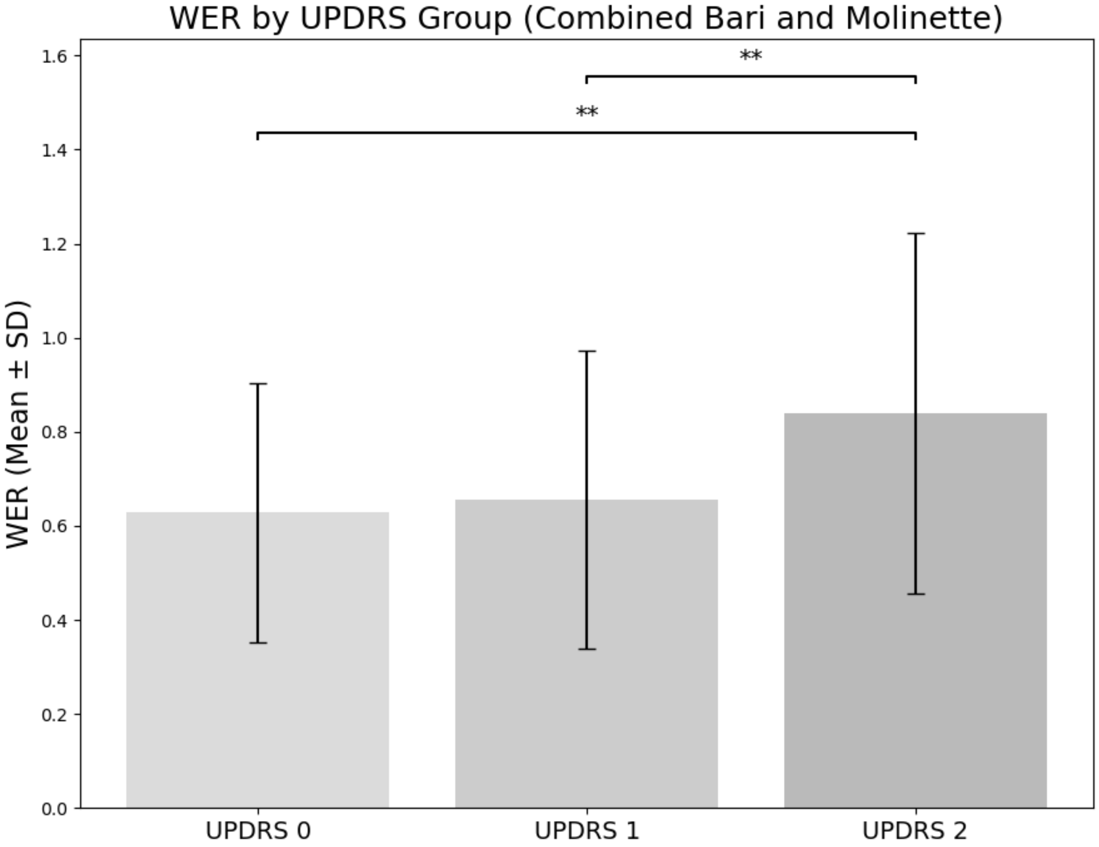


Figure 4.16: Dunn's post hoc test results for pairwise comparisons of WER across UPDRS levels in the combined Bari and Molinette dataset. Values are measured as mean ± SD. Significance levels: *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

Table 4.13: Training and Validation Metrics for the ViT Parallel Model (Averaged Across 5 Folds).

Metric	**Training Set**	**Validation Set**
Accuracy	0.83	0.67
Precision	0.83	0.68
Recall	0.83	0.67
F1-Score	0.83	0.67

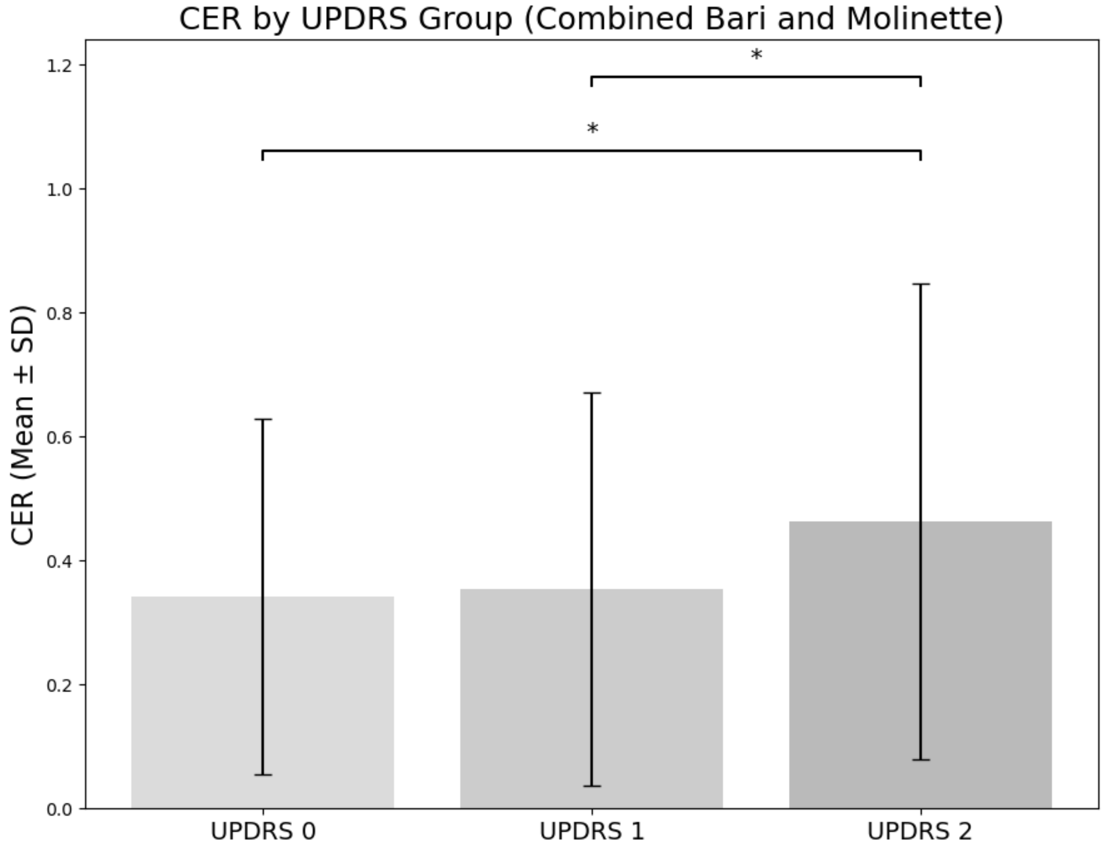


Figure 4.17: Dunn's post hoc test results for pairwise comparisons of CER across UPDRS levels in the combined Bari and Molinette dataset. Values are measured as mean ± SD. Significance levels: *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

Table 4.14: UPDRS-Based Analysis for the ViT Parallel Model.

UPDRS Level	Mean Probability (PD)	Percentage Classified as PD
0	0.77	83.6%
1	0.77	91.3%
2	0.84	97.4%
3	0.86	100.0%
4	0.88	98.5%

4.2.3 Audio Spectrogram Transformer Results

Table 4.15 summarizes the training and validation metrics for the AST model, averaged across the 5 cross-validation folds. The model achieved a training accuracy of 0.75 and a validation accuracy of 0.65. Precision, recall, and F1-score values demonstrate consistency, with the validation set metrics slightly lower than those for the training set.

The UPDRS-based analysis results are shown in Table 4.16. The mean probability of belonging to the Parkinsonian class progressively increases from 0.58 for UPDRS Level 0 to 0.64 for UPDRS Level 2, and reaches 0.73 for UPDRS Level 4.

Table 4.15: Training and Validation Metrics for the AST Model (Averaged Across 5 Folds).

Metric	Training Set	Validation Set
Accuracy	0.75	0.65
Precision	0.76	0.65
Recall	0.72	0.68
F1-Score	0.74	0.66

Table 4.16: UPDRS-Based Analysis for the AST Model.

UPDRS Level	Mean Probability (PD)	Percentage Classified as PD
0	0.58	62.9%
1	0.59	69.3%
2	0.64	79.0%
3	0.56	60.0%
4	0.73	100.0%

4.2.4 Parallel Audio Spectrogram Transformer Results

Table 4.17 summarizes the training and validation metrics for the AST Parallel model, averaged across the 5 cross-validation folds. The model achieves an accuracy of 0.98 on the training set and 0.72 on the validation set. Precision, recall, and F1-score are consistent across the training and validation sets, with values ranging from 0.93 to 0.94 for the training set and 0.70 to 0.75 for the validation set.

The UPDRS-based analysis results are provided in Table 4.18. The mean probability of belonging to the Parkinsonian class increases progressively from UPDRS Level 0 to UPDRS Level 4, ranging from 0.78 to 0.97.

Table 4.17: Training and Validation Metrics for the AST Parallel Model (Averaged Across 5 Folds).

Metric	**Training Set**	**Validation Set**
Accuracy	0.93	0.73
Precision	0.94	0.75
Recall	0.93	0.70
F1-Score	0.93	0.72

Table 4.18: UPDRS-Based Analysis for the AST Parallel Model.

UPDRS Level	**Mean Probability (PD)**	**Percentage Classified as PD**
0	0.78	78.1%
1	0.84	84.1%
2	0.84	81.6%
3	0.86	80.0%
4	0.97	94.7%

Chapter 5

Discussion and Conclusion

5.1 Word Error Rate and Character Error Rate

5.1.1 Dataset Differences and Their Impact on Word Error Rate and Character Error Rate

The results presented in this study show that the performance of the Whisper speech-to-text model varies across different datasets, with both WER and CER values differing significantly between control and Parkinsonian groups (Tables 5.1 and 5.2) . These differences can be attributed to various factors, including the characteristics of the datasets, the recording conditions, and the properties of the Whisper model itself.

In fact, the three datasets—Bari, PC-GITA, and Molinette—display notable differences in both the structure of the speech stimuli and the characteristics of the recordings. In the Bari dataset, the sentences were specifically designed to challenge Parkinsonian patients on phonemes that are often problematic for them [34, 49], while the Molinette dataset used proverbs selected to be easier to pronounce. This design difference likely contributed to the higher WER and CER values in the Bari dataset, as these sentences were intentionally chosen to highlight the phonetic impairments typical of Parkinson's disease. In contrast, the Molinette dataset, with its less challenging phonetic content, generally shows lower error rates when compared to the more challenging sentences in Bari.

Furthermore, differences in the recording equipment used across datasets may also explain some of the observed disparities in error rates. In particular, as detailed in Section 3.1.1, the Molinette dataset was recorded using a voice recorder application on a smartphone, while different, more specialized recording setups were used for the PC-GITA and Bari datasets (Section 3.1.2). These variations in recording equipment could impact Whisper's ability to accurately transcribe speech, as different setups introduce varying levels of audio quality, background noise, and clarity, which influence the model's performance.

Moreover, Whisper's performance on Spanish-language data, such as that in the PC-GITA dataset, may be inherently better due to the model being trained on a larger proportion of Spanish data compared to Italian [25]. This likely explains the lower error rates observed in the PC-GITA dataset. In contrast, Whisper's handling of Italian speech, as seen in the Bari and Molinette datasets, may be less accurate, which could account for the higher error rates in these cases.

Despite differences in the absolute values of WER and CER across the datasets, a consistent pattern emerges: in all datasets where both control and Parkinsonian groups are analyzed, the Parkinsonian group consistently shows higher error rates. This trend remains consistent regardless of variations in dataset characteristics or recording conditions, suggesting that both WER and CER are effective in capturing the speech impairments commonly associated with Parkinson's disease.

Table 5.1: Mean WER and CER Comparison Across Datasets

Dataset	Control WER Mean	PD WER Mean	Control CER Mean	PD CER Mean
PC-GITA	0.262	0.427	0.080	0.172
Bari	0.650	0.753	0.400	0.504
Molinette	N/A	0.647	N/A	0.243

Table 5.2: Median WER and CER Comparison Across Datasets

Dataset	Control WER Median	PD WER Median	Control CER Median	PD CER Median
PC-GITA	0.200	0.333	0.039	0.100
Bari	0.473	0.953	0.175	0.740
Molinette	N/A	0.600	N/A	0.135

5.1.2 Statistical Significance of Differences Between Groups

The analysis of WER and CER between control and Parkinsonian groups across the PC-GITA and Bari datasets reveals significant differences (Table 5.3). In both datasets, CER values between controls and Parkinsonian patients are significantly different. For WER, significant differences are also observed, although the results are less consistent. Notably, in the Bari dataset, the p-value for WER is $5.27e{-2}$, indicating a borderline level of significance. This suggests that while WER can distinguish control from Parkinsonian speech, it may not capture the subtleties of speech impairments as reliably as CER, especially in datasets with speech stimuli that vary in complexity. Nevertheless, WER shows a clear trend toward significance in broader comparisons.

Table 5.3: Mann-Whitney U Test p-values for WER and CER in the PC-GITA and Bari Datasets

Dataset	WER p-value	CER p-value
Bari	$5.27e{-2}$	$2.78e{-3}$
PC-GITA	$2.76e{-21}$	$4.36e{-26}$

The observed difference in statistical significance between the datasets can be partly attributed to their varying sample sizes. The PC-GITA dataset, with a larger participant pool, enhances statistical power and the robustness of observed differences. Specifically, after outlier removal, PC-GITA contains 484 control subjects and 500 Parkinsonian patients for WER, and 453 controls and 475 Parkinsonian patients for CER. In contrast, the Bari dataset is smaller, with 130 controls and 105 Parkinsonian patients for WER, and 130 controls and 106 Parkinsonian patients for CER.

In summary, both WER and CER are effective in differentiating between control and Parkinsonian speech. However, CER consistently shows greater sensitivity, likely because it captures errors at a more detailed, character-specific level. These findings highlight the potential of CER as a key metric for future studies aimed at speech-based evaluations of Parkinson's disease.

5.1.3 Statistical Differences Across UPDRS Levels

Both WER and CER show a clear upward trend with increasing UPDRS levels across all datasets, indicating their sensitivity to the progression of Parkinson's disease. As illustrated in Figures 5.1 and 5.2, both mean and median values for WER and CER rise consistently with higher UPDRS levels. This trend suggests that these metrics effectively capture the severity of speech impairment associated with Parkinson's disease.

As shown in Table 5.4, the Kruskal-Wallis test results indicate significant differences across UPDRS levels for all metrics except WER in the Bari dataset, where no significant difference was observed ($p = 0.2261$). This finding is consistent with the results of the Mann-Whitney U test for distinguishing between controls and Parkinsonians in the Bari dataset, where WER was the only metric with a borderline significance level ($p = 5.27e{-2}$).

However, when datasets are combined (e.g., Bari and Molinette), WER shows an even higher level of statistical significance than CER, with a Kruskal-Wallis p-value of $6.22e{-3}$ compared to $3.31e{-2}$ for CER. This observation suggests that, in a generalized context with larger and more diverse samples, WER may become more sensitive, potentially due to the increased variability in the combined dataset.

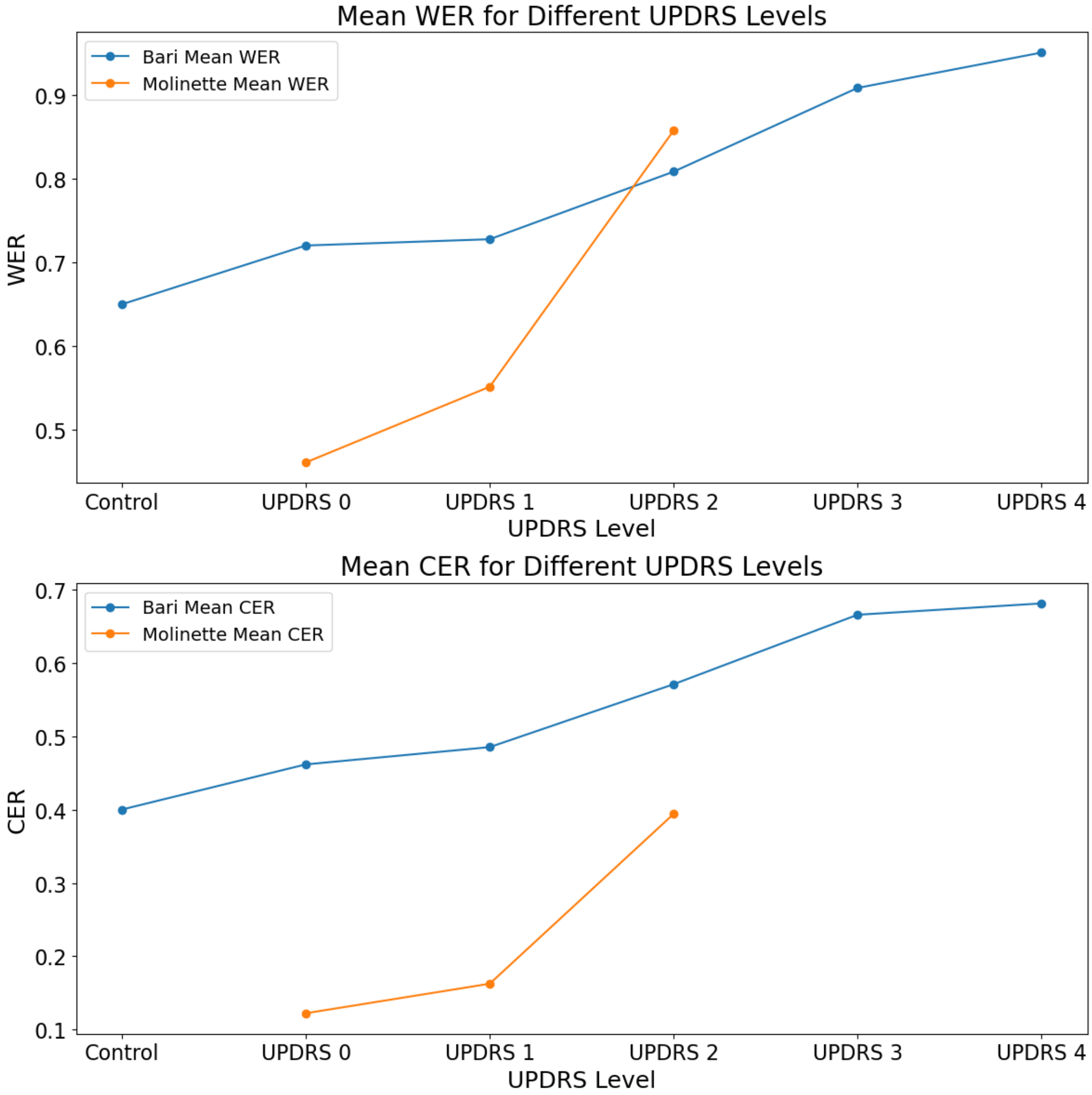


Figure 5.1: Mean WER and CER for Different UPDRS Levels across Bari and Molinette datasets. In the Molinette dataset, only UPDRS levels 0, 1, and 2 are represented, while the Bari dataset includes all UPDRS levels.

The post hoc Dunn's test for pairwise comparisons provides additional insights into the significance of differences between specific UPDRS levels. In the Bari dataset, significant differences in CER are observed between the control group and higher UPDRS levels (specifically, UPDRS 2, 3, and 4). This trend is consistent across the Molinette dataset, where Dunn's test shows significant differences in both WER and CER between UPDRS 0 and UPDRS 2, as well as between UPDRS

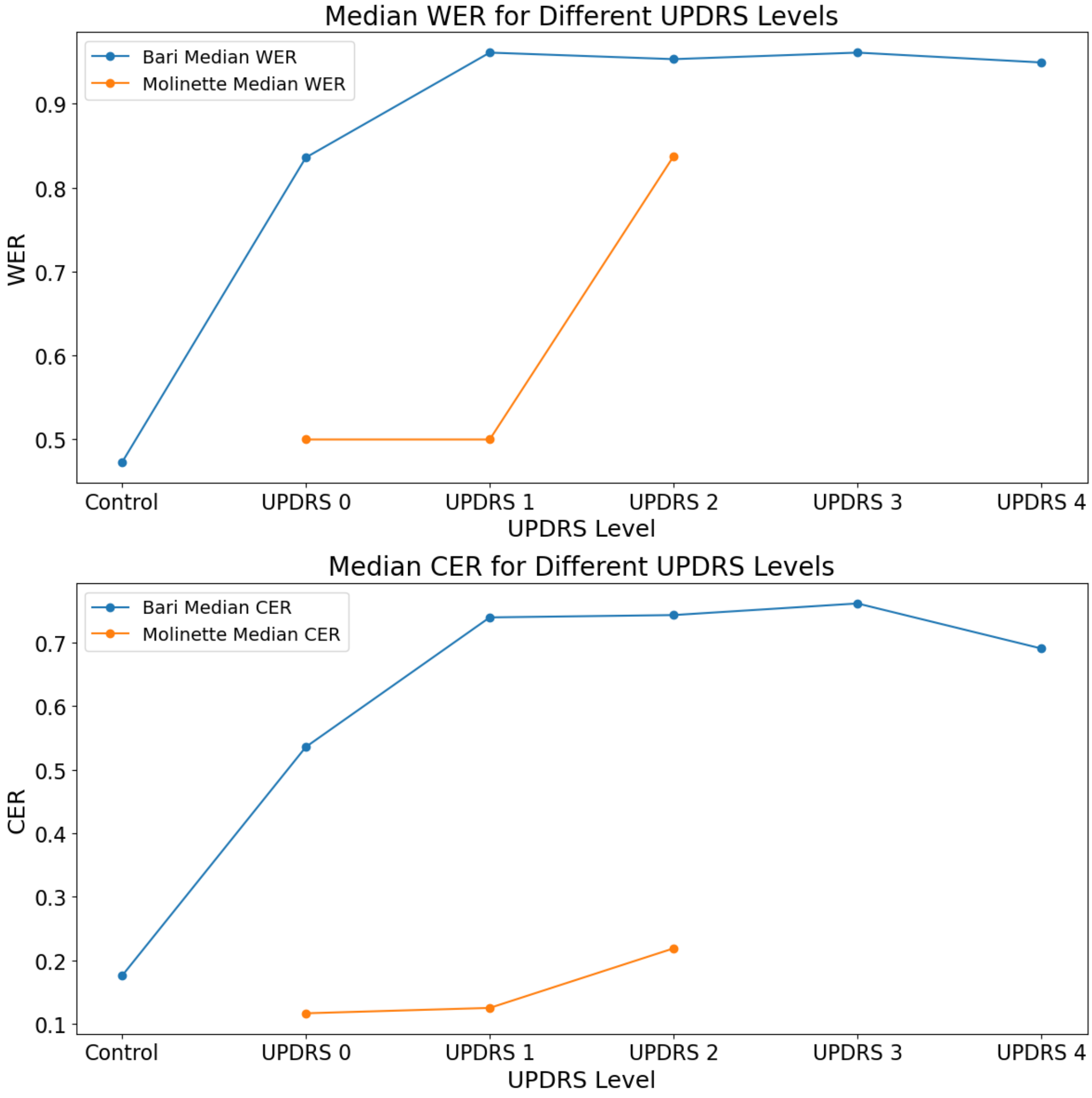


Figure 5.2: Median WER and CER for Different UPDRS Levels across Bari and Molinette datasets. The Molinette dataset includes only UPDRS levels 0, 1, and 2, whereas the Bari dataset includes the full range of UPDRS levels.

1 and UPDRS 2.

In the combined dataset, WER shows a high level of statistical significance in distinguishing between lower and higher UPDRS levels, especially in comparisons involving UPDRS 2. Figures 4.16 and 4.17 illustrate these trends, with significant differences consistently observed between the lower UPDRS levels and UPDRS 2, supporting the hypothesis that speech recognition errors escalate with disease

Table 5.4: Kruskal-Wallis Test p-values for WER and CER across UPDRS Levels in the Bari, Molinette, and Combined Datasets

Dataset	WER p-value	CER p-value	Conclusion
Bari	$2.26e{-}1$	$1.13e{-}2$	CER significant only
Molinette	$3.59e{-}4$	$1.45e{-}5$	Both significant
Combined Bari-Molinette	$6.22e{-}3$	$3.31e{-}2$	Both significant

progression.

The significant differences identified by Dunn's test between lower and higher UPDRS levels underscore the potential of both metrics for monitoring the progression of speech impairments in PD.

5.1.4 Word Error Rate and Character Error Rate as Measures of Speech Impairment

The analysis confirms that both WER and CER are effective indicators of speech impairment in Parkinsonian patients, with each metric providing distinct strengths. CER consistently achieves statistically significant results across datasets with lower p-values, indicating its sensitivity to subtle, character-level variations associated with Parkinsonian speech impairments. By contrast, WER, though showing borderline significance in some cases—such as in the Bari dataset ($p = 5.27e{-}2$)—demonstrates its utility, especially when used in conjunction with CER. While WER is less effective at highlighting differences in smaller or individual datasets, its performance improves significantly when applied to mixed datasets or larger, more diverse datasets, where it captures broader trends and variability effectively. Notably, in the combined Bari and Molinette dataset, WER showed an increased level of statistical significance, suggesting its enhanced adaptability and sensitivity in more diverse or generalized settings.

These findings align with expectations, as described earlier, that speech recognition errors increase with Parkinson's disease progression. Both WER and CER reveal statistically significant differences for most comparisons, capturing disease-related speech characteristics consistently.

It is important to interpret these metrics as relative indicators rather than absolute values, given that Whisper is not a gold standard and is influenced by variations in recording conditions and dataset structure. However, as demonstrated by the combined analysis of datasets with different characteristics, these metrics can generalize effectively across varying tasks and datasets. This versatility highlights the potential of WER and CER not only for comparing control and Parkinsonian groups but also for assessing changes in speech as the disease progresses.

In summary, WER and CER emerge as robust, complementary metrics for evaluating speech impairments in Parkinson's disease. Their statistical significance across diverse datasets and ability to generalize support their potential applications in longitudinal studies and telemonitoring, where it is important to monitor disease progression and the impact of therapeutic interventions.

5.2 Classification

5.2.1 Overview of Classification Performance

Table 5.5 provides an overview of the validation results for the ViT, AST, ViT Parallel, and AST Parallel models. These evaluations were conducted on a combined dataset that included sustained vowels from the Molinette, Bari, and PC-GITA datasets. This approach tested the models' ability to generalize across diverse data sources, posing challenges for maintaining consistent performance and robustness.

Table 5.5: Summary of Validation Metrics for All Classifiers.

Model	Accuracy	Precision	Recall	F1-Score
ViT	0.65	0.68	0.60	0.63
ViT Parallel	0.67	0.68	0.67	0.67
AST	0.65	0.65	0.68	0.66
AST Parallel	0.73	0.75	0.70	0.72

In the case of the ViT model, previous studies have applied it to PD classification [21], though not with the same datasets used in this study. The combination of sustained vowel datasets from Molinette, Bari, and PC-GITA necessitated the model to generalize across varying data sources. Compared to the reference study [21], the results obtained here are approximately 0.1 lower in terms of F1-score, recall, and precision. This difference can primarily be attributed to the use of different datasets in the two studies and, additionally, to the limited computational resources available for this work, which restricted the ability to perform extensive hyperparameter tuning and additional training experiments. Notably, the ViT Parallel configuration demonstrated a slightly better performance than the traditional ViT, as shown in Table 5.5, likely due to its ability to capture temporal variations in the data.

Unlike the ViT model, the AST architecture had not previously been applied to PD classification. As a result, there are no direct comparisons available in the literature.

Among all models evaluated, the AST Parallel configuration achieved the best performance, with a validation accuracy of 0.73, precision of 0.75, and F1-score of

0.72. As shown in Figure 5.3, this model also exhibited the highest sensitivity to changes in disease severity across UPDRS levels.

The ViT model showed the lowest validation performance, with an accuracy of 0.65 and an F1-score of 0.63, indicating its limitations in this application. The ViT Parallel model demonstrated a slight improvement, achieving an accuracy and F1-score of 0.67. Similarly, while the AST model performed on par with the ViT in its non-parallel configuration, the AST Parallel model showed a notable improvement, emphasizing the benefit of temporal segmentation in this context.

Despite these advancements, both parallel configurations (ViT and AST) exhibited signs of overfitting, as their performance metrics on the training set were notably higher than those on the validation set. Techniques such as dropout and layer normalization were applied to address this issue, resulting in partial improvements. However, further optimization, which would require more computational resources, might have enhanced the models' performance and generalization capabilities.

The experiments were conducted on Google Colab, where computational limitations posed significant challenges. These constraints restricted the ability to conduct extensive hyperparameter tuning or to train the models on larger datasets. While the dataset size used in this study was adequate for preliminary evaluations, larger and more diverse datasets could potentially improve the models' accuracy and robustness, leading to stronger conclusions.

5.2.2 UPDRS-Based Probability Analysis

The analysis of UPDRS-based probabilities is presented in Figure 5.3. For each UPDRS level, the mean probabilities of belonging to the Parkinsonian class (label 1) were calculated from the softmax-transformed logits. UPDRS Level 3 was excluded from the graph due to its limited sample size ($n = 5$). The probabilities for the remaining levels show a generally increasing trend, with higher UPDRS levels corresponding to higher probabilities of PD classification.

Notably, the parallel configurations (ViT Parallel and AST Parallel) exhibit steeper increases in probabilities compared to their non-parallel counterparts. This suggests that the parallel architectures may be more sensitive to capturing features linked to disease progression, potentially leveraging changes that occur at the beginning and end of the speech recordings.

While these results are encouraging, only a subset of the audio samples included UPDRS annotations, limiting the ability to perform detailed statistical analyses. Despite these limitations, the observed alignment between increasing disease severity and higher model probabilities is promising. This suggests that the classifiers are not only capable of distinguishing between control and Parkinsonian subjects but also show sensitivity to disease severity.

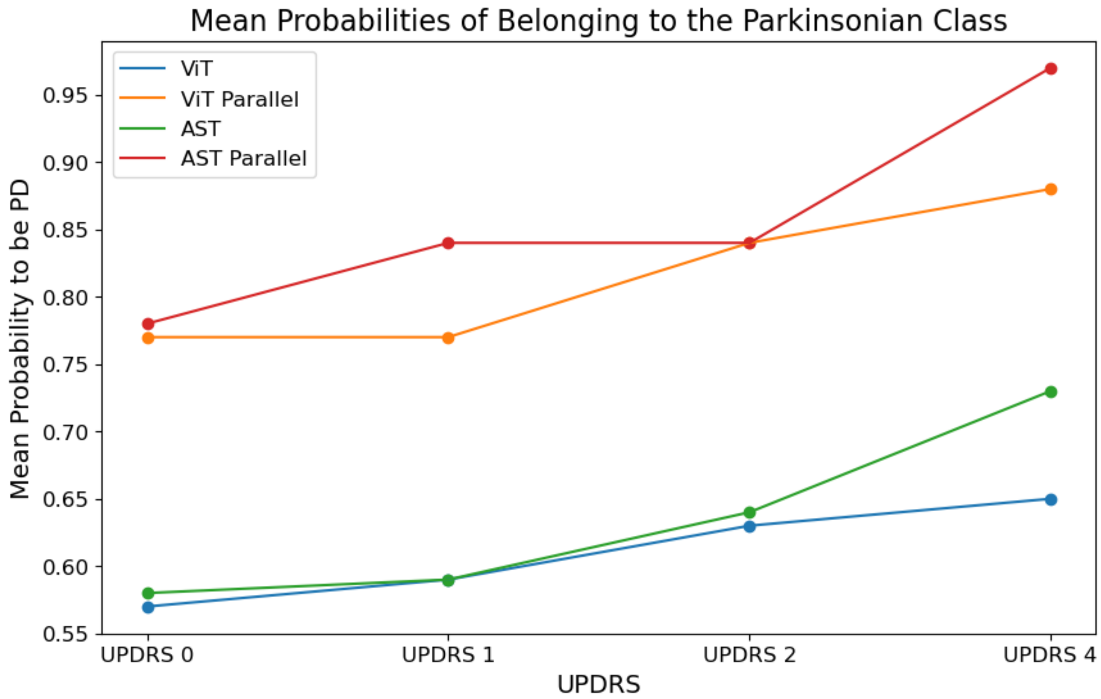


Figure 5.3: UPDRS-based mean probabilities of belonging to the Parkinsonian class (label 1) across all classifiers. UPDRS Level 3 was excluded due to its low sample size ($n = 5$). The trend shows an increase in probabilities with higher UPDRS levels, consistent with disease severity.

These findings support the potential of transformer-based architectures for telemonitoring PD. The models demonstrate both classification capability and an inherent scoring mechanism, as higher disease severity correlates with increased probabilities of Parkinsonian classification. With more extensive training and enhanced computational resources, these methods could be refined further, providing a valuable tool for remote monitoring and management of Parkinson's disease.

5.3 Limitations and Future Directions

5.3.1 Limitations of the Study

This study faced several limitations that open avenues for future research. One of the main challenges was the limited computational resources available for the experiments, especially for the classification tasks. Transformer-based architectures like ViT and AST are computationally demanding, requiring substantial resources for effective training, hyperparameter tuning, and advanced regularization. Due to these constraints, the study implemented only basic techniques, such as dropout layers and layer normalization, to reduce overfitting. While these strategies led to some improvement, the lack of computational power restricted the exploration of more advanced methods and architectures that could potentially enhance performance and generalization.

Another important limitation was the size and structure of the dataset. Although the combined dataset, including Molinette, Bari, and PC-GITA, was adequate for initial evaluations, transformer models typically perform better with larger datasets. The limited number of audio samples, particularly those annotated with UPDRS scores, reduced both the classification performance and the robustness of the statistical analyses. Increasing the dataset size and including more samples with detailed UPDRS information would allow for better generalization of the models and deeper insights into the link between vocal characteristics and PD severity.

Finally, the evaluation of WER and CER relied on the base version of the Whisper model. While this version is computationally efficient, its transcription accuracy is more limited compared to advanced versions. Consequently, the WER and CER values reported in this study should be viewed as relative metrics, useful for comparisons between control and Parkinsonian groups or across different UPDRS levels, rather than as absolute measures. Employing more advanced versions of Whisper or fine-tuning the model for this specific task could improve transcription accuracy, especially for speech samples with greater impairments, leading to more reliable evaluations.

5.3.2 Future Research Directions

Future research should aim to overcome the limitations identified in this study. One key priority is expanding the datasets, particularly by increasing the number of recordings with detailed UPDRS annotations. Expanding the datasets with more recordings and greater diversity would offer a clearer understanding of how acoustic features relate to disease severity. Additionally, a greater availability of UPDRS data would support more robust statistical analyses, especially for higher severity levels that are currently underrepresented.

Building on the UPDRS-based analysis conducted here, which revealed an increasing trend in classification probabilities with higher UPDRS levels, future work could utilize Explainable Artificial Intelligence (XAI) techniques to further explore these findings. XAI methods could offer transparent and interpretable explanations for the model's behavior, enhancing understanding of its decision-making processes [50].

Heatmaps, a well-established XAI tool, are widely used in medical imaging to highlight key regions of input data that influence model predictions. Techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) and Layer-wise Relevance Propagation (LRP) have been successfully applied in areas like histology and brain MRI analysis to improve model interpretability [51]. In this context, heatmaps could be used to identify critical time-frequency regions in spectrograms, shedding light on the features that drive classification decisions. This would not only clarify the model's reasoning but also enhance understanding of the patterns associated with disease progression and UPDRS scores.

These explainability techniques could also have direct clinical implications. By identifying the most relevant features in spectrograms, heatmaps could inform the development of targeted classification strategies and therapeutic interventions. For instance, they might pinpoint specific speech impairments linked to disease progression, helping clinicians tailor treatments to individual needs. Furthermore, the transparency provided by XAI-driven models aligns with regulatory frameworks like GDPR, which emphasize clear and interpretable decision-making processes [52]. This approach could ultimately support personalized strategies for monitoring disease progression and adapting interventions over time.

The ultimate goal of this research direction is to enable telemonitoring of Parkinson's disease using widely available devices, such as smartphones. The WER and CER metrics, along with the classifiers developed in this study, showed sensitivity to disease severity, highlighting their potential for remote monitoring. For instance, a smartphone application could analyze speech recordings over time to detect signs of worsening conditions, enabling timely clinical interventions.

Beyond telemonitoring, future research could focus on supporting Parkinsonian patients with severely impaired speech intelligibility. When a patient's speech becomes too difficult to understand, the approaches explored in this study could be extended to improve communication. For example, an advanced version of Whisper could be used to transcribe low-intelligibility speech with greater accuracy. By applying heatmaps to identify critical regions in Mel-spectrograms, it might be possible to pinpoint the acoustic features most affected by the disease. This information could then guide restorative techniques, such as using a large language model (LLM) to reconstruct missing or distorted content, followed by text-to-speech (TTS) synthesis to produce clear, comprehensible speech.

Another approach could involve enhancing the spectrograms themselves. By identifying and correcting the time-frequency regions most impacted by Parkinsonian impairments, algorithms could be developed to restore intelligibility before applying a TTS system. This type of processing pipeline could bridge the gap between impaired speech and effective communication, offering a transformative solution for patients in advanced stages of the disease.

Overall, this study demonstrated the potential of WER and CER metrics, along with the correlation between classification probabilities and UPDRS scores, to assess and monitor Parkinson's disease. Future research could refine these tools into comprehensive systems for telemonitoring and disease management, providing both remote monitoring capabilities and innovative ways to improve the quality of life for patients with severe speech impairments. With larger datasets, more advanced computational resources, and improved interpretability techniques, these models could evolve into practical tools for both clinical use and everyday applications.

Bibliography

[1] Werner Poewe, Klaus Seppi, Caroline M. Tanner, Glenda M. Halliday, Patrik Brundin, Jens Volkmann, Anette-Eleonore Schrag, and Anthony E. Lang. «Parkinson disease». In: *Nature Reviews Disease Primers* 3 (Mar. 2017), Article number 17013. DOI: 10.1038/nrdp.2017.13. URL: http://dx.doi.org/10.1038/nrdp.2017.13 (cit. on pp. 1, 2).

[2] Bastiaan R Bloem, Michael S Okun, and Christine Klein. «Parkinson's disease». In: *The Lancet* 397 (June 2021), pp. 2284–2303. DOI: 10.1016/S0140-6736(21)00218-X. URL: https://doi.org/10.1016/S0140-6736(21)00218-X (cit. on pp. 1, 2).

[3] Christina M. Lill. «Genetics of Parkinson's disease». In: *Molecular and Cellular Probes* 30 (Nov. 2016), pp. 386–396. DOI: 10.1016/j.mcp.2016.11.001. URL: https://doi.org/10.1016/j.mcp.2016.11.001 (cit. on pp. 1, 2).

[4] Shinya Koga, Naoki Aoki, Robert J. Uitti, Jon A. van Gerpen, William P. Cheshire, Keith A. Josephs, Zbigniew K. Wszolek, John W. Langston, and David W. Dickson. «When DLB, PD, and PSP masquerade as MSA: an autopsy study of 134 patients». In: *Neurology* 85.5 (Aug. 2015), pp. 404–412. DOI: 10.1212/WNL.0000000000001807. URL: https://doi.org/10.1212/WNL.0000000000001807 (cit. on p. 2).

[5] Frederick C. Church. «Treatment Options for Motor and Non-Motor Symptoms of Parkinson's Disease». In: *Biomolecules* 11.4 (Apr. 2021), p. 612. DOI: 10.3390/biom11040612. URL: https://doi.org/10.3390/biom11040612 (cit. on pp. 2, 3).

[6] Eduardo Tolosa, Alberto Garrido, Stefan W. Scholz, and Werner Poewe. «Challenges in the diagnosis of Parkinson's disease». In: *Lancet Neurology* 20.5 (May 2021), pp. 385–397. DOI: 10.1016/S1474-4422(21)00030-2. URL: https://doi.org/10.1016/S1474-4422(21)00030-2 (cit. on p. 2).

[7] Joseph Jankovic and Louis E. Tan. «Parkinson's disease: etiopathogenesis and treatment». In: *Journal of Neurology, Neurosurgery, and Psychiatry* 91.8 (Aug. 2020). Epub 2020 Jun 23, pp. 795–808. DOI: 10.1136/jnnp-2019-322338. URL: https://doi.org/10.1136/jnnp-2019-322338 (cit. on p. 3).

[8] C.D. Marsden and J.D. Parkes. «"ON-OFF" Effects in Patients with Parkinson's Disease on Chronic Levodopa Therapy». In: *The Lancet* 307.7954 (1976), pp. 292–296. DOI: 10.1016/S0140-6736(76)91416-1 (cit. on p. 3).

[9] Christopher G. Goetz, Barbara C. Tilley, Sue R. Shaftman, Glenn T. Stebbins, and et al. «Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results». In: *Movement Disorders* 23.15 (2008), pp. 2129–2170. DOI: 10.1002/mds.22340. URL: https://doi.org/10.1002/mds.22340 (cit. on p. 3).

[10] A. Ma, K. K. Lau, and D. Thyagarajan. «Voice changes in Parkinson's disease: What are they telling us?» In: *Journal of Clinical Neuroscience* 72 (Feb. 2020). Epub 2020 Jan 14, pp. 1–7. DOI: 10.1016/j.jocn.2019.12.029. URL: https://doi.org/10.1016/j.jocn.2019.12.029 (cit. on pp. 3, 4, 6).

[11] Laureano Moro-Velazquez, Jorge A. Gomez-Garcia, Julian D. Arias-Londoño, Najim Dehak, and Juan I. Godino-Llorente. «Advances in Parkinson's Disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects». In: *Biomedical Signal Processing and Control* 66 (2021). Epub ahead of print, p. 102418. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2021.102418. URL: https://doi.org/10.1016/j.bspc.2021.102418 (cit. on pp. 3, 4, 6, 14).

[12] Tetsutaro Ozawa, Kanako Sekiya, Naotaka Aizawa, Kenshi Terajima, and Masatoyo Nishizawa. «Laryngeal stridor in multiple system atrophy: Clinicopathological features and causal hypotheses». In: *Journal of the Neurological Sciences* 361 (2016), pp. 243–249. ISSN: 0022-510X. DOI: 10.1016/j.jns.2016.01.007. URL: https://www.sciencedirect.com/science/article/pii/S0022510X16300077 (cit. on pp. 4, 5).

[13] Gustavo Andrade-Miranda. «Analyzing of the Vocal Fold Dynamics Using Laryngeal Videos». PhD thesis. Universidad Politécnica de Madrid, 2017. DOI: 10.20868/UPM.thesis.47122. URL: https://doi.org/10.20868/UPM.thesis.47122 (cit. on p. 5).

[14] K. M. Smith and D. N. Caplan. «Communication impairment in Parkinson's disease: Impact of motor and cognitive symptoms on speech and language». In: *Brain and Language* 185 (Oct. 2018). Epub 2018 Aug 6, pp. 38–46. DOI: 10.1016/j.bandl.2018.08.002. URL: https://doi.org/10.1016/j.bandl.2018.08.002 (cit. on pp. 4–6).

[15] Laureano Moro-Velazquez, Jorge A. Gomez-Garcia, Juan I. Godino-Llorente, Francisco Grandas-Perez, Stefanie Shattuck-Hufnagel, Virginia Yagüe-Jimenez, and Najim Dehak. «Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's Disease». In: *Scientific*

Reports 9.1 (Dec. 2019), Article number 19066. DOI: 10.1038/s41598-019-55271-y. URL: https://doi.org/10.1038/s41598-019-55271-y (cit. on pp. 4–6, 13).

[16] Fatemeh Majdinasab, Siamak Karkheiran, Majid Soltani, Negin Moradi, and Gholamali Shahidi. «Relationship Between Voice and Motor Disabilities of Parkinson's Disease». In: *Journal of Voice* 30.6 (Nov. 2016). Epub 2015 Dec 22, 768.e17–768.e22. DOI: 10.1016/j.jvoice.2015.10.022. URL: https://doi.org/10.1016/j.jvoice.2015.10.022 (cit. on p. 6).

[17] Eleni Tsalera, Alexandros Papadakis, and Maria Samarakou. «Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning». In: *Journal of Sensor and Actuator Networks* 10.4 (2021), p. 72. DOI: 10.3390/jsan10040072. URL: https://doi.org/10.3390/jsan10040072 (cit. on pp. 7, 8).

[18] K. J. Piczak. «Environmental Sound Classification with Convolutional Neural Networks». In: *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Boston, MA, USA, 2015, pp. 1–6. DOI: 10.1109/MLSP.2015.7324337. URL: https://doi.org/10.1109/MLSP.2015.7324337 (cit. on p. 7).

[19] Laureano Moro-Velazquez, Javier Villalba, and Najim Dehak. «Using X-Vectors to Automatically Detect Parkinson's Disease from Speech». In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain, May 2020, pp. 1155–1159. DOI: 10.1109/ICASSP40776.2020.9053770. URL: https://doi.org/10.1109/ICASSP40776.2020.9053770 (cit. on pp. 7, 8).

[20] Yuan Gong, Yu-An Chung, and James Glass. «AST: Audio Spectrogram Transformer». In: *arXiv preprint arXiv:2104.01778* (July 2021). Accepted at Interspeech 2021. arXiv: 2104.01778 [cs.SD]. URL: https://doi.org/10.48550/arXiv.2104.01778 (cit. on pp. 7, 8, 28, 33).

[21] Dominik Hemmerling, Mateusz Wodzinski, Jose R. Orozco-Arroyave, Daniel Sztaho, Michal Daniol, Pawel Jemiolo, and Malgorzata Wojcik-Pedziwiatr. «Vision Transformer for Parkinson's Disease Classification using Multilingual Sustained Vowel Recordings». In: *Annu Int Conf IEEE Eng Med Biol Soc.* July 2023, pp. 1–4. DOI: 10.1109/EMBC40787.2023.10340478 (cit. on pp. 7, 8, 66).

[22] Alberto Favaro, Thuc Cao, Thomas Thebaud, Javier Villalba, Ashwin Butala, Najim Dehak, and Laureano Moro-Velázquez. «Do Phonatory Features Display Robustness to Characterize Parkinsonian Speech Across Corpora?» In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Aug. 2023, pp. 2388–2392. DOI:

10.21437/Interspeech.2023-1784. URL: https://doi.org/10.21437/Interspeech.2023-1784 (cit. on pp. 8–10).

[23] 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». In: *Biocybernetics and Biomedical Engineering* 42.2 (2022), pp. 556–574. ISSN: 0208-5216. DOI: 10.1016/j.bbe.2022.04.002. URL: https://doi.org/10.1016/j.bbe.2022.04.002 (cit. on pp. 8, 9).

[24] Tatiana Arias-Vergara, Juan C. Vásquez-Correa, and José R. Orozco-Arroyave. «Parkinson's Disease and Aging: Analysis of Their Effect in Phonation and Articulation of Speech». In: *Cognitive Computation* 9 (Dec. 2017). Received 10 October 2016, Accepted 18 July 2017, Published 04 August 2017, pp. 731–748. DOI: 10.1007/s12559-017-9497-x. URL: https://doi.org/10.1007/s12559-017-9497-x (cit. on pp. 8, 9).

[25] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. *Robust Speech Recognition via Large-Scale Weak Supervision.* 2022. arXiv: 2212.04356 [eess.AS]. URL: https://arxiv.org/abs/2212.04356 (cit. on pp. 9, 61).

[26] Davide Mulfari and Massimo Villari. «A Voice User Interface on the Edge for People with Speech Impairments». In: *Electronics* 13.7 (2024), p. 1389. ISSN: 2079-9292. DOI: 10.3390/electronics13071389. URL: https://doi.org/10.3390/electronics13071389 (cit. on p. 9).

[27] Jeong-Uk Bang, Seung-Hoon Han, and Byung-Ok Kang. «Alzheimer's disease recognition from spontaneous speech using large language models». In: *ETRI Journal* 46.1 (2024), pp. 96–105. ISSN: 1225-6463. DOI: 10.4218/etrij.2023-0356. URL: https://doi.org/10.4218/etrij.2023-0356 (cit. on p. 9).

[28] Hadil Mehrez, Mounira Chaiani, and Sid Ahmed Selouani. «Using StarGANv2 Voice Conversion to Enhance the Quality of Dysarthric Speech». In: *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC).* 2024, pp. 738–744. DOI: 10.1109/ICAIIC60209.2024.10463241 (cit. on pp. 9, 21).

[29] Jonathan Crawford. *Linguistic Changes in Spontaneous Speech for Detecting Parkinsons Disease Using Large Language Models.* 2024. arXiv: 2404.05160 [cs.CL]. URL: https://arxiv.org/abs/2404.05160 (cit. on p. 9).

[30] Laureano Moro-Velázquez, Jaejin Cho, Shinji Watanabe, Mark Hasegawa-Johnson, Odette Scharenborg, Heejin Kim, and Najim Dehak. «Study of the Performance of Automatic Speech Recognition Systems in Speakers with Parkinson's Disease». In: *Proceedings of Interspeech 2019.* Epub 2019 Aug 15. Sept. 2019, pp. 3875–3879. DOI: 10.21437/Interspeech.2019-2993. URL: https://doi.org/10.21437/Interspeech.2019-2993 (cit. on p. 9).

[31] Green Apple Studio. *Registratore Vocale.* Available on Google Play: https://play.google.com/store/apps/details?id=com.media.bestrecorder.audiorecorder&hl=it and on the App Store: https://apps.apple.com/it/app/registratore-vocale-voz/id1336782987. 2023. URL: https://play.google.com/store/apps/details?id=com.media.bestrecorder.audiorecorder&hl=it (cit. on p. 11).

[32] Juan Rafael Orozco, Julian D. Arias-Londoño, J. Vargas-Bonilla, María González-Rátiva, and Elmar Noeth. «New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease». In: May 2014 (cit. on pp. 14, 22).

[33] Giovanni Dimauro and Francesco Girardi. *Italian Parkinson's Voice and Speech.* 2019. DOI: 10.21227/aw6b-tg17. URL: https://dx.doi.org/10.21227/aw6b-tg17 (cit. on p. 16).

[34] Giovanni Dimauro, Vincenzo Di Nicola, Vitoantonio Bevilacqua, Danilo Caivano, and Francesco Girardi. «Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System». In: *IEEE Access* 5 (2017), pp. 22199–22208. DOI: 10.1109/ACCESS.2017.2762475 (cit. on pp. 16, 22, 60).

[35] Anna Favaro, Yi-Ting Tsai, Ankur Butala, Thomas Thebaud, Jesús Villalba, Najim Dehak, and Laureano Moro-Velázquez. «Interpretable speech features vs. DNN embeddings: What to use in the automatic assessment of Parkinson's disease in multi-lingual scenarios». In: *Computers in Biology and Medicine* 166 (2023), p. 107559. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2023.107559. URL: https://www.sciencedirect.com/science/article/pii/S0010482523010247 (cit. on pp. 19, 20).

[36] Georgy Shevlyakov, Kliton Andrea, Lakshminarayan Choudur, Pavel Smirnov, Alexander Ulanov, and Natalia Vassilieva. «Robust versions of the Tukey boxplot with their application to detection of outliers». In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing.* 2013, pp. 6506–6510. DOI: 10.1109/ICASSP.2013.6638919 (cit. on p. 22).

[37] Z. Hanusz, J. Tarasinska, and W. Zielinski. «Shapiro–Wilk Test with Known Mean». In: *REVSTAT-Statistical Journal* 14.1 (2016), pp. 89–100. DOI: 10.57805/revstat.v14i1.180. URL: https://doi.org/10.57805/revstat.v14i1.180 (cit. on p. 23).

[38] Prabhaker Mishra, Uttam Singh, Chandra M. Pandey, Priyadarshni Mishra, and Gaurav Pandey. «Application of Student's t-test, Analysis of Variance, and Covariance». In: *Annals of Cardiac Anaesthesia* 22.4 (Oct. 2019), pp. 407–411. DOI: 10.4103/aca.ACA_94_19 (cit. on p. 23).

[39] Thomas W. MacFarland and Jan M. Yates. «Mann–Whitney U Test». In: *Introduction to Nonparametric Statistics for the Biological Sciences Using R*. Cham: Springer, 2016. Chap. 4. ISBN: 978-3-319-30634-6. DOI: 10.1007/978-3-319-30634-6_4. URL: https://doi.org/10.1007/978-3-319-30634-6_4 (cit. on p. 23).

[40] Tae Kyun Kim. «Understanding one-way ANOVA using conceptual figures». In: *Korean journal of anesthesiology* 70.1 (2017), pp. 22–26 (cit. on p. 24).

[41] Patrick E McKight and Julius Najab. «Kruskal-wallis test». In: *The corsini encyclopedia of psychology* (2010), pp. 1–1 (cit. on p. 24).

[42] Alexis Dinno. «Nonparametric pairwise multiple comparisons in independent groups using Dunn's test». In: *The Stata Journal* 15.1 (2015), pp. 292–300 (cit. on p. 24).

[43] Librosa Development Team. *librosa.feature.melspectrogram Documentation.* Accessed: November 13, 2024. 2024. URL: https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html (cit. on p. 27).

[44] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. «librosa: Audio and music signal analysis in Python». In: *Proceedings of the 14th Python in Science Conference*. Citeseer. 2015, pp. 18–25 (cit. on p. 27).

[45] David M. W. Powers. «Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation». In: *International Journal of Machine Learning Technology* 2.1 (2011). This open access journal appears to have been discontinued. Updated and fixed formatting errors., pp. 37–63. DOI: 10.48550/arXiv.2010.16061. arXiv: 2010.16061 [cs.LG]. URL: https://arxiv.org/abs/2010.16061 (cit. on p. 31).

[46] Bichen Wu et al. *Visual Transformers: Token-based Image Representation and Processing for Computer Vision*. 2020. arXiv: 2006.03677 [cs.CV] (cit. on p. 32).

[47] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. «Imagenet: A large-scale hierarchical image database». In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on p. 32).

[48] Alexey Dosovitskiy. «An image is worth 16x16 words: Transformers for image recognition at scale». In: *arXiv preprint arXiv:2010.11929* (2020) (cit. on p. 32).

[49] Giovanni Dimauro, Danilo Caivano, Vitoantonio Bevilacqua, Francesco Girardi, and Vito Napoletano. «VoxTester, software for digital evaluation of speech changes in Parkinson disease». In: May 2016, pp. 1–6. DOI: 10.1109/MeMeA.2016.7533761 (cit. on p. 60).

[50] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. «Explainable artificial intelligence (XAI) in deep learning-based medical image analysis». In: *Medical Image Analysis* 79 (2022). Originally submitted on arXiv as arXiv:2107.10912 [eess.IV], p. 102470. DOI: 10.1016/j.media.2022.102470. URL: https://doi.org/10.48550/arXiv.2107.10912 (cit. on p. 69).

[51] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. «Explainable deep learning models in medical image analysis». In: *Preprint submitted to MDPI* (2020). arXiv preprint arXiv:2005.13799. arXiv: 2005.13799 [cs.CV]. URL: https://doi.org/10.48550/arXiv.2005.13799 (cit. on p. 70).

[52] Tribikram Dhar, Nilanjan Dey, Surekha Borra, and R. Simon Sherratt. «Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust». In: *IEEE Transactions on Technology and Society* 4.1 (2023), pp. 68–75. DOI: 10.1109/TTS.2023.3234203 (cit. on p. 70).

Acknowledgements

I would like to sincerely thank Professor Gabriella Olmo and PhD Federica Amato for their guidance and support during my thesis work. Their expertise, thoughtful feedback, and constant encouragement have been invaluable throughout this journey.

A heartfelt thanks to my family, who have always supported me and believed in me, even during the most difficult moments.

Finally, I am deeply grateful to my friends, who have shared this experience with me. Their presence and support have made this journey both rewarding and unforgettable.