

MASTER THESIS
ARTIFICIAL INTELLIGENCE

Radboud University



**Cross-Lingual Parkinson's Disease
Classification using Few-Shot Transfer
Learning with Interpretable and
Non-Interpretable Features**

Author:

Indy Dolmans (S1008515)
Artificial Intelligence dept.
Radboud University
indy.dolmans@ru.nl

Supervisor:

dr. C. Tejedor-García
Centre for Language Studies
Radboud University
cristian.tejedorgarcia@ru.nl



February, 2025

Acknowledgements

This thesis would not have been possible without the invaluable support and contributions of several individuals. I am sincerely grateful to those who helped shape this work into something I can proudly reflect upon.

First and foremost, I offer my genuine thanks to my supervisor, Cristian Tejedor-García, for his dedicated time, insightful guidance, and constructive feedback during our weekly meetings. I am particularly grateful for his rapid responses to emails and messages, which ensured the project's continuity. His expertise and encouragement assisted me throughout this project.

I would also like to express my deep appreciation to Emmy and Lisanne, my fellow students working on their theses within the same project. Their presence at the university, thoughtful advice, and willingness to brainstorm ideas were both motivating and reassuring.

Special thanks go to the RAIVD project¹ for providing the opportunity to contribute to the important field of Parkinson's Disease diagnosis. Within the project, I am particularly grateful to Terry for his insights.

Lastly, to my office mates—thank you for all the yapping, shared coffee breaks, and good times that lightened the workload. Your listening ears and thoughtful perspectives provided motivation when challenges arose. And let's not forget the delicious snacks that surely played their part in getting this project to completion!

¹This work was supported by the NWO research programme AiNed Fellowship Grants under the project Responsible AI for Voice Diagnostics (RAIVD) - NGF.1607.22.013.

Abstract

Parkinson’s Disease (PD) is the second-most prevalent neurodegenerative disorder worldwide, affecting approximately 1% of individuals over the age of 60. Speech biomarkers are among the earliest signs of PD, allowing for voice diagnostics for early detection to start a timely intervention. However, linguistic differences in PD symptoms complicate classification in new languages. Transfer learning (TL) offers a promising approach for cross-lingual classification, though it has not yet been extensively explored in the context of PD classification. Recent advancements in Artificial Intelligence (AI) in healthcare have further encouraged its adoption in clinical practice. This thesis investigates the potential of few-shot TL for PD classification across languages using speech data, focusing on the performance of interpretable (IFM) and non-interpretable (NIFM) feature models. Through a series of experiments, the impact of increasing the fine-tuning set size on classification performance was explored, with the model trained on a base dataset and fine-tuned incrementally on a target dataset in a different language. This study demonstrates the feasibility of cross-lingual few-shot learning for PD classification, with NIFM showing slight but consistent advantages. The findings reveal that the zero-shot scenario, in which no target language fine-tuning occurs, results in chance-level classification performance for both IFM and NIFM. However, when fine-tuning using data from the target set, both feature models show significant performance gains. This performance improvement flattens as the fine-tuning set exceeds 50% of the target data. There is a consistent, though modest, advantage of NIFM over IFM across all experiments. Both feature types demonstrate improvement with fine-tuning, with NIFM maintaining a slight advantage even in the zero-shot setting. Furthermore, the results indicate that the relative effectiveness of IFM and NIFM is highly dataset-dependent. While base set performance declined as fine-tuning progressed, incorporating a balanced mix of base and target samples during fine-tuning partially mitigated this effect. Visual analyses and feature importance graphs highlighted language-dependent differences in the way PD symptoms manifest in voice features, demonstrating the challenge of cross-lingual diagnosis. This study emphasized the importance of data quality for effective TL, highlighting the need for dataset developers to establish and follow standardized data collection protocols. Future work will focus on developing more generalizable feature representations and exploring advanced fine-tuning approaches for cross-lingual voice-based PD classification.

Keywords— Parkinson’s disease, speech classification, cross-lingual, few-shot transfer learning, interpretable features, non-interpretable features

Contents

1	Introduction	4
2	Background	7
2.1	Artificial Intelligence (AI)	7
2.2	Automated PD detection	8
3	Related Work	11
3.1	TL	11
3.2	Few-Shot TL	12
3.3	Cross-Lingual Classification	13
3.4	IFM and NIFM	13
4	Methods	14
4.1	Material	14
4.2	Model Design and Training	19
4.3	Experimental Setup	21
4.4	Metrics	21
5	Results	23
5.1	Data Visualization	23
5.2	Feature Importance	24
5.3	NIFM Embeddings	26
5.4	Dataset Classification	26
5.5	Mono-Lingual Results	26
5.6	Cross-Lingual Results	27
6	Discussion	29
6.1	Discussion of Experimental Results	29
6.2	Base Set Performance	32
6.3	Dataset Variation Issues	32
6.4	Limitations	32
7	Conclusion and Future Research	34
7.1	Future Work	34
7.2	Afterword	35
A	Data Visualization	44
B	Cross-Lingual Results	46

1 Introduction

Parkinson’s Disease (PD) is the second-most prevalent neurodegenerative disorder worldwide, affecting approximately 1% of individuals over the age of 60 [1]. The disease is characterized by symptoms like slowed and stiff movement and tremor, caused by the loss of dopaminergic neurons in the substantia nigra [2, 3]. Additionally, cognitive decline and depression are linked with this loss of dopamine [4]. Many PD patients experience speech impairments. These impairments, likely caused by motor decline, are known as hypokinetic dysarthria and can significantly reduce intelligibility and eventually disrupt communication [5–7]. The condition depends on age and sex, with a negligible number of cases reported under the age of 50 and a higher prevalence among men compared to women [1–3]. The chronic disease progresses slowly, with symptoms worsening over time. No cure exists, although medication can alleviate symptoms [8]. Early diagnosis allows earlier treatment, greatly improving the quality of life of those affected [9]. In addition to the heavy burden PD places on individuals and their families, PD is associated with a large economic burden due to early retirement and the loss of income from informal caregivers [10, 11]. As the population ages, the incidence of PD is expected to rise significantly, with diagnosed cases projected to double between 2005 and 2030 [12]. This increase places growing pressure on health and economic systems as costs continue to rise [13]. Currently, there are no standardized diagnostic tests for PD, and assessments suffer from high inter-reader variability [8]. Commonly used diagnostic methods are subjective and time-consuming, highlighting the potential of Artificial Intelligence (AI) to address these issues [14].

Impaired speech production is often one of the first deficits observed in individuals with early-stage PD [9], with voice disorders affecting up to 90% of patients [3]. This grants possibilities for innovative technologies that enable rapid diagnosis using vocal biomarkers, analysed through voice diagnostics [15, 16]. AI methods aim to differentiate people with Parkinson’s (PWP) from healthy controls (HC) by identifying and learning decision boundaries based on distinctive features in speech data.

The initial step in the process of automated detection is the extraction of features. Currently, two primary methods of feature extraction are being explored in voice diagnostics. Traditional approaches focus on extracting interpretable features from audio recordings, such as speech biomarkers and features derived from waveform analysis. These features typically capture the phonatory, articulatory, and prosodic characteristics of speech [3, 17]. As these features are interpretable for humans, Favaro et al. [15] defines these as the Interpretable Feature Model (IFM). More recent methods rely on deep acoustic features generated by Deep Neural Networks (DNNs) [7, 15]. These *deep features* are essentially the model’s internal data representations. Since these features are produced by black-box neural networks that lack transparency, Favaro et al. [15] refers to these as the Non-Interpretable Feature Model (NIFM). Studies have shown that for PD classification, these non-interpretable feature models outperform their interpretable counterparts [18].

However, interpretability is highly valued in clinical settings. Clinicians are less likely to trust and adopt medical decision support systems in practice if they are non-transparent [19]. Due to the heterogeneity and complexity of medical data, interpretability is crucial in decision-making. As PD diagnosis is subjective and does not involve standardized tests, understanding the rationale behind the diagnosis is essential [8, 20]. Risks associated with unknown biases can only be mitigated when the model’s output is comprehensible. Providing interpretable results and safe generalizability of outcomes in clinical settings is more important than achieving high training accuracy [21].

Various AI models and techniques are being explored for PD classification, including

end-to-end learning, deep feature extraction, and transfer learning (TL) [22]. While end-to-end learning primarily concerns model architecture, deep feature extraction influences the feature set, and TL relates to a learning concept. TL leverages knowledge from one domain or task to benefit learning in a related domain, making it particularly valuable when data from the target domain are scarce or difficult to obtain. TL is promising because of its potential to perform well across languages, addressing a gap in speech-based PD studies. Currently, most research is focused on achieving high accuracy for the diagnosis of PD in a single language [22]. While some studies have explored cross-lingual approaches [15, 23–26], diagnosing PD in new languages remains challenging due to linguistic differences [27]. This limits applicability in real-world scenarios, as models cannot generalize without extensive new data. Collecting health-related datasets, particularly speech-based PD data, is time-consuming and expensive, further limiting real-world usability. Research aimed at scaling classifiers to perform across languages is important for clinical use. Few studies have explored cross-lingual TL for PD classification, focusing either on the comparison between IFM and NIFM [28], or on few-shot TL [29]. However, no study has combined these aspects by examining IFM and NIFM performance in a few-shot context while varying the number of shots.

This study investigates the potential of cross-lingual TL for PD classification using few target language samples. It also studies how varying sample sizes influence the performance of models based on interpretable and non-interpretable features in a few-shot context, an aspect that, to the best of our knowledge, has not yet been explored in cross-lingual TL settings. The following Research Questions (RQs) guide the project toward achieving these goals:

- **RQ1.** What is the impact of the fine-tuning set size on binary PD classification performance using speech data in cross-lingual few-shot TL?
- **RQ2.** How do IFM and NIFM compare in terms of binary PD classification performance in a cross-lingual TL setting as the number of samples increases?

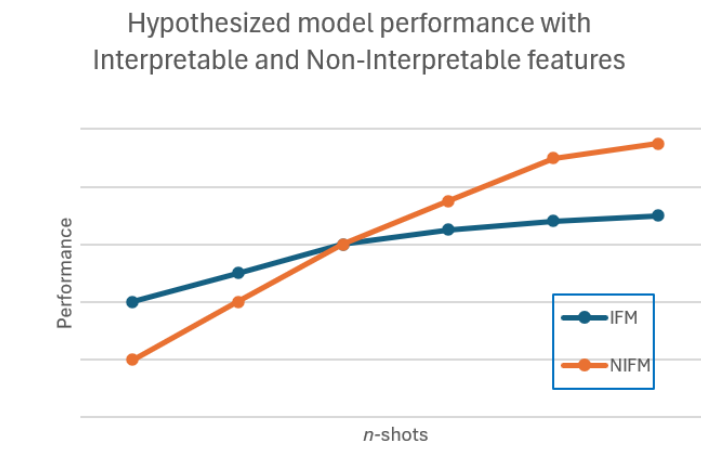


Figure 1: Hypothesis of the performance gain as the data set increases in size. IFM=Interpretable Feature Model; NIFM=Non-Interpretable Feature Model.

Regarding RQ1, we hypothesize that model performance improves as the fine-tuning set increases in size, an effect observed for Few-Shot TL across various domains [30–32]. Figure 1

illustrates an example of saturation, which means that as the fine-tuning set grows, more data results in flattening performance gains.

As for RQ2, we hypothesize that interpretable prosodic features (IFM) will outperform non-interpretable deep acoustic features (NIFM) when the fine-tuning set size is small. This is because NIFM are less likely to learn adequate representations of the new language with limited data. In zero-shot learning, where the model has not been fine-tuned for the target language, IFM is expected to perform similarly to NIFM, as observed by Hireš et al. [27] and Vásquez-Correa et al. [33]. However, as the fine-tuning set size increases, we predict a significant improvement in NIFM performance due to an adjusted fit of the internal embeddings, while IFM performance is expected to show only modest gains. This is consistent with findings that deep, non-interpretable features generally outperform interpretable acoustic features [15].

This study provides insights that are valuable for improving model generalizability and performance in multilingual settings by contributing to the integration of the following, which is novel.

1. Cross-lingual TL, particularly its application to PD classification with speech data;
 - 1.1. The impact of adding more data from new languages;
2. The performance of IFM and NIFM as the size of the dataset increases.

The outline of this thesis is as follows. Section 2 sets the project on a larger background. In Section 3, a comprehensive review of the existing literature is provided. Section 4 describes the methodology used in this study. Subsequently, Section 5 presents the findings and results of the experiments. Additionally, Section 6 provides a comprehensive discussion and interpretation of the findings. Finally, Section 7 concludes this thesis by summarizing the main findings and making suggestions for further research.

2 Background

This section provides a background covering the application of AI in healthcare, with a particular focus on its integration into speech-based PD detection. It examines various AI techniques to address the challenges of diagnosing PD using speech biomarkers. The section highlights current state-of-the-art methodologies for feature extraction and model development.

2.1 Artificial Intelligence (AI)

In recent years, AI has become a prominent research field, and the technique is already integral in modern society. Worldwide, industry and academia are investing heavily in AI, and major big-tech companies are spending substantial amounts to lead advances in AI research and development [34]. AI has the potential to reshape society, particularly through the automation of tasks. The effectiveness of AI systems depends on the quality of the data used for training [21, 35]. The use of low-quality data often results in unreliable or unusable outcomes, following the principle of ‘garbage in, garbage out’. High-quality, diverse, and representative datasets are essential for accurate and robust AI models.

AI research spans many domains, with healthcare emerging as a significant focus area. As the population ages and the shortage of caregivers increases, increasing demand for services and increasing healthcare costs pose critical challenges for society [36]. The application of AI in the healthcare sector is no longer unusual, and its adoption is increasing steadily [34]. With the growing availability of data, AI models are becoming more capable of learning patterns in the data to support clinical decision-making. Furthermore, disease diagnosis often suffers from low inter-reader agreement, where trained clinicians may interpret symptoms differently. AI can address this by its ability to process large datasets, offering consistent evaluations and uncovering patterns undetectable by human experts. With this capability, AI has the potential to enhance diagnostic accuracy and advance personalised medicine.

Although speech-based healthcare AI is less prominent compared to other healthcare AI systems, studies have been integrating auditory data in medical contexts [36, 37]. Most of the voice or speech research in AI in health is aimed at creating interactive voice assistants rather than disease detection using speech [38]. However, voice and speech as a health biomarker is gaining traction since its data collection is non-invasive [39]. This is primarily through diagnostics for pathological voice disorders, including Alzheimer’s Disease and Parkinson’s Disease.

While AI offers significant potential, it also introduces substantial risks. One big concern is the opaqueness of many models [21]. Often, it is unclear what these self-learning systems are capturing. This lack of transparency is particularly problematic in healthcare, where understanding and mitigating unwanted biases learnt from the data is crucial [19]. In clinical settings, models should be able to generalize well to ensure reliable application across diverse patient populations. Non-transparent models lack interpretability and struggle to gain trust, which limits their adoption in medical contexts. Furthermore, the restricted accessibility of high-quality healthcare data poses a barrier to scaling AI implementation, affecting broader applicability and impact [40].

As is evident in numerous medical fields, innovation in PD diagnosis is moving toward data-driven healthcare [5, 19]. Within this data-driven approach, data-centric AI is incorporated, which prioritizes improving data quality and usability by pursuing standards to develop representative, balanced, and well-annotated datasets [35, 41]. This is crucial for AI in healthcare, as the limited availability of sensitive medical data, strictly protected by

privacy regulations, presents challenges to the generalizability of results. The field is subject to many ethical considerations. It is crucial to gain insight into the biases present in the dataset [16]. Priority is given to ensuring interpretable results that can be safely generalized to clinical settings, rather than solely maximizing training accuracy [21].

2.2 Automated PD detection

PD was first described as ‘the shaking palsy’ by James Parkinson more than 200 years ago, referring to involuntary movements that look like shaking [1]. Parkinson described symptoms now known as bradykinesia, rigidity, and tremor [4]. Bradykinesia refers to slowed movements, rigidity relates to muscle stiffness, and tremor involves involuntary muscle contractions causing vibrations in the arms or legs.

In addition to the characteristic hand tremors, PD patients often experience speech disorders, which are linked to difficulties controlling speech muscle movement due to motor deficits. This condition is called hypokinetic dysarthria and is associated with increased noise production, slower speech rate, and reduced loudness [3]. Voice disorders affect an estimated 89% of PD patients, and symptoms commonly include reduced loudness, monopitch, breathiness, and a harsh vocal quality [9]. Although the exact neural cause remains unclear, deficits in precise muscle activation are often seen as the main cause of these vocal difficulties.

PD is a progressive disorder, with symptoms worsening over time. Disease severity is commonly assessed using the Hoehn and Yahr (H&Y) scale or the Unified Parkinson’s Disease Rating Scale (UPDRS). The H&Y scale, introduced by Hoehn and Yahr [42], categorizes the degree of disability into five stages, from minimal impairment in stage 1 to complete disability in stage 5. The (MDS)-UPDRS, initially developed in the 1980s and later revised by the Movement Disorder Society, consists of questions in four categories, each scored on a scale from zero (‘normal’) to four (‘severe’) [43]. Neurologists perform assessments of the H&Y scale and UPDRS based on the presence of symptoms [8].

Currently, diagnosis is made by neurologists or movement specialists based on diagnostic criteria. A recent study found a diagnostic accuracy of 97.2% for experts and 90.3% for clinicians [44]. Clinical misdiagnosis includes both false positives and false negatives [8, 20]. While false positives pose stress on patients and their families, false negatives lead to delayed treatment start. Automated diagnosis of PD using AI has the potential to reduce misdiagnosis [45].

Studies on automated diagnosis cover various aspects of automatic detection, including the creation of high-quality datasets, feature extraction, and modelling approaches using Machine Learning (ML) and AI models [22]. This project focuses specifically on speech-based detection of PD, using speech biomarkers to identify the disease. However, other modalities are also explored in the detection of PD, such as gait analysis [29, 46], electrogastrographic signals [47], and handwriting analysis [48]. As these modalities fall outside the scope of this thesis, they are not discussed further.

2.2.1 Datasets

Various datasets have been developed for speech-based PD classification, some of which are publicly available. These datasets, collected worldwide, naturally reflect different languages. Many are specifically designed for PD detection, incorporating speech tasks intended to capture the vocal changes associated with PD [49]. Several of the most recorded speech tasks are included in the following list.

- Sustained Phonation: Holding a vowel or consonant, such as /a/ or /m/, as steadily and as long as possible. PD patients are shown to have difficulty maintaining the same pitch and loudness.
- Reading task: Reading of (phonemically balanced) sentences or texts. This task highlights prosodic differences between PD patients and Healthy Controls.
- Diadochokinetic (DDK) task: Rapid repetition of the sequence /pataka/. This task highlights difficulties with rapid movement of articulators, particularly slowed lip and tongue movement in PD patients.

To ensure data quality, dataset developers must account for factors that could introduce variability. These factors typically include medication intake, as treatment effects can influence speech; a balanced distribution of age and gender, to minimize bias in the model; and the time since diagnosis, as it can significantly impact the severity of voice symptoms. In addition, environmental factors such as outside noise should be minimized. Although high-quality microphones and quiet rooms are typically used to collect datasets, this controlled setup may not replicate real-world conditions [50].

2.2.2 Features

The audio signal contains large amounts of information, which presents challenges for the models to determine the relevant characteristics without extensive data. Therefore, features are often extracted and used as input instead of the raw waveform to guide the model. These features can be either features from traditional speech analysis or features generated by neural networks. Features from traditional speech analysis are interpretable features derived from prior knowledge of PD, such as that reduced movement amplitude of the vocal cords leads to decreased pitch variability [3]. Features from neural networks, on the other hand, are obtained by passing data through a neural network and extracting the values of the nodes in a specific layer, which is the internal representation of the data in that layer [15, 22]. When using a pre-trained model, it is expected that the network has learnt to efficiently capture essential information, although this heavily depends on the training data used.

Since these features are sequential and vary over time, a decision must be made on how to aggregate their values for each recording. One common technique is data segmentation, where the recordings are divided into equal-length fragments and the features are calculated for each fragment. For example, Ferrante et al. [28] calculated the features for 4-second chunks, treating them as separate samples. An alternative approach involves calculating feature values for very short and stable chunks, such as 40-ms frames, which show minimal variability. The entire recording is then summarized using statistical descriptors of these values, such as mean, standard deviation, skewness, and kurtosis, as demonstrated by Laganas et al. [50], Vásquez-Correa et al. [29], and Scimeca et al. [51]. Another option is dimensionality reduction, like Principal Component Analysis (PCA), to scale down the feature space. While models using interpretable features may benefit from such techniques in terms of computational speed and accuracy, they are primarily applied to reduce the high dimensionality of embeddings extracted from deep neural networks [52].

Other feature types include Variational Mode Decomposition (VMD) and Empirical Mode Decomposition (EMD), non-linear transformations used for feature extraction from time series data [53, 54]. Both methods decompose the signal into simpler components via Intrinsic Mode Functions (IMFs), representing oscillatory modes intrinsic to the data. These decomposition methods were originally used for signal denoising, but were found to yield characterizing features for voice diagnostics.

Features based on time-frequency transformations, such as the Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-spectrograms, often show strong discriminating abilities [55]. Spectrograms, in particular, are widely used in PD classification from speech because of their ability to visually represent the frequency spectrum of audio signals over time. Since spectrograms are image-like representations, they are typically input into Convolutional Neural Networks (CNNs). These networks excel in finding patterns in image data through the convolution operation and have proven highly effective for PD detection, making them widely used for this task across various studies [5, 27, 33, 55–57].

2.2.3 Approaches

Researchers have applied various AI approaches to classify PD from speech data, achieving increasingly accurate results. Current studies primarily explore three directions: end-to-end models, typically based on CNN or Transformer architectures; deep acoustic feature extraction (DAFE) methods; and TL techniques [22]. Each approach has distinct advantages and limitations. While E2E models currently dominate the field, DAFE and TL have gained momentum in recent years. Studies indicate that DAFE methods outperform traditional acoustic features [15, 28]. DAFE models, known as black-box or NIFM, use embeddings from neural networks as features. In contrast, traditional IFM rely on prosodic features derived from waveform analysis [15].

End-to-end networks map a speech signal directly to an output and currently provide state-of-the-art performance in PD detection [22]. These models bypass feature extraction by directly processing the raw speech signal, learning patterns by internally processing the data.

DAFE relies on the internal representation of a pre-trained model for a specific data sample. These networks are often trained on large, diverse datasets within a specific modality, such as audio or image data. This allows them to capture generalizable aspects of data, stored in the hidden layers of the deep neural model. These features are non-interpretable for humans, as they consist entirely of neural network-derived values that lack inherent meaning [15].

TL addresses the challenge of limited data availability, a common obstacle across various domains [58]. This technique has gained traction in the classification of voice disorders, allowing the adaptation of knowledge from a trained model to a new related task [56]. In speech processing, TL often refers to the transfer of knowledge learnt on one language to another [59]. For instance, when PD characteristics in one language are understood, models require fewer samples to adapt to a new language. However, the language-dependent nature of PD characteristics makes generalizing the classification of PD across languages challenging [27].

3 Related Work

This section comprehensively describes the most recent research studies with a connection to this project. Table 1 presents an overview of the scientific works discussed in Subsection 3.1, Subsection 3.2, Subsection 3.3, and Subsection 3.4.

Study ↓	Language	Dataset	Features	Model	Performance	Connection
Vásquez-Correa et al. (2018) [29]	CO, CZ*, DE*	PC-GITA	Spectrogram	CNN	zero-shot: .5-.7 90%-shot: .6-.93	Few-shot TL
Rios-Urrego et al. (2020) [25]	CO, CZ*, DE*	PC-GITA	Spectrogram	CNN	multi-lingual: .58-.82 AUC	TL
Vásquez-Correa et al. (2020) [24]	CO, CZ*, DE*	PC-GITA	MFCC+BBE	SVM	cross-lingual: .70-.77 Acc	IFM/NIFM TL
Vásquez-Correa et al. (2021) [33]	CO, CZ*, DE*	PC-GITA	Spectrogram MFCC+BBE	CNN SVM	cross-lingual: .76-.84 AUC	TL
Karaman et al. (2021) [55]	US	mPower	Spectrogram	CNN	mono-lingual: .90 Acc	TL
Kovac et al. (2021) [26]	CZ*, US*		IFM	LR	cross-lingual: .55-.64 AUC	Cross-lingual IFM
Galaz et al. (2022) [48]	<i>Handwriting</i>		IFM/NIFM	CNN	<i>No speech class.</i>	Cross-lingual TL IFM/NIFM
Hireš et al. (2022) [57]	CO	PC-GITA	Spectrogram	FM	mono-lingual: .90 AUC	TL
Laganas et al. (2022) [50]	DE*, EN*, GR*		Onset-offset MFCC	SVM	cross-lingual: .50-.72 AUC	Cross-lingual MFCC
Favaro et al. (2023) [15]	CO, CZ*, DE*, ES, US, IT	PC-GITA, IPVS, NLS, NeuroVoz	IFM/NIFM	ML	cross-lingual: .55-.83 AUC	Cross-lingual IFM/NIFM
Ferrante et al. (2023) [28]	EN*, TE*		IFM/NIFM	CNN	cross-lingual: .40-.65 AUC	Cross-lingual IFM/NIFM
Ferrante and Scotti (2023) [60]	EN*, TE*		IFM/NIFM	CNN	Few-shot TL: .70-.90 Acc	Cross-lingual Few-shot TL IFM/NIFM
Hireš et al. (2023) [27]	CO, CZ*, EN, IT	PC-GITA, RMIT-PD, IPVS	Spectrogram	CNN	cross-lingual: .41-.70 AUC	Cross-lingual Fine-tuning
			IFM	XGboost	cross-lingual: .36-.74 AUC	Cross-lingual IFM
Scimeca et al. (2023) [51]	IT, ES, CZ*	IPVS, PC-GITA	MFCC, IFM	ML	multi-lingual: .65-.70 AUC	IFM
Tamm et al. (2023) [61]	EN*, GR*	<i>Alzheimer's disease</i>	eGeMAPS	NN	cross-lingual: 0.72-0.83 Acc	Cross-lingual TL
Tirronen et al. (2023) [62]	DE	SVD	MFCC/NIFM	SVM/FM	<i>No PD class.</i>	IFM/NIFM TL
Gimeno-Gómez et al. (2024) [18]	CO, CZ*, DE*, ES, PO	PC-GITA, NeuroVoz FraLusoPark	IFM+NIFM	FM	cross-lingual: .40-.79 F1	Cross-lingual IFM/NIFM
Kovac et al. (2024) [63]	CO, CZ*, IL*, IT, US*	IPVS, PC-GITA	IFM	XGboost	cross-lingual: .50-.79 Acc	Cross-lingual IFM
Quatra et al. (2024) [64]	CO	e-PC-GITA	Raw signal	FM	.82 AUC	Model fine-tuning
Veetil et al. (2024) [54]	IT, ES	PC-GITA, IPVS	VMD modes	CNN/ML	cross-lingual: .25-.73 Acc	Cross-lingual
Won and Kim (2024) [56]	DE	SVD	Spectrogram NIFM	FM	<i>No PD class.</i>	Few-shot TL
Zhao et al. (2024) [35]	EN, IT	MDVR, IPVS	MFCC	Triplet NN	mono-lingual: .99 Acc	Fine-tuning

Table 1: Recent studies in the field of automated PD detection with a connection to this project. BBE=Bark Band Energies; LR=Logistic Regression; FM=Foundation Model. CZ=Czech, DE=German, ES=Spanish, EN=English, GR=Greek, US=US English, CO=Colombian Spanish, IT=Italian, IL=Israeli, PO=Portuguese, TE=Telugu. * stands for unnamed private dataset.

3.1 TL

In speech-based PD classification, TL has been explored for fine-tuning foundation models or transferring knowledge across languages. Foundation models are deep, pre-trained models trained on large corpora, gaining prominence in the medical domain due to their capacity to generalize across tasks [40]. TL adapts these models for specific tasks, reducing the need for large task-specific datasets [40, 64]. Karaman et al. [55] demonstrated the potential of TL in

a monolingual context, achieving 90% accuracy on the mPower dataset by applying spectrograms as input to various pre-trained models. Similarly, Hireš et al. [57] achieved comparable performance on PC-GITA using a pre-trained ResNet50. Their multiple fine-tuning approach involved fine-tuning the model on vowel spectrograms and voice disorders as mediator sets before fine-tuning on PC-GITA. The mediator set bridged the semantic gap between the base model and the target task, improving the accuracy from 87% to 91% only when both mediator sets were included. Peng et al. [65] explored TL using the pre-trained OpenL3, fine-tuning it on a voice disorder dataset. They extracted embeddings from the fine-tuned model and applied dimensionality reduction methods to reduce embedding sizes before passing the reduced data to an SVM, achieving near 100% discrimination accuracy between four types of voice disorders. In another study, Chen and Jin [52] predicted Alzheimer’s disease from the output of automatic speech recognition (ASR) by extracting GPT-3 embeddings. These embeddings were transformed by RF-MDSA to align feature spaces across corpora and reduce dimensionality. They achieved accuracies between 56% and 73% in cross-lingual experiments involving Chinese, English, and Spanish.

TL also demonstrated its ability to transfer knowledge across languages, as seen in studies such as Rios-Urrego et al. [25] and Vásquez-Correa et al. [33]. In these works, models were trained on a base language and subsequently fine-tuned on a target language, using spectrograms as input for CNN models. Their results indicate that, while performance varies depending on the base language and target language, TL improves overall performance. For example, Vásquez-Correa et al. [53] observed that Spanish served as a good base language for Czech and German, with accuracies improving up to 14% over the baseline that did not use pre-training. Using the same features and CNN model, Vásquez-Correa et al. [29] investigated few-shot TL in cross-lingual settings. Zero-shot TL produced poor performance for most datasets, but significant improvements were observed as the training portion of the target set increased. Their study is the only one we found that applies few-shot TL to cross-lingual PD detection. Similarly, Ferrante and Scotti [60] applied CNNs to PD datasets in English and Telugu, finding cross-lingual TL results ranging from 45% accuracy (using acoustic features) to 65% (using VGGish embeddings), observing high variability between IFM and NIFM performance. This variability was also observed by Laganas et al. [50], who used datasets in English, Greek, and German with MFCC features and found that the AUC ranged from 0.42 to 0.65. In a study on Alzheimer’s disease classification, Tamm et al. [61] achieved a cross-lingual median accuracy of 76.1% in a Greek test set by fine-tuning an English-trained model with just 8 Greek samples. While both directions of TL research in PD detection are promising, no study has combined few-shot TL across languages while comparing IFM and NIFM performance as the number of shots increased.

3.2 Few-Shot TL

Few-Shot TL often involves incrementally increasing the number of target set samples added to the training set to evaluate their influence on model performance. On pathological voice data, Won and Kim [56] reported a notable improvement in accuracy when shifting from a one-shot (56.9%) to a five-shot (73.7%) approach. In other domains, few-shot TL has shown similar patterns. Sahoo et al. [31] examined social bias detection in tweets across different languages and found that adding more fine-tuning samples improved model performance by 8-51% on the F1-score compared to the zero-shot scenario, until reaching a saturation point at approximately 50% of target samples. They also observed that pre-training on one language could enhance transfer to some target languages but not others, highlighting that not all cross-lingual transfer is effective.

3.3 Cross-Lingual Classification

Various studies investigated the performance of models in a cross-lingual context by training a model on one language and testing it on another without fine-tuning. This approach assesses the model’s ability to learn language-independent characteristics. For example, Kovac et al. [26] used interpretable features to classify Czech and English speakers, reporting an AUC of 0.55 to 0.64 across languages. This indicates the model’s limited suitability for languages it was not trained on. In a follow-up study, Kovac et al. [63] extended the experiment to five datasets, finding substantial variability in performance across languages. Similarly, Veetil et al. [54] applied Variational Mode Decomposition (VMD), generally accepted to yield language-independent features, but observed accuracies between 25% and 73%, highlighting the challenges to achieve consistent cross-lingual performance. In other studies, a single model is trained on multiple languages, and its performance is evaluated across all languages. For instance, Scimeca et al. [51] combined MFCC and IFM of Italian, Spanish, and Czech datasets. When evaluating each language separately, they observed an AUC ranging between 0.65 and 0.70, demonstrating the potential for multilanguage generalization.

3.4 IFM and NIFM

Although interpretable features are preferred in medical contexts for their transparency, models using non-interpretable features often outperform them [15]. Comparing these feature types provides valuable insights into their applicability. Favaro et al. [15] applied ML models to six datasets using both interpretable and non-interpretable features, finding that in cross-lingual scenarios, AUC ranged from 0.50 to 0.83 for interpretable features and from 0.64 to 0.82 for HuBERT embeddings, which were the best performing non-interpretable features. While the results varied significantly depending on the language, the non-interpretable models seemed to better capture language-independent features. Conversely, Galaz et al. [48] investigated PD handwriting and found that interpretable features outperformed CNN-extracted features by 8-9% accuracy in 3 of 4 languages in a Leave-One-Language-Out sentence writing experiment.

Rios-Urrego et al. [25] investigated the impact of freezing different layers of a model during fine-tuning and found that TL improved cross-lingual accuracy with 2-17%. However, they also reported that the effectiveness of TL depends on factors such as the base language and the TL settings, including the number of frozen layers. Their findings highlighted that the optimal number of frozen layers differs with the variation of the base and target language, a conclusion supported by Peng et al. [65].

4 Methods

This section describes the datasets employed in the experiments and outlines the methodology applied to achieve the experimental results. The sequential steps of the project are illustrated in Figure 2, with each step discussed in detail throughout the section. Our methodology follows a typical ML pipeline and begins with data acquisition (Subsection 4.1.1), followed by data pre-processing (Subsection 4.1.3) and feature extraction (Subsection 4.1.4). Next, the process continues with model training (Subsection 4.2.2) and subsequent TL (Subsection 4.2.3).

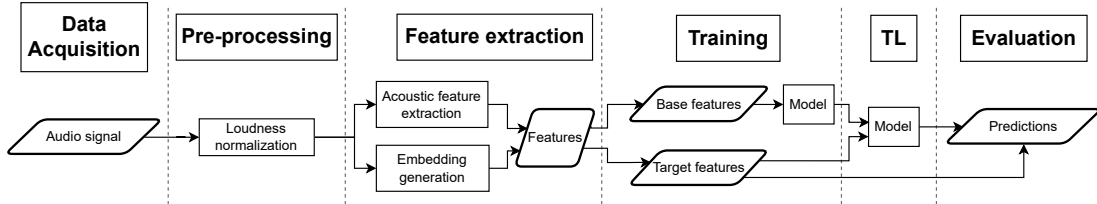


Figure 2: Pipeline of the experimental procedure.

4.1 Material

The initial steps of the pipeline involve data pre-processing, beginning with the acquisition of the datasets and the selection of relevant recordings. This is followed by pre-processing to enhance data quality. Subsequently, feature extraction is conducted, extracting two distinct types of features from the data, which are then normalized. Finally, labels corresponding to each feature map are stored, enabling the model to learn a mapping between features and labels.

4.1.1 Datasets

The data used in this study consisted of PD speech datasets recorded in various languages. These datasets include NeuroVoz (Castilian Spanish), PC-GITA (Colombian Spanish), IPVS (Italian), and MDVR-KCL (English). Each dataset was designed for speech-based PD classification, containing recordings of persons with PD and healthy controls (HC). To facilitate cross-dataset comparisons, we ensured alignment in the speech tasks. An overview of these datasets is presented in Table 2 and a brief description of each dataset is provided below.

Dataset	PC-GITA Colombian Spanish				NeuroVoz Castilian Spanish				IPVS Italian				MDVR-KCL English			
	PD		HC		PD		HC		PD		HC		PD		HC	
Group	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F
Nr of subjects	25	25	25	25	33	20	28	26	19	9	10	12	9	7	19	2
Age	62.2 (11.2)	60.1 (7.8)	61.2 (11.3)	60.7 (7.7)	71.9 (12.3)	70.9 (8.4)	66.6 (6.4)	68.4 (6.0)	68.6 (6.4)	64.3 (12.2)	69.3 (5.6)	65.3 (4.1)	-	-	-	-
UPDRS-III*	37.7 (22.0)	37.5 (14.0)			19.6 (11.8)	16.9 (11.5)			-	-					2.6*	(0.9)
H&Y	2.3	2.3			2.34	2.28			-	-					0.7 (0.9)	
Years since diagnosis	7.6 (4.7)	6.4 (6.4)			8.9 (5.9)	12.6 (11.5)			-	-					-	-

Table 2: Overview of datasets used in this project, reporting the mean values with standard deviations in parentheses. *MDVR-KCL reports score for item 18 (speech) of UPDRS-III. Other datasets report average UPDRS-III score.

NeuroVoz NeuroVoz is a private PD dataset collected in Madrid, Spain, and was released in 2024 [see 66]. The dataset was developed using data from native Castilian Spanish speakers [67]. Participants with neurological conditions other than PD or with other diseases that affect speech were excluded. The dataset consists of 53 individuals with PD and 55 HC. The mean age of the participants was 71 years in the PD group and 64 years in the HC group [67]. For the PD group, the average UPDRS score was 17.6 (± 11.7) and the mean H&Y scale score was 2.2 (± 0.7). All PD participants used medication, which was taken 2 to 4 hours before the recording session. The dataset includes recordings of the DDK speech task, 16 spoken sentences, sustained vowel phonation, and a monologue.

PC-GITA The private PC-GITA dataset was developed in Medellín, Colombia [68]. It is balanced in terms of gender and medical condition, with 25 male and 25 female speakers included in both the PD group and the HC group. All participants in the PD group took their PD medication no more than three hours prior to the recording session. The average UPDRS score for the PD group was 37.6 (± 18.0), and the mean H&Y scale score was 2.3 (± 0.5). Speech tasks included DDK, sustained vowel phonation, utterances of various sentences and words, a doctor-patient dialogue, and a monologue.

IPVS The Italian Parkinson’s Voice Dataset (IPVS) is a publicly accessible dataset², containing recordings of the DDK task, sustained vowel phonation, and various spoken sentences [69]. The dataset comprises 28 PD speakers and 22 HC. All PD participants were on PD medication during the recording session.

MDVR-KCL The Mobile Device Voice Recordings at King’s College London (MDVR-KCL) is an open dataset³, consisting of recordings from 16 PD patients (ranging from early to severe stages) and 21 HC. The speech tasks were limited to spontaneous speech and reading of a text. An expert assessed the UPDRS and H&Y scores, finding that one healthy control exhibited mild speech impairment, while one PD participant showed no signs of speech impairment [70].

4.1.2 Speech Tasks

In this thesis project, we focus on three speech tasks: the sustained phonation task, spoken sentences, and the DDK task, which were defined in Subsection 2.2.1. The sustained phonation task involves the phonation of the vowel /a/. Spoken sentences differ across datasets, requiring participants to read a predefined list of sentences or, for the NeuroVoz set, repeat spoken sentences [66]. The DDK task involves the rapid repetition of /pataka/ across all datasets.

4.1.3 Pre-processing

To ensure consistency between datasets, all audio files were downsampled to 16 kHz. This aligns with the required input format of the non-interpretable feature model. Additionally, audio normalization was performed using the open-source audio processing tool `ffmpeg`. In particular, the `ffmpeg-normalize-functionality` was applied to standardize audio files to a uniform loudness level, following the EBU R128 loudness normalization procedure [15, 18]. This adjustment noticeably enhanced volume consistency, particularly for recordings with low

²<https://iee-dataport.org/open-access/italian-parkinsons-voice-and-speech>

³https://data.niaid.nih.gov/resources?id=zenodo_2867215

amplitudes. Although we considered applying a high-low-pass filter to remove frequencies outside the natural speech band, we refrained from doing so as Zhang et al. [71] noticed that these frequencies might still carry relevant information for PD classification. Given the diverse background noise across datasets, applying this filter could remove potentially valuable information.

4.1.4 Feature Extraction

In this project, we distinguish between two feature sets: IFM and NIFM, following the terminology used in Favaro et al. [15]. The IFM set comprises interpretable prosodic features, while the NIFM set consists of deep feature embeddings extracted from a large pre-trained model. This dual approach allows us to assess the advantages of interpretable versus non-interpretable features in the context of speech-based PD classification.

To account for intra-recording variations and to limit the impact of recording duration on the classification results, both feature sets were created using 4 second long audio chunks, following the methodology used by Ferrante et al. [28]. Each audio recording was split into 4-second segments and the features were extracted from these chunks. These features served as independent samples for model training. Since it is impossible to reliably determine the values of several interpretable features from short chunks, segments shorter than half of the audio chunk length were discarded.

IFM In the context of audio analysis, interpretable features are those directly derived from the analysis of the audio signal. These features reflect the speech characteristics relevant to PD classification. Early automatic PD detection models utilized these features to differentiate between healthy speech and PD speech. Features were defined based on well-known deficits in PD voice, such as linguistic, acoustic, and cognitive impairments. Acoustic features and MFCCs have frequently appeared in the literature as key indicators. For this project, five of the most common feature sets identified in recent studies were selected (see Section 3). In total, 179 features were used in this thesis project. Table 3 groups these features into five categories. Below the table, each feature is briefly explained, along with evidence supporting its relevance for PD voice diagnostics. The selected feature categories were as follows:

- Fundamental frequency (F0)
- Formants
- Jitter
- Shimmer
- Mel-Frequency Cepstral Coefficients (MFCCs)

Feature	Applied in studies	#Features	Feature groups	Feature description
F0	[15, 18, 28, 72, 73]	6	Mean and standard deviation Minimum and maximum Mean first-order derivative Mean second-order derivative	Fundamental frequency distribution across the recording Min and max of the F0 in the recording Avg. speed of change in F0 Avg. speed of change of change in F0
Formant	[18, 51, 72]	6	Mean formants Standard deviation formants	Avg. frequency of the first, second, and third formant Standard deviation of the first, second, and third formant
Jitter	[5, 18, 26, 28, 51]	5	Absolute and percentage Jitter RAP PPQ5 DDP	Avg. difference between consecutive periods, absolute and relative Avg. difference between period and avg. of its neighbours Avg. difference between period and avg. of its four closest neighbours Avg. difference between consecutive differences between periods
Shimmer	[5, 26, 28, 51, 73]	6	Absolute and percentage Shimmer APQ3, APQ5, APQ11 DDA	Avg. difference between period amplitudes, absolute and relative Avg. difference between amplitude and its N neighbours Avg. diff. between consecutive amplitude differences between periods
MFCC	[24, 28, 35, 50, 51]	156	13 MFCC First-order derivative of 13 MFCC Second-order derivative of 13 MFCC	Spectral envelope description; distribution described by functionals: mean, standard deviation, skewness, kurtosis Speed of change in MFCC, described by above functionals Speed of change of change in MFCC, described by above functionals

Table 3: Description of interpretable features used in this project. In total, 179 features are used.

Muscle rigidity caused by PD often leads to a reduction in fundamental frequency variability (F0) during speech production [23]. Studies have shown that the variability of F0 is significantly lower for subjects with PD, attributed to a decrease in the amplitude of vocal cord movements [3]. For feature extraction, F0 was characterized by four features: the mean, standard deviation, minimum, and maximum of the F0 values found across the waveform. Furthermore, the mean values of the first and second derivatives of F0 were also used as features. All of these features are measured in Hertz.

PD is also associated with a harsh vocal quality, which affects vowel pronunciation. This quality is primarily determined by the energy present in the first formants (F1, F2, and F3) [74]. These formants reflect the position of the tongue during speech production. Since PD patients often suffer from a limited articulatory range, the bandwidth of formant frequencies tends to decrease, leading to formant centralization [72]. This means that high-frequency formants shift toward lower frequencies and vice versa. The first three formants were characterized by their mean value and standard deviation throughout the recording.

Jitter represents the variability in the pitch period between consecutive cycles, providing insight into the stability of the vocal fold vibrations [49]. Furthermore, several other measures of period perturbation, including RAP, PPQ5, and DDP, assess the difference between consecutive pitch periods. PD patients tend to show higher jitter compared to healthy controls, which can be used to distinguish between these groups [23].

Shimmer is derived from the variability in the maximum amplitude between consecutive cycles [49]. A higher shimmer value often correlates with voice disorders, as it indicates irregularities in vocal fold vibrations. Several shimmer measures are used in this thesis, such as APQ3, APQ5, APQ11, and DDA, which all capture the difference between the maximum amplitude of one cycle and that of its close neighbours.

MFCCs are frequently used in speech processing, capturing the short-term power spectrum of an audio recording [24, 28, 35, 50, 51]. For this thesis, the MFCCs were calculated as a 156-dimensional representation, derived from 13 MFCC features that describe the global shape of the frequency content in the audio signal, similar to Scimeca et al. [51] and Zhao et al. [35]. These features were extracted from multiple short time windows, allowing for a dynamic representation of the signal’s properties. Each window consisted of 512 frames, on a 16 kHz signal equal to approximately 30 ms. To characterize the distribution of these MFCCs throughout the recording, we used statistical measures that include mean, standard deviation, skewness, and kurtosis, similar to Scimeca et al. [51]. These described the overall distribution of frequencies in the audio file. Additionally, to capture the temporal changes in the frequency content, we computed the first- and second-order derivatives of the MFCC values between consecutive frames, leading to 13 + 13 additional features. The distribution

of these changes was also described by mean, standard deviation, skewness, and kurtosis.

All features, except MFCCs, were extracted using Praat⁴ software, widely recognized for voice diagnostics. Praat’s capabilities were accessed through a Python implementation that directly interacts with Praat’s internal code via the Parselmouth library⁵.

NIFM Regarding the non-interpretable feature set, we utilized embeddings from a neural network, capturing the model’s internal representation of the data. Even though the model was not specifically trained for PD classification, its pre-training on a related domain allows it to capture meaningful features. Based on a comparative study by Favaro et al. [15], the performance of HuBERT embeddings for PD classification was superior among five embeddings evaluated. Hence, we selected HuBERT embeddings as the non-interpretable feature set. This model originates from Facebook’s AI research department [75]. HuBERT processes raw audio signals to generate 1024 features for every 40 ms window. It was pre-trained on the LibriSpeech corpus, a dataset comprising more than 1000 hours of English read speech⁶.

Among the three model sizes, Base (90 million parameters), Large (300 million parameters), and X-Large (1 billion parameters), the Large model (`hubert-large-ls960-ft`) was selected to balance accuracy and embedding generation speed. The embeddings were extracted from the two internal layers of the model, and the layer that performed best in a comparative evaluation was used in our experiments. Each embedding contained 1024 features extracted from 40 ms windows. To summarize the feature distributions across each audio sample, four statistical functionals were applied to each feature: the mean, standard deviation, skewness, and kurtosis [29]. This reduced the $N \times 1024$ feature matrix to a 4×1024 matrix.

4.1.5 Feature Normalization

Normalization of features ensures that all feature scales are equal, improving the accuracy and efficiency of the ML model. One of the most commonly applied methods in PD classification is z-score normalization [76], also known as standardization, which standardizes data by subtracting the mean and dividing by the standard deviation. This project applied z-score normalization. The mean $\mu_{X_{train}}$ and standard deviation $\sigma_{X_{train}}$ were calculated based solely on the training set, since test set statistics are typically unavailable. Subsequently, both the training set and the test set were normalized using the statistics derived from the training set, as expressed formally in Equation 1.

$$x'_j = \frac{x_j - \mu_{X_{train}}}{\sigma_{X_{train}}} \quad \forall \quad x_j \in X_{train} \cup X_{test} \quad (1)$$

In TL, where models are trained incrementally as data become available, the full distribution of the training set is unknown a priori. To address this, the mean and standard deviation were incrementally updated following the method outlined in Chan et al. [77], formula (1.5a) and (1.5b). This method was implemented in `sklearn StandardScaler.partial_fit(...)`, allowing the model’s normalization parameters to adapt dynamically as new data points arrive over time. This leads to increasingly specific estimates of the normalization statistics.

⁴<https://www.fon.hum.uva.nl/praat/>

⁵<https://parselmouth.readthedocs.io/en/stable/>

⁶<https://www.openslr.org/12>

4.1.6 Classification Target

The classification performed in this thesis project is binary classification, predicting either PD or HC. This approach is consistent with most studies in the field, although some datasets include information on the severity of PD. Binary classification was chosen due to the limited data size, which provides insufficient samples across different severity levels for accurate estimation. Additionally, not all datasets include severity measures. Severity assessment also varies across datasets, using scales such as the UPDRS (ranging from 0 to 4) and the H&Y scale (ranging from stage 1 to 5) [42, 43]. These scales evaluate overall disease progression across motor, cognitive, and other symptoms, rather than being specific to speech disorders. This lack of specificity may lead to discrepancies, such as patients with severe speech symptoms but mild cognitive decline being assessed with a low overall severity score, complicating the model’s ability to learn patterns. Although item 18 of the UPDRS motor subscore (UPDRS III) specifically evaluates perceived speech disorder severity [78], this information is not available for three of the four datasets used in this study. As a result, we focus solely on binary classification.

4.1.7 Feature Visualization

To get a global idea of the spread of the data across the feature space, the data compressed into a two-dimensional space was visualized. We used t-SNE (t-distributed stochastic neighbour embedding) [79], a commonly used dimensionality reduction technique designed to represent high-dimensional data in a two-dimensional space. This method compresses complex information into two dimensions to enable visual assessment of patterns and separability within the data. This was done for all datasets combined, to observe patterns in the spread of the data, as well as for each dataset separately, to observe the spread across PD and HC groups.

A second method for assessing the distribution of features across datasets involved the automatic determination of feature contribution, applied to IFM features. Since NIFM features are non-explainable, projecting their individual contributions was not irrelevant. The `scikit-learn` implementation of the random forest classifier includes a built-in function to compute feature contribution based on impurity reduction. The computed values were accessed through the classifier’s feature importance attribute.

4.1.8 Dataset Classification

In addition to the feature visualizations, the feature space overlap among datasets can be assessed by training a classifier to distinguish between them. The sustained phoneme task was selected because it does not involve language-specific words. To eliminate the influence of language-dependent PD characteristics, only speech from healthy controls was used. The experiment focused on the classification of the NeuroVoz, PC-GITA, and IPVS datasets. The MDVR dataset was excluded because it does not contain recordings of the sustained phoneme task. Classification was performed using both IFM and NIFM to evaluate feature space differences between datasets.

4.2 Model Design and Training

The processed data were modelled using two types of models: A traditional ML model and a deep neural network. This section provides a brief overview of these models and describes the training procedures in detail.

4.2.1 Model Definition

This thesis project used two ML models for training, a traditional (baseline) model and a neural network, naturally adapted for TL and fine-tuning [25]. Specifically, the first model is a Stochastic Gradient Descent classifier, one of the few `sklearn`-implemented ML models that allows for incremental learning. Since this project is not aimed at evaluating differences between the ML and DNN models, we continue the analysis with the results of the best-performing model.

ML Model The Stochastic Gradient Descent (SGD) classifier implemented in the Python package `sklearn` is a linear model that optimizes the decision boundary between the two classes using the stochastic gradient descent algorithm. SGD is a widely used optimization algorithm that updates model weights incrementally based on the gradient of the loss function, making it suitable for handling large datasets. For the monolingual experiment, the SGD classifier was trained using the entire dataset with the `fit()` function, whereas the model in the cross-lingual experiment was incrementally trained using `partial_fit()`, allowing it to update parameters iteratively as new data points became available.

DNN Model The deep neural network (DNN) consisted of a simple four-layer architecture with ReLU activation functions. A dropout rate of 0.25 was applied to prevent overfitting. The first and second hidden layers consisted of 1024 nodes, the third hidden layer consisted of 512 nodes, and the final layer contained 256 nodes. The output layer was a fully connected layer with sigmoid activation, suitable for binary classification tasks. This DNN was designed to learn complex hierarchical patterns in the data, leveraging its architecture to handle diverse characteristics of the extracted features and speech data, providing a robust alternative to traditional models.

4.2.2 Model Training

The SGD classifier was trained using the default hyperparameters provided by the `sklearn` implementation of the `SGDClassifier`. This meant that a linear SVM was fitted using L2-regularization ($\alpha = 0.0001$).

The DNN model was trained for 10 epochs on the training set by minimizing cross-entropy loss. Training used the AdamW optimizer scheme with an initial learning rate of 1×10^{-3} , which was gradually reduced by a factor of 2 ($\gamma = 0.5$) every 100 iterations. This schedule facilitated stable convergence and reduced the risk of overfitting. The model was trained with batches of 64 samples.

4.2.3 Transfer Learning (TL)

The experiments conducted in the cross-lingual setting employed a TL approach based on the method described by Vásquez-Correa et al. [33]. In this methodology, the model was initially trained on the *training portion* of a dataset, referred to as the *base set*, and then fine-tuned on the *fine-tuning portion* of another dataset, referred to as the *target set*. Performance was evaluated on a separate partition of the *target set*.

The models are incrementally updated applying a few-shot approach, where the fine-tuning set is incrementally increased with one PD and one HC sample. Hence, a single shot is both a positive and a negative sample. The fine-tuning set was increased until only one positive or one negative sample remained in the test set, necessary to evaluate performance. During fine-tuning on data from the *target set*, we applied mixed-batch training, where data

from the *base set* was injected into the fine-tuning portion so that the model would not forget its previously learnt base set predictions, as proposed by Tamm et al. [61]. Since finding the optimal mixing parameter would fill a separate study, we fixed this value at 0.5. This means that for every 2 target set samples, one base set sample is added to the fine-tuning portion.

As for the SGD classifier, the incremental learning mode available in the `SGDClassifier`, implemented in `sklearn`, was used using the `partial_fit()` method. This allowed the classifier to adapt to the fine-tuning set incrementally.

The DNN model followed the same training scheme as previously described. After initial training, the model was fine-tuned for an additional 5 epochs on the fine-tuning set. Re-training occurred with the same hyperparameter settings as for initial training, except for a reduced initial learning rate of 1×10^{-4} . This adjustment facilitated gradual adaptation to the training set without losing the learnt patterns from the base set.

4.2.4 Few-Shot Learning

Our implementation of TL applies Few-Shot Learning, where we gradually increase the fine-tuning set size and measure the performance at every step, to obtain a graph with the performance gain as more data is used for fine-tuning [see 29, Figure 8]. Each step, one positive and one negative sample are added to the fine-tuning set. This process is repeated until we reach n -shots, where n is the lesser value of the number of PD and HC speakers, minus 1. The minus 1 is necessary to ensure one sample for performance evaluation. As the data sets are of unequal size, n differs per dataset. Therefore, the TL performance plots show a varying number of n -shots.

4.3 Experimental Setup

Initially, we started with the execution of a monolingual model, which served as a baseline for subsequent TL experiments. Cross-lingual experiments were then conducted to address the research questions. All feature extraction and experimental procedures were executed using Google Colab.

Consistency between monolingual and cross-lingual experiments was ensured by maintaining identical model architectures, cross-validation structures, and data-splitting methods. A crucial step to prevent biased results involved maintaining speaker exclusivity between training and test sets. Overlapping speakers could inflate performance metrics, as models might learn speaker-specific traits instead of those indicative of PD.

The experiments underwent k -fold cross-validation, with k defined as the smaller value of the sizes of the PD and HC groups. This effectively resulted in Leave-One-Subject-Out (LOSO) cross-validation. To account for random variations, each experiment was repeated five times. Consequently, the reported performance scores represent averages over $5 \times k$ runs. Each data split was stratified by PD status and gender, ensuring a balanced representation between training and test sets.

4.4 Metrics

Model performance was assessed using the Area Under the Receiver Operating Characteristic Curve (AUC) score, which is commonly applied in PD classification studies. For fair comparison to relevant literature, the AUC was chosen as performance metric. Additionally, this method is not misleading in imbalanced datasets, which is a major issue for the accuracy metric. When classes are imbalanced, a model can achieve high accuracy by favouring the

majority class while failing to distinguish minority class instances. Hence, all experiments report the AUC metric. A majority vote was used to aggregate the predictions at the sample level, providing a speaker-level prediction. The reported AUC scores reflect the correctness of these speaker-level predictions.

5 Results

In this section, the results are presented structurally. Before presenting our main experiment in Subsection 5.6, results of our exploratory analysis through t-SNE visualizations are presented in Subsection 5.1. This is followed by an examination of feature contributions to PD classification across datasets (Subsection 5.2). Next, we assess the performance of different NIFM model layers, supporting the choice of the final representation (Subsection 5.3). Then, results from a model trained to distinguish between datasets are presented in Subsection 5.4. The monolingual IFM and NIFM results are then briefly outlined to establish a baseline for the cross-lingual experiments (Subsection 5.5). Finally, the results of the cross-lingual few-shot TL experiments are reported in Subsection 5.6.

To compactly describe the results obtained, results reported in this section focus solely on the experiments using the TDU task, despite analysing three distinct speech tasks, because experiments using the other tasks showed similar patterns. The TDU task was selected because it most closely resembles the text-reading task of the MDVR dataset. Therefore, data visualization and cross-lingual results in this section are exemplified by the TDU task, highlighting differences between datasets and between PD and HC groups. The results for all three tasks are presented in Appendix B, Figure 9.

5.1 Data Visualization

First, to get a picture of the distribution of datasets and the distribution of PD and HC groups, an exploratory analysis was conducted through visualization of the data using t-SNE. The t-SNE visualization with all datasets combined, depicted in Figure 3, reveals notable differences between the feature types. NIFM features (Figure 3b) distinctly separate the datasets, whereas IFM features (Figure 3a) show a poorly differentiated plot with no clear patterns observable.

The t-SNE visualizations for individual datasets are displayed in Appendix A, Figure 8.

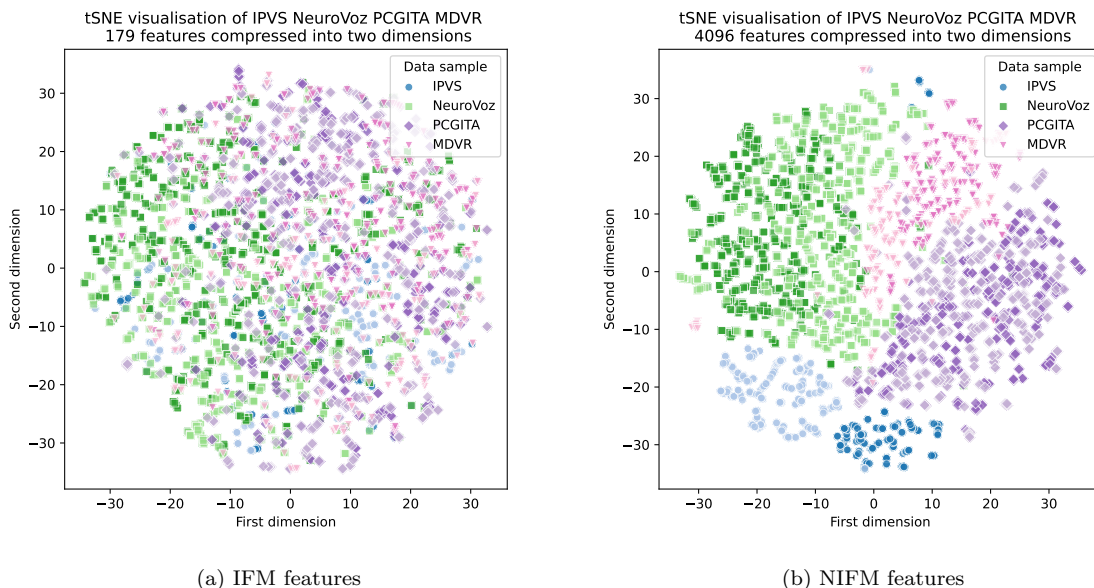


Figure 3: t-SNE visualization of IFM and NIFM extracted from four datasets. The darker shade indicates PD samples and the lighter shade indicates samples from HC.

NeuroVoz (Figure 8a) and PC-GITA (Figure 8b) visualizations show weak differentiation between the PD and HC groups, indicating minimal distinction between these speaker groups based on both IFM and NIFM. In contrast, the separation between PD and HC is visually apparent for the IPVS dataset (Figure 8c) and somewhat less pronounced for MDVR (Figure 8d), particularly for NIFM.

5.2 Feature Importance

To examine feature differences across languages, we trained a random forest classifier on IFM features for each dataset separately. The results were validated using five-fold LOSO cross-validation. The feature importance values of the TDU task are visualized in Figure 4. The y-axis lists all features, grouped by feature set, whereas the x-axis indicates their relative contributions to the model. The MFCC features are highlighted in three shades of blue: the lightest shade represents the MFCCs, with their first and second derivatives depicted in progressively darker shades. This visualization provides insight into how different features influence PD classification across different languages.

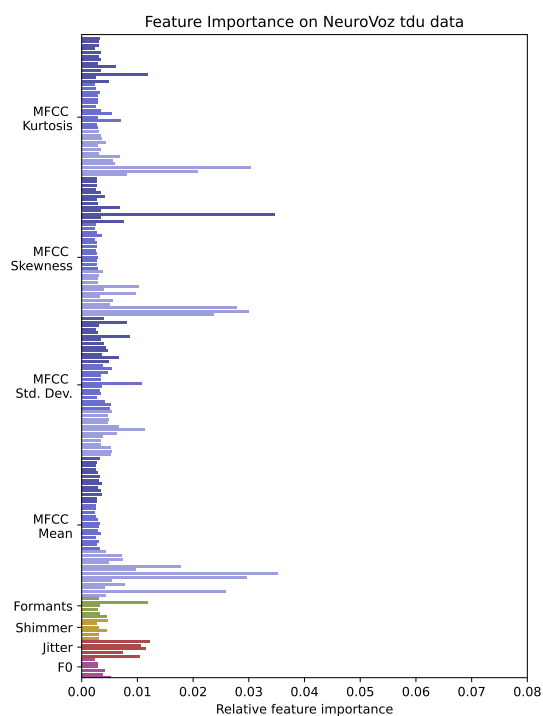
In general, the contributions of F0, jitter, shimmer, and formants varied substantially between datasets. Nevertheless, the mean MFCCs consistently played a key role, while their first and second derivatives made limited contributions. The model demonstrated a greater dependence on a small number of features for the IPVS dataset, with contributions of individual features reaching up to 8% to the model’s predictions, whereas for NeuroVoz, PC-GITA, and MDVR lower maximum individual contributions (around 4%) were observed.

In the NeuroVoz dataset (Figure 4a), the first few MFCCs, particularly their mean (first ten), skewness, and kurtosis (both first three), made the largest contribution to the classification of PD. These were followed by features related to jitter, which were relatively more important for this dataset compared to others.

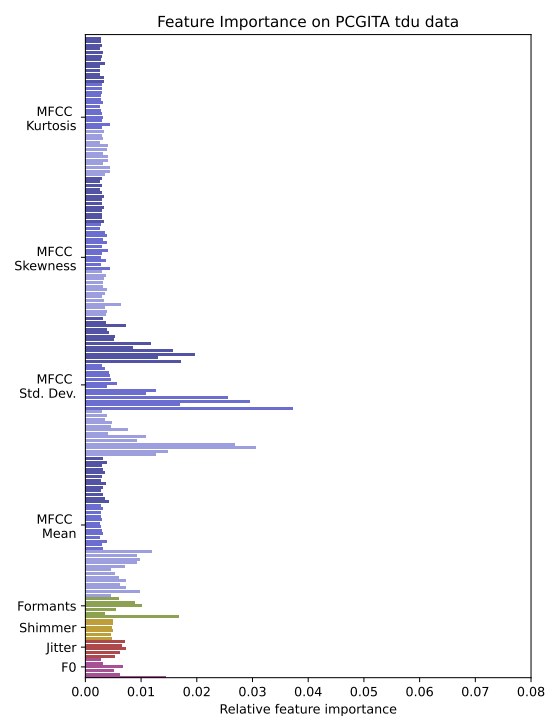
The PC-GITA dataset (Figure 4b) showed notable contributions from the standard deviations of MFCCs and their derivatives. Similar to other datasets, the mean values of the 13 MFCCs also played a significant role.

In the Italian IPVS dataset (Figure 4c), shimmer features stood out as the main contributors, which is notably different from the other datasets. In addition to these features, the mean of the 13 MFCCs contributed up to 8% per individual feature.

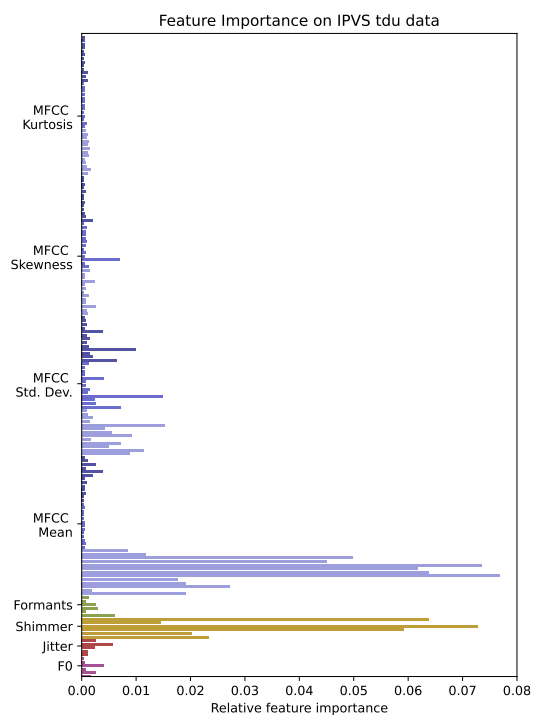
For the English MDVR dataset (Figure 4d), the model primarily relied on the mean of the first three formants and their derivatives, which contrasted with their contributions in other datasets. Features related to fundamental frequency also played a notable role, along with the consistently important MFCC mean.



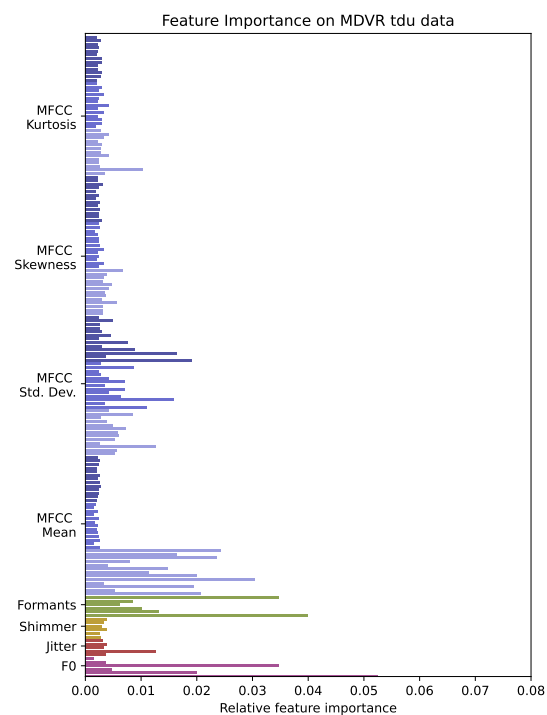
(a) Feature contribution NeuroVoz



(b) Feature contribution PC-GITA



(c) Feature contribution IPVS



(d) Feature contribution MDVR

Figure 4: Feature significance across datasets, as learnt by a Random Forest Classifier. Feature categories are represented by distinct colours. The 13 MFCCs (light blue), along with their first two derivatives (darker shades of blue), were characterized by four statistical functionals.

5.3 NIFM Embeddings

To determine the optimal representation of the HuBERT model, that provides two embedding layers, we evaluated the performance of both layers. We selected the embedding that demonstrated the best performance for PD classification. The performance was assessed across three monolingual experiments: the PC-GITA DDK task, the NeuroVoz listen-and-repeat task, and the SP task in IPVS. The results, illustrated in Figure 5, indicate an AUC between 0.08 and 0.14 higher for models trained with the first hidden representation compared to the last in two of three experiments and an insignificant difference in the third (0.007 difference). Based on these findings, subsequent experiments focused exclusively on the first hidden representation of HuBERT as NIFM.

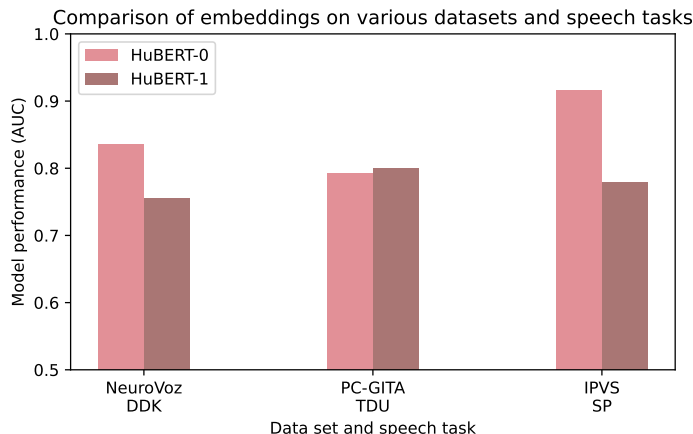


Figure 5: Performance comparison of the embeddings extracted from two HuBERT layers.

5.4 Dataset Classification

To assess dataset variation using ML, we trained a model to distinguish between datasets. The model successfully classified NeuroVoz, PC-GITA, and IPVS samples in the sustained phonation task, achieving a five-fold cross-validated AUC of 99.2% using IFM and 100% using NIFM. The evaluation was conducted on a balanced test set comprising 20% of the data, with speaker exclusivity maintained between the train and test sets. Figure 6 displays the confusion matrices for the dataset classification model using IFM (Figure 6a) and NIFM (Figure 6b) features. We observe that the model using IFM correctly assigned 100% of NeuroVoz and IPVS samples. 89% of PC-GITA samples were correctly classified as such, with the remaining 11% incorrectly classified as NeuroVoz samples. Using NIFM, the model was able to correctly classify all samples.

5.5 Mono-Lingual Results

To establish a baseline performance for the TL experiment, the models were first evaluated in a monolingual setting. In this context, the model was trained and validated within a single language, using a holdout set from the same language for testing. The cross-validated average AUC scores for both IFM and NIFM across speech tasks are presented in Table 4. For the TDU task, IFM achieved AUCs comparable to NIFM for PC-GITA (0.782 vs. 0.780) and IPVS (0.993 vs. 0.987) and better than NIFM for NeuroVoz (0.841 vs. 0.815). The opposite

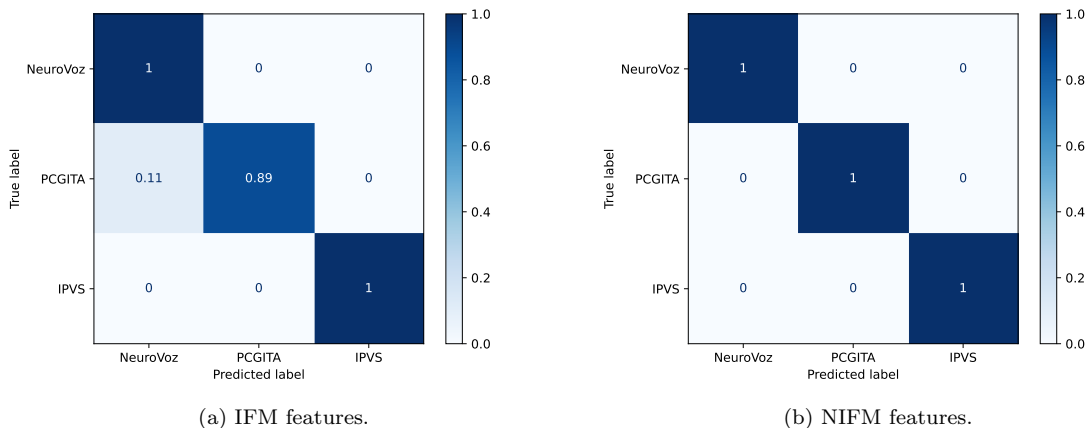


Figure 6: Normalized confusion matrices for dataset separation model.

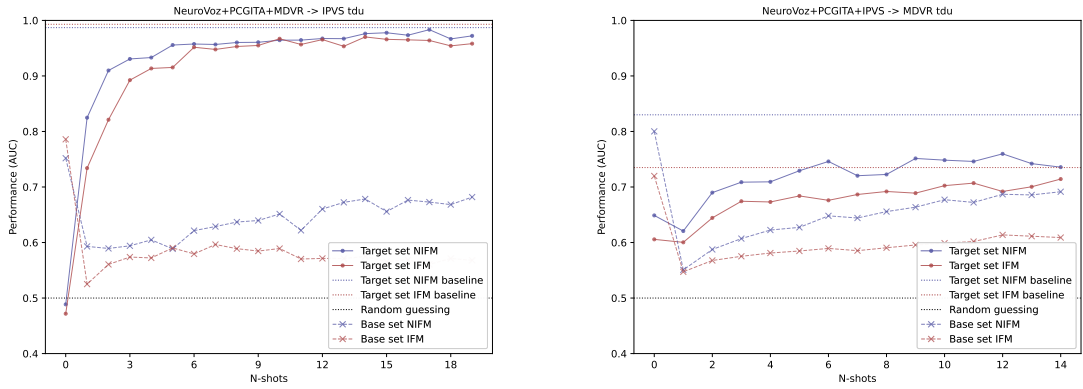
Dataset	Speech task	IFM	NIFM
NeuroVoz	SP	0.681 (0.120)	0.72 (0.130)
	DDK	0.738 (0.150)	0.853 (0.116)
	TDU	0.841 (0.112)	0.815 (0.115)
PC-GITA	SP	0.718 (0.142)	0.725 (0.138)
	DDK	0.676 (0.140)	0.739 (0.132)
	TDU	0.782 (0.128)	0.78 (0.114)
IPVS	SP	0.914 (0.116)	0.9 (0.117)
	DDK	0.963 (0.091)	0.972 (0.088)
	TDU	0.993 (0.047)	0.987 (0.045)
MDVR	RP	0.735 (0.143)	0.83 (0.122)

Table 4: 10-fold cross-validated performance (AUC metric) for IFM and NIFM models in the monolingual experiment, per speech task. Values within parentheses indicate standard deviation. SP=Sustained Phonation /a/; DDK=Diadochokinetic task; TDU=Text-Dependent Utterances; RP=Read Passage.

was observed for the SP and DDK task, where NIFM outperformed IFM for NeuroVoz (0.681 and 0.738 vs. 0.720 and 0.853) and PC-GITA (0.718 and 0.676 vs. 0.725 and 0.739). Although some performance differences were minimal, notable gaps were observed in tasks such as MDVR RP (0.735 vs. 0.830) and the DDK tasks for NeuroVoz (0.738 vs. 0.853) and PC-GITA (0.676 vs. 0.739), where NIFM showed a clear advantage.

5.6 Cross-Lingual Results

The core experiment of this thesis involves cross-lingual PD classification using few-shot TL. The results of these experiments are detailed in this section. Due to the large number of experiments and resulting plots, we present only two illustrative examples in Figure 7. The example in Figure 7a exemplifies the general pattern observed in nearly all train-test combinations, while Figure 7b depicts the sole exception to this pattern. The remaining experiments can be found in Appendix B. This appendix includes comprehensive grids of experiments. Figure 9 displays the experiments across all speech tasks, using a Leave-One-Language-Out scheme. Similarly, Figure 10 presents all combinations of training and test languages for the TDU task.



(a) Few-Shot TL of IPVS on TDU speech task, pretrained on NeuroVoz, PC-GITA, and MDVR. (b) Few-Shot TL of MDVR on TDU speech task, pretrained on NeuroVoz, PC-GITA, and IPVS.

Figure 7: Two selected examples of the Few-Shot TL experiment.

Figure 7a shows the performance on the IPVS dataset initially trained on the combined sets of NeuroVoz, PC-GITA, and MDVR. This example represents the patterns observed in most TL configurations:

- At zero-shot, both IFM and NIFM showed a chance-level (~ 0.5) AUC.
- As n -shots increased, the AUC improved for both IFM (increase between 0.13 and 0.64) and NIFM (increase between 0.14 and 0.71).
- The AUC gain flattened as n increased, with minimal improvements (between 0.02 and 0.13) beyond fine-tuning on half of the data.
- NIFM consistently outperformed IFM at all n .
- NIFM performance reached monolingual performance after fine-tuning on 90% of the data.
- IFM approached, but did not fully reach, monolingual performance for any n (see Table 4).
- Base set performance dropped when n increased from 0 to 1.
- Base set performance of NIFM was consistently higher than that of IFM across all n .

Figure 7b illustrates an exception to the pattern observed previously. Here, the zero-shot AUC was above random guessing for both IFM (0.6) and NIFM (0.65). Despite this initial advantage, the performance gain from adding samples from the target set was limited, with an increase of only 0.15 AUC for IFM and 0.17 for NIFM.

6 Discussion

In this project, we explored few-shot TL for cross-lingual PD classification, with a particular focus on comparing IFM and NIFM. Through a series of experiments, we evaluated how increasing the size of the fine-tuning set affected model performance to answer two RQs. The experiments involved training a model on a base set, composed of samples from either a single or multiple datasets. This model was subsequently fine-tuned on a target set, a separate language not part of the training data. The model’s classification accuracy was measured on a holdout set of the target language. Two model types were used: an SGD-optimized linear SVM and a four-layer DNN. The results highlighted major performance patterns and differences between IFM and NIFM. The observations presented previously will now be connected to relevant literature, followed by a discussion of methodological limitations and issues encountered during the study.

6.1 Discussion of Experimental Results

The conducted experiments involved three distinct speech tasks from four datasets. The experimental observations allow us to address the previously defined research questions and hypotheses (see Section 1). In addition to the experiments designed to address the research questions, we carried out an exploratory analysis that included visualizing the data using t-SNE and generating feature importance plots for each dataset. Dataset variation was assessed by training a model to classify dataset samples. Furthermore, the selection of the specific HuBERT embedding as the NIFM was supported by a rationale based on its superior performance.

6.1.1 RQ1: Impact of the Fine-Tuning Set Size

The first RQ addresses two few-shot learning scenarios: the zero-shot case, where no fine-tuning is applied, and the n -shot case, where the model is fine-tuned using n samples.

Initially, with an empty fine-tuning set, the zero-shot scenario showed chance-level performance, as illustrated in the various experiments in Figure 9. This suggests that the model was unable to effectively and consistently classify PD in new languages without exposure to target-language data. These findings are consistent with those of Vásquez-Correa et al. [29], who reported high variability in zero-shot TL effectiveness, ranging from 50 to 85% accuracy. Similarly, Favaro et al. [15] reported instability of performance metrics in zero-shot cross-lingual experiments (ranging from 0.50 to 0.83), linking this to inefficient data standardization and distribution shifts. Kovac et al. [63] also emphasizes the importance of including target language speakers during training. Our results align with observations by Hireš et al. [27], who investigated cross-lingual classification using the sustained phonation task and found performance ranging from 0.41 to 0.70 AUC in zero-shot settings, regardless of whether one or multiple base languages were used.

As the fine-tuning set increased, target language performance improved. This is consistent with the findings of Vásquez-Correa et al. [29]. The performance increase seen by Won and Kim [56], going from one-shot (56.9% accuracy) to five-shot (73.7%) learning correlates with these observations. Our results indicate that, although larger fine-tuning sets enhance model performance, diminishing returns are evident as n continues to grow, as illustrated in Figure 9. Where we observed minimal improvement beyond incorporating 50% of the target data, Vásquez-Correa et al. [29] reported substantial improvements after including the final 10 to 20% of the target set. However, not all of their experiments replicated

this trend, with some showing negligible benefits even from increasing beyond the zero-shot scenario.

An exception to the observed pattern was the experiment which was trained on NeuroVoz, PC-GITA, and IPVS, and fine-tuned on MDVR. The zero-shot performance exceeded random guessing for both IFM and NIFM. However, there was a limited performance gain from increasing the fine-tuning set. These results suggest that the model successfully transferred knowledge from the base set to the target set. However, it also indicates that the model learnt minimally as more samples were added, supported by the nearly flat trajectory of performance gain. Moreover, the model did not reach baseline performance, observed when pre-training was entirely bypassed.

Some studies, including Rios-Urrego et al. [25] and Vásquez-Correa et al. [33], have reported classification performance exceeding the baseline after fine-tuning models that were pretrained on different languages. They demonstrated that this effect was model- and dataset-dependent. Where certain datasets benefited from pre-training, others showed a negative effect on performance. Our findings similarly indicate a model- and dataset-dependent pattern, where pre-training occasionally led to baseline-exceeding performance. Notably, these improvements were primarily achieved by NIFM models, similar to observations by Favaro et al. [15], with the exception of the MDVR dataset. In this case, all baseline-exceeding results were observed with IFM models instead of NIFM. This discrepancy may be attributed to differences in baseline performance: NIFM typically outperformed IFM in the monolingual setting, thereby setting a lower threshold for cross-lingual IFM models to exceed.

6.1.2 RQ2: IFM vs NIFM

To address RQ2, we examined performance differences between IFM and NIFM in cross-lingual few-shot learning. The findings indicate that both feature types benefit from increasing the fine-tuning set size, with a slight performance advantage for NIFM.

In the zero-shot setting, neither feature type performed reliably across languages, often producing results at or below chance level. However, closer inspection revealed that NIFM outperformed IFM by small margins in most zero-shot experiments. This finding is consistent with Favaro et al. [15], although we did not replicate the substantial performance gaps between IFM and NIFM (ranging between 4% and 20% on F1-score) reported in their cross-lingual study. Similarly, Gimeno-Gómez et al. [18] observed that NIFM outperformed IFM by 8% to 39% on the F1-score in a cross-lingual experiment, depending on the dataset. This suggests that the performance differences between IFM and NIFM are dataset dependent, which our results support, as varying performance gaps were observed depending on the dataset and speech task.

As the fine-tuning set size increased, models using both IFM and NIFM improved, with the performance gap remaining stable throughout the experiment. This contradicts our initial hypothesis, which anticipated a greater performance gain for NIFM. No prior study on voice-based PD classification had explored few-shot learning with both types of features, making it challenging to provide expectations. Our hypothesis was instead based on findings from different few-shot learning contexts, which ultimately yielded outcomes distinct from ours.

6.1.3 Data Visualization

The project started with an analysis of the different datasets used in this study. Since our experiments were of a cross-lingual nature, large differences between datasets could be

problematic. Therefore, we applied the t-SNE method to visualize the high-dimensional data. These plots showed that dataset separation was easily done with NIFM, as opposed to IFM, where samples of the four datasets were mixed in the visualization. The finding that datasets can be easily separated using NIFM was also reported by Hireš et al. [27], who used a different NIFM model to achieve a well-separated plot among four datasets. The difference in PD and HC separability across the different datasets is remarkable, but does not reflect the classification accuracy found in the literature. For instance, it is reported by Favaro et al. [15] that the IPVS dataset is very well separable, as opposed to NeuroVoz and PC-GITA. The difference in separability between IFM and NIFM suggests that the NIFM model outperforms the IFM model due to its better separability. However, from our earlier comparison, such separability did not correlate with classification accuracy. This indicates that the model was able to learn complex patterns in the data that cannot be visualized in a two-dimensional figure.

6.1.4 Feature Importance

To provide insights into which features the model relied on for distinguishing between PD and HC samples, an ML classifier with a measure of feature importance was trained for each dataset separately. The results of the relative feature importance calculations, shown in Figure 4, indicate that PD symptoms manifest differently across datasets, as different feature sets were important for classification in different languages. This observation aligns with the known challenge of cross-lingual diagnosis due to the language-dependent characteristics of PD. However, it may also be partially attributable to variations in dataset collection, which will be discussed in detail later. Assuming a minimal influence from dataset variations, these findings suggest that the vocal symptoms of PD are expressed differently across languages, which aligns with the literature on cross-lingual PD [23, 26, 27]. Our results are consistent with those presented by Favaro et al. [23], who reported that different features contributed to PD discrimination across languages. Additionally, they found that certain features were negatively associated with PD in one language, while showing a positive association in another.

6.1.5 NIFM: HuBERT Embedding Layer

The choice for the HuBERT representation for the NIFM model was established by Favaro et al. [15], but the specific embedding layer used in their study was not specified. To address this, we compared the two HuBERT representations on our own datasets and selected the layer that generated the highest classification accuracy. The comparison between the two HuBERT embeddings, displayed in Figure 5, revealed a preference for the first-layer representation in the NeuroVoz and IPVS datasets, while no significant difference was observed for PC-GITA. Although the speech tasks varied across datasets and could have influenced performance, there is no clear reason to believe that these differences affected layer preference, as each audio signal was split into short fragments, minimizing task-specific influences. Since the model was pre-trained on the LibriSpeech corpus, it is likely that the last layer embedding was optimized for speech recognition. As a result, the features valuable for distinguishing PD from HC from voice signals were better captured in the first-layer embedding. This finding aligns with the observations of Tirronen et al. [62], who found that the first layer of the HuBERT model outperformed other layers in distinguishing healthy from disordered speech in women.

6.2 Base Set Performance

Maintaining model robustness includes that training on a new dataset should not compromise the base set performance. We could not find any studies on cross-lingual PD classification that report base set performance, and therefore the impact of extending their model to a new language is unclear. However, it seems impractical to add new languages if it compromises performance on the original language. Our experiments, displayed in Figure 9, demonstrated a clear decline in performance on a holdout portion of the base set as the fine-tuning size increased, which occurred already after fine-tuning on a single target sample. Fine-tuning was performed with a balanced mix of base and target samples, a technique proposed in Tamm et al. [61]. Without including base samples during fine-tuning, the performance decline on the base set would likely have been more severe.

6.3 Dataset Variation Issues

Differences in data collection and curation present significant challenges for PD classification across multiple datasets. Each dataset is constructed using distinct procedures, leading to variations in recording conditions and patient characteristics. These differences include the types of speech tasks conducted, which may vary from sustained phonation exercises to story reading tasks or patient-doctor dialogues. This issue was acknowledged by Rusz et al. [41], who proposed guidelines for speech recordings for dysarthria.

Notably, when we trained a model to distinguish between datasets, the model performed remarkably well, which is demonstrated in Subsection 5.4. In an experiment with three datasets and the sustained phonation of the vowel /a/ task, an AUC of ≥ 0.99 was achieved, using both IFM and NIFM. This finding aligns with the results of Botelho et al. [36], who demonstrated that a model could easily distinguish between six datasets in the same language.

A further complication is the unrealistic balance between HC and PD patients often present in these datasets, which does not reflect real-world conditions where PD is less prevalent. Additionally, datasets curation varies, with some including PD-positive patients who do not exhibit speech symptoms, potentially affecting model performance. Medication status also differs across datasets; while some samples are recorded from patients not using PD medication, others include only patients on medication. Moreover, the time since diagnosis varies across participants, introducing heterogeneity in symptom severity and disease progression, a challenge previously identified by Moro-Velazquez et al. [80]. These factors complicate the development of robust models capable of generalizing across datasets.

6.4 Limitations

During the execution phase of this study, we encountered several limitations that could not be easily addressed due to time or resource constraints. One major limitation was the limited amount of public datasets, which clearly restricted the possibilities to train the base model. Access to a broader range of datasets could potentially force the model to focus on language-independent features, enhancing its cross-lingual robustness.

Our experiments also centred on a single set of IFM and NIFM. The IFM set was constructed based on features frequently used in recent studies. However, other prosodic, linguistic, and cognitive features associated with PD have been identified in the literature. Including those additional features might improve the model performance.

For the NIFM, we selected a single pretrained model based on findings from a comparative study in the literature. We then adopted the embedding from that model that performed

best across a series of experiments. However, this does not guarantee that it was the most suitable model for our datasets or cross-lingual experiments. Alternative embeddings might have produced better results.

In the cross-lingual experiments, we interchangeably used the baseline and DNN models, selecting whichever performed best in the monolingual experiment. However, the model that excels in monolingual contexts may not necessarily be optimal for cross-lingual scenarios. This selection approach may have influenced the results. Due to computational constraints, it was not feasible to optimize both models for all experimental runs.

Finally, in this project, MFCCs were categorized as part of the interpretable feature model. However, the interpretability of these features can be questioned. Although they are not generated by deep neural models and are derived through well-defined transformations of the time-frequency representation, they are not inherently interpretable. Therefore, it could be argued that MFCCs do not align with the goal of interpretability. However, feature importance plots revealed that MFCCs contribute significantly to the model's classification performance, and excluding them would result in a substantial decline in classification accuracy.

7 Conclusion and Future Research

This study explored the application of few-shot TL within cross-lingual contexts for the classification of speech-based PD, focusing on the contributions of IFMs and NIFMs. Through experiments with various publicly available datasets, the impact of the fine-tuning set size on performance using TL in cross-lingual contexts was evaluated (RQ1). Furthermore, to address RQ2, a comparative evaluation of IFM and NIFM performance was carried out in the experiments.

In particular, regarding RQ1, the results showed that in zero-shot classification, the models were unable to effectively transfer knowledge from the base language to the target language (AUC ~ 0.5). This pattern was observed across all experiments, with the exception of the MDVR target set (AUC 0.6-0.65). Fine-tuning on the target language improved performance; however, this came at the cost of forgetting the PD-defining characteristics of the base language. By incorporating both the base language and the target language during fine-tuning, this issue was partially mitigated, retaining stable base set performance.

Regarding RQ2, we found that NIFM generally outperformed IFM in both monolingual (NIFM advantage between 0.08-0.12) and fine-tuned cross-lingual contexts (NIFM advantage between 0.05-0.21). However, the performance gap varied between datasets and speech tasks, and in certain cases, IFM performed better than NIFM. Importantly, neither feature set demonstrated language-independent characteristics, as both performed poorly (AUC ~ 0.5) when transferred to a new dataset without fine-tuning. Although IFMs have the benefit of being interpretable, their performance is generally inferior to that of NIFMs. Consequently, it is a crucial consideration of whether clinical settings should emphasize explainability over effectiveness or whether there is a way to achieve a balance that ensures both transparency and reliability.

The poor cross-lingual generalization observed in this study suggests that PD exhibits different vocal changes across languages or that inconsistencies in dataset collection processes impacted the results. Visualizations of the data and the model trained to distinguish datasets indicated that variations in recording procedures, quality, and curation may have contributed to the limited performance of cross-lingual generalization. These findings emphasize the importance of standardizing data collection processes for future research on speech-based PD classification, particularly in multilingual settings.

In conclusion, while few-shot TL shows potential for cross-lingual PD diagnosis, achieving robust performance across languages requires attention to standardizing dataset collection procedures, optimizing fine-tuning strategies, and enhancing feature representations to better capture language-independent PD characteristics.

7.1 Future Work

The field of automated speech-based PD classification using AI models is rapidly evolving, with many promising directions for future research. By addressing these areas, future research could overcome current limitations and contribute to the development of robust and generalizable models for the diagnosis of cross-lingual PD.

This thesis has highlighted the need to address cross-lingual differences and dataset inconsistencies in speech-based PD classification. Both the data visualizations and the model trained to separate datasets indicated that variations in recording procedures, quality, and curation likely impacted performance. These findings reinforced the importance of standardizing data collection processes, emphasizing the potential for advancements in multilingual PD classification research. This aligns with the guidelines for recording speech datasets pro-

posed by Rusz et al. [41]. In addition to adhering to a global standard for data collection, dataset developers could incorporate additional data modalities in datasets, such as gait or handwriting, which are known biomarkers of PD.

Future research could address the challenge of capturing language-independent characteristics. Exploring alternative feature sets, such as Variational Mode Decomposition (VMD), may offer a pathway to derive more generalizable features, as was posed by Veetil et al. [54]. Another promising direction involves reducing the feature space gap between datasets through domain adaptation techniques, which could improve cross-lingual generalizability. Given the often reported effectiveness of spectrograms in monolingual PD classification [22], investigating their performance in cross-lingual contexts could be beneficial.

Further investigation could also focus on identifying the most effective hidden layers in deep network embeddings for PD diagnostics, such as presented in Tirronen et al. [62] for HuBERT and Wav2Vec 2.0. Our comparison of the two HuBERT layer representations revealed significant performance differences for two out of the three investigated datasets and speech tasks. This suggests that the optimal layer may depend on the type of data used for model pretraining. Speech-specific models might favour lower-level layers, unlike models trained on general audio. Another possibility is to aggregate the representations and use both embeddings as model features. Identifying models capable of capturing language-independent features could further enhance cross-lingual performance, as our selected representation demonstrated limited success in this area.

The performance of the NIFM could be further improved by fine-tuning the pretrained model, allowing its internal representations to better align with the classification task. A multistage fine-tuning approach may enhance performance, where the model is initially fine-tuned on related data, such as vowel phonation, followed by fine-tuning on the base language, before being optimized on the target dataset. This procedure was followed by Hireš et al. [57], who fine-tuned on a pathological voice dataset as an intermediate step. Strategies like freezing specific layers during fine-tuning could help mitigate forgetting of the base language. We observed that mixing base and target samples partially addressed this issue, as the base set performance increased by increasing the fine-tuning set size (see Appendix B, Figure 10). Optimizing the base-to-target sample ratio presents a potential solution to inefficient cross-lingual learning.

7.2 Afterword

During the past semester, I dedicated my efforts to this project as the final step toward completing my Master’s degree in Artificial Intelligence. When I first encountered the RAIVD project on the University’s website, I immediately knew that I wanted to contribute through my thesis. I believe that AI has the potential to benefit society significantly, and the intersection of AI and healthcare represents one of the most promising areas of research. The automated diagnosis of Parkinson’s disease fits seamlessly in that field.

This journey has been filled with valuable learning experiences. One of the key personal lessons was realizing my growing confidence in presenting a topic in which I had engaged myself. Letting go of unnecessary tension allowed me to deliver more engaging and effective presentations. Having presented at the RAIVD annual meeting alongside fellow thesis students and at the monthly research meeting, I developed both communication skills and confidence in delivering presentations. Naturally, the project also posed challenges. The models did not initially yield the expected results, forcing me to rethink my approach and explore out-of-the-box solutions. This experience taught me the importance of adaptability and creativity in solving problems.

I started this project with basic knowledge on voice diagnostics, gained through various courses and a previous internship. However, applying it to PD classification was entirely new to me, as well as incorporating TL. Over the past semester, I have developed expertise in all three domains and I am now able to explain how they can intersect. The field of speech technology has long interested me, and this thesis project, along with my visit to the 3rd Dutch Speech Tech Day in Hilversum (February 2025) out of personal interest, has further strengthened my enthusiasm.

Looking back, I recognize that the initial scope of the project was defined too broadly for a Master's thesis. Although I made early efforts to narrow it down by limiting datasets and feature types, I learned that being even more selective would have been beneficial. This insight will surely inform my approach to future projects. Despite these challenges, the journey has been greatly rewarding. Working on this meaningful topic has deepened my understanding of both AI and its potential applications in healthcare. I hope that my work will provide valuable insights and contribute positively to the RAIVD project. All in all, while the path was not without its difficulties, I thoroughly enjoyed the process and take great pride in the knowledge and skills I have gained along the way.

References

- [1] Tysnes, O.B. and Storstein, A., (2017). Epidemiology of parkinson’s disease. *Journal of Neural Transmission*, 124:901–905. doi:10.1007/s00702-017-1686-y.
- [2] Poewe, W., Seppi, K., Tanner, C., Halliday, G.M., Brundin, P., Volkman, J., Schrag, A.E., Lang, A.E., Edmond, S.J., Morton, C., and Shulman, G., (2017). Parkinson disease. *Nature reviews Disease primers*, 3:1–21. doi:10.1038/nrdp.2017.13.
- [3] Rusz, J., Tykalová, T., Novotný, M., Růžička, E., and Dušek, P., (2021). Distinct patterns of speech disorder in early-onset and late-onset de-novo parkinson’s disease. *npj Parkinson’s Disease*, 7. doi:10.1038/s41531-021-00243-1.
- [4] Jenkinson, C., Fitzpatrick, R., Peto, V., Greenhall, R., and Hyman, N., (1997). The parkinson’s disease questionnaire (pdq-39): development and validation of a parkinson’s disease summary index score. *Age and ageing*, 26(5):353–357.
- [5] Ngo, Q.C., Motin, M.A., Pah, N.D., Drotár, P., Kempster, P., and Kumar, D., (2022). Computerized analysis of speech and voice for parkinson’s disease: A systematic review. *Computer Methods and Programs in Biomedicine*, 226:107133. doi:10.1016/J.CMPB.2022.107133.
- [6] Ho, A.K., Iannsek, R., Marigliani, C., Bradshaw, J.L., and Gates, S., (1999). Speech impairment in a large sample of patients with parkinson’s disease. *Behavioural neurology*, 11(3):131–137. doi:10.1155/1999/327643.
- [7] Harel, B., Cannizzaro, M., and Snyder, P.J., (2004). Variability in fundamental frequency during speech in prodromal and incipient parkinson’s disease: A longitudinal case study. *Brain and Cognition*, 56:24–29. doi:10.1016/J.BANDC.2004.05.002.
- [8] Pahwa, R. and Lyons, K.E., (2010). Early diagnosis of parkinson’s disease: recommendations from diagnostic clinical guidelines. *Am J Manag Care*, 16(4):94–99.
- [9] Lorraine O Ramig, C.F. and Sapiro, S., (2008). Speech treatment for parkinson’s disease. *Expert Review of Neurotherapeutics*, 8(2):297–309. doi:10.1586/14737175.8.2.297.
- [10] Keränen, T., Kaakkola, S., Sotaniemi, K., Laulumaa, V., Haapaniemi, T., Jolma, T., Kola, H., Ylikoski, A., Satomaa, O., Kovanen, J., et al., (2003). Economic burden and quality of life impairment increase with severity of pd. *Parkinsonism & related disorders*, 9(3):163–168.
- [11] Huse, D.M., Schulman, K., Orsini, L., Castelli-Haley, J., Kennedy, S., and Lenhart, G., (2005). Burden of illness in parkinson’s disease. *Movement Disorders*, 20:1449–1454. doi:10.1002/mds.20609.
- [12] Dorsey, E.a., Constantinescu, R., Thompson, J., Biglan, K., Holloway, R., Kieburtz, K., Marshall, F., Ravina, B., Schifitto, G., Siderowf, A., et al., (2007). Projected number of people with parkinson disease in the most populous nations, 2005 through 2030. *Neurology*, 68(5):384–386.
- [13] Chen, J.J., (2010). Parkinson’s disease: health-related quality of life, economic cost, and implications of early treatment. *The American journal of managed care*, 16 Suppl Implications:S87—93.

- [14] Prabhod, K.J. et al., (2024). The role of artificial intelligence in reducing healthcare costs and improving operational efficiency. *Quarterly Journal of Emerging Technologies and Innovations*, 9(2):47–59.
- [15] Favaro, A., Tsai, Y.T., Butala, A., Thebaud, T., Villalba, J., Dehak, N., and Moro-Velázquez, L., (2023). 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. doi:10.1016/j.compbiomed.2023.107559.
- [16] Fagherazzi, G., Fischer, A., Ismael, M., and Despotovic, V., (2021). Voice for health: The use of vocal biomarkers from research to clinical practice. *Digital Biomarkers*, 5: 78–88. doi:10.1159/000515346.
- [17] Favaro, A., Motley, C., Cao, T., Iglesias, M., Butala, A., Oh, E.S., Stevens, R.D., Villalba, J., Dehak, N., and Moro-Velazquez, L., (2023). A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders. pages 532–539. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/SLT54892.2023.10022435.
- [18] Gimeno-Gómez, D., Botelho, C., Pompili, A., Abad, A., and Martínez-Hinarejos, C.D., (2024). Unveiling interpretability in self-supervised speech representations for parkinson’s diagnosis. *arXiv preprint arXiv:2412.02006*. doi:10.1109/JSTSP.2025.3539845.
- [19] Vellido, A., (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 32:18069–18083. doi:10.1007/s00521-019-04051-w.
- [20] Pagán, F.L., (2012). Improving outcomes through early diagnosis of parkinson’s disease. *American Journal of Managed Care*, pages 176–182.
- [21] Zha, D., Bhat, Z.P., Lai, K.H., Yang, F., Jiang, Z., Zhong, S., and Hu, X., (2025). Data-centric artificial intelligence: A survey. *ACM Comput. Surv.*, 57(5). doi:10.1145/3711118.
- [22] van Gelderen, L. and Tejedor-García, C., (2024). Innovative speech-based deep learning approaches for parkinson’s disease classification: A systematic review. *Applied Sciences*, 14(17):7873. doi:10.3390/app14177873.
- [23] Favaro, A., Moro-Velázquez, L., Butala, A., Motley, C., Cao, T., Stevens, R.D., Villalba, J., and Dehak, N., (2023). Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in parkinson’s disease. *Frontiers in Neurology*, 14:1142642. doi:10.3389/fneur.2023.1142642.
- [24] Vásquez-Correa, J.C., Arias-Vergara, T., Rios-Urrego, C.D., Schuster, M., Rusz, J., Orozco-Arroyave, J.R., and Nöth, E., (2020). Convolutional neural networks and a transfer learning strategy to classify parkinson’s disease from speech in three different languages. doi:10.1007/978-3-030-33904-3_66.
- [25] Rios-Urrego, C.D., Vásquez-Correa, J.C., Orozco-Arroyave, J.R., and Nöth, E., (2020). Transfer learning to detect parkinson’s disease from speech in different languages using convolutional neural networks with layer freezing. volume 12284 LNAI, pages 331–339. Springer Science and Business Media Deutschland GmbH. doi:10.1007/978-3-030-58323-1_36.

- [26] Kovac, D., Mekyska, J., Galaz, Z., Brabenec, L., Kostalova, M., Rapcsak, S.Z., and Rektorova, I., (2021). Multilingual analysis of speech and voice disorders in patients with parkinson’s disease. In *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pages 273–277. doi:10.1109/TSP52935.2021.9522597.
- [27] Hireš, M., Drotár, P., Pah, N.D., Ngo, Q.C., and Kumar, D.K., (2023). On the inter-dataset generalization of machine learning approaches to parkinson’s disease detection from voice. *International Journal of Medical Informatics*, 179. doi:10.1016/j.ijmedinf.2023.105237.
- [28] Ferrante, C., Menon, B., Pillai, A.S., Sbattella, L., and Scotti, V., (2023). Analysis of features for machine learning approaches to parkinson’s disease detection. *Machine Learning and Deep Learning in Natural Language Processing*, pages 169–183.
- [29] Vásquez-Correa, J.C., Arias-Vergara, T., Orozco-Arroyave, J.R., Eskofier, B., Klucken, J., and Nöth, E., (2018). Multimodal assessment of parkinson’s disease: a deep learning approach. *IEEE journal of biomedical and health informatics*, 23(4):1618–1630. doi:10.1109/JBHI.2018.2866873.
- [30] Park, H., Kang, Y., and Kim, J., (2023). Enhancing structure-property relationships in porous materials through transfer learning and cross-material few-shot learning. *ACS Applied Materials and Interfaces*, 15:56375–56385. doi:10.1021/acsami.3c10323.
- [31] Sahoo, N.R., Mallela, N., and Bhattacharyya, P., (2023). With prejudice to none: A few-shot, multilingual transfer learning approach to detect social bias in low resource languages. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13316–13330.
- [32] Wu, J., Zhao, Z., Sun, C., Yan, R., and Chen, X., (2020). Few-shot transfer learning for intelligent fault diagnosis of machine. *Measurement: Journal of the International Measurement Confederation*, 166. doi:10.1016/j.measurement.2020.108202.
- [33] Vásquez-Correa, J.C., Rios-Urrego, C.D., Arias-Vergara, T., Schuster, M., Rusz, J., Nöth, E., and Orozco-Arroyave, J.R., (2021). Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages. *Pattern Recognition Letters*, 150:272–279. doi:10.1016/j.patrec.2021.04.011.
- [34] Mou, X., (2019). Artificial intelligence: Investment trends and selected industry uses. *International Finance Corporation*, 8(2):311–320.
- [35] Zhao, A., Wang, N., Niu, X., Chen, M., and Wu, H., (2024). A triplet multimodel transfer learning network for speech disorder screening of parkinson’s disease. *International Journal of Intelligent Systems*. doi:10.1155/2024/8890592.
- [36] Botelho, C., Schultz, T., Abad, A., and Trancoso, I., (2022). Challenges of using longitudinal and cross-domain corpora on studies of pathological speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022-September, pages 1921–1925. International Speech Communication Association. doi:10.21437/Interspeech.2022-10995.
- [37] Feng, K. and Chaspari, T., (2024). A pilot study on clinician-ai collaboration in diagnosing depression from speech. *arXiv preprint arXiv:2410.18297*.

- [38] Chen, M. and Decary, M., (2020). Artificial intelligence in healthcare: An essential guide for health leaders. In *Healthcare management forum*, volume 33, pages 10–18. SAGE Publications Sage CA: Los Angeles, CA.
- [39] Bélisle-Pipon, J.C., Powell, M., English, R., Malo, M.F., Ravitsky, V., Consortium, B.V., and Bensoussan, Y., (2024). Stakeholder perspectives on ethical and trustworthy voice ai in health care. *Digital Health*, 10:20552076241260407.
- [40] Zhang, Y., Gao, J., Tan, Z., Zhou, L., Ding, K., Zhou, M., Zhang, S., and Wang, D., (2024). Data-centric foundation models in computational healthcare: A survey. *arXiv preprint arXiv:2401.02458*.
- [41] Rusz, J., Tykalova, T., Ramig, L.O., and Tripoliti, E., (2021). Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. *Movement Disorders*, 36(4):803–814. doi:<https://doi.org/10.1002/mds.28465>.
- [42] Hoehn, M.M. and Yahr, M.D., (1967). Parkinsonism: onset, progression, and mortality. *Neurology*, 17.
- [43] Goetz, C.G., Tilley, B.C., Shaftman, S.R., Stebbins, G.T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M.B., Dodel, R., et al., (2008). Movement disorder society-sponsored revision of the unified parkinson’s disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, 23(15):2129–2170.
- [44] Virameteekul, S., Revesz, T., Jaunmuktane, Z., Warner, T.T., and De Pablo-Fernández, E., (2023). Clinical diagnostic accuracy of parkinson’s disease: where do we stand? *Movement Disorders*, 38(4):558–566.
- [45] Wang, M., Ge, W., Apthorp, D., Suominen, H., et al., (2020). Robust feature engineering for parkinson disease diagnosis: New machine learning techniques. *JMIR Biomedical Engineering*, 5(1):e13611. doi:10.2196/13611.
- [46] Di Biase, L., Di Santo, A., Caminiti, M.L., De Liso, A., Shah, S.A., Ricci, L., and Di Lazzaro, V., (2020). Gait analysis in parkinson’s disease: An overview of the most accurate markers for diagnosis and symptoms monitoring. *Sensors*, 20(12):3529. doi:10.3390/s20123529.
- [47] Araki, N., Yamanaka, Y., Poudel, A., Fujinuma, Y., Katagiri, A., Kuwabara, S., and Asahina, M., (2021). Electrogastrography for diagnosis of early-stage parkinson’s disease. *Parkinsonism & Related Disorders*, 86:61–66. doi:10.1016/j.parkreldis.2021.03.016.
- [48] Galaz, Z., Drotar, P., Mekyska, J., Gazda, M., Mucha, J., Zvoncak, V., Smekal, Z., Faundez-Zanuy, M., Castrillon, R., Orozco-Aroyave, J.R., Rapcsak, S., Kincses, T., Brabenec, L., and Rektorova, I., (2022). Comparison of cnn-learned vs. handcrafted features for detection of parkinson’s disease dysgraphia in a multilingual dataset. *Frontiers in Neuroinformatics*, 16. doi:10.3389/fninf.2022.877139.
- [49] Rusz, J., Cmejla, R., Ruzickova, H., and Ruzicka, E., (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson’s disease. *The journal of the Acoustical Society of America*, 129(1):350–367. doi:10.1121/1.3514381.

- [50] Laganas, C., Iakovakis, D., Hadjidimitriou, S., Charisis, V., Dias, S.B., Bostantzopoulou, S., Katsarou, Z., Klingelhoefer, L., Reichmann, H., Trivedi, D., Chaudhuri, K.R., and Hadjileontiadis, L.J., (2022). Parkinson’s disease detection based on running speech data from phone calls. *IEEE Transactions on Biomedical Engineering*, 69: 1573–1584. doi:10.1109/TBME.2021.3116935.
- [51] Scimeca, S., Amato, F., Olmo, G., Asci, F., Suppa, A., Costantini, G., and Saggio, G., (2023). Robust and language-independent acoustic features in parkinson’s disease. *Frontiers in Neurology*, 14:1198058. doi:10.3389/FNEUR.2023.1198058/BIBTEX.
- [52] Chen, G. and Jin, Y., (2024). Cascaded transfer learning strategy for cross-domain alzheimer’s disease recognition through spontaneous speech. doi:10.21437/Interspeech.2024-627.
- [53] Vásquez-Correa, J.C., Rios-Urrego, C.D., Rueda, A., Orozco-Aroyave, J.R., Krishnan, S., and Nöth, E., (2019). Articulation and empirical mode decomposition features in diadochokinetic exercises for the speech assessment of parkinson’s disease patients. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, October 28-31, 2019, Proceedings 24*, pages 688–696. Springer.
- [54] Veetil, I.K., Sowmya, V., Orozco-Aroyave, J.R., and Gopalakrishnan, E.A., (2024). Robust language independent voice data driven parkinson’s disease detection. *Engineering Applications of Artificial Intelligence*, 129. doi:10.1016/j.engappai.2023.107494.
- [55] Karaman, O., Çakın, H., Alhudhaif, A., and Polat, K., (2021). Robust automated parkinson disease detection based on voice signals with transfer learning. *Expert Systems with Applications*, 178. doi:10.1016/j.eswa.2021.115013.
- [56] Won, J.H. and Kim, D.H., (2024). Metric-based few-shot transfer learning approach for voice pathology detection. *IEEE Access*. doi:10.1109/ACCESS.2024.3480523.
- [57] Hireš, M., Gazda, M., Drotár, P., Pah, N.D., Motin, M.A., and Kumar, D.K., (2022). Convolutional neural network ensemble for parkinson’s disease detection from voice recordings. *Computers in biology and medicine*, 141:105021. doi:10.1016/j.combiomed.2021.105021.
- [58] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q., (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76. doi:10.1109/JPROC.2020.3004555.
- [59] Wang, D. and Zheng, T.F., (2015). Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237. IEEE.
- [60] Ferrante, C. and Scotti, V., (2023). Cross-lingual transferability of voice analysis models: a parkinson’s disease case study.
- [61] Tamm, B., Vandenberghe, R., and Hamme, H.V., (2023). Cross-lingual transfer learning for alzheimer’s detection from spontaneous speech. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. doi:10.1109/ICASSP49357.2023.10096770.

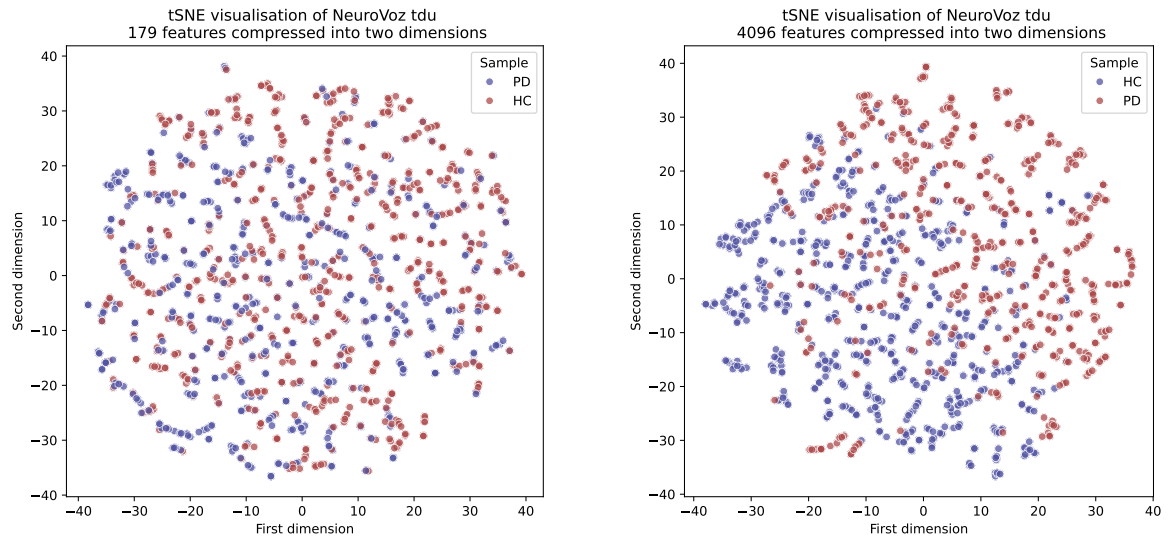
- [62] Tirronen, S., Kadiri, S.R., and Alku, P., (2023). Hierarchical multi-class classification of voice disorders using self-supervised models and glottal features. *IEEE Open Journal of Signal Processing*, 4:80–88. doi:10.1109/OJSP.2023.3242862.
- [63] Kovac, D., Mekyska, J., Aharonson, V., Harar, P., Galaz, Z., Rapcsak, S., Orozco-Arroyave, J.R., Brabenec, L., and Rektorova, I., (2024). Exploring digital speech biomarkers of hypokinetic dysarthria in a multilingual cohort. *Biomedical Signal Processing and Control*, 88:105667. doi:10.1016/j.bspc.2023.105667.
- [64] Quatra, M.L., Turco, M.F., Svendsen, T., Salvi, G., Orozco-Arroyave, J.R., and Siniscalchi, S.M., (2024). Exploiting foundation models and speech enhancement for parkinson’s disease detection from speech in real-world operative conditions. *arXiv preprint arXiv:2406.16128*. doi:10.21437/Interspeech.2024-522.
- [65] Peng, X., Xu, H., Liu, J., Wang, J., and He, C., (2023). Voice disorder classification using convolutional neural network based on deep transfer learning. *Scientific Reports*, 13. doi:10.1038/s41598-023-34461-9.
- [66] Mendes-Laureano, J., Gómez-García, J.A., Guerrero-López, A., Luque-Buzo, E., Arias-Londoño, J.D., Grandas-Pérez, F.J., and Godino-Llorente, J.I., (2024). Neurovoz: a castillian spanish corpus of parkinsonian speech [data set]. doi:10.5281/zenodo.10777657.
- [67] Mendes-Laureano, J., Gómez-García, J.A., Guerrero-López, A., Luque-Buzo, E., Arias-Londoño, J.D., Grandas-Pérez, F.J., and Godino-Llorente, J.I., (2024). Neurovoz: a castillian spanish corpus of parkinsonian speech. *arXiv preprint arXiv:2403.02371*.
- [68] Orozco, J.R., Arias-Londoño, J.D., Vargas-Bonilla, J.F., González-Rátiva, M.C., Orozco-Arroyave, J.R., No, J.D.A.L., Vargas-Bonilla, J.F., González-Rátiva, M.C., and Nöth, E., (2014). New spanish speech corpus database for the analysis of people suffering from parkinson’s disease.
- [69] Dimauro, G., Di Nicola, V., Bevilacqua, V., Caivano, D., and Girardi, F., (2017). Assessment of speech intelligibility in parkinson’s disease using a speech-to-text system. *IEEE Access*, 5:22199–22208. doi:10.1109/ACCESS.2017.2762475.
- [70] Jaeger, H., Trivedi, D., and Stadtschnitzer, M., (2019). Mobile device voice recordings at king’s college london (mdvr-kcl) from both early and advanced parkinson’s disease patients and healthy controls [data set]. doi:10.5281/zenodo.2867215.
- [71] Zhang, H., Wang, A., Li, D., and Xu, W., (2018). Deepvoice: A voiceprint-based mobile health framework for parkinson’s disease identification. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 214–217. IEEE.
- [72] Jeong, S.M., Song, Y.D., Seok, C.L., Lee, J.Y., Lee, E.C., and Kim, H.J., (2024). Machine learning-based classification of parkinson’s disease using acoustic features: Insights from multilingual speech tasks. *Computers in Biology and Medicine*, 182:109078. doi:10.1016/J.COMPBIOMED.2024.109078.
- [73] Lahmiri, S., (2017). Parkinson’s disease detection based on dysphonia measurements. *Physica A: Statistical Mechanics and its Applications*, 471:98–105. doi:https://doi.org/10.1016/j.physa.2016.12.009.

- [74] Rusz, J., Cmejla, R., Tykalova, T., Ruzickova, H., Klempir, J., Majerova, V., Picmausova, J., Roth, J., and Ruzicka, E., (2013). 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.
- [75] Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A., (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460.
- [76] Di Cesare, M.G., Perpetuini, D., Cardone, D., and Merla, A., (2024). Machine learning-assisted speech analysis for early detection of parkinson’s disease: A study on speaker diarization and classification techniques. *Sensors*, 24(5):1499. doi:<https://doi.org/10.3390/s24051499>.
- [77] Chan, T.F., Golub, G.H., and LeVeque, R.J., (1983). Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, 37(3): 242–247. doi:<https://doi.org/10.1080/00031305.1983.10483115>.
- [78] Rusz, J., Bonnet, C., Klempír, J., Tykalová, T., Baborová, E., Novotný, M., Rulseh, A., and Ružička, E., (2015). Speech disorders reflect differing pathophysiology in parkinson’s disease, progressive supranuclear palsy and multiple system atrophy. *Journal of neurology*, 262:992–1001.
- [79] Van der Maaten, L. and Hinton, G., (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11):2579.
- [80] Moro-Velazquez, L., Gomez-Garcia, J.A., Arias-Londoño, J.D., Dehak, N., and Godino-Llorente, J.I., (2021). Advances in parkinson’s disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects. *Biomedical Signal Processing and Control*, 66. doi:[10.1016/j.bspc.2021.102418](https://doi.org/10.1016/j.bspc.2021.102418).

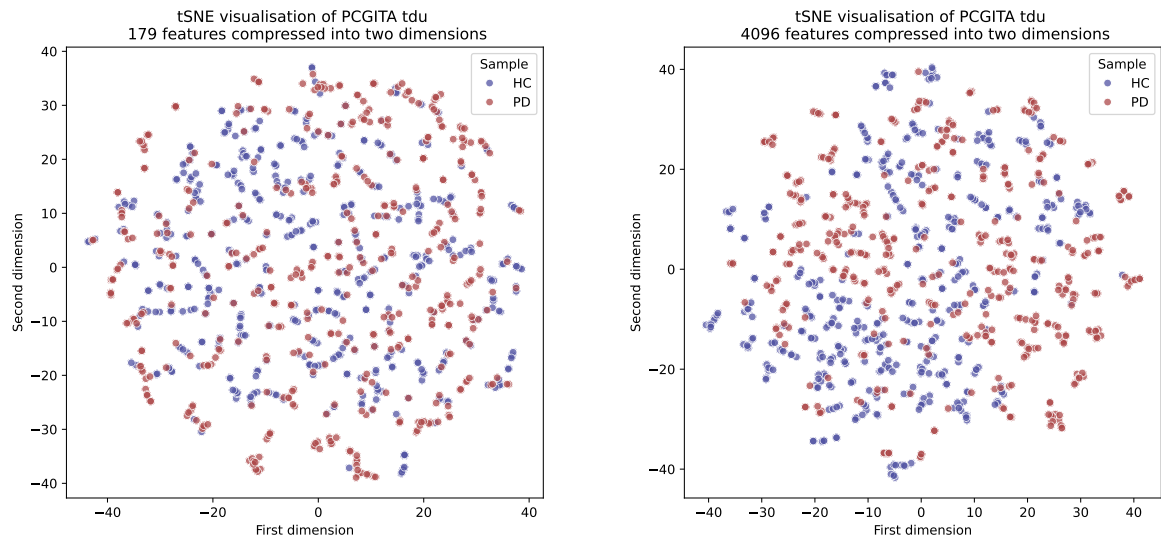
Appendix

A Data Visualization

The datasets were visualised using t-SNE, a method that reduces data dimensionality to two, enabling visualisation on a two-dimensional graph. Figure 8 presents t-SNE plots for each dataset, distinguishing between PD and HC.

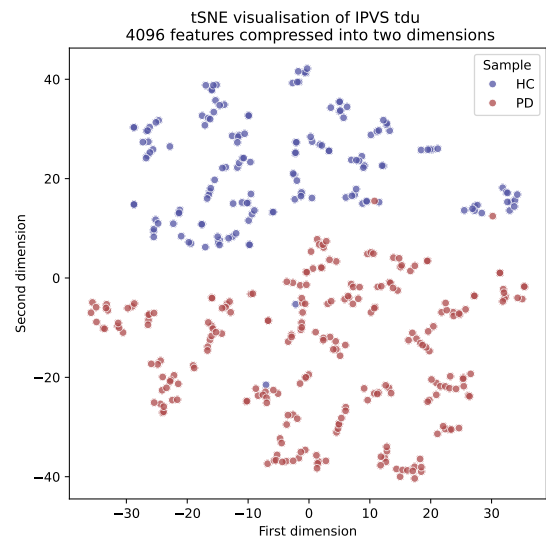
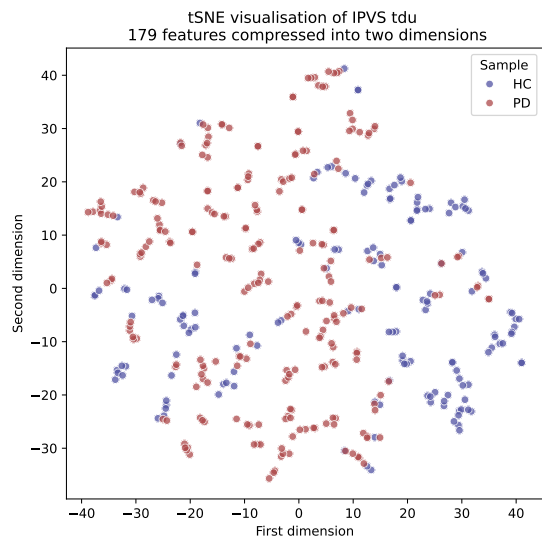


(a) NeuroVoz t-SNE visualizations

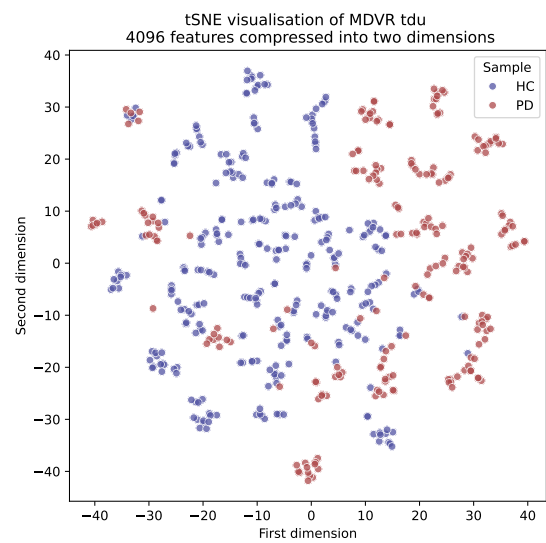
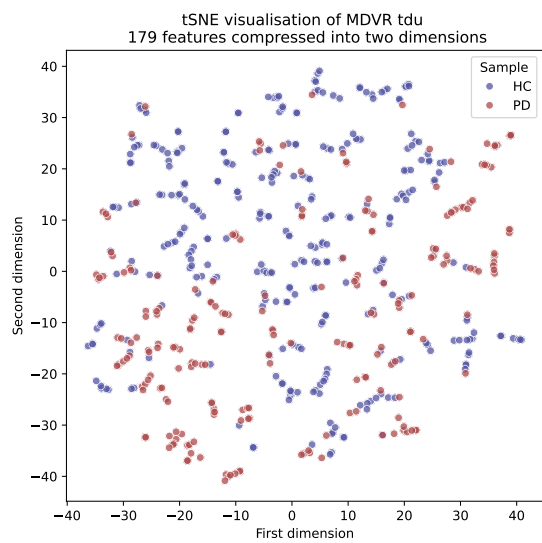


(b) PC-GITA t-SNE visualizations

Figure 8: t-SNE visualizations of the TDU task per dataset. The plots show the data represented by IFM (left) and NIFM (right), transformed using t-SNE.



(c) IPVS t-SNE visualizations



(d) MDVR t-SNE visualizations

Figure 8 (cont.): t-SNE visualizations of the TDU task per dataset. The plots show the data represented by IFM (left) and NIFM (right), transformed using t-SNE.

B Cross-Lingual Results

The main experiment involved few-shot TL for PD classification in a cross-lingual context. It was conducted using various languages and speech tasks, each serving as either a base or target set in all possible combinations. The results are presented in the following grids of plots.

Figure 9 shows the first set of TL plots, where each row represents a different speech task, and each column indicates a target dataset. The final column is incomplete because the MDVR dataset does not include sustained phonation or DDK tasks.

The TL protocol was executed by training on multiple languages as the base set or on a single language as the base set. Figure 9 displays the results for models trained on all datasets except the target set.

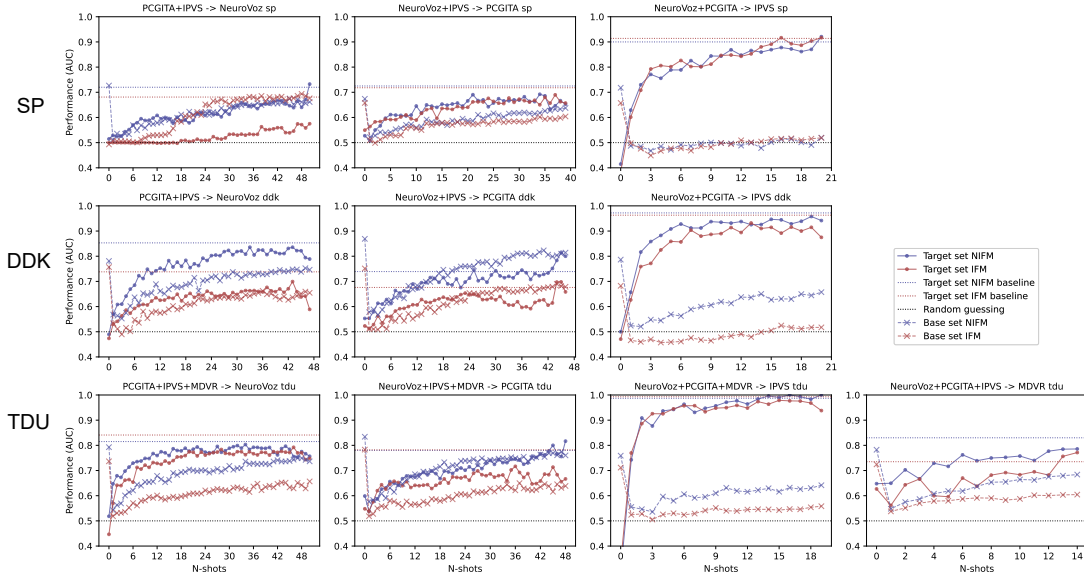


Figure 9: Cross-lingual experiments on the three speech tasks. Each row presents one speech task, marked on the left.

Figure 10 presents the results for training on one dataset and testing on another using the TDU task. For easier comparison, the lower row replicates the results of the previous plot for the TDU task, showing models trained on all datasets except the target language.

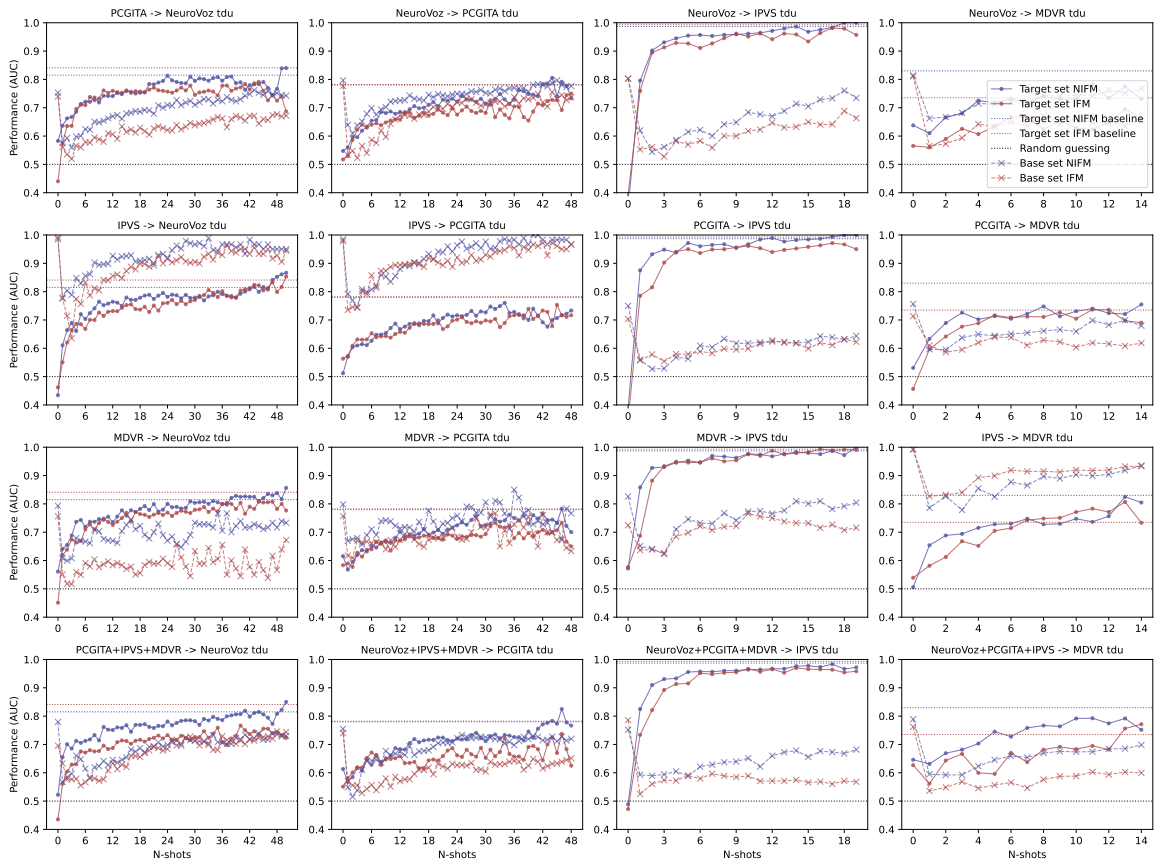


Figure 10: Cross-lingual experiments using every combination of base and target datasets on TDU task.