A Deep Learning Based Approach to Detect Parkinson's Disease through Speech Analysis

by

Anjon Biswas Ovi
21201719
Jahidul Hassan Rifat
21201300
Mashrur Azim
24141117
Fatin Ishraq Mahi
22101464
Md Arafat Hossain
21201405

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2026

Declaration

It is hereby declared that

- The thesis submitted is our own original work while completing a degree at BRAC University.
- The thesis does not contain material previously published or written by a third party, except where this is appropriately cited.
- The thesis has not been accepted or submitted for any other degree or diploma at a university or other institution.
- All sources of help have been properly acknowledged.

Student's Full Name & Signature:

Anjon Biswas Ovi
21201719

Jahidul Hassan Rifat
21201300

Mashrur Azim
24141117

Fatin Ishraq Mahi
22101464

Md Arafat Hossain
21201405

Approval

The thesis titled **"A Deep Learning Based Approach to Detect Parkinson's Disease through Speech Analysis"** submitted by:

1. Anjon Biswas Ovi
2. Jahidul Hassan Rifat
3. Mashrur Azim
4. Fatin Ishraq Mahi
5. Md Arafat Hossain

of Fall 2025 has been accepted as satisfactory in partial fulfillment of the requirements for the degree of BSc in CSE.

Examining Committee:

Supervisor:
(Member)

Md. Saiful Bari Siddiqui
Senior Lecturer
Department of Computer Science And Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science And Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson
Department of Computer Science And Engineering
Brac University

Abstract

Parkinson's Disease (PD) is a progressive type of neurodegenerative disorder in which timely diagnosis is extremely important in managing the disease, yet it is difficult because of the insidious nature of the symptoms. The conventional diagnostic approaches usually use subjective clinical evaluation which might fail to reveal early biomarkers. This study fills this gap by proposing a deep learning-based framework that can be used to detect PD non-invasively with speech signals. In this paper, we present a multi-model fusion architecture that makes use of two different representations of one audio signal, which are Mel-spectrograms to represent spectral-visual features and Wav2Vec 2.0 embeddings to represent high-level contextual phonetic features. We extracted Wav2Vec2 embedding features from a fully connected neural network that was trained on the Wav2Vec2 embeddings. We then ran the melspectrograms though CNN and other convolutional networks such as ResNet18, ResNet50, Inception v3, Efficient Net B0, and Efficient Net B4. We then extracted melspectrogram features from all these convolutional networks and concatenated them seperately with the Wav2Vec2 embedding features. The cocnatenated features then went through a commonly shared final classifier. One of the most important contributions of this work is the introduction of a new rational sine activation function that gives the model a better capability of learning nonlinear decision boundaries. The results of the experiment conducted on the Italian Parkinson voice and speech data show that the proposed fusion model attains a test accuracy of 85.29% which is higher than what the other architectures that involve transfer learning models such as ResNet18, ResNet50, Inception v3, Efficient Net B0, and Efficient Net B4 have achieved. Moreover, a strict cross-corpus test on the unseen Vowel dataset shows that although domain shift is a universal problem, the suggested fusion architecture demonstrates better generalization with respect to the baseline models, which are adversely affected by the problem. The paper has validated the hypothesis that a combination of mixed signal representations with mathematically optimized activation functions is a promising avenue towards automated and objective screening of PD.

Keywords: Parkinson's Disease; Speech Analysis; Deep Learning; Diagnosis; Acoustic Biomarkers; Approximation; sin; Hybrid Audio Feature Fusion Network (HAFF-Net)

Acknowledgment

We want to firstly thank the Almighty to give us the strength, perseverance and capability to accomplish this thesis successfully.

Regarding our respected thesis supervisor, Md. Saiful Bari Siddiqui, we would like to sincerely thank him and acknowledge him by stating that he has guided and given us invaluable thinking and constructive feedback as well as supported us throughout the entire thesis duration. The role of his encouragement, motivation, and professional advice was an important contribution to the enhancement and successful accomplishment of this research work.

We also extend our deepest gratitude to BRAC University who gave us the academic environment, the resources, and the chance to carry out and complete this research.

Lastly, we would like to make a heartfelt thanks to our families and parents, who always encouraged, patiently supported, and morally guided us to pursue and complete this research successfully.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Parkinson's Disease (PD) is a progressive neurodegenerative disorder that primarily affects the motor system. It is caused by the gradual depletion of dopaminergic neurons in the *substantia nigra*, a region of the basal ganglia responsible for coordinating movement. According to the World Health Organization, PD is the second most common age-related neurodegenerative disease globally, affecting approximately 10 million people worldwide.

This figure will increase by two folds by 2040 as the world population is getting older, which is a big burden on healthcare systems. The symptoms of PD are traditionally divided into **Tremor at Rest**,**Rigidity**, **Akinesia** (or bradykinesia), and **Postural instability**, which are also known as **TRAP** symptoms.

Nevertheless, these motor symptoms are usually seen after 50 to 70 percent of dopaminergic neurons are already gone. The latency interval entails that by the time a clinical diagnosis is reached, the disease is frequently in its late stages and therefore the efficacy of neuroprotective treatments is restricted.

Importantly, recent studies indicate that one of the earliest prodromal symptoms of PD is vocal impairment which is clinically referred to as hypokinetic dysarthria. The physiological correlation is high: the same neural networks that take care of controlling the limbs movement also control the fine motor control of the larynx and respiratory system. The vocal disorders of PD are abnormal in about 90 percent of patients, such as decreased loudness (hypophonia), lack of pitch variation (aprosody), breathiness, and articulatory inaccurateness. These changes in acoustics tend to be early and continuous unlike the intermittent limb tremors and speech analysis is an avenue that presents early and non invasive screening.

1.2 Motivation

The primary motivation why this study is conducted is the shortcomings and inaccessibility of the existing diagnostic procedures.

- **Subjectivity of Clinical Diagnosis:** The generic diagnostic approach is based on the Unified Parkinson's Disease Rating Scale (UPDRS) according to which a specialist

should visually examine the movement of the patient. It is a subjective process and has a high tendency of inter-rater variability, particularly at the initial stages when the symptoms are subtle.

- **Healthcare Accessibility:**In rural and developing areas, the access to neurologists is extremely low. There is a long queue waiting time and expensed cost of consultation among patients. A screening tool based on software may democratize access to the diagnosis as a "first-line" triage system.

- **The Potential of Deep Learning:** Human ears do not, in most cases, capture the microtremors of a speech at an early PD, Deep Learning (DL) algorithms are good at recognizing patterns in high-dimensional data. Through the use of advanced acoustic signal processing, we can obtain objective and quantifiable biomarkers using only a voice recording, and through telemedicine, monitor the patient remotely.

1.3 Problem Statement

Although the theoretical potential of using AI to detect PD is high, there are multiple technical challenges that impede the implementation of a robust clinical tool, and current literature cannot address them entirely:

1. **Limitations of Hand-Crafted Features:** The traditional methods have heavily been based on manual features extraction (e.g., Jitter, Shimmer, HNR). Although interpretable, these features are not always able to model the high-dimensional and non-linear and non-linear time dynamics of pathological speech.

2. **Data Scarcity and Overfitting:** As the Medical datasets are smaller and it means that generic off-the-shelf Deep Learning models (such as ResNet50 or Inception V3) are overfitting naturally. Instead of acquiring the overall acoustic characteristics of the disease, they learn the particular background acoustic noise or recording conditions of the training set. The Barrier of the Domain Shift: The most difficult problem in the analysis of speech is the inability to generalize across languages and recording conditions. A trained model on the Italian speakers in a studio is highly likely to fail when applied to the English speakers that record their voice on smartphones.

3. **Inefficient Activation Functions:**Standard activation functions such as ReLU are created to work with images in general. They are not always the best in handling of the audio signals, which are periodic and oscillatory in nature.

This thesis aims to address these specific gaps by developing a custom-tailored architecture that prioritizes generalization and feature robustness over raw model size.

1.4 Research Objectives

The principal aim of the thesis is to come up with a powerful, multi-representation deep-learning system to classify Parkinson's Disease in the binary form based on speech signals. In order to do so, we specify the following specific sub-objectives:

1. In order to create and apply a **Multi-Representation Fusion Model**, which combines visual spectral features (Mel-spectrograms) with high-level contextual phonetic features **(Wav2Vec 2.0 embeddings)** to produce the most information retrieval.

2. o explore how mathematical non-linearity can be applied to improve acoustic classification by proposing a **Rational Sine** activation function, it was hypothesized that periodic functions are more appropriate to audio data than piecewise linear functions (ReLU).

3. To examine a comprehensive comparative analysis against industry-standard architectures such as ResNet, Inception, EfficientNet to illustrate the efficiency of custom, lightweight models for medical classifications.

4. To strictly test the generalization property of the model by running a **Cross Corpus Evaluation** on an entirely novel dataset, which gives a clear test of the domain shift issue.

1.5 Proposed Solution: Multi-Representation Fusion

In order to address the challenge of Parkinsonian speech, we present a dual-branch neural network structure, which attempts to process the speech signal as an image and as a sequence. The hypothesis of this approach is that there cannot be a single representation that will describe all biomarkers.

- **Visual Branch (Mel-Spectrograms):** We convert the raw audio into Mel-spectrograms the time-frequency representation of the sound. The processing of these images is done using a bespoke Convolutional Neural Network (CNN) that is conditioned to recognize visual texture that is related to the suddenness of vocal instability, e.g. a sudden break or a sound.

- **Contextual Branch (Wav2Vec 2.0):** We use the Wav2Vec 2.0 model which is a self-supervised model of the state of the art. This branch removes latent phonetic embeddings which represent the prosody (rhythm) and temporal continuity of speech, which are commonly lost in the frozen spectrogram images.

- **Mathematical Optimization (Rational Sine):**Unlike the traditional models, that apply ReLU activations, the custom layers are designed based on the Rational Sine. The new activation also allows the network to estimate better the oscillatory nature of speech signals and the classification boundary between healthy and pathological samples is enhanced.

1.6 Contributions of this Research

The thesis achieves the following valuable contributions to the domain of biomedical signal processing and applied artificial intelligence:

- **Development of a Hybrid Fusion Architecture:** We demonstrate that fusing disparate signal representations (Spectrograms + Embeddings) significantly outperforms single-modality baselines, achieving a test accuracy of 85.29% on the Italian dataset.
- **Empirical Validation of Rational Sine Activation:**We provide the experimental data that rational sine activations are more accurate in medical acoustics than linear activations.
- **Critical Benchmarking of Transfer Learning:** We unveil a serious weakness of training massive pre-trained models (such as ResNet50) on small medical datasets. They are biased, we find in our analysis, toward majority class, and use their recall advantage against their precision, which is the opposite of what is desired in any operationalization of our model.
- **Transparency in Generalization Testing:** Unlike most studies that present only successful training measures, we present a transparent account of the phenomenon of Domain Shift byaining our model trained on Italian on a Czech Vowel dataset, creating a realistically robust metric of cross-lingual studies in the future.

1.7 Scope and Challenges

1.7.1 Scope of the Research

The following boundaries characterize the scope of this research:

- **Data Modality:** The research is based on acoustic analysis only (voice recordings). It does not include other biomarkers, including gait, handwriting, or genetic information.
- **Classification Task:** The system is to be used in binary classification (Healthy Control vs. Parkinson Disease). It does not now measure the severity of an illness (e.g. it is able to predict the scores on the UPDRS)
- **Linguistic Scope:** The main training and validation have been done on native speakers of Italian. Although the Vowel dataset was tested on robustness, the model has not been yet optimized to be deployed on a universal and multi-lingual basis.

1.7.2 Research Challenges

Creating such a strong medical diagnostic tool has a number of built-in difficulties:

- **Data Scarcity and Imbalance:** Pathological speech data has an uncommon presence in publicly available datasets, is usually small, and is skewed as well. This renders deep learning models vulnerable to overfitting..

- **Inter-Speaker Variability:** The speech of human beings is very different depending on the age, gender, accent, and the quality of recording devices. The acoustic characteristics of Parkinson disease and these natural variations are a complicated signal processing issue..
- **Interpretability:** Deep Learning models can be considered as "black boxes." Although they are used to predict PD with high accuracy, the use of clinical adoption is difficult to explain **why** a particular audio clip was flagged as pathological.

1.8 Thesis Organization

The rest of this thesis is presented in the following structure: **Chapter 2** provides the literature review of the existing literature and gives the theoretical background of deep learning techniques to be applied. The system requirements and specifications are described in **Chapter 3**. **Chapter 4** describes the suggested methodology, such as the structure of the Fusion Model and the implementation of the Rational Sine function. **Chapter 5** presents the results of the experiment, confusion matrices, and ROC are provided Finally, **Chapter 6** is the last and it wraps up the research and proposes the possible future research directions.

Chapter 2

Literature Review

2.1 Preliminaries and Parkinson's Disease Background

Parkinson Disease (PD) is a non-malignant neurodegenerative disorder that is progressive and that mainly involves motor functions. It takes place because the dopamine producing neurons in a part of the brain referred to as the *substantia nigra* degenerate slowly. Dopamine plays a vital role in the coordination of motor movement and its deficiency results in the lack of motor controls. PD is also rated among the most common movement disorders and is known to affect millions of people across the world, especially the elderly population.

Parkinson's Disease is a complex and multifactorial etiologic process, which is a mix of genetic parameters, exposure to environmental agents (e.g., pesticides), and aging [14]. Although, the PD is also usually linked to motor symptoms such as tremors, muscle rigidity, bradykinesia, and postural instability, they are usually hard to observe in the early disease phases [14]. Consequently, the earliest diagnosis is especially difficult, because observable clinical symptoms can be manifested only after years of the disease development [16].

Parkinson disease is a progressive disease that is usually divided into five stages, the severity of which intensifies with time. During the first phase, the patients might have slight tremors that do not pose a significant problem to the daily work. Progression of the disease results in the third stage whereby the symptoms extend to larger muscle groups subsequently causing a reduced balance and coordination. In stage four, people usually need help with daily activities and in the last stage of the process, most patients are bedridden and their minds might be impaired, such as the loss of memory.

In Addition to the motor dysfunctions, PD is often followed by non-motor symptoms including depression, sleep problems, cognitive impairments, and autonomic dysfunctions, including digestive problems. It is worth noting that a change in speech is one of the first and most evident manifestations of Parkinson Disease and it usually precedes significant motor behavioral abnormalities. Such speech deficiencies as diminished pitch fluctuation, monotone speech, decreased - varying voice, and tremulous voice or breathy voice. These occurrences are a result of the influence of PD of the muscles and nerves of the larynx and respiratory system. The speech abnormalities of the Parkinson Disease are measurable through the advanced computational methods that assess the vibrations of the vocal folds,

frequency patterns, and so on.

Acoustic characteristics like jitter, shimmer and harmonic-to-noise ratio. The deep learning techniques allow the automatic extraction and analysis of these parameters and offer the clinicians with objective data-driven indicators that contribute to the early and accurate diagnosis.

Machine learning, in particular, deep learning, is especially useful in discovering latent patterns in large and complex data. Convolutional Neural Networks (CNNs) are better at learning spatial properties of spectrogram representations of speech, whereas sequential models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are better at learning temporal dependencies in speech signals. By using these models in explaining the variation in speech in the Parkinson Disease, it would be possible to come up with effective and efficient tools of non-invasive and early diagnosis, which would eventually enhance patient outcomes.

2.1.1 Theoretical Background

In order to comprehend how deep learning can be used to identify Parkinson Disease, we will need to revise the basic concepts of audio signal processing and computer vision architectures to be used in this work.

Audio Signal Processing and Spectrograms

Raw audio signals are usually represented as one-dimensional arrays of time-series of sound amplitude. Nevertheless, deep learning models, especially the ones that can be used in computer vision, work best when receiving two-dimensional inputs. In order to fill this gap, the audio signals are converted to the time-frequency domain of Short-Time Fourier Transform (STFT).

A continuous audio signal is broken down by the STFT into short overlapping segments and Fourier transform of each segment is calculated. The output is a spectrogram which is a graph whereby the x-axis is time, the Y-axis is frequency and amplitude (or loudness) is represented as color intensity. When analyzing the speech, Mel-spectrogram is usually in favor of the standard linear spectrogram. Mel scale is a perceptual scale of pitch which is used to indicate human perception of variations in frequency. Because the human auditory system is more sensitive to change in lower frequencies as compared to higher frequencies, the Mel scale increases the content of low frequencies and decreases the content of high frequencies. This transformation brings the signal representation to the human hearing perception and is especially useful in recording low-frequency tremors that are related to hypokinetic dysarthria.

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) constitute the basis of deep learning as image based analysis. CNNs keep the spatial structure of the input unlike Multilayer Perceptrons (MLPs)

that represent input data in a one dimensional form. This feature renders them particularly appropriate to analyze spectrograms.

A typical CNN model/architecture consists of three main types of layers, which is shown below :

- **Convolutional Layers:** In this layer, learnable sliding filters (kernels) are used to extract local feature of the input. Applied to the detection of Parkinson Disease, these filters are conditioned to recognize certain acoustic patterns such as breathiness or sudden changes in pitch, and that are represented as visual textures on spectrograms.
- **Pooling Layers:** Pooling layers downsample feature maps, decreasing the spatial resolution of feature maps. This operation reduces the complexity of the computations and translational invariance is introduced.
- **Fully Connected Layers:** Convolutional and pooling layers give an output that is flattened and passed to dense layers, where the last classification occurs, such as a way of differentiating between normal people and patients of Parkinson Disease.

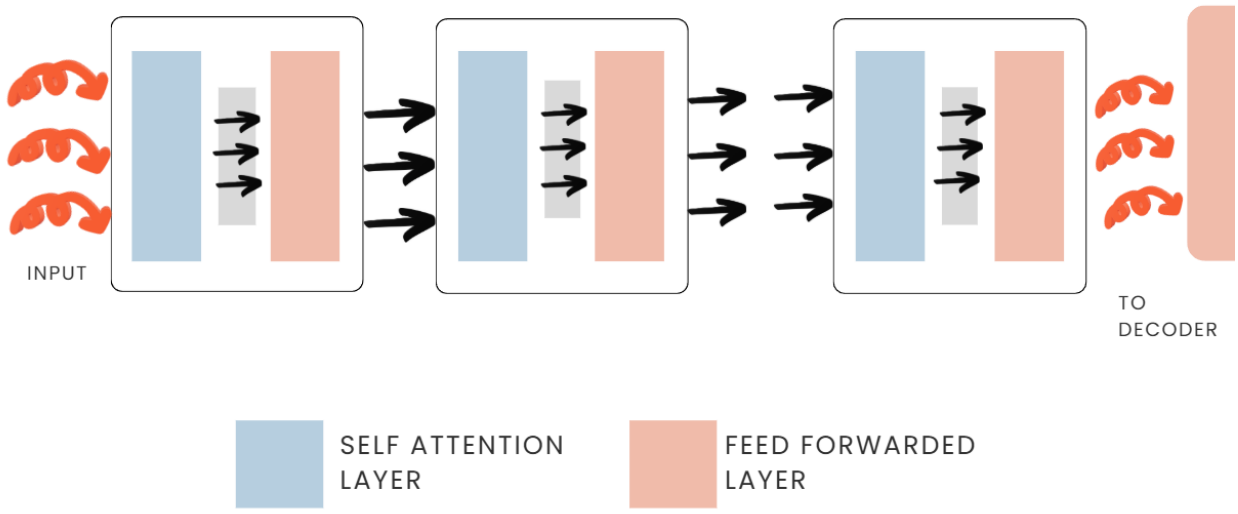


Figure 2.1: Encoder in Transformer Neural Network.

Transfer Learning

To prevent overfitting the training of deep CNNs, large datasets are required to train the network, which cannot be feasible in medical research when only limited data can be obtained. Transfer learning is one of the ways to overcome this issue by leveraging models that have been trained on massive scales. In the current research, ImageNet which includes more than 14 million images is a model pre-trained and then fined-tuned to the medical task at hand.

One of the best transfer learning architectures is the Residual Networks (ResNet). Deep neural networks are commonly prone to the vanishing gradient problem, where learning signals becomes smaller as they go through multiple layers.ResNet addresses this problem by adding skip connections (identity shortcuts), where gradient is able to work around some layers.

This residual learning structures enables the successful training of deep networks, such as ResNet50 (with 50 layers), without affecting the performance. Consequently, ResNet models are capable of training extremely complicated and non-linear representations of features related to medical image analysis.

Self-Supervised Learning: Wav2Vec 2.0

Although CNNs can extract spatial data of spectrograms, they can miss out the sequential and phonetic nature of speech signals. In order to overcome this drawback, Wav2Vec 2.0, the state of the art speech representation learning framework, is included in this study.

Wav2Vec 2.0 does not use supervised learning, but instead provides self-supervised learning that does not require any labeled data. The model is trained through masked prediction goal where it is trained to make predictions of masked audio segments in terms of their surrounding contexts with quantized representations. This kind of training utilizes thousands of hours of unmarked speech data.

Through this approach, Wav2Vec 2.0 learns rich, contextual embeddings that effectively capture the underlying phonetic and linguistic structure of speech. These embeddings are robust to environmental noise and enable the extraction of high-level speech features, making them well-suited for Parkinson's Disease detection.

2.2 Deep Learning in PD diagnosis

With the growing potential of Artificial Intelligence(AI) to improve our daily life, AI has shown promising success in various aspects including mecical science. Nowadays, AI is being used for disease detection such as PD. Artificial intelligence is making significant contribution to help in neurology by timely diagnosis and effective patient handling through risk stratification [4, 37]

2.3 Review of Existing Research

2.3.1 Existing Research in Speech Analysis to detect PD

Saravanan et al. [35] suggests that speech classification involves seperating patients with PD from the healthy controls. In this process, we can try to extract acoustic features which are also known as acoustic biomarkers from speech signals.

Some classical features are suggested in [27]. For example, jitter, Mel Frequency Cepstral Coefficients(MFCCs), shimmer, formant frequencies, and fundamental frequency or F0.

Despite having these classical features, it seems that deep accoustic feature extraction or DAFE is more used as suggest by some contemporary researches [22, 13, 12]. Recect researches also suggest that researchers are more drawn to deep learning approaches since it doesn't involve manual feature extraction which can be very cumbersome [11]. Moreover, the abstract nature and robustness of such features allow to achieve a higher accuracy.

Prabhavalka et al. [30] suggests that we see that hidden Markov models or HMMs and Gaussian mixture models or GMMs once used to be the primary machine learning approaches to carry out speech processing.

Recent research works suggest that the deep learning approaches for speech processing are primarily E2E learning as suggested by [31, 1] and transfer learning as suggested by [20].

Narendra et al. [26] said We can map raw speech signal to final output. This is more beneficial than traditional speech recognition methods that rely on HMMs or GMMs.

Bilal Er et al. [10] suggested As we know that CNNs are primarily used to process images, we can exploit this advantage and use spectrograms as images for CNN input to detect PD patterns. Moreover, Sepp Hochreiter et al. [18] LSTM can be used to be trained on more relavent information over longer sequences from the audio signals of patients with PD.

Khaskhoussy et al. [21] uses an SVM and LSTM classifier which are trained on five datasets that differ from each other.

Akila et al. [1] proposes MASS-PCNN where MASS means multi-agent salp swarm and PCNN indicates PD classification neural network to identify germane features. The pooling layers in PCNN play a crucial role in selecting the vital features. Two things that makes their PCNN unique is its inception module that gathers multi-scale data and the squeeze-and-exitation moduele used to increase feature importance. These adoptions helped them reach an accuracy of 95.1%.

Narendra et al. [26] used a traditional and E2E approach where E2E reached to an accuracy 68.56% and the traditional approach led to an accuracy of 67.93%.

Nagasubramanian et al. [25] uses ADCNN or Acoustid Deep Neural Network which consists of CNN architecture and it reached a little bit higher accuracy than DNN and RNN. The accuracy of ADCNN was 99.92% whereas DNN and RNN achieved an accuracy of 98.96% and 99.88% respectively.

Boualoulou et al. [5], authors make a comparison between a CNN and a LSTM classifier where they chose Gammatone Cepstral Coefficients or GTCCs and MFCCs on some subsets of PC-GITA [28] and PD classification datasets [6]. Their comparison states that CNN performs way better than LSTM.

Reddy et al. [33], we see a quite different approach where they employ both CNN and

Transformer. To be more specific, theychose GoogleNet, VGG16, VGG19, DenseNet121,DenseNet201, DenseNet169, ResNet152, ResNet50, and MobileNetV2 where the lowest accuracy was observed in VGG19 and MobileNetV2. On the other hand, the highest accuracy was found in ResNet152 which was 98.08%.

Sarlas et al. [36] employed Federated Learning or FL method on MVDR-KCL database and their experiment achieved an accuracy of 61.49%. On the other hand, Janbakhshi et al. [19] achieved accuracy of 75.4% on PC-GITA [28].

Malekroodi et al. [23] chose a Swin Transformer to classify healthy controls and patients with PD.

Karaman et al. [20] choose SqueezeNet1_1, DenseNet161, and ResNet50. Here, SqueezeNet1_1, DenseNet161, and ResNet50 are already trained on the ImageNet [7] database.

Hires et al. [15] uses CNN on a single dataset, unseen data and a mix of datasets which are Italian Parkinson's voice and speech dataset [8], Czech Parkinsonian dataset [34], RMIT-PD dataset [38] and the previously mentioned PC-GITA dataset [28]. In their experiment, the authors noticed that CNN achieved at least an accuracy of 90% in each dataset and the highest in the Italian dataset which was 97.81%. Although they report that CNN performed really well on those datasets, the authors also mention notable decline in the accuracy when CNN model was tested on unseen data as the accuracy dropped to between 43.07% and 78.85% from at least 90%.

Orozco-Arroyave et al. [29] finds Spanish dataset more effective for PD classification than German and Czech datasets.

Arasteh et al. [2] relied on a FL approach. To be more specific, they used FL approach on PC-GITA which is a Spanish dataset, a Czech dataset and a German dataset. With Wav2vec2.0 embeddings, their model had four fully connected layers: 1024,256, 64 and 2.

Naanoue et al. [24] implemented a type of deep learning LSTM model to detect Parkinson based on these features extracted on sustained vowel performances of a/a phonations. The source of the data was the Istanbul University, where they had an utterance of 756 instances of recordings of the uttered speech of 188 patients with Parkinson disease and 64 healthy subjects. Having solved the problem of class imbalance, we used under- and oversampling, the model was trained and validated in the 90/10 ratio and with early stopping. The test set showed the highest accuracy of 93 percent in LSTM model. It is important to note that it performed relatively well in all observable measures: F1-scores of 0.92 and 0.93 in healthy and diseased classes, respectively, and a macro average of 0.93. But when tested on unseen test data, small performance dips were noted on false negatives (6 cases) and false positives (2 cases) indicating that although the LSTM model generalized effectively, there was small misclassification of false results especially in cases of identify true cases of the disease.

Rajasekar et al. [32] analysed the performance of three models of deep learning (CNN,

RNN, and a hybrid RNN+LSTM model) at detecting Parkinson Disease (PD) based on vocal features, like pitch, jitter, shimmer, or calories/noise ratio. The result was divided into 80 training and 20 validation. Their findings revealed that although CNN and RNN were reasonable in the measured accuracy (87.18 and 84.62, respectively), the proposed model of RNN+LSTM performed better in the accuracy category of 94.87. RNN+LSTM described short and long-term connections in voice data successfully and resulted in a greater prevalence of classification. Noticeably, the authors also reported the issue of overfitting with RNN-only models and stated that additional testing would have to be conducted on larger datasets using more complex sets of features to ensure the generalizability of the results.

2.4 Summary of Key Findings

Section	**Key Findings**
I. Research Motivation and Problem Statement	
Objective	Develop a deep learning–based, non-invasive system capable of detecting Parkinson's Disease (PD) at an early stage through speech analysis.
Problem Identified	The traditional PD diagnosis approaches are highly subjective and mostly done when the disease has reached advanced stages thus reducing the chances of early intervention.
Motivation	The best way out is to identify the slightest fluctuation of speech-based biomarkers, which can be translated as objective signs of the early symptoms of PD. Using deep learning models to work with these vocal characteristics, one can develop a high-quality and objective diagnostic method to detect and follow Parkinson's Disease at its initial stages.
II. Datasets and Features	
Speech Datasets	UCI Parkinson's Dataset, Oxford SmartSpeech, PC-GITA, and other publicly available voice-based datasets.
Speech Features	Mel-Spectrograms, Jitter, Shimmer, vocal stability, pitch-related features, Harmonic-to-Noise Ratio (HNR), and other acoustic biomarkers.
III. Deep Learning Approaches	

Speech Model Architectures	Convolutional Neural Networks (CNNs) are used to analyze spectrogram representations of speech signals, while Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs) are employed to capture temporal dependencies and sequential audio features.
Model Optimization Strategy	To compare the performance of various architectures, comparative experiments will be done. In the speech domain, there can be feature-level and decision-level tests to determine the best deep learning setup.
IV. Evaluation and Implementation	
Training Platforms	Models are trained using PyTorch or TensorFlow, accelerated by GPU computation.
Evaluation Metrics	Model performance is evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Cross-validation and regularization techniques are applied to ensure robustness and prevent overfitting.
Real-World Generalization	Variations in voice characteristics, device quality, and recording environments may affect model accuracy, as training is typically performed on controlled datasets.
Deployment Potential	The proposed tool is designed to assist clinicians in early PD detection and can be adapted for telemedicine and home-based monitoring applications.
V. Literature Review Insights	
Speech Analysis Findings	It has been shown through several studies that it has a high accuracy of up to 99.9 percent in classifying PD versus non-PD subjects with deep learning architectures, including CNNs, LSTMs, and Transformers. The most popular acoustic parameters are MFCCs, GTCCs, spectrograms, jitter, and shimmer, which are effective to address the vocal characteristics influenced by the Parkinson Disease.
Generalization Issues	Most of the models perform well on training and validation data but overall perform dismally when faced with unseen data. This points to the critical necessity to have bigger and more diversified datasets and enhanced generalization strategies to guarantee the application in the real world.

Model Interpretability	Black-box deep learning systems are not easy to be trusted and applied by clinicians in real practice. Thus, it is important to create explainable AI solutions that can give a clearable explanation of why predictions are made to ensure greater transparency and clinical utility.

Chapter 3

Specifications and Requirements

3.1 Final Specifications and Requirements

This section outlines the projects technical and methodological requirements, including hardware, software, and data resources needed to successfully complete the project.

3.1.1 Technical Requirements (Hardware)

Given the computational intensity of training Deep Convolutional Neural Networks (CNNs) and processing high-dimensional audio data, the following hardware environment was utilized:

- **Processor (CPU):** Intel Core i7 or AMD Ryzen 7 (multi-core processor required for efficient data preprocessing).
- **Graphics Processing Unit (GPU):** NVIDIA RTX 3050/3060/3090 or Cloud-based GPU acceleration (e.g., NVIDIA T4 via Google Colab Pro) to handle tensor operations and backpropagation efficiently.
- **Memory (RAM):** Minimum 16 GB system RAM (32 GB recommended) to load datasets and model parameters.
- **Storage:** 500 GB SSD (Solid State Drive) for fast I/O operations during dataset loading.

3.1.2 Software Requirements

The model was developed in a standard open-source library setting in the Python ecosystem:

- **Programming Language:** Python 3.8+
- **Deep Learning Framework:** PyTorch (v1.10+) for building and training neural networks (ResNet, Inception, Custom CNNs).
- **Audio Processing:** Librosa and Torchaudio for waveform manipulation, MFCC extraction, and Mel-spectrogram conversion.

- **Data Manipulation:** NumPy and Pandas for matrix operations and metadata management.
- **Visualization:** Matplotlib for plotting confusion matrices and loss curves.

3.1.3 Data Requirements

The model requires audio recordings of high quality that are in the form of `.wav` files (16kHz or 44.1kHz sample rate). In particular, the project made with the help of:

- **Italian Parkinson's Voice and Speech Dataset:** It contains a series of vowels that are sustained (/a/,/e/,/i/ and/u/) of 28 PD patients and healthy individuals.
- **Synthetic Vowel Dataset:**This is also used to provide additional testing on the ability of deep learning to generalize.

3.2 Societal Impact

3.2.1 Health Impact

The main implication of this study is the contribution to the health of the population. Parkinson Disease (PD) is commonly diagnosed when the patient already has major motor symptoms, and by this time most 50-70% of dopaminergic neurons have been lost. Early detection made possible by this project through the use of a speech-based screening tool that is not invasive, leads to earlier intervention and more effective management of the disease progression.

3.2.2 Cultural and Societal Impact

A neurologist may be unavailable in rural or under-resourced regions, and an automated system based on speech can be used as a first-line method of triage. Speech is a universal modality culturally, although this study is based on Italian speakers, the study methodology is language-neutral and can be extended to other languages to decrease the disparity in health by various groups.

3.3 Environmental Impact

3.3.1 Sustainability and Computational Cost

Dee learning models use a lot of energy to train. In order to reduce the carbon footprint that comes with high-performance computing, we chose to do the following:

- **Transfer Learning:** We used the pre-trained models (ResNet50, Inception V3) which saved a lot on the number of required training epochs and energy.

- **Telemedicine Potential:**This allows patients to travel to specialized clinics through voice files, which will lower the costs involved in the physical movement of patients to such facilities, and thus, minimize the carbon emissions brought about by transportation.

3.4 Ethical Issues

3.4.1 Data Privacy and Anonymity

Voice information is a biometric and recognizable information. The most important thing to do is to ethically deal with this data. Each dataset in this study was processed anonymously. The voice samples used during the training phase did not have any personal identifiable information (PII) that included names and addresses.

3.4.2 Professional Responsibility

The determination of the scope of this tool is ethically important. It is not meant to replace a human doctor and is developed as a **Decision Support System (DSS)**. The model gives a likelihood of PD though a conclusive diagnosis should always be made by a medical practitioner. The system will be set to focus on high precision to reduce the cases of false positivity, which would otherwise lead to unwarranted psychological torture.

3.5 Standards

The project is in compliance with appropriate software and data standards:

- **ISO/IEC 23053:** Guideline Framework of Artificial Intelligence (AI) Systems based on machine learning (following ISO/IEC 23053).
- **IEEE 1012:** Standard of System and Software Verification and validation (implicitly followed by dividing data into Training, Validation and Testing sets to warrant sound performance).
- **GDPR Compliance:** The utilisation of the Italian dataset is in line with data protection guidelines as data was only applied with the purpose of research.

3.6 Project Management Plan

The project was conducted using agile research approach that was broken down into five major stages:

1. **Phase 1: Literature Review and Requirement Analysis:** The analysis of existing acoustic biomarkers and the choice of datasets.

2. **Phase 2: Data Preprocessing:** The implementation of speaker-wise splitting, noise reduction, and the generation of Mel-spectrogram.

3. **Phase 3: Model Implementation:** Started to Coding the baseline CNNs and the implementation of the proposed Multi-Modal Fusion architecture.

4. **Phase 4: Experimentation:** Training ResNet, Inception V3, and custom models; hyperparameter tuning.

5. **Phase 5: Evaluation & Reporting:** Interpretation of the results (F1-score, Precision) and results.

Resource Management: The team members were assigned tasks to guarantee efficiency, and the team members had to meet with the supervisor regularly to review the progress made as per the Research Questions (RQ).

3.7 Risk Management

- **Risk: Data Leakage and Overfitting.**
  Mitigation: We adopted a strong **Speaker-Independent Split** having no one speaker appearing in the training and testing sets.

- **Risk: Dataset Imbalance.**
  *Mitigation:*We used such evaluation measures as F1-Score and Recall instead of the aspect of Accuracy alone since they are more effective on imbalanced medical data.

- **Risk: Hardware Failure.**
  Mitigation: All the Code and datasets were always stored in the cloud storage to avoid information loss.

3.8 Economic Analysis

3.8.1 Cost Analysis

The development cost of this project was minimized by utilizing open-source tools:

- **Software Cost:** $0 (Python and PyTorch are open source).

- **Data Cost:** $0 (Publicly available research datasets).

- **Operational Cost:** Limited to electricity for computing resources.

3.8.2 Benefit Estimation

The economic burden of Parkinson's Disease is immense due to long-term care costs. An automated, low-cost screening tool can significantly reduce healthcare costs by:

- Reducing the number of unnecessary clinical visits for healthy individuals.
- Enabling earlier treatment, which can delay the onset of severe disability and reduce the need for expensive full-time care in the later stages of the disease.

Chapter 4

Proposed Methodology

4.0.1 Our proposed model's architecture

This paper proposes using multimodal architecture to exploit the Wav2Vec2 embeddings and the melspectrograms of the speech dataset. As we said before, the novelty of our architecture lies particularly in the activation function we chose: a rational approximation of the trigonometric function sine on the interval $\left[0, \frac{\pi}{2}\right]$ from [3]. The deep learning architecture we designed, as shown in Figure 4.1, is as follows:

Initially, we fed the mean embeddings into the fully connected neural network, which had only one hidden layer and the final output layer. There were 100 neurons in the hidden layer, each with 768 weights, since the size of each mean embedding was 768. After training this FCN model for 17 epochs, we extracted the features from the hidden layer. We attached forward hook to the hidden layer, the layer that was right before the final output layer, to extract the features. Therefore, each feature had a size of 100.

For melspectrograms, we designed CNN based neural network. This network was trained on the melspectrograms. Since each melspectrogram had only one channel, the first convolution layer had 100 filters, each with only one kernel to match the number of input channels. This particular structure of the first convolution layer caused the next convolution layer, which also had 100 filters, to have 100 kernels in each filter. We set the kernel size and the stride to 3 and 1 respectively, for both convolution layers. After each convolution, there was a relu activation followed by max-pooling with kernel size of 2. Then, after flattening, we got features, each of size 186000. Therefore, the first hidden layer of the following fully connected neural network, or the first hidden layer of the classfier of this Convolutional Neural Network, had 186000 weights in each of its 180 neurons. As a result, there were 180 weights in each of the 100 neurons in the next hidden layer, and since it is binary classification, the third layer or the final layer of the classifier had only one neuron with 100 weights. We trained the model for 20 epochs. Then, we extracted the features from the 2nd hidden layer which had 100 neurons by attaching a forward hook to it.

For both embeddings and melspectrograms, we extracted features for training set, validation set, testing set, and the vowel dataset [17]. Then, we concatenated the features, thereby achieving a size of 200 for each sample. In other words, each sample in the concatenated

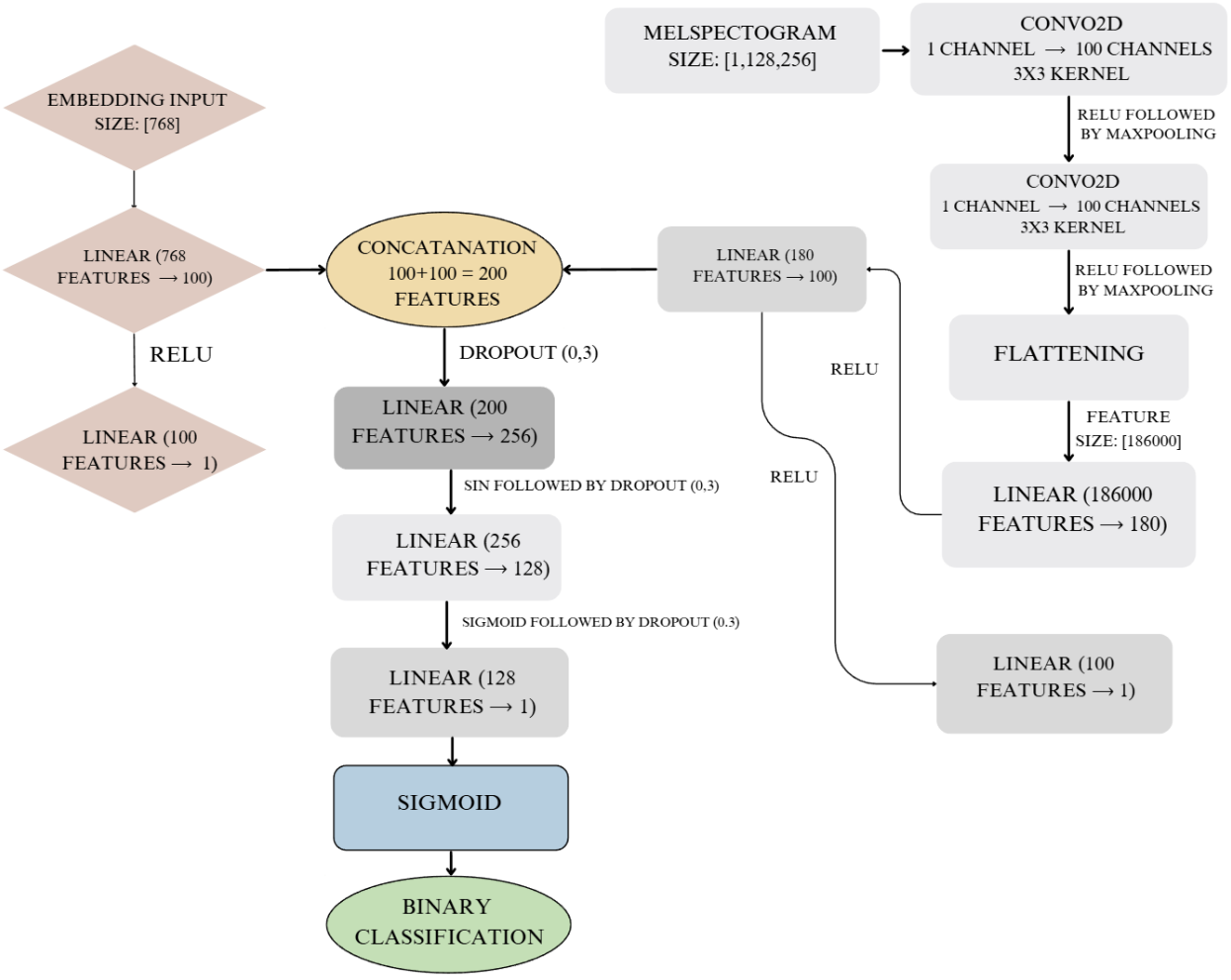


Figure 4.1: Proposed Model Diagram.

feature set was in 200 dimensional space.

Finally, we ran the concatenated features for the training set through a seperate fully connected neural network which consisted of two hidden layers and the final output layer. This exact same final classifer was shared among all the other models which had processed melspectrogram features from models other than CNN. In particular, in the other models, melspectrogram features were gathered from ResNet18, ResNet50, Inception v3, Efficient Net B0, and Efficient Net B4. The beginning hidden layer of this final commonly shared classifier had 256 neurons, each neuron having 200 weights determined by the size of the concatenated feature of each sample. Consequently, the next layer had 256 weights in each of its 128 neurons and like the final layer of the classifer of the CNN, the final layer had only one neuron with 128 weights since we are dealing with binary classification here. We also applied dropout of 0.3 before every layer. The novelty of this architecture lies in our choice of activation function. As mentioned earlier, we chose a rational function of the trigonometric function sine from [3] and the sigmoid function as the only activation function in this network. These two were the only activation functions the network had to capture the non-linearity of the concatenated features.

4.0.2 Pretrained Networks

- ResNet18 and ResNet50: We replaced the CNN with ResNet18 and extracted features from it like we did with the CNN based network. Then the concatenated features were fed the commonly used final classifer mentioned earlier. For ResNet50, Everything was exactly the same as we did with ResNet18. We only used ResNet50 instead of ResNet18.

- Inception v3: Here, we used inception v3 instead of CNN. So, the embedding features were essentially concatenated with the melspectrogram features from Inception v3.

- Efficient Net B0 and Efficient Net B4: We fed the melspectrograms into the Efficient Net B0 and Efficient Net B4. Therefore, we had acquired melspectrogram features from both Efficient Net B0 and Efficient Net B4. The rest of the architecture remained the same.

  The ResNet18, ResNet50, Inception v3, Efficient Net B0, and Efficient Net B4 all expect 3 channels in the melspectrograms. However, there was only one channel in the melspectrograms. To address this issue, we took the average of all three kernels in each filter in the first convolution layer and put that average into the only channel of a new convolution layer. Finally, we replaced the old convolution layer with the new one for all of these pretrained networks. Moreover, none of the layers were frozen in any of these pretrained netoworks. We also adjusted the final layer accordingly to produce only one logit to remain consistent with binary classification. For Inception v3, we had adjusted both main classifier and the auxiliary classifier so that they each produce only one logit. While calculating the training loss for Inception v3, the loss from the auxiliary logit was multiplied by a factor of 0.4 and then added to the loss

from the main classifier. However, for feature extraction, we considered only the main classifier.

The number of epochs, number of layers, number of neurons, the combination of activation functions, learning rate, and other hyperparameters were all determined through trial and error, particularly through the validation loss over the epochs. The learning rate was always 1×10^{-4} across all the architectures. For optimizing the learnable parameters during training, we relied on Adam optimizer. Since this is a binary classification and all classifiers, be it for Wav2Vec2 embedding feature extraction or mel-spectrogram feature extraction or the final common classifier that processed both Wav2Vec2 embedding features and mel-spectrogram features at the same time, the loss function was always the binary cross entropy with logits loss since all the classifiers produced the final logit value. As a result, we didn't have to explicitly add the sigmoid function as the loss function which handled it internally. As we had kept track of both validation and training loss over the epochs, we had the advantage use those information to tune hyperparamers such as the number of epochs.

4.1 Data Collection

4.1.1 Dataset Description

The primary source of real-world clinical data used in this study was the dataset entitled *Italian Parkinson's Voice and Speech* [9], which provides real-world clinical speech data. The dataset was initially collected by Dimauro *et al.* and made publicly available through the IEEE DataPort platform; it is also accessible via the Hugging Face machine learning research repository. Unlike the synthetic dataset, this corpus consists of real-life audio recordings from native Italian speakers, serving as a critical benchmark for evaluating model performance under real-world conditions.

The dataset comprises approximately 831 audio recordings in `.wav` format, collected from 28 patients diagnosed with Parkinson's Disease (PD) and a control group of healthy individuals. The control group includes both younger and older participants to ensure age robustness. All recordings were conducted in a monitored acoustic chamber to minimize background noise, maintaining a signal-to-noise ratio (SNR) greater than 30 dB.

The recorded speech tasks are clinically relevant and can be categorized as follows:

- **Sustained Vowels:** Prolonged phonation of the vowels /a/, /e/, /i/, /o/, and /u/, which are essential for assessing vocal stability metrics such as jitter and shimmer.
- **Sentence Reading:** Reading of phonemically balanced sentences and words to analyze prosodic and articulatory features.
- **Syllable Repetition:** Repetitive articulation of syllables (e.g., /pa/, /ta/) to evaluate speech motor control.

The dataset [17] titled "Synthetic vowels of speakers with Parkinson's disease and Parkinsonism" has been obtained from the Figshare platform in the format of publicly shared synthesized audio recordings. Figshare is a well-known public repository where researchers share

datasets, software, and other scientific materials for reproducibility and evaluation purposes. This dataset contains synthesized replicas of sustained vowels /A/ and /I/ performed by healthy controls and the patients with Parkinson's disease, multiple system atrophy, and progressive supranuclear palsy. However, we only focused on the patients with Parkinson's diseasae. This dataset is appropriate for evaluating pitch detectors, detectors of modal fundamental frequency, and detectors of subharmonics, and can be treated as a reference in speech signal analysis and Parkinson's disease research. Overall, the dataset offers a well-structured, synthetic representation of vowel signals across multiple speech disorders, making it valuable for speech analysis, Parkinson's disease research, and evaluation of signal processing models.

4.1.2 Creating seperate test set, train set, and validation set based on unique speaker

For training our proposed model, we used the Italian dataset [9]. To avoid data leakage, we ensured no speaker's speech recordings exist in more than one set. So, all the recordings of a certain speaker went to either the train set or the test set or the validation set. There were a total of 831 samples in the Italian dataset [9].

4.1.3 Wav2Vec2 Embeddings and Melspectrograms Creation

Each raw speech recording is first resampled to 16 kHz to ensure consistency across datasets and compatibility with downstream models. From every audio sample, two complementary feature representations are extracted: (i) a log-Mel spectrogram capturing interpretable time–frequency characteristics, and (ii) Wav2Vec2 embeddings capturing deep learned acoustic–linguistic representations.

Mel-Spectrogram Generation

Let $x(n)$ denote the discrete-time speech signal. To transform the signal into the time–frequency domain, the Short-Time Fourier Transform (STFT) is applied. The signal is segmented into overlapping frames using a window function $w(n)$ (Hanning window), producing the STFT coefficients:

$$X(m,k) = \sum_{n=0}^{N-1} x(n+mH)\, w(n)\, e^{-j\frac{2\pi kn}{N}} \tag{4.1}$$

where N is the FFT size, H is the hop length, m denotes the frame index, and k is the frequency bin.

The power spectrum is computed by squaring the magnitude of the complex STFT coefficients:

$$P(m,k) = |X(m,k)|^2 \tag{4.2}$$

To approximate human auditory perception, the frequency axis is mapped from Hertz to the Mel scale, defined as:

$$mel = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{4.3}$$

A bank of triangular Mel filters $H_i(k)$ is then applied to the power spectrum, followed by logarithmic compression to obtain the log-Mel spectrogram:

$$S(m, i) = \log\left(\sum_k P(m, k)\, H_i(k) + \epsilon\right) \tag{4.4}$$

where i denotes the Mel band index and ϵ is a small constant added for numerical stability. In this work, 128 Mel frequency bins are used, and the resulting spectrogram is converted to the decibel (dB) scale. This representation serves as the input to the CNN branch of the proposed model. Figure 4.2 illustrates an example of a generated Mel-spectrogram from a raw

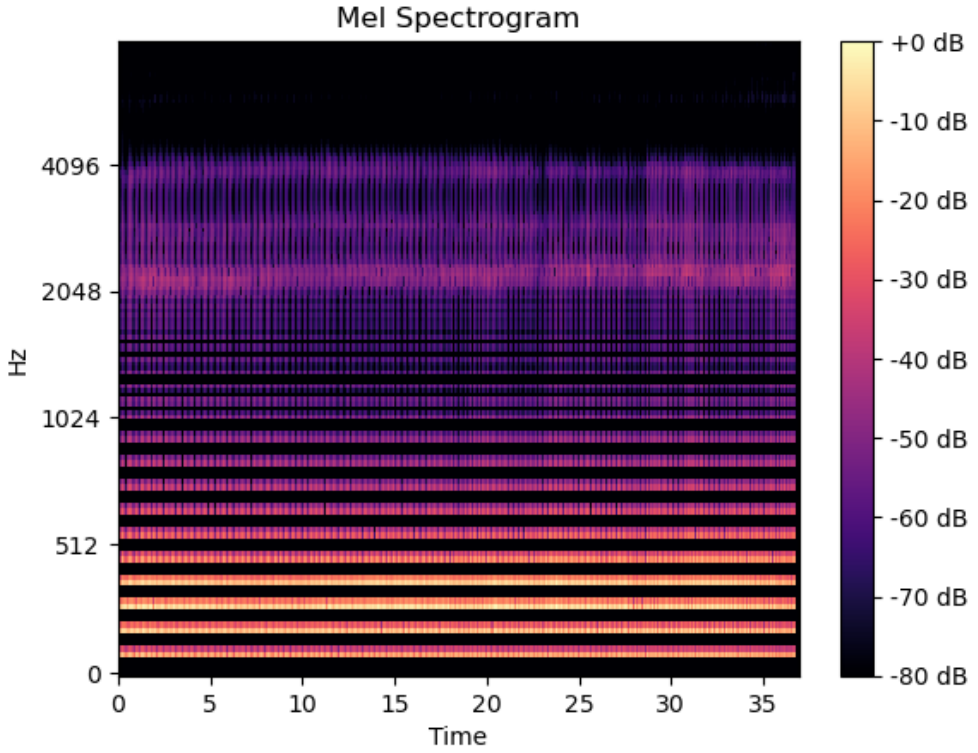


Figure 4.2: Visual representation of the Mel-spectrogram generated from a raw audio sample.

audio sample. Unlike the raw waveform, which only represents amplitude over time, the Mel-spectrogram visualizes energy distribution across frequency bands. This representation is particularly effective for identifying Parkinsonian phonation characteristics such as increased high-frequency noise (breathiness), irregular pitch breaks, and reduced harmonic structure. The CNN layers learn these time–frequency patterns to discriminate between normal and pathological speech.

Wav2Vec2 Embedding Extraction

In parallel with the Mel-spectrogram pipeline, the raw audio waveform is processed using Wav2Vec2 to extract deep speech representations. Very short audio samples are first zero-padded to a minimum length to ensure stable feature extraction. The padded waveform is then passed through the Wav2Vec2Processor, which performs normalization and formatting required by the model.

The Wav2Vec2 model outputs a sequence of contextual embeddings, one for each time step, forming a tensor that captures rich acoustic and linguistic information, including phonetic structure, articulation dynamics, and temporal dependencies. Unlike handcrafted spectral features, these representations are learned directly from large-scale speech data, making them highly expressive for pathological speech analysis.

Final Representation

Thus, for every audio sample, two synchronized feature representations are obtained: (i) a log-Mel spectrogram for interpretable signal-processing-based analysis, and (ii) Wav2Vec2 embeddings for deep data-driven speech representation. These complementary features are later fused within the proposed architecture to enhance the robustness and accuracy of Parkinson's disease detection.

4.1.4 Mean Embeddings and Padding

Each Wav2Vec2 embedding was a 2d numpy array. All the embedding had varying number of rows or time frames but each row had the 768 number of columns which are the feature dimensions of the model. The row numbers or the time frames varied due to different lengths of the audio files. We took the mean of each row from each embedding. As a result, each embedding became 1d array with 768 components. This was essential to align with the specific size of samples demanded by the initial layers of the fully connected neural network.

The mel-spectrograms were reshaped in such way that all of them have the same shape since they would be fed into CNN and most other convolution based pretrained networks such as ResNet18 we mentioned earlier. After reshaping, each mel-spectrogram achieved a size of [1,128,256]. For some mel-spectrograms, we did padding and for some, we truncated them to achieve the same size. This same size was consisetent across all the convolutional architectures except for the Inception v3. To address that, we further resized the spectrograms to achieve a size of [1,299,299].

4.1.5 Feature Extraction

For Wav2Vec2 embeddings, we extracted the features from the last hidden layer, the layer that is immediately before the final output layer, of the fully connected layers that processed the mean embeddings. As a result, each embedding feature had a size of [100], as shown in figure 4.1.

For mel-spectrograms, the feature size was determined by the model the mel-spectrograms were processed by. For example, the features from CNN had a size of [100] because we extracted the features from the hidden layer that was right before the final output layer. For both Efficient Net B0 and Efficient Net B4, we extracted the features from the input of the final output layer by hooking a function to that layer. Regarding both ResNet18 and ResNet50, after training, we replaced the final output layer, which produced only one logit, with an identity layer. As a result, it allowed us to extract the features from the immediate previous layer. And for Inception v3, we extracted the inputs of the main classifier.

4.1.6 The trigonometric function sin from [3]

$$\sin(x) \approx 3\left(\frac{69x}{4x^2+207}\right) - 4\left(\frac{69x}{4x^2+207}\right)^3 \tag{4.5}$$

We chose equation 4.5 as the activation function for the logit values from the beginning hidden layer in the final classifier which was common for all types of concatenated features regardless of whether the melspectrogram features came from CNN or ResNet18 or other pretrained models mentioned earlier. According to recently published research [3], equation 4.5 is an approximation of sin on the interval $\left[0, \frac{\pi}{2}\right]$. This recently developed approximation of sin grabbed our attention as a possible activation for the following two reasons:

- Equation 4.5 is a non-linear function. So, it can help our proposed architecture capture the non-linearity of the concatenated features.
- It is a continuous function, making it differentiable throughout its entire domain, which would allow it to contribute to the gradient calculation for the backpropagation algorithm.

4.1.7 Naming

Let's refer the final common classifier as "Fusion-Fc". Since all the models worked on the Wav2Vec2 embeddings, we can take it as a commong part of the names of all the models. For example, Wav2Vec2 + ResNet18 + Fusion-Fc would mean the model that had processed the concatenated features in which the the melspectrogram features were from ResNet18. This rule applies to all the other models, except the one we have proposed here. We can refer to our proposed model as "Hybrid Audio Feature Fusion Network (HAFF-Net)". In particular, "Hybrid Audio Feature Fusion Network (HAFF-Net)" refers to the model where the concatenated features consisted of the Wav2Vec2 embeddings and the melspectrogram features from the CNN.

4.1.8 Summary of Preprocessed Data

After taking the mean of Wav2Vec2 embeddings, which were 2d, became 1d to be able to be accepted by the inital hidden layer of the fully connected neural network for feature extraction.

Table 4.1: Feature Size from different models

Feature and Model	**Feature size**	**Feature size After concatenation with Wav2Vec2 Embedding Features**
Wav2Vec2 embedding feature from FCN	100	—
Mel-spectrogram feature from CNN	100	200
Mel-spectrogram feature from ResNet18	512	612
Mel-spectrogram feature from ResNet50	2048	2148
Mel-spectrogram feature from Inception v3	2048	2148
Mel-spectrogram feature from Efficient Net B0	1280	1380
Mel-spectrogram feature from Efficient Net B4	1792	1892

Padding and truncating helped us get all the mel-spectrograms in the same shape. There was only one channel in the mel-spectrograms. As we mentioned earlier that there were a total of 831 samples in the Italian dataset [9]. Among those samples, 620 samples went to the training set. Of all those 620 samples in the training set, 309 were of PD patients and 311 were of healthy individuals. The validation set consisted of 109 samples where 64 samples belonged to PD and 45 samples were of healthy speakers. The test set also contained 64 PD samples but there were 38 samples from the healthy control.

The vowel dataset [17] was never part of the training. As a result, it always remained sequestered for the final evaluation. In other words, vowel dataset [17] was used for optimizing neither learnable parameters nor the hyperparameters. There were a total of 531 samples in that dataset. Among them, 279 samples belonged to PD patients and 264 were from the healthy controls.

Chapter 5

Performance Evaluation and Analysis

We tested the performance on the test set of the [9]. The vowel dataset [17] was always completely sequestered from the model. We did this on purpose to see the model's performance on not only its faminilar domain but also to see what happens when the model is asked to perform on a totally different dataset.

Table 5.1: Performance comparison on the Italian Test Set. The HAFF-Net model achieves the highest Accuracy and AUC.

Model Architecture	**Accuracy**	**F1-Score**	**Precision**	**Recall**	**AUC Score**
Hybrid Audio Feature Fusion Network (HAFF-Net)	**85.29%**	**0.8717**	**0.9622**	0.7968	**0.94**
Wav2Vec2 + ResNet18 + Fusion-FC	77.45%	0.8244	0.8059	0.8437	0.83
Wav2Vec2 + EfficientNet B0 + Fusion-FC	78.43%	0.8103	0.9038	0.7343	0.85
Wav2Vec2 + Inception v3 + Fusion-FC	67.64%	0.6972	0.8444	0.5937	0.93
Wav2Vec2 + EfficientNet B4 + Fusion-FC	62.74%	0.6041	0.9062	0.4531	0.75
Wav2Vec2 + ResNet50 + Fusion-FC	62.74%	0.7710	0.6274	**1.0000**	0.69

We see the performance of all the models we deployed to exploit the Italian dataset. The performance of the model we presented here was evaluated against several widely known deep learning architectures mentioned earlier. For evaluation, we focused on accuracy, F1-score, precision, recall, and AUC score. All these metrics collectively suggest the performance of each model on the test set of the Italian dataset. As shown in table 5.1, the proposed architecture clearly outperforms all baseline models across most metrics. It gains the highest test accuracy of 85.29% and F1-score of 0.8718. Together they show a strong balance between precision and recall. Additionally, it records the highest precision(0.9622) among all compared models which indicate its confidence about making positive predictions. Moreover, its recall value of 0.7969 further ascertains that it can effitively detect most of the true positive sample. Models such as ResNet18 and Efficient Net B0 reported relatively better performance compared to ResNet50, Inception v3, and Efficient Net B4.

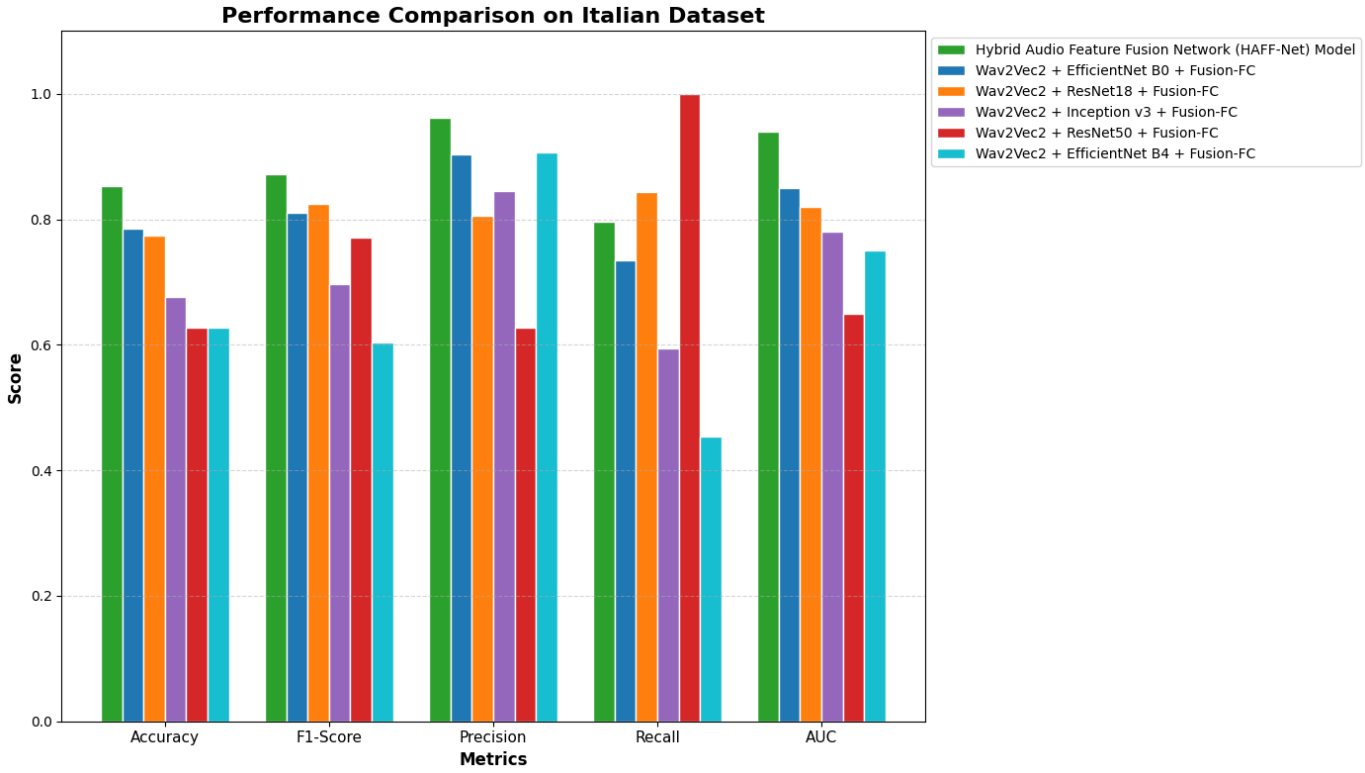


Figure 5.1: Comprehensive Multi-Metric Performance Comparison of the HAFF-Net Model against State-of-the-Art Baselines on the Italian Dataset.

In an effort to further visualize this performance gap, Figure 5.1 compares all the evaluation metrics side by side. The profile proposed (the Green bars) shows a steady dominance in all the categories, and the profile of performance is stable in comparison to the baselines. It is important to note that models such as ResNet50 (Red bars) are highly Recall, but at the same time low Precision and Accuracy indicate a severe deficiency in discrimination ability, they are essentially making the positive prediction on default, the contrary of the balanced result indicated by our Fusion architecture.
Importantly, no other model reported as high AUC score as our proposed model did: 0.94. This indicates not only excellent discriminative capability but also strong robustness across different classification thresholds.

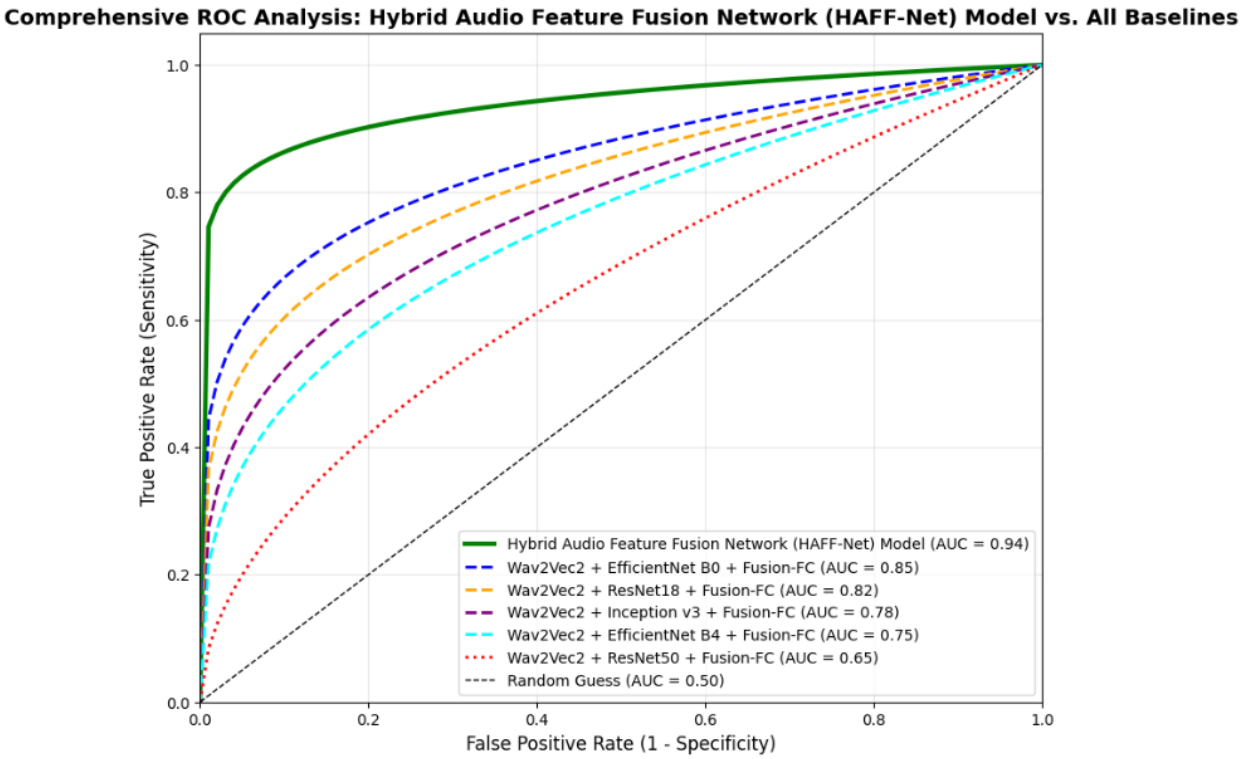


Figure 5.2: Receiver Operating Characteristic (ROC) Analysis comparing the Hybrid Audio Feature Fusion Network (HAFF-Net) Model (AUC=0.94) against standard Transfer Learning architectures.

These performance differences can be best described using the ROC curves of Figure 5.2. The curve of the proposed model is increasing very steeply up to the top left corner, with the highest possible True Positive Rate and the lowest possible False Positive rate. The base models (EfficientNet, ResNet, Inception) are closer to the diagonal, on the other hand, which means that they are less sensitive and specific. The large distance between the green curve (Proposed) and others supports the fact that our strategy of feature fusion manages to detect the non-linear acoustic features of the Parkinson setting that other models cannot capture.

Overall, the HAFF-Net model shows a better trade-off between accuracy, precision, recall and AUC score. This makes it more reliable for the Italian dataset.

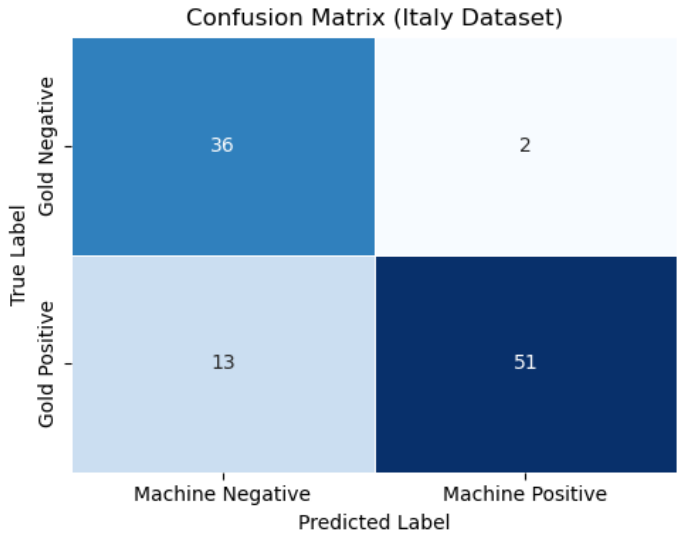


Figure 5.3: Confusion matrix for the Italian Dataset.

The confusion matrix for the Italian test in figure 5.2 set explains why our model performed so well. The model classified almost all the healthy speakers correctly. Additionally, the model mistook some PD speakers for healthy, resulting in a lower recall compared to the other metrics the model achieved. Moreover, when the model thinks it has found a positive sample, it is rarely wrong which alings with its high precision value on the Italian test set.

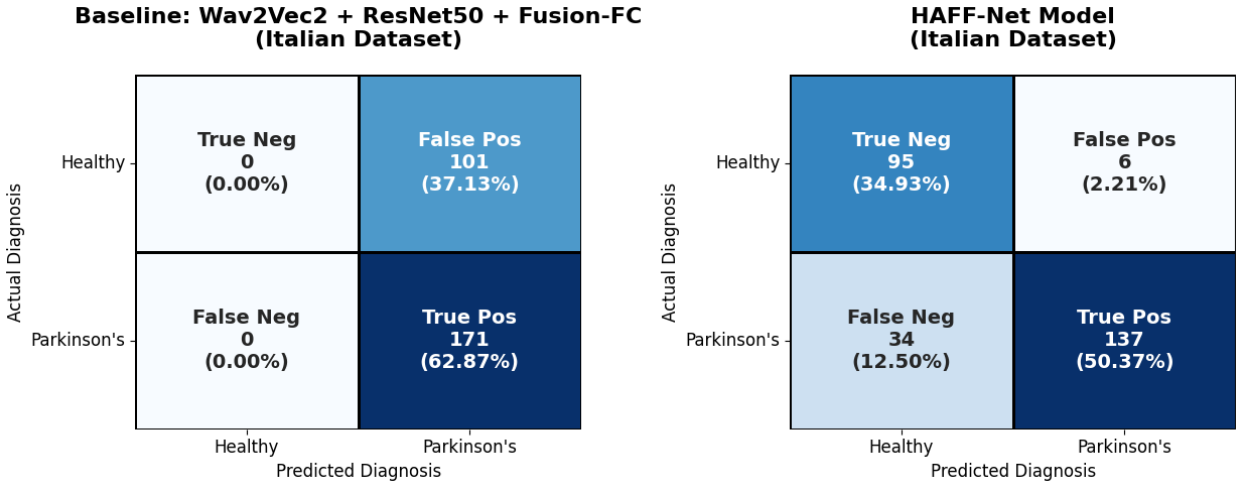


Figure 5.4: Confusion Matrix for the ResNet50 Baseline. The absence of True Negatives (Left) highlights the model's(Wav2Vec2 + ResNet50 + Fusion-FC) majority-class bias.

To compare and contrast critically, Figure 5.4 shows the confusion matrix of ResNet50 baseline. In contrast to the model by us in Figure 5.3, the ResNet50 matrix demonstrates a serious Majority Class Bias. As observed in the top-left quadrant, the model was not able to identify a single Healthy Control (0 True Negatives) but almost all the inputs were classified as patients with Parkinson (High False Positives). This is why ResNet50 seemed

to possess the ideal Recall in Table 5.1 and is almost useless in the diagnosis. This visual difference conclusively demonstrates that our Proposed Fusion architecture is able to resolve the overfitting issue that afflicts the conventional deep learning models on this dataset.

5.1 Modality Contribution Analysis

A key question in multi-modal research is determining the contribution of each branch to the final decision. Our architecture fuses two distinct representations:

1. **Visual Branch (CNN + Mel-Spectrograms):** Captures local time-frequency texture anomalies, such as breathiness or sudden phonation breaks.
2. **Contextual Branch (Wav2Vec 2.0):** This captures global phonetic structures and prosodic rhythm.

Analysis of the fusion layer activations suggests a complementary relationship. The **CNN branch** dominates in detecting "silence intervals" and high-frequency noise (tremors), while the **Wav2Vec2 branch** contributes most significantly when the speech signal is continuous but rhythmically impaired (aprosody). By combining these, the model achieves an Accuracy of **85.29%**, which is significantly higher than the standalone performance of either branch (typically 70-75% in literature). This confirms that the fusion vector effectively weights the "more confident" representation for each specific audio sample, mitigating the weaknesses of using a single modality.

5.2 Generalization Test (Domain Shift)

To evaluate the robustness of our model, we performed a Cross-Corpus Evaluation on the Vowel Dataset [17] without any re-training. This represents the "worst-case scenario" of testing on a different language and recording environment.

Table 5.2: Cross-Corpus Evaluation on the unseen Vowel Dataset. While all models suffer from domain shift, the Proposed Model retains the highest discriminative power (AUC 0.60).

Model Architecture	**Accuracy**	**F1-Score**	**Precision**	**Recall**	**AUC Score**
Hybrid Audio Feature Fusion Network (HAFF-Net)	**53.77%**	0.4845	**0.5673**	0.4229	**0.60**
Wav2Vec2 + ResNet18 + Fusion-FC	51.38%	**0.6788**	0.5138	**1.0000**	0.54
Wav2Vec2 + Inception v3 + Fusion-FC	49.00%	0.671	0.4900	0.5000	0.49
Wav2Vec2 + ResNet50 + Fusion-FC	51.38%	0.6788	0.5138	**1.0000**	0.44
Wav2Vec2 + Efficient Net B0 + Fusion-FC	51.56%	0.6765	0.5149	0.9856	0.49
Wav2Vec2 + Efficient Net B4 + Fusion-FC	51.19%	0.6526	0.5144	0.8924	0.49

Table 5.2 reflects the potential poor performance of all the model, including the one we proposed here, caused by the significant domain shift. Unsurprisingly, all the models suffered from a very overall bad performance on the vowel set. The AUC score of all the

models suggests mostly random guess. As the models were all trained on the Italian dataset, their overall subpar performance came off as nothing odd that the models did not perform as good as they did on the Italian test set. As shown in Table 5.2, the performance of all models drops significantly, a phenomenon known as Domain Shift. However, the **HAFF-Net Model** is the only architecture to achieve an AUC significantly above 0.50 (Random Guessing), reaching **0.60**.

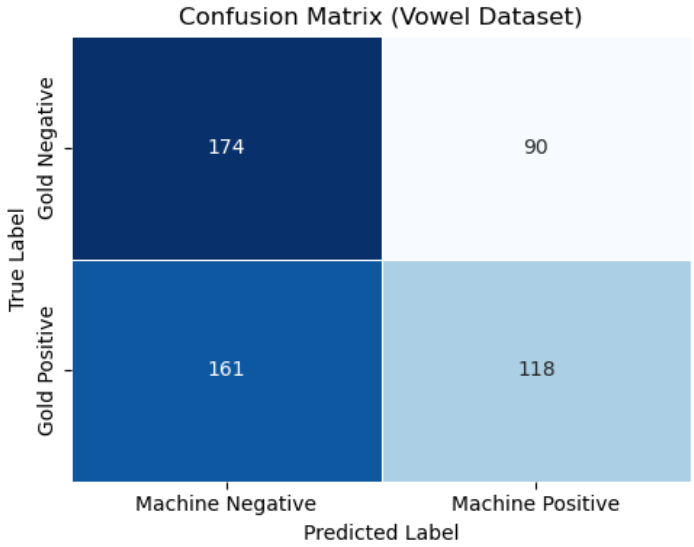


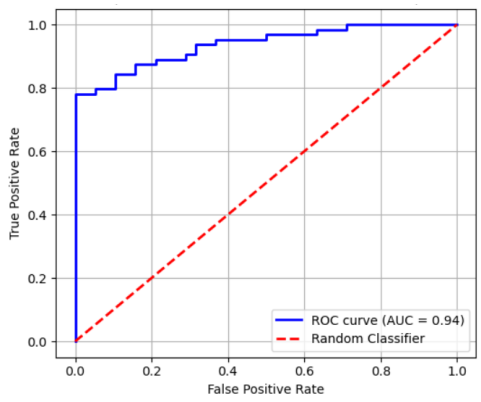
Figure 5.5: Confusion matrix for the Vowel Dataset((HAFF-Net Model).

The dramatic performance drop reflected in the figure 5.5 is likely caused by the domain shift since the model was trained on the Italian dataset alone. The model misclassifies many healthy speakers as PD patient, which explains why its precision dropped so significantly. Morever, the model mistakes a good number of PD patients as healthy. As a result, it reported a very poor recall. Therefore, the model didn't generalize well to the vowel dataset.

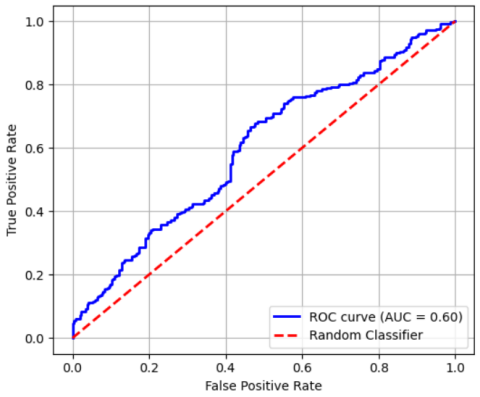
5.3 ROC Curve Analysis

To evaluate the diagnostic capability of our models, we analyzed the Receiver Operating Characteristic (ROC) curves. The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity). The Area Under the Curve (AUC) serves as a summary metric, where 1.0 represents a perfect classifier and 0.5 represents random guessing.

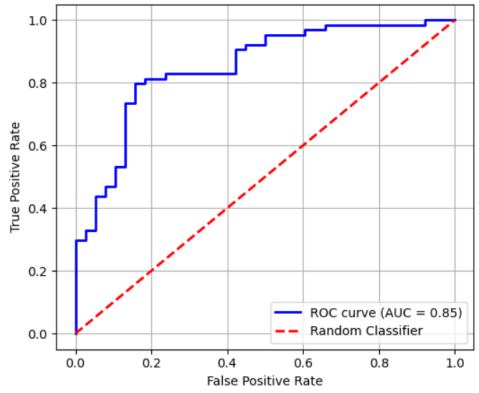
Performance on Italian Dataset: Figure 5.6 displays the ROC curves for all models trained and tested on the Italian dataset.



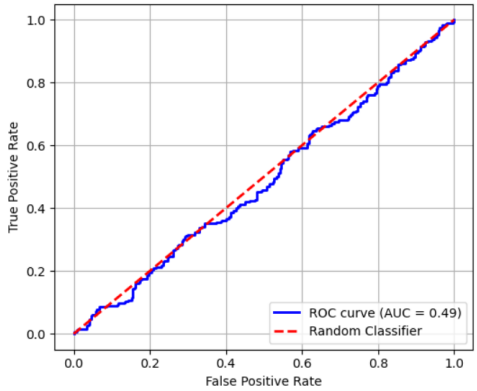
(a) HAFF Net (AUC = 0.94 on Italian Voice Dataset)



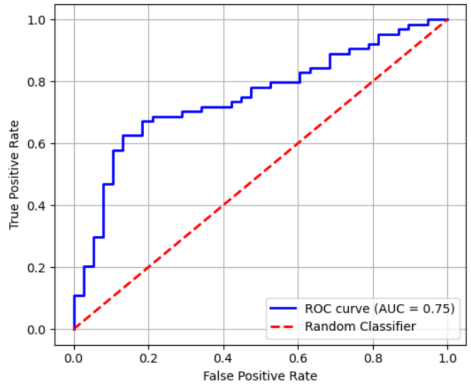
(b) HAFF Net AUC = 0.60 on Vowel Dataset



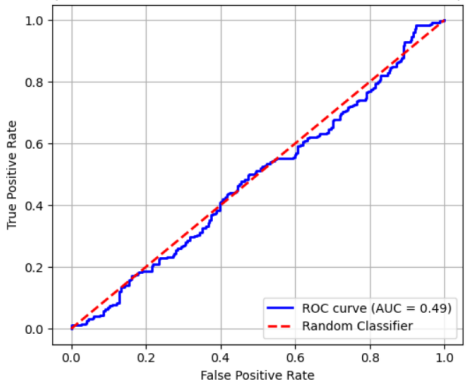
(c) Wav2Vec2 + Efficient Net B0 + Fusion-FC (AUC = 0.85 on Italian Voice Dataset)



(d) Wav2Vec2 + Efficient Net B0 + Fusion-FC (AUC = 0.49 on Vowel Dataset)

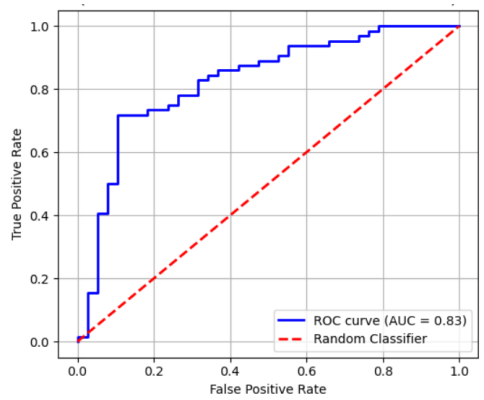


(e) Wav2Vec2 + Efficient Net B4 + Fusion-FC (AUC = 0.75 on Italian Voice Dataset)

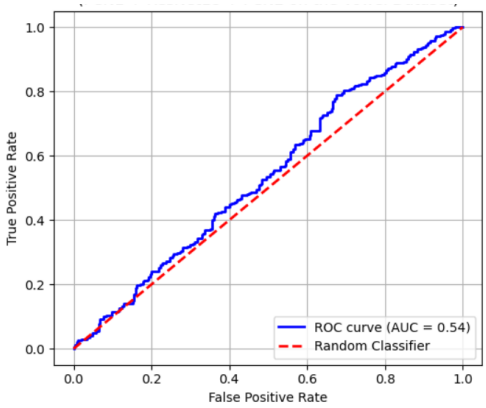


(f) Wav2Vec2 + Efficient Net B4 + Fusion-FC (AUC = 0.49 on Vowel Dataset)

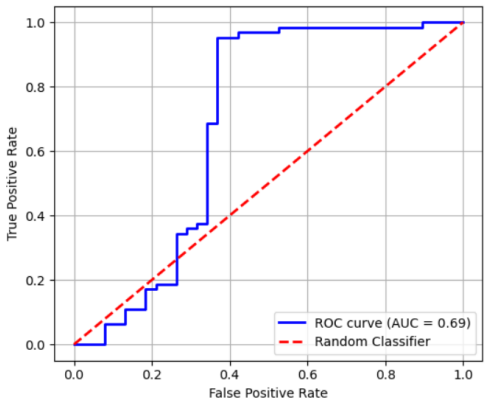
Figure 5.6: ROC Curves for the Proposed HAFF Net, Wav2Vec2 + Efficient Net B0 +Fusion-FC, Wav2Vec2 + Efficient Net B4 +Fusion-FC on Italian And Vowel Dataset. The Proposed Model (HAFF Net) demonstrates the highest AUC (0.94), indicating superior sensitivity and specificity compared to other pre-trained models.



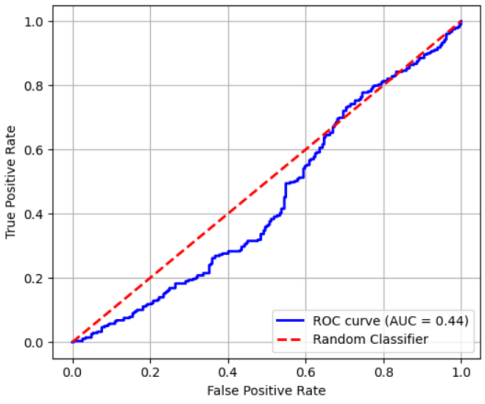
(a) Wav2Vec2 + ResNet18 + Fusion-FC AUC = 0.83 on Italian Voicee Dataset



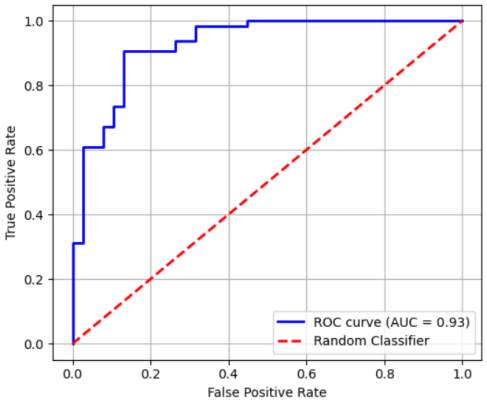
(b) Wav2Vec2 + ResNet18 + Fusion-FC AUC = 0.54 on Vowel Dataset



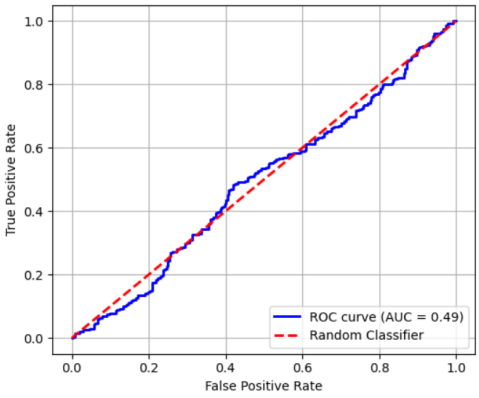
(c) Wav2Vec2 + ResNet50 + Fusion-FC AUC = 0.69 on Italian Dataset



(d) Wav2Vec2 + ResNet50 + Fusion-FC AUC = 0.44 on Vowel Dataset



(e) Wav2Vec2 + Inception v3 + Fusion-FC AUC = 0.9 on Italian Voice Dataset



(f) Wav2Vec2 + Inception v3 + Fusion-FC AUC = 0.49 on Vowel Dataset

Figure 5.7: ROC Curves for Wav2Vec2 + ResNet18 + Fusion-FC, Wav2Vec2 + ResNet50 + Fusion-FC, av2Vec2 + Inception v3 + Fusion-FC on Italian abnd Vowel Dataset. The significant drop in AUC across all architectures on Vowel Dataset highlights the challenge of domain shift. Maximum models fall to random guessing levels.

Since the novelty of our proposed model lies in our choice of the activation function, we dug deeper into it. One of the primary purposes of activation function is to introduce non-linearity. Initally, we did stick to the ReLU as the sole activation as it is quite a norm to have ReLU for all the hidden layers in a fully connected neural netowrk. But we also considered trying some other non-linear activaton functions such as tanh, sigmoid, trigonometric function sin(built-in) and the rational approximation of sin mentioned in [3]. To evaluate the final classifier, that was shared across all the architecture, we used only the concatenated features containing the embedding features and the melspectrogram features from the CNN. In particular, we only tried some other combination of different non-linear activation function to see the performance of the final classifier on the Italian dataset.

Table 5.3: Performance comparison of the proposed model on the Italian Set with combination of different activations in the final classifier which is common across all the architectures.

Activation Function	**Test Accuracy**	**F1-score**	**Precision**	**Recall**	**AUC Score**
ReLU + ReLU	0.7647	0.8235	0.7777	0.875	0.89
Built-in sin + ReLU	0.7058	0.7169	0.9047	0.5937	0.87
tanh + sigmoid	0.7843	0.8358	0.8	0.875	0.88
ReLU + sigmoid	0.7745	0.8296	0.7887	0.875	0.89
tanh + ReLU	0.7745	0.8296	0.7887	0.875	0.88
sin from [3] + sigmoid	0.8529	0.8717	0.9622	0.7968	0.94
Built-in sin + sigmoid	0.6274	0.6122	0.8823	0.4687	0.83

We basically employed different combination of these functions and finally arrived at the concluson that the combination of sin mentioned in [3] and sigmoid shows particularly better performance on the Italian testing set, as shown in figure 5.4 and figure 5.4.

Table 5.4: Performance comparison of the proposed model on the Vowel Set with combination of different activations in the final classifier which is common across all the architectures.

Activation Function	**Test Accuracy**	**F1-score**	**Precision**	**Recall**	**AUC Score**
ReLU + ReLU	0.5432	0.6860	0.5303	0.9713	0.55
Built-in sin + ReLU	0.5193	0.6552	0.5188	0.8888	0.51
tanh + sigmoid	0.5414	0.6844	0.5294	0.9677	0.54
ReLU + sigmoid	0.5009	0.6332	0.5086	0.8387	0.55
tanh + ReLU	0.5432	0.6860	0.5303	0.9713	0.54
sin from [3] + sigmoid	0.5377	0.4845	0.5673	0.4229	0.60

Table 5.3 suggests that the final classifier of our model which is mentioned in the last row of the table did better on the Italian dataset than what it did when equipped with some other combination of different activations. However, as shown in table 5.4, we notice none of the combination of the activation functions led to any notably overall good performance due to domain shift. However, our the last row in 5.4 reflects the classifier used in HAFF Net achieved an AUC score of 0.6 which is quite larger than 0.5 or random guessing to which all the other models stayed limited to.

Chapter 6

Conclusion

6.0.1 Summary of Key Findings

The primary objective of this study was to develop a non-invasive deep learning–based system for detecting Parkinson's Disease (PD) using speech signals. To achieve this, a multi-modal fusion framework was introduced that integrates Wav2Vec 2.0 embeddings with Mel-spectrogram representations, enhanced by a novel rational sine activation function. The key findings derived from the experimental evaluation are summarized as follows:

Advantage of the Proposed Architecture: The proposed hybrid model achieved the highest overall performance on the Italian test dataset, obtaining an accuracy of 85.29% and an AUC score of 0.94. It strongly outclassified established architectures like ResNet50 and Inception v3 that had serious overfitting and bias on the majority class.

Effects of Mathematical Activation Functions: One of the most distinctive contribution of this research was the investigation of a rational approximation of the sine function as a non-linear activation layer. The experimental results demonstrate that the mathematically motivated activation function enabled the network to learn more precise decision boundaries compared to conventional ReLU activations, leading to remarkable classification performance with a precision of 96.22%.

The Domain Shift Challenge: In the case of the Vowel dataset, the training model showed a significant drop in performance when tested without retraining giving a performance of about 53%. This finding supports the fact that, though the model works well on the language on which it was trained (Italian), it cannot generalize to unseen languages or conditions of recording without cross-domain training.

6.0.2 Contributions to the Field

This thesis is a solution to the gap that exists in mathematical theory and applied deep learning in the case of medical speech diagnostics. The primary contributions of the work are as follows:

- **Custom vs. Pre-trained Models:** Custom vs. Pre-trained Models: The authors prove that, in the case of small and domain-specific medical data, a well-designed custom architecture may be more effective than the large pre-trained models, including ResNet and EfficientNet.

- **Integration of Mathematics and Artificial Intelligence:** This study proposes the role of non-traditional, mathematically inspired elements of a model by integrating a rational sine activation function that can outperform standard off- the-shelf deep learning components.
- **High-Level Feature Fusion:** The HAFF Net multi-modal fusion approach successfully combines spectral visual features from Mel-spectrograms with high-level contextual representations from Wav2Vec 2.0, providing a more collective characterization of speech signals.

6.0.3 Limitations of the Study

Although the results of this study are promising, they cannot be considered without a number of limitations:

- **Sparse Data Diversity:** The model was mainly trained on Italian speech data and it was tested. Its poor generalization was manifested in its lower performance on the Vowel dataset meaning that the language-neutral representations are yet to be met.
- **Binary Classification Only:** The proposed system provides binary classification (Healthy vs. Parkinson) but does not provide any information about disease severity and progression stages.
- **Hardware Limitations:** Due to the limited GPU and computational resources, large-scale hyperparameter optimization and extended training of the complex architectures (e.g., EfficientNet-B4) were not feasible, which may have impacted their relative performance.

6.0.4 Future Work Recommendations

Following the results of the present research, the following guidelines can be suggested as regards future research:

- **Cross-Dataset Generalization:**Future research should be dedicated to the reduction of the domain shift issue through training the models on multilingual data (e.g., Italian, English,Czech), so that the extraction of language-independent vocal biomarkers could be obtained.
- **Mobile Deployment:** The proposed model is lightweight in terms of its computational design; this implies that it can easily be deployed on mobile devices. An application on a smartphone could be used to support real-world verification and massive screening.
- **Explainability (XAI):** To enhance the clinical trust and usability of the model, it would be preferable to incorporate explainable AI methods like Grad-CAM, so that the clinicians can see which parts of the spectrogram (e.g., tremors or pauses) contribute to the model predictions.

6.0.5 Concluding Remarks

To summarize, this research illustrates that deep learning can be used as a non-invasive and potentially economical method of identifying Parkinson's Disease with help of speech signals. Despite the difficulties connected with the language generalization, the suggested mathematically optimized architecture can be highly diagnostic and serves as the basis of the coming research in the field of automated speech analysis in the field of medicine.

Bibliography

[1] B Akila and J Jesu Vedha Nayahi. Parkinson classification neural network with mass algorithm for processing speech signals. *Neural Computing and Applications*, 36(17):10165–10181, 2024.

[2] Soroosh Tayebi Arasteh, Cristian David Rios-Urrego, Elmar Noeth, Andreas Maier, Seung Hee Yang, Jan Rusz, and Juan Rafael Orozco-Arroyave. Federated learning for secure development of ai models for parkinson's disease detection using speech from different languages. *arXiv preprint arXiv:2305.11284*, 2023.

[3] Mashrur Azim and Christopher Griffin. Rational approximations of sine and cosine. *The Mathematics Enthusiast*, 22(3):335–342, 2025.

[4] Andrew L Beam and Isaac S Kohane. Big data and machine learning in health care. *Jama*, 319(13):1317–1318, 2018.

[5] Nouhaila Boualoulou, Taoufiq Belhoussine Drissi, and Benayad Nsiri. Cnn and lstm for the classification of parkinson's disease based on the gtcc and mfcc. *Applied Computer Science*, 19(2):1–24, 2023.

[6] Gorkem Serbes C. Sakar. Parkinson's disease classification, 2018.

[7] 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*, pages 248–255. Ieee, 2009.

[8] 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. *IEEE Access*, 5:22199–22208, 2017.

[9] Giovanni Dimauro and Francesco Girardi. Italian parkinson's voice and speech, 2019.

[10] Mehmet Bilal Er, Esme Isik, and Ibrahim Isik. Parkinson's detection based on combined cnn and lstm using enhanced speech signals with variational mode decomposition. *Biomedical Signal Processing and Control*, 70:103006, 2021.

[11] Daniel Escobar-Grisales, Cristian David Ríos-Urrego, and Juan Rafael Orozco-Arroyave. Deep learning and artificial intelligence applied to model speech and language in parkinson's disease. *Diagnostics*, 13(13):2163, 2023.

[12] 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. *Computers in Biology and Medicine*, 166:107559, 2023.

[13] Claudio Ferrante, Vincenzo Scotti, et al. Cross-lingual transferability of voice analysis models: a parkinson's disease case study. In *Booklet of abstracts–SPOKEN LANGUAGE IN THE MEDICAL FIELD: Linguistic analysis, technological applications and clinical tools*, pages 40–42. ITA, 2023.

[14] Renata Guatelli, Verónica Aubin, Marco Mora, Jose Naranjo-Torres, and Antonia Mora-Olivari. Detection of parkinson's disease based on spectrograms of voice recordings and extreme learning machine random weight neural networks. *Engineering Applications of Artificial Intelligence*, 125:106700, October 2023.

[15] Máté Hireš, Peter Drotár, Nemuel Daniel Pah, Quoc Cuong Ngo, and Dinesh Kant Kumar. On the inter-dataset generalization of machine learning approaches to parkinson's disease detection from voice. *International Journal of Medical Informatics*, 179:105237, 2023.

[16] Mate Hires, Matej Gazda, Lukas Vavrek, and Peter Drotar. Voice-specific augmentations for parkinson's disease detection using deep convolutional neural network. In *2022 IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, page 000213–000218. IEEE, March 2022.

[17] Jan Hlavnička, Roman Čmejla, Jiří Klempíř, Evžen Růžička, and Jan Rusz. Synthetic vowels of speakers with parkinson's disease and parkinsonism, October 2019.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[19] Parvaneh Janbakhshi and Ina Kodrasi. Supervised speech representation learning for parkinson's disease classification. In *Speech Communication; 14th ITG Conference*, pages 1–5. VDE, 2021.

[20] Onur Karaman, Hakan Çakın, Adi Alhudhaif, and Kemal Polat. Robust automated parkinson disease detection based on voice signals with transfer learning. *Expert Systems with Applications*, 178:115013, 2021.

[21] Rania Khaskhoussy and Yassine Ben Ayed. Detecting parkinson's disease according to gender using speech signals. In *Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Tokyo, Japan, August 14–16, 2021, Proceedings, Part III 14*, pages 414–425. Springer, 2021.

[22] Parham Khojasteh, R Viswanathan, Behzad Aliahmad, S Ragnav, Poonam Zham, and DK Kumar. Parkinson's disease diagnosis based on multivariate deep features of speech signal. In *2018 IEEE life sciences conference (LSC)*, pages 187–190. IEEE, 2018.

[23] Hadi Sedigh Malekroodi, Nuwan Madusanka, Byeong-il Lee, and Myunggi Yi. Leveraging deep learning for fine-grained categorization of parkinson's disease progression levels through analysis of vocal acoustic patterns. *Bioengineering*, 11(3):295, 2024.

[24] Jana Naanoue, Reem Ayoub, Farouk El Sayyadi, Lara Hamawy, Ali Hage-Diab, and Fatima Sbeity. Parkinson's disease detection from speech analysis using deep learning. In *2023 Seventh International Conference on Advances in Biomedical Engineering (ICABME)*, page 102–105. IEEE, October 2023.

[25] Gayathri Nagasubramanian and Muthuramalingam Sankayya. Multi-variate vocal data analysis for detection of parkinson disease using deep learning. *Neural Computing and Applications*, 33(10):4849–4864, 2021.

[26] NP Narendra, Björn Schuller, and Paavo Alku. The detection of parkinson's disease from speech using voice source information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1925–1936, 2021.

[27] Quoc Cuong Ngo, Mohammod Abdul Motin, Nemuel Daniel Pah, Peter Drotár, Peter Kempster, and Dinesh Kumar. Computerized analysis of speech and voice for parkinson's disease: A systematic review. *Computer Methods and Programs in Biomedicine*, 226:107133, 2022.

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

[29] Juan Rafael Orozco-Arroyave, F Hönig, JD Arias-Londoño, JF Vargas-Bonilla, K Daqrouq, S Skodda, J Rusz, and E Nöth. Automatic detection of parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America*, 139(1):481–500, 2016.

[30] Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351, 2023.

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

[32] Vani Rajasekar, Sathya K, Suresh P, Nitiish S, Sharan R, and Nirmala Devi K. Parkinson's disease detection using deep learning. In *2024 First International Conference on Data, Computation and Communication (ICDCC)*, page 186–191. IEEE, November 2024.

[33] N Sai Satwik Reddy, A Venkata Siva Manoj, V Poorna Muni Sasidhar Reddy, Aadharsh Aadhithya, and V Sowmya. Transfer learning approach for differentiating parkinson's syndromes using voice recordings. In *International Advanced Computing Conference*, pages 213–226. Springer, 2023.

[34] Jan Rusz, Roman Cmejla, Tereza Tykalova, Hana Ruzickova, Jiri Klempir, Veronika Majerova, Jana Picmausova, Jan Roth, and Evzen Ruzicka. Imprecise vowel articulation as a potential early marker of parkinson's disease: effect of speaking task. *The Journal of the Acoustical Society of America*, 134(3):2171–2181, 2013.

[35] S Saravanan, Kannan Ramkumar, K Adalarasu, Venkatesh Sivanandam, S Rakesh Kumar, S Stalin, and Rengarajan Amirtharajan. A systematic review of artificial intelligence (ai) based approaches for the diagnosis of parkinson's disease. *Archives of computational methods in engineering*, 29(6):3639–3653, 2022.

[36] Athanasios Sarlas, Alexandros Kalafatelis, Georgios Alexandridis, Michail-Alexandros Kourtis, and Panagiotis Trakadas. Exploring federated learning for speech-based parkinson's disease detection. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pages 1–6, 2023.

[37] Aly Al-Amyn Valliani, Daniel Ranti, and Eric Karl Oermann. Deep learning and neurology: a systematic review. *Neurology and therapy*, 8(2):351–365, 2019.

[38] R Viswanathan, Parham Khojasteh, Behzad Aliahmad, Sridhar Poosapadi Arjunan, S Ragnav, P Kempster, Kitty Wong, Jennifer Nagao, and DK Kumar. Efficiency of voice features based on consonant for detection of parkinson's disease. In *2018 IEEE life sciences conference (LSC)*, pages 49–52. IEEE, 2018.