Sex Differences in the Detection of Parkinson's Disease from Speech

Hiba Akhaddar

A Thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Master of Computer Science (Computer Science) at
Concordia University
Montréal, Québec, Canada

November 2025

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Hiba Akhaddar**

Entitled: **Sex Differences in the Detection of Parkinson's Disease from Speech**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Marta Kersten-Oertel

_____ Examiner
Dr. Yang Wang

_____ Thesis Supervisor
Dr. Tristan Glatard

_____ Co-supervisor
Dr. Mirco Ravanelli

Approved by _____
Dr. Denis Pankratov, Graduate Program Director

November 10, 2025 _____
Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Sex Differences in the Detection of Parkinson's Disease from Speech

Hiba Akhaddar

Parkinson's disease (PD) is the second most common neurodegenerative pathology in the world. It has been shown that 70% to 75% of PD patients would suffer from speech disorders at some stage. PD affects men more frequently than women, which has resulted in previous works using unbalanced datasets that potentially introduce sex-related biases in the models. In this paper, we investigate the sex differences in the detection of PD from speech using various models. We extract features using WavLM, Wav2vec2.0, Whisper, and FBanks and feed them into ECAPA-TDNN, deep neural network architecture, and then a binary classification layer. We use ComParE2016 features with Random Forest as well. We also conducted sex-specific experiments to assess generalization across sexes. In our main experiments, we use a large subset of the mPower open dataset, where sex and disease status (PD vs Healthy Controls (HC)) are balanced. In subsequent experiments evaluating dataset size and sex-specific trainings, we used 3 additional datasets, mPower matched, mPower small, and PC-GITA, where age was also matched between PD and HC groups. All the datasets include sustained phonation task. We observe that on the large dataset, female speakers are more easily detected than males. However, on smaller datasets, none of the classification results were found to exceed chance performance. Sex-specific results show that when the models are trained on one sex, they fail to generalize to the other one especially when using traditional Random Forest pipeline. Our results highlight the importance of including sex as a variable in the development of fair PD detection systems.

Acknowledgments

First of all, I would like to express my deepest gratitude to my parents for their love, constant support, presence, and patience throughout my life. They instilled in me a curiosity about the world and nurtured my passion for science from an early age. Their encouragement to ask questions, explore ideas, and never stop learning has been the foundation of this journey.

I am profoundly thankful to my siblings and grandparents for their continuous encouragement and support. Their belief in me and constant check-ins have been a source of strength and motivation.

This research project would not have been possible without the supervision of Dr. Tristan Glatard and Dr. Mirco Ravanelli whose guidance and feedback have been invaluable throughout this project. Their insightful advice and rigorous approach to research have helped me learn technically and grow intellectually.

I would like to extend my sincere thanks to my labmates from both research groups for their collaboration, helpful discussions, and willingness to share their time and expertise.

Many thanks as well to all those who share their knowledge generously through books, videos, and lectures. The open-source community, educational content creators, and researchers who make their work accessible have been essential in my learning. As Newton said, “If I have seen further, it is by standing on the shoulders of giants.”

Finally, I am deeply grateful to my family and friends for their encouragement and understanding throughout this journey. A special thanks to my two aunts, whose kindness

and support helped me adapt to life in a new country.

Contribution of Authors

This research project was conducted in collaboration with several co-authors. The work presented in this thesis will be submitted for publication in a peer-reviewed journal. Preliminary results were presented as a poster at the Women in Machine Learning Workshop, co-located with NeurIPS 2024 [1].

The project was co-supervised by Dr. Tristan Glatard and Dr. Mirco Ravanelli who provided guidance, expertise, and feedback throughout all stages of the research and writing processes. I elaborated the methodology, conducted the experiments, and drafted the paper and the thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Parkinson's disease (PD) is a neurodegenerative disease that affects more than 10 million people in the world [2], making it the second most common neurodegenerative pathology [3]. In Canada, out of every 100,000 people who are 40 years and older, 530 individuals live with parkinsonism which includes Parkinson's disease and other conditions with movement-related symptoms [4]. PD is caused by a deficiency in dopamine production that results from the degeneration of neurons in the substantia nigra of the midbrain region. The low level of dopamine causes symptoms like involuntary motion, muscle rigidity, and alterations in speech and handwriting [5].

The UK Parkinson's Disease Society Brain bank and the Movement Disorders Society's diagnostic criteria showed that the clinical diagnosis of PD can be done based on bradykinesia and one of the following features which are rigidity, rest tremor, and postural instability [6]. However, definitive diagnosis remains challenging as these criteria require the presence of motor symptoms which typically appear only after almost 50-70% of dopaminergic neurons in the substantia nigra have already degenerated [7, 8].

Beyond motor impairment, PD influences the prosody of speech with near-mutism, hesitancy, and dysfluency [9]. Speech impairment is a common early symptom experienced by people with Parkinson's Disease as up to 75% of patients experience dysarthria that is

characterized by reduced loudness, monopitch, imprecise articulation, and altered prosody [10]. The neurodegenerative processes cause motor symptoms that affect the muscles used for speech which results in changes in voice and speech patterns [11].

Importantly, changes in voice and speech can appear as early as 5 years before major motor symptoms [12], but the diagnosis is often delayed until the presence of motor symptoms [13]. This highlights the potential of speech-based approaches for early and accessible PD detection.

Recent works have investigated the use of Deep Learning models for PD detection from speech because they can capture complex information directly from raw data [14]. These models are more affordable compared to traditional models and they can be deployed remotely through smartphones and telemedicine platforms. These advantages are important given the growing rate of PD which requires scalable diagnostic tools.

Also, according to [15, 16], neural circuits for speech are different between men and women, potentially leading to sex-specific vocal impairments among PD patients. These sex-specific characteristics for speech production in addition to the different progression of dopaminergic degeneration may result in acoustic biomarkers that appear differently among men and women.

Although there is currently no effective treatment for the pathology, providing an early and accurate diagnosis is crucial [3] for better disease monitoring, optimizing patient care, and advancing research efforts.

This thesis is motivated by the need to understand and address sex differences in PD diagnosis from speech. By analyzing multiple datasets and models, this work aims to study sex-related disparities in PD detection performance and contribute to the building of fair and reliable speech-based diagnostic systems.

1.1 Contributions

This work contributes to the topic of sex differences found when using speech-based models to detect Parkinson's Disease. The contributions are mentioned below.

1. We conduct an analysis of sex differences found when detecting Parkinson's disease from speech across multiple datasets and model architectures.

2. We compare self-supervised models (WavLM, Wav2Vec2.0, Whisper) and traditional acoustic features (FBanks, ComParE2016) when combined with ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation in Time-Delay Neural Network) and Random Forest classifiers.

3. We construct and use balanced subsets of the mPower and PC-GITA datasets, controlling for sex and age, to eliminate confounding demographic effects.

4. We analyze how model performance varies with dataset size and when training on one sex and testing on both sexes to study the generalization challenges in current PD detection pipelines.

5. Based on our findings, we provide recommendations for dataset construction and model evaluation to promote fairness in future clinical speech AI systems.

1.2 Thesis Outline

In Chapter 2, we present the background of this research which includes a literature review of previous studies on PD detection from speech using machine learning and deep learning as well as the sex differences reported in some works on the same topic. Chapter 3 describes the methodology adopted in this work including models, datasets, and experimental

protocol. Chapter 4 discusses the results and provides an analysis of the findings. Finally, Chapter 5 concludes the thesis and outlines future works.

Chapter 2

Background

This background chapter aims to provide an overview of the key concepts, methodologies, and advancements in the topic of Parkinson's detection from speech using deep learning and the sex differences reported.

2.1 Speech production

Speech is produced as a result of a sensorimotor process that involves a high coordination of respiratory, phonatory, and articulatory subsystems that is regulated by a network of cortical and subcortical brain structures [17].

The process of speech production begins in the brain with motor planning and intention, where cortical and subcortical regions formulate the linguistic and motor commands necessary for articulation [18]. This initial stage is particularly relevant for PD because the disease affects the neural control mechanisms before manifesting in speech.

Speech production then proceeds with respiratory control in which the diaphragm and intercostal muscles provide enough subglottal air pressure to initiate phonation [19]. Then, through vocal fold vibration, the laryngeal system transforms the aerodynamic energy into

an acoustic one [20]. Finally, the tongue, lips, jaw, velum, and related muscles that constitute the articulatory system modulates the acoustic signal to generate phonemes and speech [21]. This process is regulated by cortical motor regions, basal ganglia, cerebellum, and brainstem nuclei that ensure the production of a clear and fluent speech [18].

It is the striatum in the basal ganglia that modulates the activity of thalamocortical circuits and governs the coordination, timing, and sequencing of speech movements [22, 23]. The basal ganglia contribute to the internal cueing and automatic execution of learned motor sequences, like the rapid and coordinated articulatory movements required for the production of fluent speech [24, 25].

The pathophysiological mechanisms underlying speech impairment in Parkinson's disease are multifactorial [26, 27]. At the respiratory level, PD patients present reduced vital capacity, decreased subglottal pressure, and impaired coordination between respiratory and phonatory subsystems. This results in diminished vocal loudness and shortened breath groups [28, 29].

Phonatory dysfunction appears as vocal fold bowing, incomplete glottal closure, and reduced vocal fold mobility, contributing to breathy voice quality and hypophonia [30, 31]. These symptoms can be considered measurable biomarkers that can be objectively quantified through acoustic analysis.

2.2 Sex differences

Although, based on epidemiological data, PD is approximately 1.5 times more prevalent in men [32] and the disease tends to progress at a faster rate in females [32] with some studies reporting faster cognitive decline and greater functional disability in women [33]. Women also show greater propensity for tremor-dominant phenotypes, earlier start of motor fluctuations and dyskinesias which is the involuntary and often repetitive movements affecting the face, limbs, or trunk, in addition to severe non-motor symptoms like depression and

anxiety [32].

Studying the sex difference is also crucial since natural speech patterns differ between healthy men and women like fundamental frequency, which typically averages around 120 Hz for males and 220 Hz for females [34], and formant frequencies [35].

In [19], the authors found that age was affecting the production of speech. Older adults use more abdominal movement in loud speech. Also, regarding sex, younger men produced speech at lower lung volumes compared to older ones. However, there were no significant differences reported between younger and older females.

These variations in acoustic properties could influence how PD may manifest differently across sexes and thus the performance of the models. In [36], the authors showed that speech characteristics associated with dysphagia differ by sex. Hypophonia was more prevalent among men, while imprecise articulation was more common among women. However, many existing studies overlook these differences, by usually relying on unbalanced datasets that include more male patients which could lead to biased models and less generalizability [37].

2.3 Detecting Parkinson's Disease from Speech using Machine and Deep Learning

Several works showed that it is possible to detect PD from speech using different Machine Learning (ML) and Deep Learning (DL) algorithms. They used several modeling designs such as traditional handcrafted features with classical ML classifiers and self-supervised learning features with neural processing.

2.3.1 Self-Supervised Learning Approaches

In [38], La Quatra M et al. developed a PD detector based on speech enhancement and foundational models. They outperformed prior approaches by fine-tuning WavLM Base and HuBERT Base on the standard PC-GITA dataset [39]. PC-GITA is balanced in terms of sex and age between PD patients and Healthy Controls (HC). The authors included diadochokinetic (DDK) exercises that involve the rapid repetition of syllables, reading sentences, and monologues as speech tasks. However, when tested on the extended PC-GITA that has recordings from real-world conditions, the accuracy dropped. After using a test set with speech enhancement techniques, they were able to regain strong performance. They combined the extended speech dataset to WavLM Base and HuBERT Base which yielded the best results with an accuracy of around 88.33%, F1-score of 88.13%, and ROC-AUC of 88.23%. This work shows the importance of preprocessing recordings and the failure of pretrained models when tested on real-world scenarios.

In another work, La Quatra M et al. proposed a dual-head deep learning architecture that uses self-supervised speech embeddings to detect PD. They used PC-GITA that includes 100 Spanish speakers and EWA-DB that has 50 Polish speakers, where PD and HC classes were equal in terms of size. In the architecture, one of the heads was used for diadochokinesis (DDK) tasks where the participants had to repeat /pa-ta-ka/ rapidly and clearly. The other head was for natural speech. Each branch extracted features from the raw audio files using WavLM combined with a wavelet-based temporal modeling component to capture multi-scale temporal dynamics. Their best F1-scores and accuracy, respectively, were 89.4% and 89.1% on PC-GITA and 83.1% and 82.7% on EWA-DB. They were also able to generalize the performance over cross-lingual datasets and outperform the monolingual baselines [40]. This validation across different languages shows that there are characteristics from languages that affect the performance of the model as shown by the 6% decrease from Spanish to Polish.

Simone et al. [14] integrated an attention mechanism into a deep learning model based on self-supervised embeddings to interpret the speech regions responsible for the predictions of PD using the Italian Parkinson's Voice and Speech Dataset. They achieved an accuracy of 99.14 ± 1.60 on the extended speech signals that consists of two readings of a phonemically balanced text. The attention weights revealed that the model focused on sustained vowel segments and voice consonants.

Gimeno-Gomez et al. proposed a transformer-based model that combines self-supervised embeddings with interpretable attention layers. They compared the performance of their model on multiple datasets. Regarding the vowels task on PC-GITA, they achieved an F1-score of 65.4 ± 0.6 [41]. La Quatra et al. achieved a better result on the same dataset but on a different speech task (89.4%) [40]. We conclude that the type of the speech task influences the classification performance.

Chronowski et al. employed a dataset of 2141 samples from 48 people for their experiments. They used a pretrained Wav2Vec2.0 model and finetuned it on their samples for classification. They achieved a 97.92% cross-validated accuracy [42]. The small number of unique speakers may result in overfitting which leads to good results on small datasets but may not be generalizable to larger ones.

2.3.2 Traditional Machine Learning and Feature Engineering Approaches

Almeida et al. used phonation task from the mPower dataset and compared it to a speech task from Lithuanian language. They found out that the sustained vowel task was better to detect PD with an accuracy of 94.55% and AUC of 0.87 when using acoustic cardioid microphone channel and 92.94%, AUC 0.92 when using the smartphone microphone channel. They used 18 feature extraction techniques and 4 machine learning algorithms k-Nearest Neighbour (kNN), Multi-Layer Perceptron (MLP), Optimum-Path Forest (OPF), and Support Vector Machine (SVM) [43].

The reviewed literature shows different performances from 65 % to 99% depending on datasets, speech tasks, and algorithms used. The datasets are usually small in size and the lack of information on data splits and cross validation strategies may lead to inflated reported performance. Our work addresses these gaps by conducting a comparison of different machine learning and deep learning approaches on different datasets with proper data splits and cross validation to avoid overfitting and data leakage.

2.4 Sex differences in the detection of Parkinson's Disease from Speech using machine and deep learning

Several studies have reported conflicting results regarding the impact of sex on the accuracy of PD detection from speech. While some works show that PD is easily detected in males, others report a better performance for female speakers. This suggests inconsistent patterns across, datasets, languages, and models.

2.4.1 Studies Reporting Female Performance Advantages

Houle N. et al. used a dataset of 68 speakers where sex and disease status were balanced. They ran logistic regression on acoustic features and found that women had more distinguishable acoustic patterns. According to the authors, sex affected measures like smoothed cepstral peak prominence, net syllables per second, percent pause ratio, and articulatory-acoustic vowel space [44]. In addition, the authors noticed that release burst precision was differentially affected by sex in PD. Individuals with PD produced fewer plosives with a single burst. Females more frequently produced multiple bursts, whereas males more frequently produced no burst at all.

Rusz et al. used logistic regression to assess the accuracy of PD detection from speech. They extracted speech features related to phonation, articulation, and prosody of 60 men

and 40 women de novo PD patients and matched HC. The authors found out that the prevalence of speech abnormalities in the PD cohort was higher for women compared to men (based upon the ROC curve, 65% vs 56%). The same prevalence pattern was present when discriminating between PD and HC with an AUC of 0.93 in women and 0.86 in men. They also observed that the voice quality, articulation of consonants, and production of pauses were better for women, but the loudness variability was better for men. They concluded that the speech differences were associated to nigro-putaminal dopaminergic deficits in addition to sex differences [33].

Adnan et al. proposed a novel architecture to detect PD from speech where they combined Wav2Vec2.0, WavLM, and ImageBind embeddings and fed them to a neural classifier. Their dataset included 1306 participants with 392 PD patients and 53% females. Their results showed a better performance for women with an accuracy of 86.9% vs 84.3 % for men [45]. They explain that by the fact that female PD patients have more consistent and pronounced acoustic deviations that enables their model to better detect PD for women. The authors did not investigate the causes of the sex difference, even if small.

2.4.2 Studies Reporting Male Performance Advantages

In their survey paper, van Gelderen et al. [46] reviewed 33 deep learning studies about the detection of PD from speech. They discussed the sex bias that appears in certain works where the performance by sex is included. Notably, publications by Jeancolas et al. [47] and Khaskhoussy et al. [48] where the F1-scores are higher for men.

Jeancolas et al. used a sex-balanced dataset of 221 French speakers with more PD patients than HC participants. They proposed two methods. The baseline was a MFCC-GMM system that uses a log-likelihood ratio classification. The second one used x-vector [49] to extract speaker embeddings that were classified using cosine similarity, LDA (Linear Discriminant Analysis), or PLDA (Probabilistic Linear Discriminant Analysis). The

EER (Equal Error Rate) was lower which means better when using x-vector, with values of 22% for males and 30-39% for females. This indicates a superior performance on male participants. In Jeancolas et al. the sex difference in the results could be attributed to the lower inter-speaker acoustic variability in the MFCC distributions of men.

Khaskoussy et al. used 4 datasets from which they extracted acoustic, prosodic, and phonetic features from speech recordings. They trained SVM (Support Vector Machines) and LSTM (Long Short-Term Memory) classifiers. LSTM achieved the best performance on all datasets where F1-score are ranging from 81% to 99%. In Khaskhoussy et al., the sex difference could be because the datasets used had more men than women which could lead to decision boundaries optimized for the majority class of men.

The inconsistent direction of sex differences across studies indicates the complexity of the phenomenon and its causes being context-dependent. In our work, we conducted several evaluations investigating both dataset-level and methodological factors contributing to the sex disparity.

Table 1 shows a summary of studies investigating sex differences in PD detection from speech.

Table 1: Summary of studies investigating sex differences in PD detection from speech.

Author	**Dataset**	**Methodology**	**Result**	**Justification of Sex Difference**	**Sex with best performance**
Houle et al. [44]	68 speakers, balanced by sex and disease status	Logistic regression on acoustic features	Women had more distinguishable acoustic patterns; sex affected CPPS, net syllables/sec, pause ratio, vowel space	Acoustic measures differed by sex; release burst precision differed: females produced multiple bursts, males no burst	**Female**

Author	Dataset	Methodology	Result	Justification of Sex Bias	Bias Direction
Rusz et al. [33]	60 male and 40 female de novo PD patients with matched HC	Logistic regression; phonation, articulation, prosody features	Prevalence of speech abnormalities higher in women (65% vs. 56%); AUC: women 0.93, men 0.86	Voice quality, consonant articulation, pause production more discriminative in women; loudness variability more discriminative in men	**Female**
van Gelderen et al. [46]	Review of 33 DL studies	Systematic review of deep learning studies on PD speech detection	Sex differences observed; some studies report higher performance for men, others for women	Bias may arise from dataset composition, sex imbalance, or feature distributions	**Mixed**

Author	Dataset	Methodology	Result	Justification of Sex Bias	Bias Direction
Jeancolas et al. [47]	221 French speakers, sex-balanced, more PD than HC	MFCC-GMM baseline; x-vector embeddings with cosine similarity, LDA, PLDA	EER better for men: 22% vs. 30–39% for women (lower EER = better)	Lower inter-speaker acoustic variability in male MFCC distributions may facilitate classification	**Male**
Khaskhoussy et al. [48]	4 datasets with acoustic, prosodic, phonetic features with different speech tasks and number of participants	SVM and LSTM classifiers	LSTM achieved best performance; F1-scores 99% on men in the training dataset that has 16 women and 24 men	Datasets contained more males than females, leading to majority class bias	**Male**

Author	**Dataset**	**Methodology**	**Result**	**Justification of Sex Bias**	**Bias Direction**
Adnan et al. [45]	1,306 participants, 392 PD, 53% females	Combined Wav2Vec2.0, WavLM, ImageBind embeddings with neural classifier	Better performance for women: accuracy 86.9% vs. 84.3% for men	Female PD patients exhibit more consistent and pronounced acoustic deviations enabling better detection	**Female**

Chapter 3

Methodology

Our experimental approach, shown in Figure 1, evaluates sex differences in the detection of PD from speech using machine learning and deep learning. We use 3 self-supervised learning models in a frozen configuration in addition to FBanks as feature extractors followed by ECAPA-TDNN, and a binary classifier with each extractor making a specific pipeline. We also use ComPar2016 features followed by Random Forest as a baseline.

3.1 Models

We compared four distinct feature extraction approaches for PD detection, each coupled with identical downstream architecture for fair comparison: Wav2Vec2.0 [50], WavLM [51], Whisper [52], and traditional 80-dimensional log-mel filterbank energies (FBank) that constitute a conventional baseline.

Deep learning embeddings were extracted from each pretrained model and fed to an ECAPA-TDNN network [53], that was originally designed for speaker recognition, but used for many other speech processing tasks like speaker diarization [54], language classification [55], and text to speech [56]. ECAPA-TDNN was followed by a fully connected binary classification layer to predict disease status.

The pretrained self-supervised learning (SSL) models were chosen because they capture nuanced acoustic and phonetic patterns that are potentially relevant to classify between PD and HC. These embeddings encapsulate speech dynamics such as articulation rate, voice tremor, and phonatory instability, which are subtle and could be indicative of PD [57].

The ComParE2016 feature set, that includes 6373 handcrafted acoustic features, was extracted using OpenSMILE toolkit [58]. The ComParE2016 features which include foundational frequency parameters such as F0, jitter, and shimmer, which are different between men and women in healthy speech influence the sex specific pattern learning. It was used with Random Forest (RF) as a baseline to benchmark against deep learning architectures.

This comparison allows assessment of whether large-scale pretrained models offer representational benefits over handcrafted acoustic descriptors.

Wav2Vec2.0 [50] learns speech representations by masking portions of the input and using a contrastive loss to identify the correct features. It was pretrained on approximately 60,000 hours of unlabeled speech from the LibriSpeech [59], Libri-Light[60], and CommonVoice [61] datasets. The model uses a convolutional encoder that processes raw audio waveforms at 16 kHz sampling rate, producing latent speech representations at 50 Hz frame rate. The context network, implemented as a Transformer architecture [62], generates contextualized representations that capture temporal dependencies across the utterance.

WavLM [51] extends this approach by adding gated relative position bias and utterance mixing to capture phonetic, acoustic, and prosodic information. The model was pretrained on 94,000 hours of public audio data using a combination of masked prediction and denoising objectives. These enhancements allow better modeling of speaker characteristics and acoustic variations making it suitable for tasks involving pathological speech [63]. We used WavLM-Large.

Whisper [52] was trained on approximately 680,000 hours of multilingual and multitask supervised data for automatic speech recognition. The encoder produces rich cross-lingual

embeddings that capture acoustic representations across diverse recording conditions and speaker populations. We specifically used the Whisper-large-v2 models which contains approximately 1.55 billion parameters and is based on a Transformer-based encoder-decoder architecture.

In contrast, FBank captures the spectral content of the speech waveform through short-time Fourier transform followed by mel-filters applied on the frequency axis.

All pretrained SSL encoders were used in a frozen configuration to preserve the representational quality of pretrained speech embeddings without introducing task-specific adaptation and to investigate the transferability of pretrained features to pathological speech data. This design isolates the contribution of the learned speech representations themselves. We ensure that differences in downstream performance come only from representation quality and not from fine-tuning differences. This allows us to get a better comparison across models. Moreover, freezing the encoder weights allows us to significantly reduce the computational requirements during training and mitigate the risk of overfitting which is essential considering the limited availability of labeled pathological speech data.

The embeddings from each feature extractor were fed in turn to ECAPA-TDNN network, configured with an input size of 1024 for Wav2Vec2.0, WavLM, and FBank, and 1280 for Whisper to accommodate its higher-dimensional representations, 5 convolutional blocks with channel configuration of [1024, 1024, 1024, 1024, 3072], kernel sizes of [5, 3, 3, 3, 1], and dilations of [1, 2, 3, 4, 1]. The network employed 128 attention channels and output a 192-dimensional utterance-level embedding. These architectural parameters were selected based on prior work on robust speaker representation learning [54] and our experimental results.

ECAPA-TDNN is a neural network architecture designed for speaker recognition that uses Emphasized Channel Attention, Propagation, and Aggregation mechanisms within a

Time-Delay Neural Network (TDNN) framework. The architecture employs Squeeze-and-Excitation (SE) blocks [64] combined with Res2Net-style [65] multi-scale feature extraction to model speaker-discriminative patterns. ECAPA-TDNN transforms variable-length audio recordings into fixed-dimensional vectors for classification. The network applies one-dimensional convolutional layers to extract temporal patterns from the audio, using residual connections [66] to preserve information across layers and dilations to capture long-range temporal dependencies. The attentive pooling mechanism is used to summarize the sequence into a single representation and it gives higher weight to frames that contain the most informative speech characteristics for detecting PD. The resulting fixed-dimensional vectors were passed to a fully connected classifier with log-softmax activation that converts the logits to log-probabilities for binary prediction. We trained using the negative log-likelihood loss that maximizes the likelihood of the correct class.

3.2 Datasets

We ran experiments on two publicly available datasets namely mPower [67] and PC-GITA [68]. We obtained ethical approval through Concordia University and by getting the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2: CORE 2022) certificate in order to access them.

mPower is an open dataset that was collected through a mobile health application where volunteers recorded themselves doing the sustained phonation tasks by producing the vowel /a/ ('Aaah') for 10 seconds. The recordings were obtained using the built-in microphones of the participants' iPhones at a sampling rate of 44.1 kHz.

The dataset collection was conducted by Sage Bionetworks in collaboration with the University of Rochester as part of a study. They had over 65,000 participants who contributed voice recordings remotely through the mPower mobile application [67].

Although the dataset is large and has multiple tasks, it is imbalanced with respect to

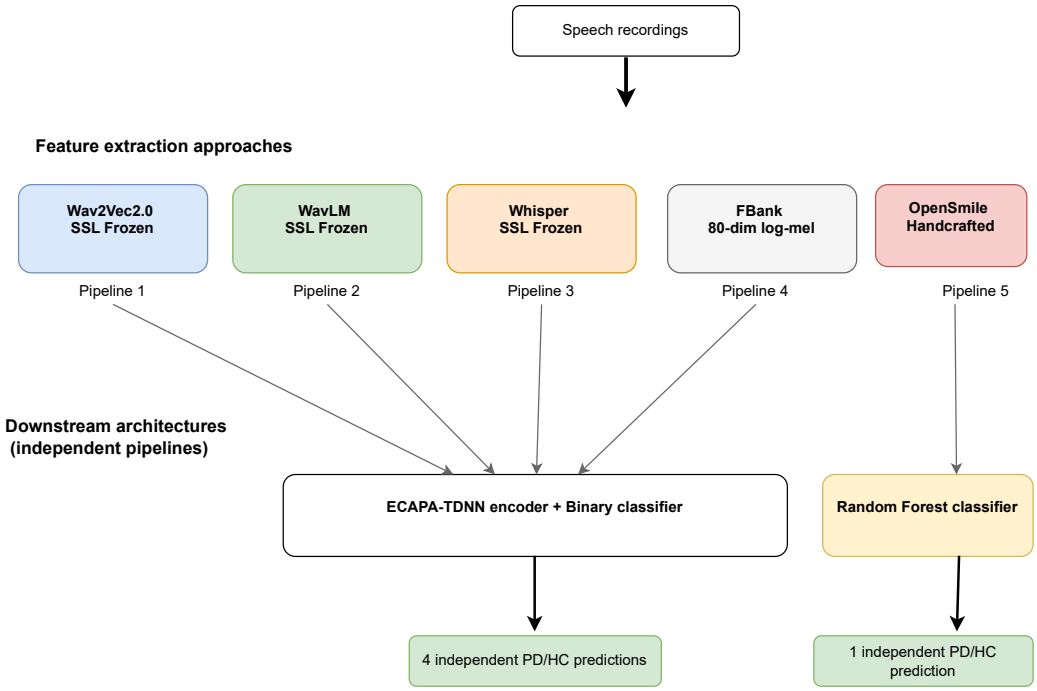

Figure 1: Comparative feature extraction approaches for Parkinson's disease detection.

age, sex, and health status. Also, there is a lot of variability in recording conditions, background noise levels, and device characteristics which makes it challenging for robust model development [69].

For fair comparison, from the mPower dataset, we extracted three subsets which are presented in Table 2.

The first subset (mPower large) has the largest number of samples but lacks age matching between males and females, which reflects real-world conditions where demographic imbalances are common in health dataset [70].

However, the second (mPower matched) is smaller in size but is age-matched, which constitutes a suitable set for direct comparison with the PC-GITA dataset. Age matching was performed by restricting participants to the 50 to 70 age range and ensuring the mean age difference was less than 1 year. The controlled demographics allow us to study the effect of model architecture and feature representation from potential age-related effects.

The third subset (mPower small) with 100 participants to match the sample size of PC-GITA makes us evaluate the performance of the model under identical data constraints. The difference with mPower matched is only the dataset size which means both sex and age are matched as well.

PC-GITA [39] is a clinically validated corpus of Colombian Spanish speakers that is balanced by both sex and age. It includes several speech tasks which are sustained phonation, text reading, and spontaneous speech. It has 100 participants, 50 are men and 50 are women, with equal numbers of PD and HC in each group. The age is matched and the severity of the disease was standardized using UPDRS-III scores [71]. The audio recordings were collected in a soundproof room at the Clínica Noel in Medellín, Colombia using a dynamic omnidirectional microphone (AKG C520). Recordings were sampled at 44.1 kHz with 16-bit resolution.

In our work, we looked at the sustained phonation task from mPower because it is the

only speech task in the dataset. Therefore, on PC-GITA, we restricted sustained phonation to the vowel /a/ to be consistent with the mPower data and because it is widely studied in voice analysis as it reflects motor impairements in PD related to laryngeal dysfunction, reduced phonatory control, and vocal fold instability [72].

The professional equipment and controlled recording environment guaranteed consistent acoustic conditions across all participants making the dataset suitable for evaluating model performance under optimal data quality conditions [68].

	mPower large	**mPower matched**	**mPower small**	**PC-GITA**
n samples	690	452	100	100
Male/Female (%)	50/50	50/50	50/50	50/50
Age (male, mean $\pm$ SD)	44.34 ± 8.63	59.53 ± 7.06	59.53 ± 7.06	61.08 ± 7.05
Age (female, mean $\pm$ SD)	51.38 ± 8.26	59.55 ± 6.86	59.55 ± 6.86	61.92 ± 11.35
UPDRS (male, mean $\pm$ SD)	—	—	—	37.76 ± 22.09
UPDRS (female, mean $\pm$ SD)	—	—	—	37.56 ± 14.03

Table 2: Demographic and clinical characteristics of the datasets used in this study. Values represent mean $\pm$ standard deviation for continuous variables. Age is expressed in years.

3.3 Experimental Protocols

We designed an evaluation framework to study sex differences by comparing the performance of feature extraction methods while controlling for demographic and methodological aspects. This design allows the assessment of how sex-related neurobiological variations in PD manifestation influence acoustic model performance.

3.3.1 Feature Extraction and Classification

We used each model as a feature extractor by freezing its parameters and extracting the embeddings from the raw audio files that we fed to ECAPA-TDNN and a binary classification layer to detect PD and HC. For the SSL models (Wav2Vec2.0, WavLM, and Whisper), features were extracted from the final encoder layer which is used to capture the most discriminative acoustic representations for downstream tasks [73].

ComParE2016 feature set [74] were also fed to Random Forest [75] for binary classification.

We implemented two comparison approaches. First, deep learning embeddings were extracted from Transformer-based models and fed to a neural network and a binary classification layer. Second, Fbanks features were also fed to ECAPA-TDNN and 6373 acoustic features extracted using the openSMILE toolkit [58] were fed to RF. This allows us to evaluate if the pretrained SSL features offer advantages over traditional spectral features (FBank) and handcrafted acoustic features (ComParE2016).

3.3.2 Training

We implemented the deep learning models using PyTorch [76] and SpeechBrain [77] frameworks. We used the Adam optimizer [78] with a learning rate of 1×10^{-4} and weight decay of 2×10^{-5}. Learning rate scheduling used a cyclic approach with exponential range mode, cycling between base learning rate 1×10^{-6} and maximum learning rate of 1×10^{-4} with exponential decay ($\gamma = 0.9998$) to prevent overfitting and maintain the stability of the training. These values were determined empirically through experimentation.

All models were trained for 30 epochs as we found that this was the optimal number of epochs with a batch size of 16 and a gradient accumulation factor of 2 to simulate an effective batch size of 32 while fitting within GPU memory constraints. The final test evaluation used used the model checkpoint with the lowest validation error to mitigate

overfitting and ensure generalization performance.

To ensure reproducibility of our results, we fixed all random seeds at the beginning of the experimental run. All pretrained models were downloaded from the HuggingFace Transformers library.

3.3.3 Evaluation and Statistical Analysis

Performance was evaluated using stratified k-fold cross-validation with k=10 folds for the mPower large dataset and k=5 folds for mPower matched and PC-GITA. We maintained equal proportions of sex and disease status across all folds to avoid demographic bias. We repeated the experiments across 5 random seeds for mPower large and 3 random seeds for the smaller datasets to account for the variability coming from random weight initialization and data shuffling. The final reported metrics are the mean and standard deviation across all folds and random seeds which constitute robust results.

We then reported the overall and sex-specific F1-scores, accuracy, precision, and recall on the test set.

Since WavLM-large achieved the strongest performance among SSL feature extractors on the mPower large dataset, we used it as the only feature extractor on the matched datasets to focus on the most promising approach. The same evaluation metrics were reported for all the datasets and experiments. This allows us to evaluate if performance gains observed on large datasets generalize to smaller and demographically controlled ones.

We also used a random classifier that randomly assigns class labels to the test set to evaluate if our models perform above chance. In addition, we computed the p-value from comparing the null distribution to the test results we got after training our models to see if the differences were significant. Statistical significance was defined as $p < 0.05$ or $p < 0.01$ or $p < 0.001$, with Bonferroni correction applied to adjust for multiple comparisons.

3.3.4 Sex-Specific and Cross-Dataset Analyses

An important component of our analysis is sex-related differences in PD speech because of the inherently existing neurobiological differences in PD between men and women. We conducted two complementary analyses which are separate model training on male-only and female-only cohorts and evaluated on both sexes to test transferability and the performance of the model to detect a sex more than the other. This strategy reveals if the embeddings learned from one sex can transfer effectively to the other.

To study the effect of age, we used PC-GITA and a subset from mPower. We implemented a matched-pairs algorithm to get a subset where age, sex, and disease status were matched to PC-GITA subjects. The two cohorts are different in terms of context as PC-GITA was clinically validated and recorded in a laboratory setting while mPower samples were collected through volunteers which creates a real world setting. This allows us to compare the robustness of features across different recording conditions.

Chapter 4

Results and Discussion

The results of the experiments are shown below. All reported metrics represent mean ± standard deviation across cross-validation folds and random seeds. Statistical significance was assessed using permutation tests [79] comparing model performance against a random baseline. Significance levels are denoted as: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, after using the Bonferroni correction.

4.1 Sex difference towards women on mPower large dataset

Table 3 shows the performance of different feature extractors when combined with ECAPA-TDNN and a classification layer on mPower large dataset. We reported the F1-score on average and per sex after running the experiments across 10 different folds and 5 random seeds. Additionally, we computed accuracy, precision, and recall to provide comprehensive evaluation (Appendix A).

To investigate the effect of sex differences in the detection of PD from speech, we trained and tested different feature extractors on sex-balanced mPower large datasets.

We observed from Table 3 that overall, none of the models were performing better than chance when applied to the entire dataset.

Models	F1	F1 male	F1 female
WavLM	68.61 ± 6.61	53.65 ± 8.39	83.12 ± 9.36 ***
FBanks	63.36 ± 5.87	43.30 ± 7.73	82.63 ± 10.57 ***
Wav2Vec2.0	55.66 ± 8.40	49.30 ± 7.39	60.79 ± 13.05 ***
Whisper	66.54 ± 6.38	56.00 ± 7.40	76.38 ± 9.36 ***
Random Forest	66.61 ± 3.36	46.80 ± 5.83	85.02 ± 3.79 ***

Table 3: Performance of 5 feature extractors on mPower large. Grey cells show no significant difference according to permutation test.

However, F1 scores related to females were statistically significant ($p < 0.001$). This suggests a reliable performance above chance level. We noticed that all models consistently achieve a better performance on women, which shows a sex-based performance disparity. The highest F1 score for female is 85.02% ± 3.79 when using Random Forest with ComParE2016 features.

The lowest performing model overall was Wav2Vec2.0 (57.20% ± 7.28) but the F1 score for women was still higher than for men (60.79% ± 13.05 vs. 49.30% ± 7.39).

Despite the lower performance of FBanks and Random Forest compared to WavLM, there was a substantial difference between male and female performances with (82.63% ± 10.57 vs. 43.30% ± 7.73) and (85.02% ± 3.79 vs. 46.80% ± 5.83) respectively.

We see from the table that all F1 scores of men were not better than chance and not significant at all. The table shows consistently higher classification performance in women relative to men on the mPower large dataset which shows the challenge in learning robust representations for male groups.

We observe that among the SSL-based frozen pretrained models used with mPower large, Random Forest with ComParE2016 features is even better than the SSL models for the female-specific F1 score (85.02% ± 3.79 vs. 83.12% ± 9.36 for WavLM). This indicates that handcrafted features on mPower large dataset may capture female acoustic characteristics while failing to generalize to male PD patients patterns.

Among the SSL models, WavLM demonstrated the best detection result followed by

Whisper, then Wav2Vec2. This could be attributed to the fact that WavLM's design captures acoustic, phonetic, and prosodic information explicitly which makes it more suitable for detecting the subtle articulatory and prosodic signals in the voice characteristic of hypokinetic dysarthria [80].

The results yielded by Whisper were lower than those by WavLM even if the first one was trained on 680,000 hours of multilingual speech indicates that the size of the pretrained dataset does not guarantee a better performance on clinical tasks. In addition, since Whisper was developed to transcribe speech in different acoustic conditions and with diverse languages, it may not be the best model to detect paralinguistic and voice quality features related to PD.

Wav2Vec2's poor classification ability can be related to its pretraining that was done on clean and read speech from LibriSpeech. The model training encourages the discrimination between discrete phonetic units which may be less relevant for focusing on voice quality and monotonous prosody that is present in PD speech [81, 80]. Also, the sustained phonation task is different from connected speech on which it was pretrained.

The performance of FBanks (63.36%) is between Whisper and Wav2Vec2 which shows that traditional methodologies can still be used for pathological speech detection. It could be that the datatsets on which the SSL models were trained on are different than the pathological audio recordings which makes it hard for them to generalize. However, the more general speech features extracted with FBank are more optimized for pathology recognition.

Overall, the datasets on which the SSL models were trained on were predominantly healthy, fluent speech from audiobooks, podcasts, and conversational recordings which is different from pathological data. This represents a fundamental challenge in transfer learning for clinical applications.

The sex difference across fundamentally different model architectures including deep

learning with SSL models and traditional machine learning with Random Forest and simple spectral features with FBanks shows that the disparity is not an artifact of any particular algorithmic design. The consistent presence of sex difference towards women cannot be attributed to a single model architecture but may be more related to datasets characteristics, especially the size and recording quality and collection methodology.

The model can be able to correctly classify the female recordings whose higher F0 values are between 165 and 255 Hz and thus learn decision boundaries related to male specific F0 ranges which are between 85 and 180 Hz [82]. This is consistent with previous work that indicate that fundamental frequency is one of the most important features when it comes to diagnosing PD [83].

4.2 Sex differences on different datasets and size effect

Since the sex difference observed in the previous section is not related to the models themselves, we assessed if it was a result of dataset-specific characteristics including size and demographic composition. For this purpose, we used WavLM-Large as a feature extractor since it achieved the best performance based on accuracy, (69.04% $\pm$ 5.86) as seen in A.1. We run the experiments on 3 random seeds and 5 folds, as the datasets are smaller. We present the results from using WavLM pipeline on mPower matched, and mPower small, in addition to PC-GITA datasets.

Datasets	F1	F1 male	F1 female
mPower matched	56.54±1.47	51.71±3.93	61.16±3.39 ***
mPower small	50.01 ±8.99	48.81±12.66	49.45±16.25
mPower large	68.61 ± 6.61	53.65 ± 8.39	83.12 ± 9.36 ***
PC-GITA	51.40±6.69	50.57±12.58	51.18±7.95

Table 4: Performance of WavLM on mPower (matched and large) and PC-GITA. Asterisks indicate significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Grey cells show no significant difference.

We observe that mPower large gave the best performance with a sex difference towards women (F1 female 83.12 $\pm$ 9.36). It is followed by mPower matched where the dataset is smaller than mPower large but sex and age are matched. We see that women are better detected (F1 female 61.16 $\pm$ 3.39 vs. F1 male 51.71 $\pm$ 3.93) as well. For PC-GITA and mPower small, where age and sex are matched, and the dataset size is 100 samples for each, the performance is approximately at chance level and it is not statistically significant ($p > 0.05$ for all metrics compared to random baseline).

Both PC-GITA and mPower small datasets showed a high variance in F1 scores with standard deviations ranging from 6.69% to 16.25%. This indicates that the models learning is unstable when trained on limited number of samples. The mPower small dataset showed a large standard deviation for female participants which shows that small sample sizes affect the reliability of the sex differences.

Table 4 shows that sex related differences in the performances were statistically significant in the large datasets only and the direction of the effect was always towards women. The results demonstrate that the sex difference could be dataset dependent with stronger performance for women in mPower datasets. The size of the dataset is also affecting the results. On mPower, the larger the dataset, the better the performance, and the larger the gap between men and women. The dataset demographic information is provided in Table 2.

The performance on PC-GITA where the recordings were collected in a laboratory and mPower small is interesting as it is very similar to each other. The absence of statistical significance in the results of small datasets shows that there was insufficient statistical data to detect the subtle speech differences between PD and HC groups when the dataset size is limited.

4.3 Performance of sex-specific training datasets

In order to further investigate the sex differences in the datasets, we conducted more experiments when we train on one sex only and test on both sexes. We evaluated WavLM and RF across 5 different folds and 3 random seeds because WavLM and RF were yielding the best performance on mPower large. The results are reported in Table 5. For each sex-specific training experiment, we used 50% of the available data which means only male or only female from the existing datasets, but we kept the same testing sets.

Model	Subset	F1	F1 female	F1 male
WavLM	male mPower matched	47.99 ± 3.08	49.06 ± 8.04	46.39 ± 9.42
WavLM	female mPower matched	54.70 ± 5.96	55.99 ± 8.47	53.23 ± 7.08
WavLM	male PC-GITA	40.03 ± 1.79	42.57 ± 3.58	37.50 ± 0.00
WavLM	female PC-GITA	62.80 ± 1.98	65.90 ± 5.37	58.30 ± 0.00
Random Forest	male mPower matched	48.19 ± 4.31	49.15 ± 2.59	45.36 ± 7.75
Random Forest	female mPower matched	45.26 ± 4.04	47.07 ± 7.17	42.49 ± 4.19
Random Forest	male PC-GITA	45.81 ± 3.75	23.08 ± 0.00	65.60 ± 7.25
Random Forest	female PC-GITA	52.91 ± 6.16	53.81 ± 11.43	42.22 ± 9.50

Table 5: Comparison of WavLM and Random Forest on female and male subsets from mPower matched and PC-GITA datasets. Values are averaged across folds and seeds in percentage. Asterisks indicate significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Grey cells show no significant difference.

From Table 5, we notice that, overall, the performance on sex-specific training datasets is lower compared to Table 4 when training was done on both sexes. This degradation shows that reducing training set diversity when excluding one sex affects model generalization which is consistent with domain adaptation theory [84].

The permutation tests show that we lost statistical significance for females. The results are not statistically significant overall and for men and women on both datasets and with both training pipelines.

From Table 5, we notice that WavLM and RF perform similarly on mPower matched and PC-GITA sustained vowel task in terms of not getting significant outcomes. Both datasets are size, age, sex, and health status matched. Therefore, the sex difference could

be coming from the dataset size, the quality of the dataset, and the disease severity. We only have the UPDRS III score matched on PC-GITA, but not on the mPower subsets as we do not have access to the scores.

In [85], the authors noted that across their analyses, each patient showed distinct speech patterns which reflects the heterogeneity of the disease.This variability can be amplified by differences related to sex, hormonal influences on dopaminergic systems, and anatomical variations in vocal tract structure [86, 87].

Chapter 5

Conclusion and future work

5.1 Conclusion

From our results, we observe that the model choice, dataset characteristics, and demographics are factors that highly influence the detection of PD from speech. These elements show complex interactions between model design, data quality, and population characteristics that have important effects on the development of clinically accurate speech-based diagnostic systems.

In this study, we investigated the sex differences when detecting Parkinson's Disease from speech using Deep Learning and traditional machine learning models. We concluded that there are statistically significant differences for PD diagnosis from speech when the datasets are large enough to provide statistical power with female participants consistently achieving higher classification performance compared to males across all evaluated models and feature extraction methods. This suggests that the cause of the differences is not algorithm specific but more general and likely attributed to dataset characteristics and biological differences in PD effects across sexes [88, 89].

However, small datasets exhibited no significant results which shows that the dataset size are critical factors in detecting sex differences based on our results.

In addition, the fluctuating standard deviations observed throughout the results from 5.81% to 9.36% for overall F1-scores show a high variability in the performance across folds and seeds which means that the conditions of the recordings and demographic differences can introduce a complexity that challenges the performance.

WavLM-Large led to the best overall performance among SSL models while Random Forest achieved the highest F1-score on women. Traditional spectral features with FBank were giving competitive results which shows that simpler approaches could still be used for pathological speech detection.

In addition, experiments on one sex only and testing on both can lead to failure in the generalization depending on the model used which shows the risks of deploying models trained on demographically biased datasets and with no statistical evaluation of the results.

5.2 Implications for Fairness and Clinical Deployment

From our findings, the sex differences observed in large datasets raise critical concerns for clinical applications. These differences are not statistical artifacts as they reflect real disparities that if unaddressed can lead to inequitable healthcare outcomes.

We designed our study with explicit attention to fairness considerations since we used balanced datasets in terms of sex and disease status, in addition to cross-validation to ensure equal representation across folds.

Despite these methodological choices, our results showed persistent sex differences. This suggests that demographic balance is necessary but not sufficient. Other datasets characteristics like variability in symptoms, recording conditions, presence of other neurodegenerative diseases and medical knowledge must be considered carefully in order to ensure algorithmic fairness.

A model that correctly identifies 85% of female cases, but chance level for males would result in diagnostic failure for men which represents a potential safety issue and ethical

violation of the principle of equitable care as it would disproportionately disadvantage one sex.

In addition, a performance without statistical significance is not enough to decide on the validity of the results. Furthermore, statistical significance alone is insufficient to determine the clinical readiness. In fact, the inability to get good performance on male participants makes our models unreliable for deployment despite the fact that the results on females is high. This indicates the importance of evaluating models across all relevant subpopulations and integrating fairness evaluation into the development of these models.

5.3 Future Work

Future work should include experimenting with other speech tasks and not only sustained vowel phonation to decide if the sex difference generalize to the other tasks. Specifically, investigating reading tasks, DDK exercises, and spontaneous speech to see if the disparities are task-dependent.

Furthermore, incorporating UPDRS scores from mPower datasets is another important direction. This would allow us to investigate if the sex difference is related to disease severity which would provide insights into the clinical aspects of model biases. By stratifying participants by disease stage and analyzing performance across severity levels, we could determine whether sex differences are more present in early stage or advanced PD.

Detailed interpretability experiments are also essential. Identifying which features are influencing the classification is necessary for a better understanding of those sex differences and inform the design of more equitable models. Feature importance analysis and attention visualization could show whether models rely on sex-specific characteristics or disease markers.

Also, implementing and evaluating fairness techniques such as demographic parity constraint and equalized odds optimization are important to assess the trade-offs between overall performance and demographic fairness. Exploring domain adaptation techniques and transfer learning approaches that account for sex differences could improve model generalization across demographic groups.

Extending our work to additional languages and cultural contexts would help identify if the sex difference are universal or specific to certain conditions.

Finally, creating sex balanced and demographically diverse speech datasets and detailed metadata is crucial to enable future fairness research. These datasets should include detailed demographic information, clinical variables, and recording metadata.

Bibliography

[1] Hiba Akhaddar, Tristan Glatard, and Mirco Ravanelli. Detection and prediction of progression of parkinson's disease from speech. Poster presented at NeurIPS 2024 — Affinity Event, 2024. Available online: https://neurips.cc/virtual/2024/109036.

[2] E. Ray Dorsey, Thomas Sherer, Michael S Okun, and Bastiaan R Bloem. The emerging evidence of the parkinson pandemic. *Journal of Parkinson's Disease*, 8(s1):S3–S8, 2018.

[3] Werner Poewe, Klaus Seppi, Caroline M. Tanner, Glenda M. Halliday, Patrik Brundin, Jens Volkmann, Anette-Eleonore Schrag, and Anthony E. Lang. Parkinson disease. *Nature Reviews Disease Primers*, 3(1):17013, 2017.

[4] Public Health Agency of Canada, Centre for Surveillance and Applied Research. Parkinsonism in canada, including parkinson disease: Data blog. `https://health-infobase.canada.ca/datalab/parkinson-blog.html`, 2025.

[5] Joseph Jankovic. Parkinson's disease: clinical features and diagnosis. *Lancet Neurology*, 7(6):465–474, 2008.

[6] Anu Iyer, Aaron Kemp, Yasir Rahmatallah, Lakshmi Pillai, Aliyah Glover, Fred Prior, Linda Larson-Prior, and Tuhin Virmani. A machine learning method to process voice samples for identification of parkinson's disease. *Scientific Reports*, 13(20615):1–10, nov 2023.

[7] John M Fearnley and Andrew J Lees. Ageing and parkinson's disease: substantia nigra regional selectivity. *Journal of Neurology, Neurosurgery & Psychiatry*, 54(6):507–513, 1991.

[8] William T Dauer and Serge Przedborski. Parkinson's disease. *Archives of Neurology*, 60(1):24–32, 2003.

[9] Christiane Arnold, Johannes Gehrig, Suzana Gispert, Carola Seifried, and Christian A. Kell. Pathomechanisms and compensatory efforts related to parkinsonian speech. *NeuroImage: Clinical*, 2014.

[10] Stefan Skodda, Ulrich Schlegel, and Wilfried Gronheit. Speech rate and rhythm in parkinson's disease. *Journal of Neurology*, 257(4):550–556, 2010.

[11] Alice K Ho, Robert Iansek, Carlotta Marigliani, John L Bradshaw, and Stephen Gates. Speech characteristics in parkinson's disease and their relation to the underlying neurodegenerative processes. *Brain and Language*, 81(2):234–248, 2012.

[12] Kiran Reddy Mittapalle and Paavo Alku. Automatic detection of parkinsonian speech using wavelet scattering features. *JASA Express Letters*, 2025.

[13] R. A. Hauser. Unmet needs in parkinson's disease. *Parkinsonism & Related Disorders*, 19:S3–S6, 2013.

[14] Lorenzo Simone, Mauro Giuseppe Camporeale, Vito Marco Rubino, Vincenzo Gervasi, and Giovanni Dimauro. Interpretable early detection of parkinson's disease through speech analysis. *arXiv preprint arXiv:2504.17739*, 2025.

[15] C. F. de Lima, R. A. Teixeira, A. M. Reis, C. Ferreira, and O. F. Gonçalves. Sex differences in the human speech production network. *NeuroImage*, 202:116092, 2019.

[16] Hee-Sun Jung, Ji-Hye Kim, Jae-Hyung Park, Sang-Hoon Lee, Jae-Seung Oh, and Jun Soo Kwon. Sex differences in the human voice and its neural control. *Cerebral Cortex*, 29(4):1769–1781, 2019.

[17] Elaine Kearney and Frank H. Guenther. Articulating: The neural mechanisms of speech production. *Language, Cognition and Neuroscience*, 34(9):1214–1229, 2019.

[18] Frank H. Guenther and Tony Vladusich. A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, 25:408–422, 2012.

[19] Jessica E. Huber. Respiratory control for speech production: A review of the literature. *Journal of Voice*, 22(5):489–499, 2008.

[20] Kenneth N. Stevens. *Acoustic Phonetics*. MIT Press, 2000.

[21] Raymond D. Kent. Research on speech motor control and its disorders: A review and prospective. *Journal of Communication Disorders*, 33(5):391–427, 2000.

[22] Hermann Ackermann, Steffen R. Hage, and Wolfram Ziegler. Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective. *Behavioral and Brain Sciences*, 37(6):529–546, 2014.

[23] Kristina Simonyan, Hermann Ackermann, Edward F. Chang, and Jeremy D. Greenlee. New developments in understanding the complexity of human speech production. *Journal of Neuroscience*, 36(45):11440–11448, 2016.

[24] Ann M. Graybiel. Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience*, 31, 2008.

[25] J. C. Houk. Agents of the mind. *Biological Cybernetics*, 92(6), 2005.

[26] Sabine Skodda. Speech and voice in parkinson's disease. *Handbook of Clinical Neurology*, 117, 2013.

[27] Ismail Midi, Melahat Doğan, Meryem Köseoğlu, Duygu Canbaz, Muhammet A. Şehitoğlu, and Dilek I. Günal. Voice disorders in parkinson's disease: A review of the literature. *Journal of Voice*, 22(6), 2008.

[28] Nancy P. Solomon and Thomas J. Hixon. Speech breathing in parkinson's disease. *Journal of Speech and Hearing Research*, 36(2), 1993.

[29] Bruce E. Murdoch, Helen J. Chenery, S. Bowler, and J. C. L. Ingram. Respiratory function and speech production in parkinson's disease: A review and preliminary results. *Journal of Neurology, Neurosurgery & Psychiatry*, 52(8), 1989.

[30] D. G. Hanson, B. R. Gerratt, and P. H. Ward. Laryngeal function in parkinson's disease. *Laryngoscope*, 94(3), 1984.

[31] K. S. Perez, L. O. Ramig, M. E. Smith, and C. Dromey. Dysphonia in parkinson's disease: Clinical and experimental observations. *Journal of Speech and Hearing Research*, 39(1), 1996.

[32] C A Haaxma, B R Bloem, G F Borm, W J Oyen, and K L Leenders. Gender differences in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 78(8):819–824, 2007.

[33] Jan Rusz, Tereza Tykalová, Michal Novotný, David Zogala, Evžen Růžička, and Petr Dušek. Automated speech analysis in early untreated parkinson's disease: relation to gender and dopaminergic transporter imaging. *European Journal of Neurology*, 29(1):81–90, 2022.

[34] Eric J. Hunter and Ingo R. Titze. Variations in intensity, fundamental frequency, and voicing for teachers in occupational versus nonoccupational settings. *Journal of Speech, Language, and Hearing Research*, 2010.

[35] M. P. Gelfer and Q. E. Bennett. Speaking fundamental frequency and vowel formant frequencies: effects on perception of gender. *Journal of voice : official journal of the Voice Foundation*, 2013.

[36] J. H. Ahn, D. Shin, M. K. Suh, Y. E. Huh, J. Youn, J. W. Cho, and J. S. Lee. Sex-specific speech correlates of dysphagia in parkinson's disease. *Movement Disorders*, 2023.

[37] Enea Traini Hossain, Mohammad Amran and Francesco Amenta. Machine learning applications for diagnosing parkinson's disease via speech, language, and voice changes: A systematic review. *Inventions*, 2025.

[38] Moreno La Quatra, Maria Francesca Turco, Torbjørn Svendsen, Giampiero Salvi, Juan Rafael Orozco-Arroyave, and Sabato Marco Siniscalchi. Exploiting foundation models and speech enhancement for parkinson's disease detection from speech in real-world operative conditions. In *Proceedings of Interspeech 2024*, 2024.

[39] Juan Rafael Orozco-Arroyave, Juan F. Vargas-Bonilla, Juan D. Arias-Londoño, Sebastián Murillo-Rendón, Gloria Castellanos-Domínguez, and Javier Garcés. Pc-gita: A spanish-language speech corpus for parkinson's disease detection, 2013.

[40] Moreno La Quatra, Juan Rafael Orozco-Arroyave, and Marco Sabato Siniscalchi. Bilingual dual-head deep model for parkinson's disease detection from speech. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025.

[41] David Gimeno-Gómez, Catarina Botelho, Anna Pompili, Alberto Abad, and Carlos-D. Martínez-Hinarejos. Unveiling interpretability in self-supervised speech representations for parkinson's diagnosis. *IEEE Journal of Selected Topics in Signal Processing*, 2025.

[42] Maurycy Chronowski, Maciej Kłaczynski, Małgorzata Dec-Cwiek, and Karolina Porębska. Parkinson's disease diagnostics using ai and natural language knowledge transfer, 2022.

[43] J. S. Almeida. Detecting parkinson's disease with sustained phonation. *Journal of Neuroscience Methods*, 323:1–8, 2019.

[44] Nichole Houle, Taylor Feaster, Amna Mira, Kirsten Meeks, and Cara E. Stepp. Sex differences in the speech of persons with and without parkinson's disease. *American Journal of Speech-Language Pathology*, 2024.

[45] Tariq Adnan, Abdelrahman Abdelkader, Zipei Liu, Ekram Hossain, Sooyong Park, Md Saiful Islam, and Ehsan Hoque. A novel fusion architecture for detecting parkinson's disease using semi-supervised speech embeddings. *npj Parkinson's Disease*, 11, 2025.

[46] Lisanne van Gelderen and Cristian Tejedor-García. Innovative speech-based deep learning approaches for parkinson's disease classification: A systematic review. *Applied Sciences*, 2024.

[47] Laetitia Jeancolas, Dijana Petrovska-Delacrétaz, Graziella Mangone, Badr-Eddine Benkelfat, Jean-Christophe Corvol, Marie Vidailhet, Stéphane Lehéricy, and Habib Benali. X-vectors: New quantitative biomarkers for early parkinson's disease detection from speech. *Frontiers in Neuroinformatics*, 15, 2021.

[48] Rania Khaskhoussy and Yassine Ben Ayed. Detecting parkinson's disease according to gender using speech signals. In Han Qiu, Cheng Zhang, Zongming Fei, Meikang Qiu, and Sun-Yuan Kung, editors, *Knowledge Science, Engineering and Management*, pages 414–425, Cham, 2021. Springer International Publishing.

[49] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[50] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[51] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, others, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[52] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[53] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Proceedings of INTERSPEECH 2020*, pages 3830–3834, 2020.

[54] Nauman Dawalatabad, Mirco Ravanelli, François Grondin, Jenthe Thienpondt, Brecht Desplanques, and Hwidong Na. Ecapa-tdnn embeddings for speaker diarization. In *Proceedings of INTERSPEECH 2021*, pages 3560–3564, 2021.

[55] M. Chaitra, Anupam Mandal, and Sourya Mukherjee. Study of ECAPA-TDNN models for spoken language identification task. In *Proceedings of the IEEE Conference*, pages 233–237, 2023.

[56] Jinlong Xue, Yayue Deng, Yichen Han, Ya Li, Jianqing Sun, and Jiaen Liang. Ecapa-tdnn for multi-speaker text-to-speech synthesis. In *Proceedings of the 13th International Symposium on Chinese Spoken Language Processing (ISCSLP 2022)*, pages 230–234, 2022.

[57] Athanasios Tsanas, Max A Little, Patrick E McSharry, Jennifer Spielman, and Lorraine O Ramig. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 59(5):1264–1271, 2012.

[58] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462, 2010.

[59] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[60] Joseph Kahn, Anna Silnova, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *Proceedings of Interspeech 2020*, pages 766–770, 2020.

[61] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Henretty, E. Morais, L. Saunders, F. Saurous, B. Pang, and S. Potapov. Common voice: A massively-multilingual speech corpus. In *Proceedings of Language Resources and Evaluation 2020 (LREC 2020)*, pages 4218–4222, 2020.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS) 30*, pages 5998–6008, 2017.

[63] Juan C Vásquez-Correa, Tomás Arias-Vergara, Juan R Orozco-Arroyave, Jesús F Vargas-Bonilla, and Elmar Nöth. Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease. *Journal of Communication Disorders*, 76:21–36, 2018.

[64] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[65] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 43, pages 652–662. IEEE, 2021.

[66] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[67] Brian M. Bot, Christine Suver, Emilio C. Neto, Michael Kellen, Adrian Klein, Joshua C. Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, Ray Dorsey, and Stephen H. Friend. The mPower study, parkinson disease mobile data collected using iphone sensors. *Scientific Data*, 3:160011, 2016.

[68] Juan R Orozco-Arroyave, Juan D Arias-Londoño, Jesús F Vargas-Bonilla, Miguel C González-Rátiva, and Elmar Nöth. New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 342–347, 2014.

[69] James Prince, Sumit Arora, and Maarten De Vos. Big data in parkinson's disease: Using smartphones to remotely detect longitudinal disease phenotypes. *Physiological Measurement*, 39(4):044005, 2018.

[70] A Coravos, S Khozin, and KD Mandl. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ Digital Medicine*, 2(1):1–5, 2019.

[71] Christopher G Goetz, Barbara C Tilley, Steven R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, others, and Nancy LaPelle. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results. *Movement Disorders*, 23(15):2129–2170, 2008.

[72] Athanasios Tsanas, Max A Little, Patrick E McSharry, Joel Spielman, and Lorraine O Ramig. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 59(5):1264–1271, 2012.

[73] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. In *Proc. Interspeech*, pages 1194–1198, 2021.

[74] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *Interspeech*, pages 2001–2005, 2016.

[75] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[76] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[77] Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Ha Nguyen, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaëlle Laperrière, Mickael Rouvier, Renato De Mori, and Yannick Estève. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333):1–11, 2024.

[78] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[79] Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863, 2010.

[80] Joseph R Duffy. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Elsevier Mosby, 3rd edition, 2013.

[81] Sabine Skodda, Wiebke Visser, and Uwe Schlegel. Vowel articulation in parkinson's disease. *Journal of Voice*, 25(4):467–472, 2011.

[82] Ronald J. Baken and Robert F. Orlikoff. *Clinical Measurement of Speech and Voice*. Singular Publishing Group, 2nd edition, 2000.

[83] Betül Erdoğdu Sakar, M. Erdem Isenkul, C. Okan Sakar, Ahmet Sertbaş, Fikret Gürgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834, 2013.

[84] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

[85] Paula Andrea P'erez-Toro, Juan Camilo Vásquez-Correa, Tom'as Arias-Vergara, Jes'us Francisco Vargas-Bonilla, Juan Rafael Orozco-Arroyave, and Elmar N"oth. Unveiling interpretability in self-supervised speech representations for parkinson's diagnosis. *Frontiers in Neuroinformatics*, 17:1—15, 2023.

[86] Irina N Miller and Alice Cronin-Golomb. Gender differences in parkinson's disease: clinical characteristics and cognition. *Movement Disorders*, 25(16):2695–2703, 2010.

[87] Ingo R Titze. Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, 85(4):1699–1707, 1989.

[88] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

[89] Agustin J. Larrazabal, Natalia Nieto, Vanessa Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences (PNAS)*, 117(23):12592–12594, 2020.

Appendix A

Appendix

The following tables show other metrics computed in addition to F1 scores reported in the main text.

Models	Accuracy	F1	F1 male	F1 female
WavLM	69.04 ± 5.86	68.61 ± 6.61	53.65 ± 8.39	83.12 ± 9.36 ***
FBanks	63.52 ± 5.81	63.36 ± 5.87	43.30 ± 7.73	82.63 ± 10.57 ***
Wav2Vec2.0	57.2 ± 7.28	55.66 ± 8.40	49.30 ± 7.39	60.79 ± 13.05 ***
Whisper	67.76 ± 5.95	66.54 ± 6.38	56.00 ± 7.40	76.38 ± 9.36 ***
Random Forest	66.64 ± 3.36	66.61 ± 3.36	46.80 ± 5.83	85.02 ± 3.79 ***

Table A.1: Performance of 5 feature extractors on mPower large. Asterix indicates significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Grey cells show no significant difference.

Datasets	Accuracy	F1	F1 male	F1 female
mPower matched	56.87 ± 1.39	56.54±1.47	51.71±3.93	61.16±3.39 ***
mPower small	53.00 ± 7.02	50.01 ±8.99	48.81±12.66	49.45±16.25
mPower large	69.04 ± 5.86	68.61 ± 6.61	53.65 ± 8.39	83.12 ± 9.36 ***
PC-GITA	53.33 ±6.99	51.40±6.69	50.57±12.58	51.18±7.95

Table A.2: Performance of WavLM on mPower (matched, small, and large) and PC-GITA. Asterix indicates significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Grey cells show no significant difference.

Subset	Accuracy	F1	Precision	Recall
male mPower matched	49.47 ± 3.04	47.99 ± 3.08	49.49 ± 3.27	49.47 ± 3.04
female mPower matched	55.40 ± 5.83	54.70 ± 5.96	55.72 ± 5.99	55.40 ± 5.83
male PC-GITA	48.85 ± 3.63	38.54 ± 6.86	35.11 ± 12.14	48.85 ± 3.63
female PC-GITA	62.33 ± 7.29	60.97 ± 7.70	64.79 ± 8.84	62.33 ± 7.29

(a) Overall metrics

Subset	Female			Male		
	F1	Prec.	Rec.	F1	Prec.	Rec.
male mPower matched	49.06 ± 8.04	51.50 ± 9.04	51.06 ± 7.96	46.39 ± 9.42	47.15 ± 9.64	47.90 ± 8.51
female mPower matched	55.99 ± 8.47	56.66 ± 8.70	56.38 ± 8.50	53.23 ± 7.08	54.93 ± 7.67	54.40 ± 6.80
male PC-GITA	38.50 ± 6.36	35.62 ± 12.35	49.23 ± 2.77	38.35 ± 7.90	34.71 ± 12.59	48.46 ± 5.55
female PC-GITA	66.89 ± 7.64	70.97 ± 7.98	68.00 ± 6.76	54.27 ± 11.58	58.30 ± 15.68	56.67 ± 10.47

(b) Gender-specific metrics

Table A.3: WavLM on female and male subsets from mPower matched and PC-GITA in percent averaged across 5 folds and 3 seeds. Grey cells show no significant difference.

Subset	Accuracy	F1	Precision	Recall
male mPower matched	49.68 ± 4.45	48.19 ± 4.31	49.80 ± 4.89	49.68 ± 4.45
female mPower matched	45.49 ± 3.92	45.26 ± 4.04	45.39 ± 4.03	45.49 ± 3.92
male PC-GITA	48.00 ± 3.56	45.81 ± 3.75	47.60 ± 4.26	48.00 ± 3.56
female PC-GITA	54.67 ± 5.91	52.91 ± 6.16	55.51 ± 6.78	54.67 ± 5.91

(a) Overall metrics

Subset	Female			Male		
	F1	Prec.	Rec.	F1	Prec.	Rec.
male mPower matched	45.36 ± 7.75	50.80 ± 11.01	49.79 ± 7.90	49.15 ± 2.59	49.60 ± 2.82	49.58 ± 2.69
female mPower matched	42.49 ± 4.19	42.76 ± 4.19	43.14 ± 3.97	47.07 ± 7.17	48.15 ± 8.93	47.84 ± 7.48
male PC-GITA	23.08 ± 0.00	18.75 ± 0.00	30.00 ± 0.00	65.60 ± 7.25	66.74 ± 7.36	66.00 ± 7.12
female PC-GITA	53.81 ± 11.43	54.85 ± 12.44	54.67 ± 10.87	42.22 ± 9.50	49.63 ± 26.33	54.67 ± 4.99

(b) Gender-specific metrics

Table A.4: Random Forest on female and male subsets from mPower matched and PC-GITA in percent. Grey cells show no significant difference.