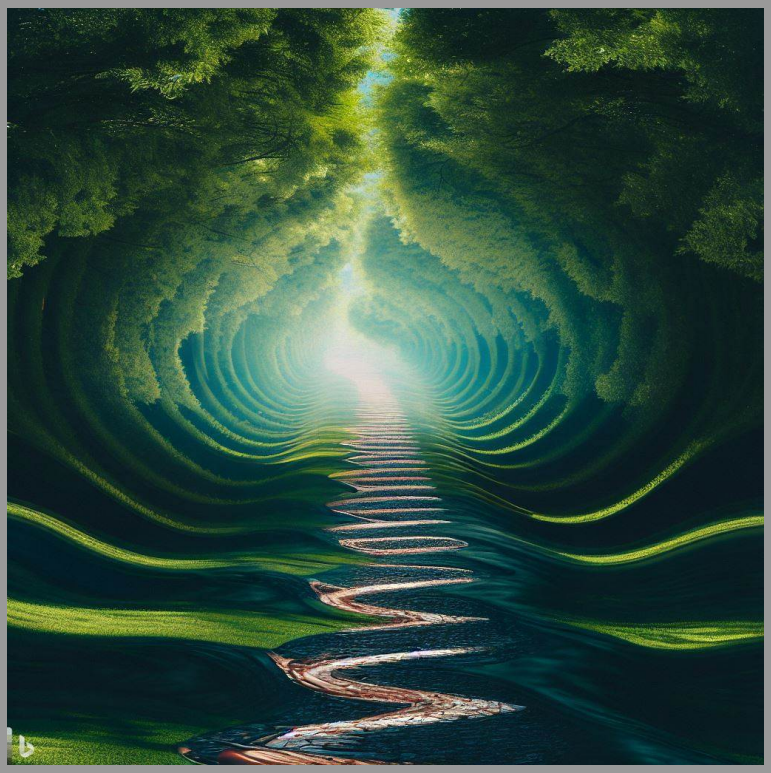
Investigating Pre-Trained Self-Supervised Deep Learning Models for Disease Recognition



Julius J. Bijkerk

Layout: typeset by the author using LaTeX.
Cover illustration: Microsoft Bing (https://www.bing.com/images/create/), suggesting a larynx formed out of raw wav forms, leading to a healthy environment. Prompt on request.

Investigating Pre-Trained Self-Supervised Deep Learning Models for Disease Recognition

Julius J. Bijkerk
12219967

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Prof. dr. P.P.G. (Paul) Boersma

Institute for Logic, Language and Computation
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

June 30, 2023

Abstract

This thesis evaluated the efficiency of pre-trained self-supervised deep learning models for disease recognition from speech, focusing on the wav2vec 2.0 model and comparing it with traditional ML methods, specifically MFCC + 1D CNN and MFCC + MLP models, on the TORGO database for dysarthria. The traditional MFCC + 1D CNN model outperformed the others in terms of accuracy (90.10%), precision (85.09%), recall (97.49%), and F1-score (90.87%), while the MFCC + MLP model achieved the highest AUC-ROC score (93.79%). Wav2vec 2.0 achieved lower, but still competitive scores: accuracy (86.25%), precision (83.49%), and recall (90.55%). The results suggest that traditional methods can sometimes outperform advanced models for specialized tasks. The study also underscores the importance of ethical considerations, including data privacy and potential biases in deploying ML models in healthcare. Future research is recommended to maximize the potential of self-supervised deep learning models in disease recognition tasks and healthcare applications in general.

Acknowledgements

First of all, I would like to thank my supervisor Prof. Dr. P.P.G. (Paul) Boersma. Paul initiated the topic "Anything on speech and AI", which was one of the possible topics during the prephase of this thesis. This topic directly catched me, because it suited my interests and I already had some ideas in mind for narrowing down the topic. Together with my fellow students Anastasia Shchupak and Nilansha Dargan, both from the Bachelor Linguistics, we had weekly meetings with Paul to discuss our progression and receive criticism as well as new insights from each other. These appointments were very valuable to me, Therefore I want to thank you, Paul, for your time to get us to understand new concepts. But also, Anastasia and Nilansha, I enjoyed the close contact we had during this period and learning about your topics was really informative. Secondly, there is one friend and fellow student from the Bachelor Artificial Intelligence who went to the same processes of writing this thesis, namely Floris Kotterink. We discussed a lot together, either during coffee meetings at science park or at all times via phone. Besides that it was fun, this way of sharing of ideas and suggesting directions for each other, even though we had completely different topics, was very valuable. Lastly, I had a couple of people proofreading some parts, if not all, of this thesis. These are my girlfriend Julia, friends Sebastiaan, Broos, Tanne, Nathan, Nathan's father Michel, and my own father, Jan Jacob. Thank you for your time, I learned a lot from your comments.

Keywords: Spectrogram; Mel-frequency cepstral coefficients (MFCCs); Dysarthria; Pre-training; Self-supervised Learning; Deep Learning; Deep Neural Network (DNN); Transformer; Encoder-decoder; Masking; wav2vec; Librosa; Scikit-learn; PyTorch

Contents

1 Introduction

Human speech could be influenced by certain health conditions. For example, the pitch and loudness of someone's speech stabilize as the first signs of Parkinson's Disease [Ma et al., 2020]. Also, even in the early prodromal stages of Alzheimer's Disease, the temporal characteristics of spontaneous speech, such as speech tempo, number of pauses in speech, and the length of these pauses are affected [Ahmed et al., 2013, Szatloczki et al., 2015]. Additionally, brain damage or neurological conditions like Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS), or again Parkinson's disease could be the cause of dysarthria [Darley et al., 1969]. Dysarthria is a motor speech disorder, which implies that it is difficult for the speaker to move the muscles necessary for speech production due to the weakness of these muscles or the neurological failure of coordinating them [Yunusova et al., 2008]. Therefore, dysarthria, as well as the mentioned diseases that could cause dysarthria, may affect the overall communicative ability of a patient.

The process of diagnosing these aforementioned, speech-affecting, diseases or disorders is normally performed by medical specialists, for example, psychologists, trained physicians, or speech pathologists [Braun et al., 2022, Schu et al., 2023]. It could consist of medical examinations, such as a full blood count, liver function tests, a chest x-ray, a brain scan, a combination of medical and psychological history taking, cognitive testing, the use of rating scales, or auditory assessments [Cooper and Greene, 2005, Schu et al., 2023]. Waiting times for appointments with medical experts frequently exceed 6 months [Braun et al., 2022], the appointments themselves can be very time-consuming [Schu et al., 2023] and the assessments may suffer from subjectivity [Kent, 1996, Connaghan et al., 2021].

However, early and adequate diagnosis and therefore the possibility of a rapid start of treatment could be key to slowing down the progression of such diseases and disorders. For example, someone with dysarthria could remarkably benefit from speech and language therapy. Starting such a therapy as early as possible means maintaining a high quality of life for as long as possible, for patients and their relatives [Enderby, 2013].

These are some of the main reasons why there has been a growing interest in developing methods for automatically detecting these diseases or disorders in their early stages mechanically, all with the aim of assisting healthcare professionals. This is where machine learning (ML) will be very convenient. Previous research in this field showed that ML could serve as a quick, objective, and non-invasive assessment of an individual's health condition [Balagopalan and Novikova, 2021]. This is because classification algorithms are capable of learning distinguishable acoustic patterns [Kotsiantis et al., 2007] for people with and without a specific disease or disorder by training on relevant data.

Modern techniques for automatically classifying dysarthria based on speech audio can be broadly grouped into two categories: either they use hand-crafted features and traditional machine learning classifiers [Jeancolas et al., 2019, Kodrasi et al., 2020] or they use deep learning methods that are trained to automatically extract discriminative speech representations and classify accordingly [Millet and Zeghidour, 2019, Joshy and Rajan, 2022]. As hand-crafted features, this thesis will be focused on the Mel-frequency Cepstral Coefficients (MFCCs) [Tiwari, 2010]. MFCCs are audio features that have been mathematically configured to emphasize lower frequencies, comparable to human auditory perception. While, for the 'latent' deep learning features, Facebook's wav2vec [Baevski et al., 2020] – a pre-trained, self-supervised deep learning model, will be used. These two approaches mainly differ in how the input features are composed and in the type of ML classifier.

The primary objective of this thesis is to investigate the pre-trained self-supervised deep learning

model, wav2vec, for the purpose of disease recognition. Therefore, this model, as well as the traditional MFCC based classifier for comparison, will be developed, trained and ultimately tested on the publicly available TORGO database on dysarthria [Rudzicz et al., 2012]. The Torgo database consists of the raw speech audio files of patients and their demographically matched, healthy counterparts with the label of either one of the groups: "dysarthria" or "control". The assessment criteria of their performance will consist of recall, precision, f1-score, accuracy, and AUC-ROC, providing an extended evaluation for each approach.

To summarize, in this thesis the following will be compared:

1. **Classifying on conventional handcrafted MFCC features:** extracting the MFCC features and composing a suitable classification algorithm to train.

2. **Classifying on deep learning features from Facebook's Wav2vec model:** Replicating the wav2vec implementation of the following research paper: *'On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches'*, by Schu, Guilherme and Janbakhshi, Parvaneh and Kodrasi, Ina [Schu et al., 2023].

The trend of open-sourcing has revolutionized the field of machine learning [Sonnenburg et al., 2007]. A global community of researchers and practitioners has emerged who contribute to the rapid advancement of artificial intelligence. This open-source ideology has democratized access to innovative ML models, encouraging widespread experimentation and innovation. Platforms such as Github [1], HuggingFace [2], Kaggle [3] and Papers With Code [4] represent this trend. Here, researchers can share their academic papers along with the associated code, enabling others to reproduce, validate, and build upon their work. This idea of transparency and collaboration stimulates problem-solving in AI, as it allows for a collective understanding and the ability to tackle complex challenges together. This leads to the second objective of this thesis, which is to create a detailed, understandable, and efficient deployment pathway for these models. By creating a "guide" or "manual" based on the encountered bottlenecks and workarounds, this will contribute to the field in a way that enables others to easily build upon this work, which is a fundamental aspect of academia. In alignment with this, the Python programming work of this thesis, in the form of Jupyter Notebooks, will be made publicly available on GitHub at `https://github.com/JungCesar/bscaithesis`.

This thesis is structured in the following manner: as the final part of the introduction, the main research question, the associated sub-questions, and the hypothesis will be presented directly hereafter. The theoretical background will be covered in section 2, this includes the current state of research and relevant technical explanations as part of understanding the problem. In section 3, an elaboration of the used methodology will be presented, which covers the replication of the wav2vec implementation from the aforementioned paper and the development of an MFCC-based classifier. This will include an explanation of the TORGO dataset and statements about certain obstacles that were discovered while developing the different models. In section 4, the results of the method will be presented and evaluated. Section 5 covers an illustration of how this could be implemented in the real world, the ethical considerations it brings, possible further research directions within this area, and a reflection of the overall process. Additionally, the sub-questions will be answered

[1] `https://github.com/`
[2] `https://huggingface.co/`
[3] `https://www.kaggle.com`
[4] `https://paperswithcode.com/`

in sections 2, 3, 4, and 5. The conclusion, which contains the answer to the main research question, is presented in section 6.

1.1 Main Research Question and Sub-Questions

The topic of this research is the following: "Investigating Pre-Trained Self-Supervised Deep Learning Models for Disease Recognition". This topic will be narrowed down to the following main research question:

> *How effective are pre-trained self-supervised deep learning models in recognizing diseases or disorders, comparing to traditional ML methods on the TORGO database on dysarthria?*

This main research question is broken down into separate sub-questions, the answers of which will be fundamental to answering the main question. These sub-questions are the following:

1. Which types of diseases and disorders can be recognized through speech, and how do these conditions alter speech patterns?

2. What is the traditional approach to diagnosing these diseases using ML, and how does it compare to the intended use of the modern pre-trained self-supervised deep learning models?

3. What are the key principles of self-supervised deep learning models, and how are they pre-trained?

4. Which steps are involved in developing the different ML models for disease recognition and what are the major obstacles here?

5. How well do pre-trained self-supervised models perform in disease recognition tasks compared to traditional methods?

6. What are the ethical considerations when using AI for disease recognition through speech?

7. What are the overall current limitations and the potential future improvements for using pre-trained self-supervised deep learning models in disease recognition?

1.2 Hypothesis

In this study, it is proposed that dysarthria, among other speech-affecting diseases and disorders, could be classified by machine learning algorithms. While traditional machine learning approaches for disease recognition rely on handcrafted features and traditional classifiers, modern pre-trained

self-supervised deep learning models are expected to offer a significant improvement since they are pre-trained on an extensive amount of unlabeled data. The development of these models typically involves data collection, preprocessing, model development, training, and evaluation. These steps will contain notable challenges including data scarcity, an effective representation of feature vectors, and other programming conflicts. If being implemented into real-life healthcare, it is crucial to understand the ethical perspective, with important issues like patient privacy, consent, and potential biases in model performance across different demographic groups. Despite the difficulties and challenges, ML methods could be very promising for real-life healthcare applications. The hypothesis is that pre-trained self-supervised models would perform better than conventional approaches because they could capture complicated representations.

2 Theoretical Background

In this chapter, an overview will be given of the concepts that are necessary to understand the origin of the research question, its relevance, and how to develop an answer to it. Earlier research will be discussed to outline the area of research and show the current state of the problem. Therefore, this part also covers a deeper technical explanation of the key concepts involved in the ML approach to recognizing diseases. Finally, in the last section, a research gap analysis will be presented.

2.1 Diseases Affecting Speech

Human voice production, in particular speech, occurs through complex movements of different physical systems in combination with the control of sophisticated neurological systems. Both the physical parts and the neurological parts can be affected by a speaker's health condition [Gómez-García et al., 2019]. When referring to physical systems within this context, one should consider the lungs, providing the airflow necessary for speech, while the larynx, or so-called "voice box" houses the vocal cords that vibrate to create sounds. The airflow is then shaped into specific sounds by the articulators, which include the tongue, lips, and teeth. The pharynx and nasal cavity serve as resonating chambers that enhance the quality of the sounds produced. These main physical parts of human speech production are shown in Figure 1.

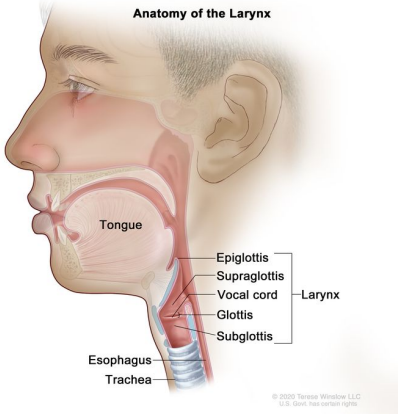


Figure 1: The anatomy of the Larynx [Winslow, 2020]

When referring to neurological control systems within the process of speech production, one should take into account regions in the brain such as Broca's area, Wernicke's area, and Geschwind's Territory as critical instances. Broca's and Wernicke's Areas play critical roles in speech production and comprehension respectively. While Geschwind's Territory plays a role in connecting and coordinating the functions of Broca's and Wernicke's areas [López-Barroso et al., 2013]. These regions in the human brain that play a main role in human speech production are shown in Figure 2.

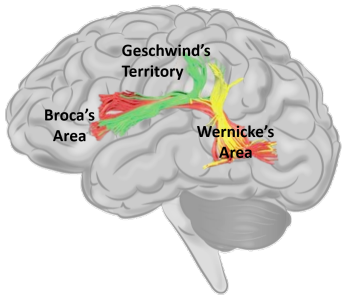


Figure 2: Different language areas in the brain [Ramoo, 2021]

The complex system of speech production, involving physical and neurological parts which are previously summarized within the confines of this thesis, can be disrupted by disease. This disruption is called a speech disorder. It is important to note that these disruptions can manifest in various ways, making it a challenging task for humans to detect and diagnose such conditions based solely on changes in speech. This underscores the significance of developing the application of artificial intelligence (AI), to analyze and interpret subtle changes in speech patterns. The potential for employing speech-based disease detection covers a broad variety of medical conditions, but the research area is still in its very early stages [Milling et al., 2022].

Voice-based AI solutions have been experimentally researched in a broad variety of medical fields, of which an overview is given in Figure 1. The conditions can roughly be categorized in the following umbrella terms: acute and chronic respiratory illnesses, pulmonary (lung) diseases, neurodegenerative diseases, developmental disorders, and psychiatric disorders.

Condition List	
Condition	Study
Dysphonia	[Costantini et al., 2021], [Cesarini et al., 2021]
COVID-19	[Costantini et al., 2022], [Han et al., 2020]
Common cold and flu	[Albes et al., 2020]
Asthma	[BT et al., 2020]
Anxiety disorder	[Baird et al., 2020]
Bipolar disorder	[Ren et al., 2019]
Depression	[Rejaibi et al., 2022]
Autism spectrum disorder	[Pokorny et al., 2017]
Alzheimer's disease	[Cummins et al., 2020]
Parkinson's disease	[Narendra et al., 2021]

Table 1: Table of speech-affecting conditions for which AI applications showed promising results in research, as summarized in [Costantini et al., 2023] and [Milling et al., 2022].

Consider Alzheimer's Disease (AD) as an example. Alzheimer's symptom progression is shown in Figure 3.

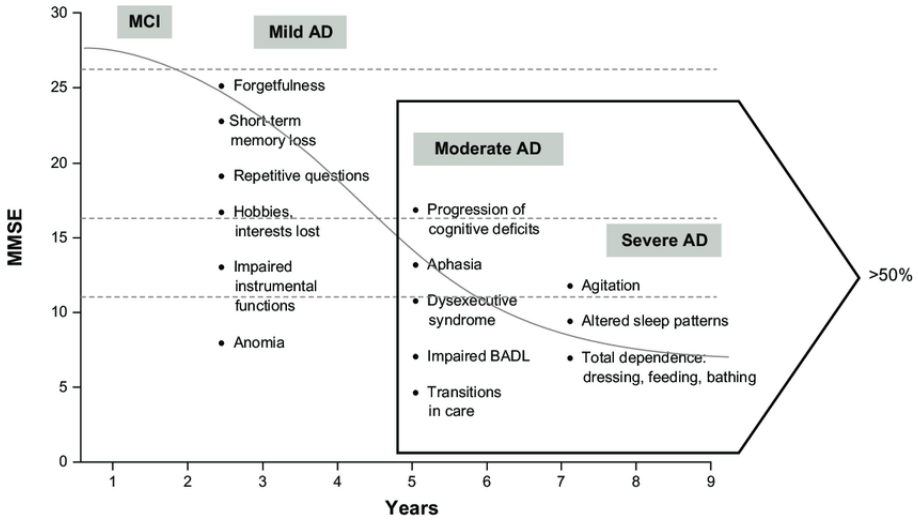


Figure 3: Alzheimer's symptom progression [Feldman and Woodward, 2005]

Aphasia is a language disorder which makes patients unable to communicate effectively with others. Even early on in the progression of AD aphasia could be observed. Therefore the criteria for the clinical diagnosis of AD also include language impairment [McKhann et al., 2011]. The moment of diagnosis is generally somewhere in the middle of Figure 3. The overall idea is to move this moment of diagnosis to the left in this figure using pre-trained self-supervised deep learning models.

Dysarthria, a motor speech disorder, can occur due to damage to the nerves that control the speech muscles, resulting from conditions such as Multiple sclerosis (MS), cerebral palsy (CP), or

amyotrophic lateral sclerosis (ALS). Patients suffering from either CP or ALS can be found in the TORGO database, which will be introduced subsequently as a practice dataset within this thesis. It is important to note that these disruptions can manifest in various ways, making it a challenging task for humans to detect and diagnose such conditions based solely on changes in speech. This underscores the significance of developing advanced techniques, such as the application of artificial intelligence, to analyze and interpret these subtle changes in speech patterns. This thesis' overall objective is to investigate how AI, in particular pre-trained self-supervised deep learning models, might help with the challenging problem of early disease recognition using voice analysis.

2.2 Traditional Speech Features

Traditional speech processing methods often involve the extraction of specific features from audio signals. These features provide a mathematical representation of the speech signal and can be used to identify patterns and inconsistencies. One of the most commonly used features in speech processing are the Mel-frequency cepstral coefficients (MFCCs), which will be discussed in the next subsection. But there are also others: pitch, tone, f1, f2, jitter, and shimmer. These features could be extracted in different ways, for example in Python, by making use of the Librosa library, or by using a computer program like PRAAT.

2.2.1 MFCC

Mel Frequency Cepstral Coefficients (MFCCs) are a type of handcrafted audio features that attempt to mimic the human auditory system. The human ear perceives different frequencies of sound in a non-linear manner. In particular, this means that humans are more sensitive to changes in lower frequencies than in higher frequencies. The calculation of MFCCs can be described as the following 7-step process:

1. **Frame the audio signal:** Audio files need to be split into frames of 20-40ms. The idea behind this step is that the frequencies in an audio signal may vary over time. If the following steps would be taken on an audio file of, for example, two seconds, the frequency contours of the signal over time would be lost. But, if taken on a shorter frame, for example, 10ms, there are not enough samples to get a reliable spectral estimate.

2. **Apply a window function:** A frame is just a chopped-off part from an audio file. However, when just cut off at certain points, there will be sudden shifts in amplitude at the edges of the frame. The step that follows consequently, taking the Discrete Fourier-Transform (DFT), expects that the data is infinite, not finite. Therefore, the amplitude should gradually drop off near the edges of a frame, instead of suddenly dropping. The window function does this, by maintaining the values in the middle of the frames but decreasing gradually to zero towards the edges.

3. **Apply the Discrete Fourier Transform (DFT):** Apply a Fourier Transform on each windowed frame to calculate the frequency spectrum. This operation transforms the signal from the time domain, i.e. the short windowed frames, to the frequency domain. Here, the different present frequencies and their magnitude could be observed for that specific time frame.

4. **Square DFT output :** Square the output of the DFT, to get the DFT power spectrum.

5. **Apply the Mel filter bank:** The Mel scale is constructed after research on human sound perceiving. It tries to mimic the non-linear frequency response of the hairs in human ears to the vibration of sound. In particular, humans are better at distinguishing small changes in pitch at low frequencies than they are at high frequencies.

   The formula for converting from frequency to Mel scale is:

   $$M(f) = 1125ln(1 + f/700)$$

   And to go from Mel scale back to frequency:

   $$M - 1(m) = 700(exp(m/1125) - 1)$$

   The Mel scale is used to create the Mel filter bank, a series of triangular band-pass filters. After applying this filter bank, the energy from a number of frequency bands that are covered will be presented.

6. **Take the logarithm:** Take the log of the output of the Mel filter bank.

7. **Discrete Cosine Transform (DCT):** Take the Discrete Cosine Transform of the log filter bank energies to receive the Mel-frequency cepstral coefficients.

MFCCs have been found to be very effective for many speech and audio analysis tasks, including speech recognition and speaker identification. Therefore, within this thesis, MFCCs will be used to compare against the latent features obtained with deep learning, which will be described subsequently.

2.3 Pre-trained Self-Supervised Deep Learning

Pre-trained self-supervised deep learning models are a relatively new approach within the field of artificial intelligence. These models have gained significant traction due to their effectiveness across a variety of tasks. Examples of models include all OpenAI's GPT versions [Brown et al., 2020], Google's BERT [Devlin et al., 2018], and Facebook's wav2vec [Baevski et al., 2020]. They have proven to be incredibly effective at tasks like speech recognition and natural language interpretation. The mentioned models were deployed by the largest companies in the field, which also shows the relevance of the topic. The individual parts of the overarching concept, pre-training, self-supervision, and deep learning, will be explained more in-depth in the following sections.

2.3.1 Deep Learning

Deep learning is the subdomain of machine learning that concerns deep neural networks (DNNs). A neural network (NN), without specifically being deep, is a structure of connected nodes, or "neurons", that are inspired by the biological neurons found in the human brain. Similar to how real neurons receive, process, and send information, each neuron in an artificial neural receives input, processes it, and forwards the output to the next neuron. A neural network could learn from experience by adjusting its parameters, which is comparable to the strengthening or weakening of synapses in the brain during a learning process.

A parameter is a broader term in machine learning that includes both weights and biases. In the context of neural networks, each connection between neurons has an associated weight. Weights

are essentially the strength of the connection between two neurons, and can be thought of as the amount of influence one neuron has on another. Weights are adjusted during the learning process to optimize the prediction capability of the network. On the other hand, a bias is an additional parameter that allows the model to adjust its output independently of its input.

The artificial neurons are usually vertically represented into so-called layers. DNNs contain at least three of these layers. A classical example of a DNN structure is shown in Figure 4.

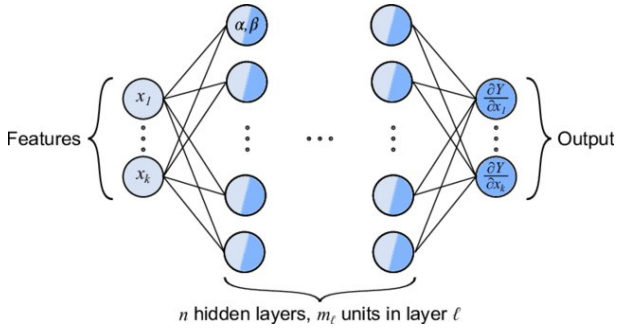


Figure 4: Deep Neural Network (DNN) [Teichert et al., 2020]

This multiple of layers is what makes DNNs "deep", it causes their ability to learn increasingly abstract patterns and representations in data. The first layer learns simple features, like edges in an image, but when moving deeper into the network, the features become more complex and abstract and could consist of color combinations or rare shapes.

DNNs receive data as input, which could essentially by anything represented by a number, so for example, images as pixel values, or text or audio represented in a numerical form. In this thesis the input consists of audio features, either the MFCCs or the deep learning features (wav2vec). Since these deep learning features do not represent some specific mathematical construction, as opposed to the traditional speech features described in section 2.2, they are sometimes called 'hidden' or 'latent' features.

A **Transformer** is a specific type of DNN. The architecture was introduced in the paper "Attention Is All You Need", by Vaswani et al. [Vaswani et al., 2017]. Transformers are especially unique in terms of their attention-mechanism and their capability for parallel processing. In Transformers, the "depth" comes from stacking multiple self-attention and feed-forward layers to form the encoder and the decoder, which could be observed in Figure 5. The encoder processes the input sequence and compresses it into a context vector. The decoder then takes this context vector and produces the output sequence. The Transformer is a type of **encoder-decoder** model, but with a specific attention-based architecture. The so-called self-attention enables to weigh the importance of different parts of the input sequence when making predictions, which is particularly useful for tasks requiring understanding of long-range dependencies in the data. Processing the inputs in a parallel manner, makes transformers very efficient for handling large sequences. Other architectures, like Recurrent Neural Networks, process sequences step-by-step and therefore require significant computational time for long sequences. Transformers can process all parts of a sequence at once. This characteristic makes Transformers well-suited for parallel computation, significantly speeding up the training process.

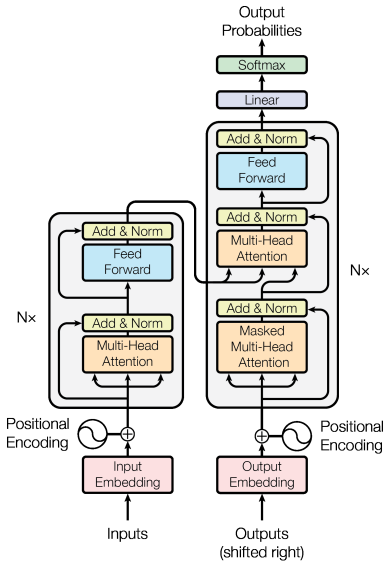


Figure 5: Transformmer network [Vaswani et al., 2017]

2.3.2 Self-Supervised Learning

Deep learning could be supervised, semi-supervised, or unsupervised and that depends on the training data. The training data for deep learning models can either be labeled or unlabeled, being labeled means that the data itself is accompanied by a certain explanation. Labels differ per task, but think of examples as images for an image recognition model along with an explanation of what can be seen on the image, like “cat”, “dog”, or “car”. For an automatic speech recognition model, the data audio could be accompanied with the transcription of the audio. For the problem of this thesis, however, the labels corresponding to the audio fragments could simply state that the speaking subject is healthy or suffers from a specific disease.

Self-supervised learning falls under the category of unsupervised learning. This means that the data is not accompanied with any labels. However, the algorithm is designed in such a way that it extracts labels from the data itself, reducing the need for extensive manual labeling. This is achieved by creating tasks where the model is trained to predict certain parts of the data given the other parts, called **masking**. For instance, in the context of language modeling, a word in a sentence could be masked, or removed. Then, the model is forced to predict or ”fill in” the masked part(s) based on the surrounding unmasked parts. For speech data this would come down to masking a certain frame of the audio, for example by replacing the numerical values in the frame with ‘0’, which means that the corresponding audio segment is silenced.. This approach to learning enables models to extract a valuable understanding of the structure and patterns in the data.

2.3.3 Pre-Training

Pre-training in this case, means training the deep learning model on a large corpus of data in a self-supervised manner. During the training phase of a DNN, the model learns the structure and patterns in the data by adjusting its parameters. These parameters are associated with connections

between neurons (nodes) and are adjusted to minimize the difference between their predictions and actual values (target labels). In this way, the classification performance of the model is being improved. The rationale is that the model can learn general patterns and representations from this extensive dataset, which may not be specific to any particular task. Following this, the model is adapted to more specific tasks via a process called "transfer learning", which involves fine-tuning the pre-trained model on a smaller, task-specific dataset. This two-step process allows models to leverage general knowledge from pre-training and apply it effectively to specific tasks. For instance, a language model could be pre-trained on a large corpus of text from the internet to learn the grammar, syntax, and semantics of a language. This approach enables models to learn rich representations of the data. The learned "knowledge" of these models can be transferred to other models, to solve more specific task. This could be done by fine-tuning the model on a smaller labeled for a specific task, like disease detection.

2.3.4 Fine-Tuning

Fine-tuning is a subsequent stage that follows pre-training. The learned representations from the previous section are adapted to a more specific task through fine-tuning, a process that involves further training the pre-trained model on a task-specific labeled dataset. Fine-tuning is another training process and therefore involves updating the models' parameters. The advantage of this approach is that it allows for the application of the knowledge gained during pre-training to the specific task, even if the task-specific dataset is relatively small. This process of adapting the knowledge of a previous model is also called **transfer learning** and is clarified in Figure 6

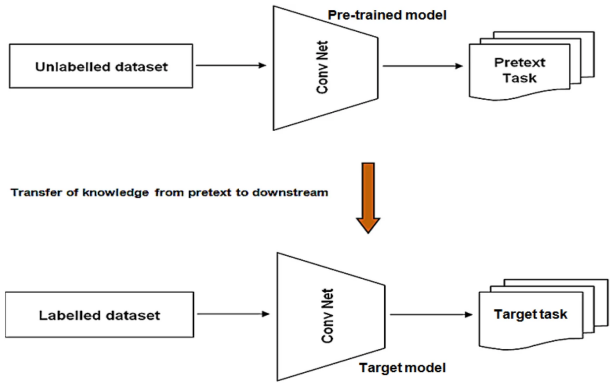


Figure 6: Transfer Learning

2.3.5 Wav2Vec

Wav2Vec and its sub-models, Wav2Vec 2.0 and XLSR-Wav2Vec, are pre-trained self-supervised deep learning models originally developed by Facebook AI [Baevski et al., 2020] to tackle the problem of automatic speech recognition, or, in other words, to recognize what is being said in any fragment of audio. In this thesis there will be a focus on the wav2vec 2.0 model.

Speech audio is a continuous signal, it captures many aspects of the recording with no clear segmentation, for example into units of words. Therefore, the model first applies a multilayer convolutional neural network (CNN) on the speech raw audio waveform to obtain latent audio

representations that are 25 ms long. Then, these representations are fed into a quantizer as well as a transformer. From a list of learned units, the quantizer selects a speech unit for the latent audio representation. Before being supplied into the transformer, half of the audio representations are masked. Information from the entire audio sequence is added by the transformer. Finally, the output of the transformer is used to solve a contrastive task. This task requires the model to identify the correct quantized speech units for the masked positions.

2.4 Research Gap

While there is extensive research done on the application of pre-trained self-supervised deep learning models for image and text related problems, the audio domain has seen less attention. Specifically working on ML audio problems in a healthcare setting could involve several challenges and constraint. Some of these challenges are very related. Te most relevant will be described in this section.

To start, speech signals are complex and vary widely between individuals. Speech signals are captured with an audio recording device, like a microphone. Microphones are fragile to capture unintentionally irrelevant environmental noise or to entail possible unwanted vibrations themselves. This complexity can make it challenging to compile datasets of great quality. A consideration could be made here. Either all training data could be recorded in the exact same environmental setting with the exact same kind of recording device. Then the recording of a new instance should be recorded with the exact same conditions for a model to make an optimal classification, because the training data did not prepare it to generalize well outside of these conditions. On the other hand, there could be a focus on acquiring as many recordings from different settings as possible. In that way a ML model would be more able to generalize well on different kinds of new, unseen audio data, most independent of recording environment or used microphone. One thing is central, which is to collect as much audio data as possible from as much different patients as possible in order for the ML model to learn features of a disease or disorder the best, in the training phase.

Secondly, there is a lack of standardized datasets. Unlike in the domains of image or text analysis, there are fewer large, standardized speech datasets available, let alone datasets concerning patient's health. This is due to various factors, including the data collection process. Patient speech data for healthcare applications often requires appointments in a controlled environment, like a hospital, and the involvement of medical professionals to instruct the patient and label the data. This is a process is particularly time-consuming and expensive. Besides that, the process is even slower due to the fact that patient data is highly confidential. In many countries, patient data is heavily protected under privacy laws (e.g., HIPAA in the United States, GDPR in the European Union). This means it can be quite challenging to obtain the necessary permissions to access and use or share this data, even for research purposes.

Lastly, while deep learning models have shown promise in some research settings, their application in clinical settings is still limited. Translating the results of research into practical tools that can be used by healthcare professionals remains a significant challenge.

These constraints underline the complexity and challenge of developing deep learning models in healthcare settings, specifically in the context of diagnosing patients through their speech. Nevertheless, this also highlights the potential impact and value of successful research in this area. Research in this area could potentially provide an efficient, effective, and non-invasive tool for early disease detection, which benefits patients, relatives and pressure on the health care system. Also, it could potentially partly eliminate the need for costly and time-consuming medical proce-

dures. Overall, despite the challenges, the promising opportunities provide a path to significant advancements in medical technology.

This thesis contributes to the understanding of machine learning methodologies in disease detection by comparing traditional MFCC feature-based classification against the more recent techniques using pre-trained self-supervised deep learning models like wav2vec. By that, obstacles could be addressed and advised on to overcome, the area of maximum potential improvement could be identified and it overall paves a way for more efficient implementations of related strategies.

3 Method and Approach

The start of this approach of this thesis consisted of finding a relevant research that performed disease detection in the way that matches this thesis objective, i.e., using a pre-trained self-supervised deep learning model, preferably wav2vec. There is a considerable amount of research published within this area, but not much that directly suited the intention of this thesis. Most obstructions come in the form of datasets that are being used, these are often hardly accessible or not of the desired format. Sometimes research focuses on applying ML techniques on spectrograms as images instead of raw audio or derived features. Or multi-model approaches were tested, which are out of the scope of this thesis.

During the research phase, it was discovered that most medical data sets are only available on request. Often institutions state that their datasets are open-source, i.e. free to use and easy accessible, unfortunately this is not really true. As a bachelor's student it is often not even possible to get access to institutional databases by yourself. Access would be possible through a request of an academically recognized supervisor. This does not necessarily have to be a problem. The supervisor of this thesis, Prof. dr. P.P.G. Boersma, was able to arrange access very quickly and conveniently. But it still makes the research more complicated and could potentially withhold others without the privilege of such a supervisor.

The TORGO database on the other hand, emerged as the most promising database for this study, partly due to its open-sourceness, by even being available in a pre-processed version on Kaggle, and partly due to the relevant speech audio it contains, namely of already labeled affected speech as well as unaffected speech. Therefore, the following article is most overlapping with the objective of this thesis: *'On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches'*, by Schu, Guilherme and Janbakhshi, Parvaneh and Kodrasi, Ina [Schu et al., 2023]. This article is about comparing the difference in results between disease classification on raw audio, the audio after applying a noise filter and the other way around, i.e. the audio without the speech, noise only. Therefore, one of the aspects of this paper, is fine-tuning and testing wav2vec on the TORGO database. Their implementation of wav2vec will be used as an example for this thesis.

This section provides a description of the research methodology employed in this study. It presents in detail how each step of the methodology is conducted. The primary focus of this research is to develop and compare the recent pre-trained self-supervised deep learning techniques for disease recognition to the traditional classification on the basis of Mel-frequency Cepstral Coefficients (MFCCs). To classify based on the MFCC features, two different classifiers will be tested, namely a multi-layer perceptron (MLP) and a 1-dimensional convolutional neural network (1D CNN). The methodology is based on a systematic approach for the three different methods. This includes data collection and pre-processing, feature extraction, training, and the evaluation of model performance. Hopefully, this section offers more than just transparency about the research process,

but also provides a foundation for future studies seeking to explore and build upon the work. The following subsections elaborate on each phase of the methodology.

An overlapping part of the method, is the overall workflow, which consisted of setting up Google Colab for coding in Python, Google Drive for data storage, and GitHub for version control. Google Colab comes with a number of pre-installed packages, such as numpy, scipy, pandas, tensorflow, and pytorch, which makes it specifically clever for machine learning research. In addition, Colab is providing free access to computing resources including GPUs. GPUs can perform multiple computations in parallel. This enables significantly faster training of ML models than on normal CPUs. When running a Jupyter notebook in Google Colab it is possible to upload data manually directly to the platform, however this data is only stored temporarily. Specifically, this uploaded data is retained only for the duration of a single continuous session, and will be automatically discarded if there are long pauses or interruptions. Uploading the data to Google Drive is one way to solve this issue, since it is possible to 'mount', or connect, your Google Colab notebook to your Google Drive in the beginning of each session, in this way you do not have to upload all data manually each session. Ultimately, saving notebooks in Google Drive is relatively standardized. However, to efficiently maintain version control and keep track of changes, it is advantageous to upload the final notebook to GitHub. Overall, this setup is ideal for effective version control, secure storage, and successful coding.

3.1 The TORGO database

This thesis focuses primarily on the dysarthric speech audio files from the TORGO database. However, as stated before, one of the objectives is to develop a robust methodology for the use of the described machine learning models, particularly the pre-trained self-supervised deep learning model, wav2vec. Hopefully, this paves a path for other researchers in this area to start developing with the knowledge of which model performs best, what the obstacles are and how to tackle them. Other researchers could apply these notebooks then to other databases. These other databases might have a wider range of applications for other disease or disorders that might affect speech. For this moment, the focus remains on leveraging the TORGO database to better understand the problem of using pre-trained self-supervised deep learning models for disease recognition.

The publicly available TORGO database is exuberantly described in [Rudzicz et al., 2012] [5]. It consists of male and female speakers with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). As a consequence of their disease, these individuals all suffer from dysarthria. Dysarthria is caused by disruptions in the neuro-motor interface. These disruptions bend the motor commands to the vocal articulators, resulting in an atypical and relatively incomprehensible speech in most cases [Kent, 2000]. For each of the individuals with dysarthria, there is a matched individual in the control group (except for one dysarthric male). The strucutre of the TORGO database is displayed in Figure 7.

[5] http://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html

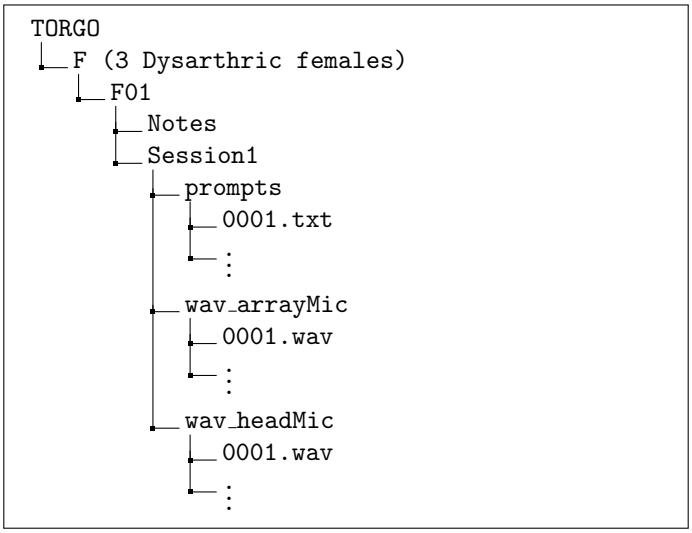


Figure 7: Structure of the TORGO database

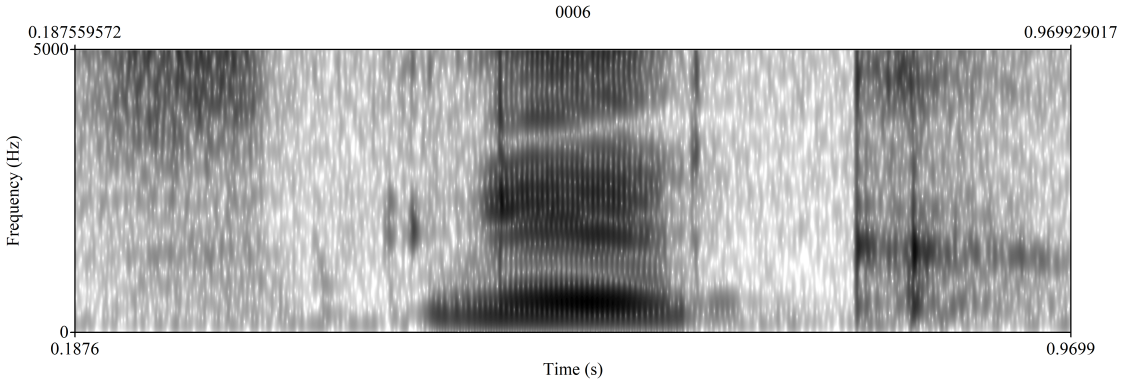
All the tasks, or speech tests, that the subjects had to perform are recorded with two separate microphones. These are called the 'head' and the 'array' microphones in the database. For this the method of this thesis, the pre-processed Kaggle version of the TORGO database will be used. [6]

3.2 Spectrogram Analysis

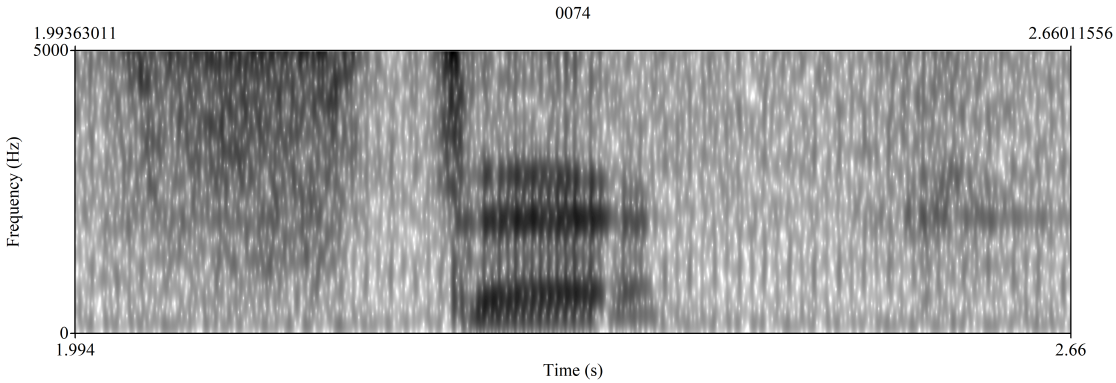
Following, a comparison between the spectrograms of a dysarthric speaker and a non-dysarthric speaker will be presented for five examples of speech tests, these include male-male and female-female comparisons. Specifically the speech test consists of pronouncing the word 'stick'. These individuals were carefully selected as a procedure of composing the TORGO database, to ensure they were comparable in personal factors such as age and gender. For this step, the original TORGO database is being used, instead of the already pre-processed one available on Kaggle. The reason for this is that the Kaggle version reduced the amount of audio files. This was done by removing duplicates, but without keeping track of which microphone was being used. Therefore this version may contain the same contents for different speakers, but the microphone could be different, for example the task might be to pronounce the word 'yes', and is performed by a female with dysarthria and her matched healthy control subject, but then one instance is recorded with the array-microphone and the other with head-microphone, which do substantially differ in quality. Since the original version contains all the available audio files from the two different microphones, this brought the opportunity of a more equal comparison. During this procedure, it was discovered that the TORGO database is not at all consistently structured. The 'prompts' sub-folder in each participant's main-folder contains text files with the prompts that were given to the participants which they were asked to speak out loud. These prompts were sometimes written differently, for example, one prompt was the following: "tear [as in "tear up that paper"]", while there also existed a ""tear [as in tear up that paper]"". Besides that, some tasks that some subjects performed, were not performed by others. So, overall, the database is not consistent.

[6] https://www.kaggle.com/datasets/iamhungundji/dysarthria-detection

At first, this caused the inability to find corresponding prompts for the matched speakers of the dysarthric and non-dysarthric groups. After writing a Python script to ignore these types of differences, as well as ignoring spaces and new lines in the prompts, some interesting comparable speech audio files were found. The word 'stick' will be used for all five comparison since this tasks was represented widely across all subject-folders and the recordings were clear. To retrieve the spectrograms, the computer program, PRAAT, was installed. The first step is to open the desired audio file, next you it is possible to select the relevant part of the recording with 'View & Edit', then 'Paint visible spectrogram' will show the spectrogram in another window, and finally an image of the spectrogram could be saved with the corresponding name as a 600 dpi png-file. The spectrograms provide a visual representation of frequency over time.



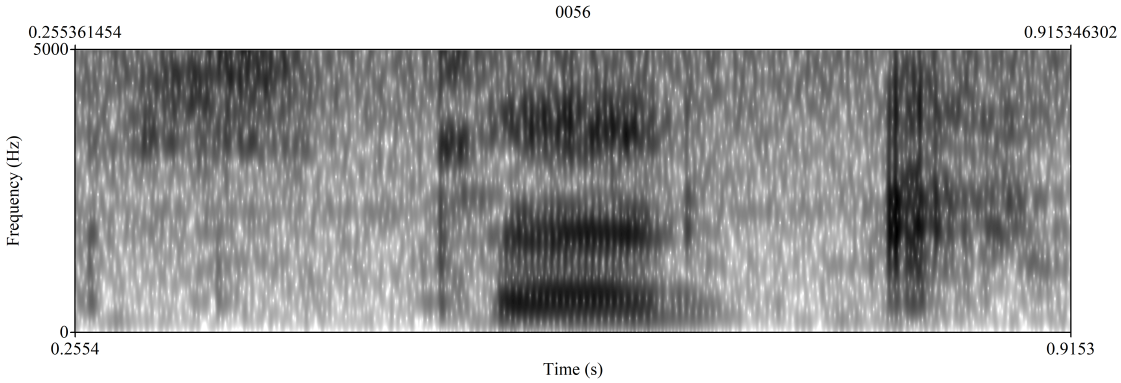
(a) F01-Session1-0006-head



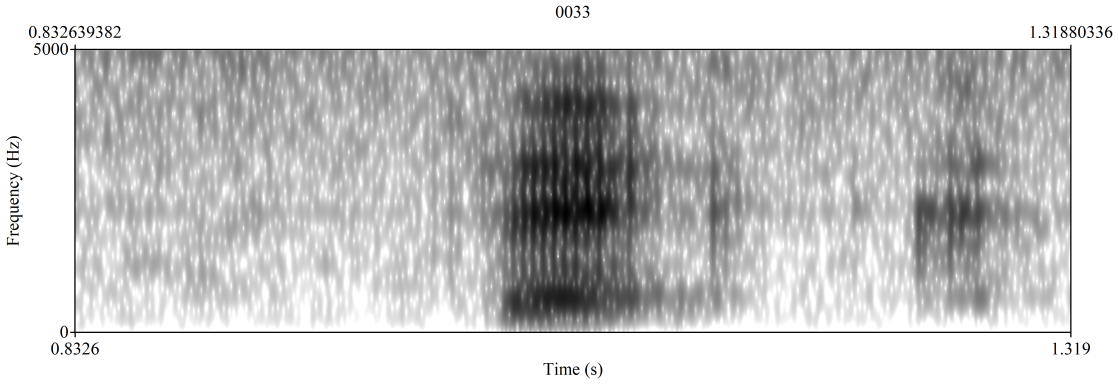
(b) FC01-Session1-0074-head

Figure 8: Spectrogram Comparison of Female Speakers: Dysarthric (F1, session 1, audio file 0006) vs. Healthy Control (FC1, session 1, audio file 0074)

The representation highlights greater frequency bands in the dysarthric speech (Figure 8a), indicative of dysfluencies. Also, the frequency bands look bended, which could be due to intonational patterns (rising or falling pitch) or other variations in the speaker's voice.



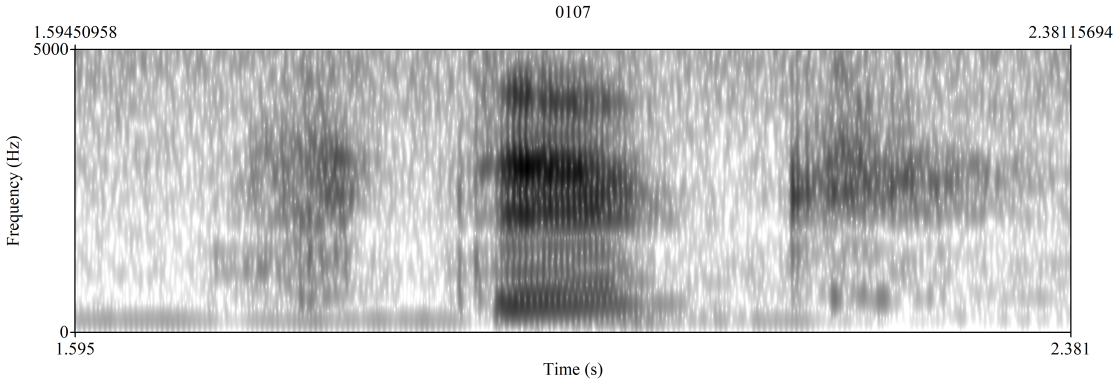
(a) F03-Session1-0056-array



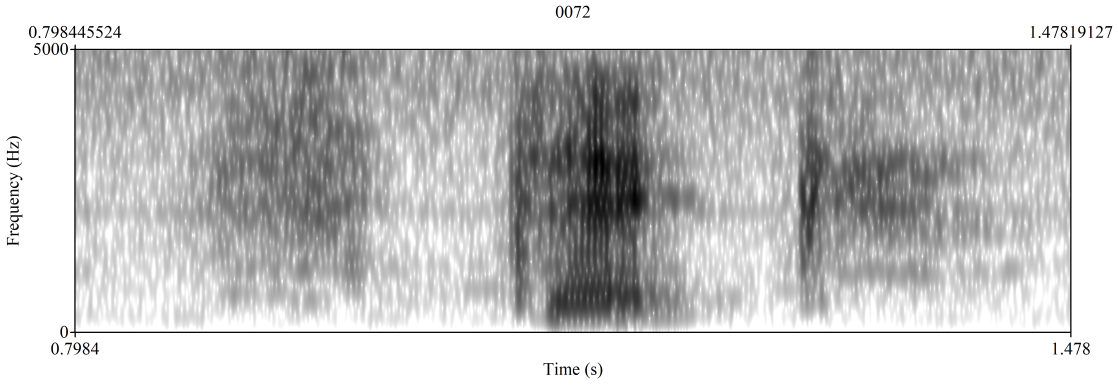
(b) FC02-Session2-0033-array

Figure 9: Spectrogram Comparison of Female Speakers: Dysarthric (F03, session 1, audio file 0056) vs. Healthy Control (FC02, session 2, audio file 0033)

In general, the spectrograms of dysarthric speech (Figure 9a) appear less smooth compared to the non-dysarthric speech (Figure 9b. This might be a result of the irregular and imprecise articulatory movements that are characteristic of dysarthria. These articulatory movements can lead to stronger variations in pitch and volume. Therefore, the frequency bands in the spectrogram may appear more rough or chaotic. When examining speech spectrograms, this fom of anti-smoothness might be a possible dysarthria sign.



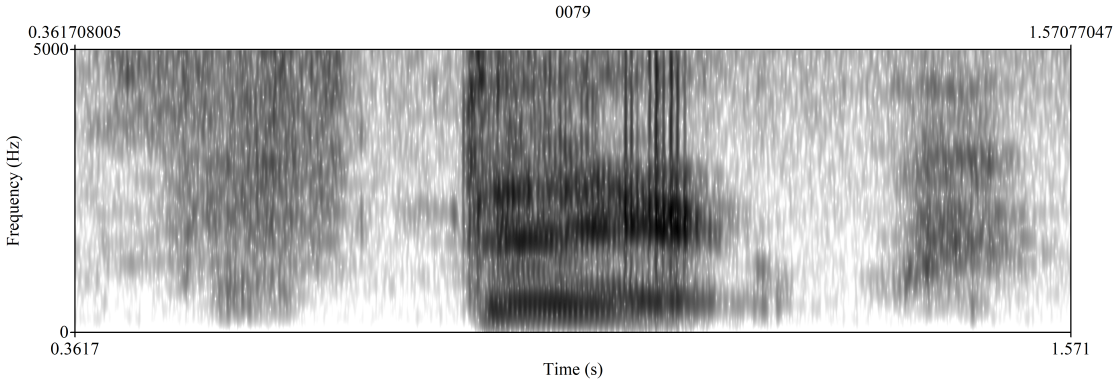
(a) F04-Session1-0107-array



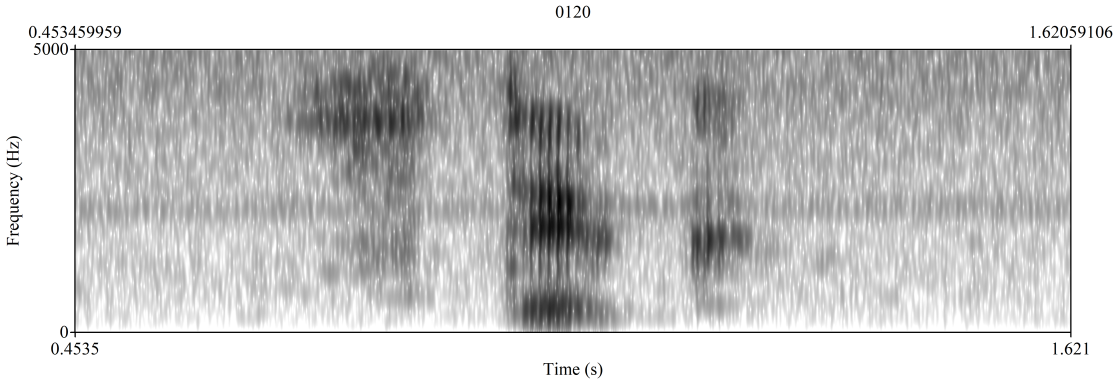
(b) FC03-Session1-0072-array

Figure 10: Spectrogram Comparison of Female Speakers: Dysarthric (F04, session 1, audio file 0107) vs. Healthy Control (FC03, session 1, audio file 0072)

This comparison illustrates distinct patterns in the spectrograms of a female speaker who is dysarthric and one who is not. The dysarthric spectrogram shows an identifiable horizontal black stripe towards the bottom that corresponds to a frequency band. On the non-dysarthric spectrogram, this is contrasted with a notably white, or silent, region in the same area. This can also be noticed in a horizontal line at the middle of the spectrogram, but then the other way around. There, the non-dysarthric speaker's spectrogram has a distinct black stripe, while the dysarthric speaker's spectrogram displays a white stripe. These differences could be illustrative of how individuals with and without dysarthria have different speaking patterns.



(a) M02-Session1-0079-array

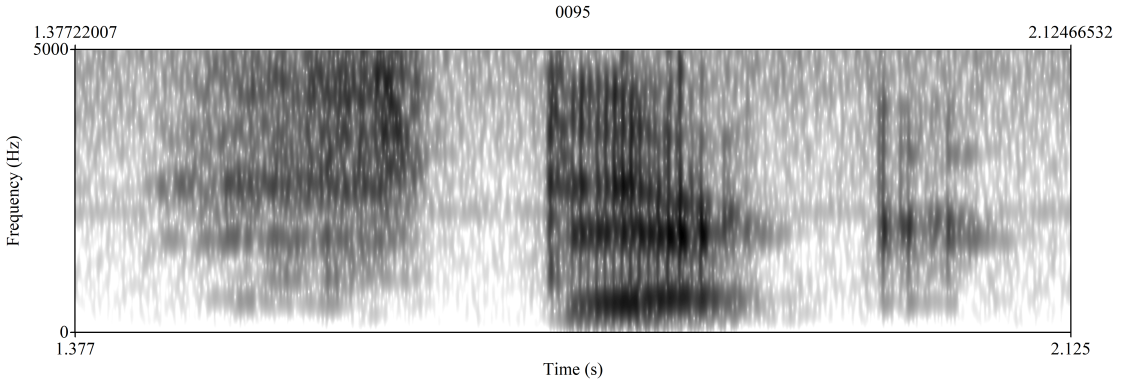


(b) MC02-Session1-0120-array

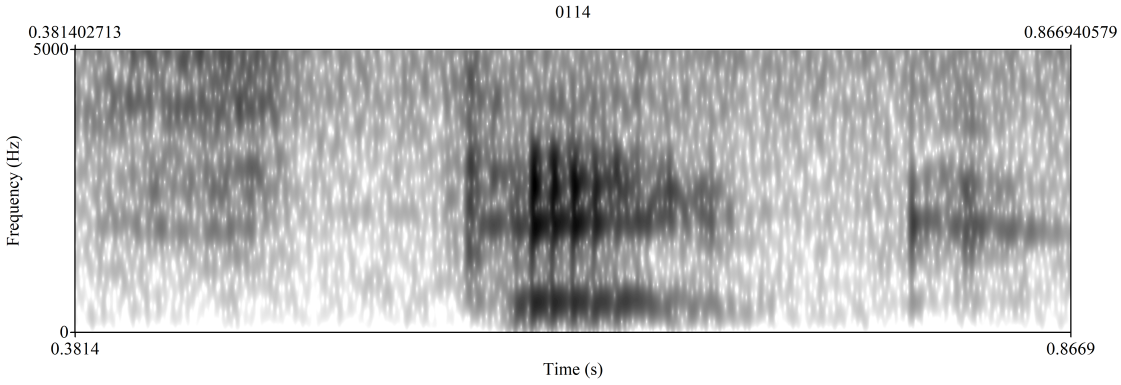
Figure 11: Spectrogram Comparison of Male Speakers: Dysarthric (M02, session 1, audio file 0079) vs. Healthy Control (MC02, session 1, audio file 0120)

11a and 11b show respectively 1,2093 and 1,1675 seconds of audio. As can be observed, 11a has more horizontally stretched vertical black lines. This presence of more stretched vertical lines in the dysarthric speech spectrogram could suggest elongated phonemes or words, which is a common characteristic of dysarthric speech. Dysarthria often results in slower, slurred, or stretched speech, as the speaker may have difficulty articulating certain sounds or controlling the muscles used in speech production.

On the other hand, the more centered and horizontally compact vertical lines in the non-dysarthric spectrogram could represent more distinct and rapid transitions between phonemes or words, which is more typical of normal, healthy speech.



(a) M03-Session1-0095-array



(b) MC03-Session1-0114-array

Figure 12: Spectrogram Comparison of Male Speakers: Dysarthric (M03, session 1, audio file 0095) vs. Healthy Control (MC03, session 1, audio file 0114)

During the process of creating the spectrogram, I listen to the audio files very well. In most samples, you hear a great difference between dysarthric speech and non-dysarthric speech (at least for this database), but within this concrete sample of dysarthric speech, I was not able to distinguish it from non-dysarthric speech. That would probably be the reason for not being able to see any notable deviations or differences in the two spectrograms in figure 12. precisely that is where the power of artificial intelligence could show itself.

3.3 Wav2vec Replication

As stated in Section 1, the research paper *'On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches'*, by Schu, Guilherme and Janbakhshi, Parvaneh and Kodrasi, Ina [Schu et al., 2023] will be used mainly as the basis for the wav2vec implementation of this thesis. However, the approach of replicating the wav2vec model was guided by a convergence of multiple resources rather than a single paper. The primary point of reference was the paper mentioned above, since it performed wav2vec on the TORGO database, but it did not

provide a comprehensive or clear methodology. The following research was discovered where they also implemented wav2vec, but then on a different dataset: *Comparing Acoustic-based Approaches for Alzheimer's Disease Detection* [Balagopalan and Novikova, 2021]. Despite the difference in the dataset, the methodology of this study offered insightful perspectives on the wav2vec implementation. This paper elaborated more on the method, but unfortunately it still was not exactly clear how wav2vec was implemented. Therefore contact has been made with the authors. One of the two responded quickly with some very helpful information about their implementation. It was indicated that their implementation was mainly based on the work of PhD student Mehrdad Farahani from Chalmers University of Technology, Gothenborg [7]. Mr Fahrahani created a notebook for recognizing emotions on Greek speech, which was published on GitHub. This notebook provided clear steps for that particular problem, which have been used the basis for the classification of dysarthria in this thesis. This approach consists of the following steps:

3.3.1 Downloading, Storing and Loading Data

The data was downloaded from Kaggle [8]. Afterwards, the folder was uploaded to Google Drive to be easily accessible when working in the Google Colab environment. In the Jupyter notebook like this:

```
# Get access to Google Drive
from google.colab import drive
drive.mount('/content/drive')
```

After 'mounting' to Google Drive accessing files that are stored there happens the same as if they were local.

3.3.2 Exploratory Data Analysis

This step starts with aligning the data in a Pandas DataFrame. Then parts of this DataFrame will be shown to get familiar with the kind of textual and numerical data. Next, one random sample will be played by using Librosa and IPython.display to get familiar with the kind of audio data.

3.3.3 Dataset Pre-processing

The pre-processing step may consists of many different sub-steps. One important step may be resampling the raw audio. Wav2vec expects 16000 kHz frequency as input, while some recording devices record at another frequency. Then files can be resampled with the use of Librosa. Secondly, it may be desired to redistribute files over different folders. The TORGO database that is used in this thesis distinguished between male and female speakers and between dysarthric and non-dysarthric. Since the goal is to recognize the presence or the absence of dysarthric features in the speech audio, it would be more efficient to redistribute the files into new folders, only distinguishing between health condition, and not on gender.

[7] https://github.com/m3hrdadfi/soxan/blob/main/notebooks/Emotion_recognition_in_Greek_speech_using_Wav2Vec2.ipynb

[8] https://www.kaggle.com/datasets/iamhungundji/dysarthria-detection

3.3.4 Dataset Splitting

After redistributing the files between two folders: dysarthric and non-dysarthric, in the previous step, it is now time to split the whole dataset into a train and test set. The dataset will be partitioned as follows, 80% of the data will be in the training set and 20% will be in the test set. This will be done using scikitlearn's train_test_split() function [9].

3.3.5 Fine-tuning Wav2vec

This process of transfer learning encompasses loading the Facebook's self-supervised pre-trained model, wav2ves 2.0, from HuggingFace [10]`https://huggingface.co/docs/transformers/model_doc/wav2vec2`. Now, the training happens in a supervised matter, since the TORGO data is labeled instead of the unlabeled data from the pre-training phase. For training the more complicated wav2vec model, it is suggested to switch to the use of GPU in Google Colab, to speed up the process.

3.3.6 Feature Extraction

Wav2Vec is capable of transforming raw audio samples into meaningful features. This step involves passing the audio data through the model to get a set of feature vectors.

3.3.7 Training the Classifier

The extracted features from the previous step are then used to train a classifier. While Wav2Vec can be fine-tuned end-to-end for a classification task, you might choose to train a separate classifier.

3.3.8 Evaluation of Model on Testset

After training, the model is evaluated on the test set to determine its performance. It's important to use various metrics for evaluation to get a comprehensive understanding of the model's performance.

3.4 Classification on MFCC Features

Since the results by themselves do not mean a lot, it is a logical procedure to compare it to the results of another model. Therefore the following two classifiers were developed for the classification on the traditional MFCC features.

3.4.1 Scikit-Learn's MLPClassifier

A practical and user-friendly multilayer perceptron (MLP) implementation is available in the scikit-learn library as MLPClassifier [11]. It has several advantageous features, for example: it automatically picks up biases, and supports a variety of activation functions. However, the MLPClassifier does have some disadvantages which may restrict its applicability in certain cases. Its lack of transparency and adaptability in comparison to a self-designed network is one of its primary limitations. When trying to create a network tailored to a given problem or dataset, the MLPClassifier's limited

[9] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

[10]

[11] `https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html`

customization possibilities might be an important obstacle. Another important disadvantage is that the MLPClassifier is designed to work with fully connected layers only. This design choice can make it inefficient for dealing with certain types of data, where local patterns and spatial hierarchies are important. In contrast, a custom-designed convolutional neural network offers far greater flexibility and control over the architecture of the model. This makes it possible to tailor the model more closely to the specific characteristics and requirements of the problem at hand, which can potentially lead to better performance. Given these factors, it is an effective first-try classifier, but in the end, the use of a custom-designed CNN in the next section would be preferred over the MLPClassifier for the development of an MFCC-based classifier.

3.4.2 1D CNN

A 1D Convolutional Neural Network (CNN) will be implemented using PyTorch [12]. 1D CNNs could be quite effective in processing time series data, and audio signals are essentially a type of time series. In this implementation, the network architecture is defined within the 'Net' class, which is a subclass of the 'nn.Module' class, a base class for all neural network modules in PyTorch. The network consists of two convolutional layers ('self.conv1' and 'self.conv2'), each followed by a Rectified Linear Unit (ReLU) activation function and a max pooling operation. The first convolutional layer 'self.conv1' has 30 input channels (which corresponds to the dimensionality of the input data) and 32 output channels. The second convolutional layer 'self.conv2' has 32 input channels (corresponding to the output of the previous layer) and 64 output channels. Both convolutional layers use a kernel size of 3. After the second pooling operation, the data is reshaped, or flattened, to fit into a fully connected layer ('self.fc1'), which is used to map the features extracted by the convolutional layers to the final output classes. The fully connected layer is defined with an input size of 384 (which should match the total number of elements in the flattened output from the previous layer) and an output size of 2, corresponding to the two classes in the classification task. A dropout layer ('self.dropout') is included before the fully connected layer to help prevent overfitting by randomly setting a fraction (p=0.5) of input units to 0 at each update during training time. The forward propagation logic, defined in the 'forward' method, applies these layers and operations in the correct sequence to the input data 'x', resulting in the final output.

4 Results and Evaluation

There are important differences to consider about which has the greatest negative impact: False Positives or False Negatives. E.g. is it preferred to classify someone as 'having a certain disease', and potentially treating like it, while the person does not have the disease, compared to the other way around? If both types of errors are equally important, the F1 score or AUC-ROC might be the preferred evaluation measurement.

[12] `https://pytorch.org`

Performance scores on metrics per model					
Model Name	Accuracy	Precision	Recall	F1-Score	AUC-ROC
WAV2VEC 2.0	86.25	83.49	90.55	86.87	-
MFCC + 1D CNN	**90.10**	**85.09**	**97.49**	**90.87**	90.87
MFCC + MLP	83.75	78.24	93.50	85.19	**93.79**

Table 2: Table of the three different models and their corresponding scores on the evaluation metrics ($\times 100$).

Due to the fact that the AUC-ROC metric is missing in the HuggingFace Evaluate package, it did not work out to compute that metric for the wav2vec model. Here the best performing model for each metric will be outlined. **Accuracy**: MFCC + 1D CNN model performs best with an accuracy of 90.10%. **Precision**: Again, the MFCC + 1D CNN model outperforms the others with a precision of 85.09%. **Recall**: MFCC + 1D CNN model wins with a recall score of 97.49%. **F1-Score**: MFCC + 1D CNN model continues to lead with an F1-Score of 90.87%. **AUC-ROC**: MFCC + MLP model tops the list with an AUC-ROC of 93.79%. As stated above, it is worth noting that the AUC-ROC for WAV2VEC 2.0 has not been provided.

Considering all these metrics, it appears that the MFCC + 1D CNN model generally performs the best across most of the evaluation metrics. The differences in performance between the models could be attributed to a variety of factors. First of all, MFCCs remain a well-established method for speech and audio analysis, capturing key frequency characteristics of the sound. These may be more beneficial for the task of disease detection than the representations learned by the Facebook's wav2vec 2.0, which may be more generalized for broader speech recognition tasks.

Secondly, 1D CNN models have been shown to be highly effective in processing sequential data, like audio, capturing local dependencies in the data. This might be contributing to its superior performance over the MLP model which does not inherently capture these temporal dependencies.

Lastly, the performance of WAV2VEC 2.0 might be limited by the amount and nature of fine-tuning performed. As a pre-trained model, it's critical that it is adequately adapted to the specific task at hand. A self designed classification method on top of the wav2vec model could make a difference here.

Overall, while wav2vec 2.0 is a state-of-the-art model for audio processing, these results emphasize that simpler, traditional approaches like MFCC coupled with 1D CNN can still be highly competitive, if not superior. Especially when tailored to a specific task such as disease recognition from speech.

4.1 Metrics

Metrics are used in machine learning to measure the performance of models, providing a quantitative method to compare different models or different configurations of the same model. The outcomes of the metrics should provide insight into a model's performance and potential areas for improvement. The choice of metric often depends on the task of the model or the type of data. A majority of these metrics utilize the concepts of true positives, false positives, true negatives, and false negatives. In the context of this thesis, being labeled 'positive' corresponds to belonging to the 'dysarthric' class, while being labeled 'negative' corresponds to the 'non-dysarthric', or 'healthy control' group. **True Positives (TP)** are the cases where the model correctly predicts the positive class. Dysarthric audio is being identified as dysarthric. **False Positives (FP)** are the cases where the model

incorrectly predicts the positive class. So, while audio was actually from the healthy control group, the model predicted it as being dysarthric. **True Negatives (TN)** are the cases where the model correctly predicts the negative class. Non-dysarthric is being classified as non-dysarthric, i.e. the healthy control group. **False Negatives (FN)** are the cases where the model incorrectly predicts the negative class. In this scenario, the actual audio came from a dysarthric person, but the model predicted it as being from a healthy person.

4.1.1 Accuracy

This is one of the most straightforward classification metrics. It is the ratio between the number of correct predictions and the total number of predictions and the formula is the following:

$$Acccuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

While accuracy is easy to understand and interpret, it may not be the best metric if the classes are imbalanced. After pre-processing the data, to receive 1 second of usable audio from each file, the two classes are not anymore equally represented. The Dysarthric-class contains 970 audio segments, while the Non-dysarthric class contains 997 audio segments, so it is not significantly unbalanced. However, other metrics will be used as well for the evaluation, especially since false negatives are important in this context.

4.1.2 Precision

Precision measures the proportion of positive predictions that were actually correct. Low precision indicates a high number of false positives. This is the formula:

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

4.1.3 Recall

Recall is in other words, the ability of a classifier to detect all the positive samples. It is represented by this formula:

$$Recall = \frac{TP}{(TP + FN)}$$

Low recall indicates a high number of false negatives. Which could be specifically dangerous when deploying a ML model in a healthcare setting, it can withhold someone from early treatment with with many possible consequences.

4.1.4 F1-Score

The F1 score is the harmonic mean of precision and recall. It measures the balance of precision and recall and is particularly useful in the case of imbalanced datasets. It can be calculated as follows:

$$F1 = 2 \times (\frac{Precision \times Recall}{Precision + Recall})$$

An F1 score reaches its best value at 1 (perfect precision and recall) and its worst at 0.

4.1.5 AUC-ROC

Receiver Operator Characteristic (ROC) curve is being used as a evaluation metrics for binary classification problems. The Area Under this Curve (AUC) of the ROC can be described as a single number that shows the performance of a classifier over all possible classification thresholds. A model whose predictions are 100% correct will have an AUC of 1, while a model that makes random predictions will have an AUC of 0.5.

5 Discussion

In the preceding chapters, various machine learning models were developed and evaluated. This demonstrated their application to disease recognition, in the specific form of dysarthria classification, through speech data. In this section these empirical findings form the basis of a broader perspective. Herein, several related aspects to this work will be discussed.

5.1 Difference in Results

The papers that were used as a basis for the implementations in this thesis showed overall lower results on (partly) the same metrics. This could be due to several reasons. The most probable will be discussed here.

The choice of frame size when computing MFCCs could have a significant impact on the resulting features. A frame size of 500ms is mentioned in the article *'On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches'*. This is quite long and could potentially lead to a loss of important temporal detail in the audio signal. This is because MFCCs are a kind of 'summary' of the spectral content of a frame. If the frame is too long, it would average out more detail. For the purpose of this thesis it was considered more optimal to let go of this aspect of the replication and instead choose for a frame size of 25ms.

In *'On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches'* classifying they state something about classifying on MFCCs as well as classifying on spectrograms. Using a spectrogram as input to a classifier is a valid approach, but it may not be as effective as using raw audio (wav2vec) or MFCCs. Since their procedure does not clearly state their approach, they might be performing the classification on spectrograms instead of MFCCs. This could be a factor of lower results.

In *Comparing Acoustic-based Approaches for Alzheimer's Disease Detection* they use cross-validation. Cross-validation is a technique for evaluating a model's ability to generalize to unseen data. This paper are uses both the leave-one-speaker-out method and the 10-fold cross-validation method separately. This means that they are testing their models on data from speakers that the models have not seen at all during the training phase. This is a stricter evaluation method and can often lead to lower performance scores. Since the methods described in this thesis are not using a similar cross-validation technique, they might be benefiting from having seen data from all speakers during training, leading to higher scores.

Lastly, there could be other yet unknown differences in implementation. Therefore, it would be beneficial if the other methods were published on one of the previously mentioned platforms to find out what the real differences are.

5.2 Real-World Application

5.3 Ethics

When applying AI technologies ethical considerations should always be taken into account. This is especially important for developing and deploying machine learning (ML) models in healthcare settings. The ethical considerations arise in almost all different stages. For example, critical considerations arise already when gathering the data, but most importantly when implementing a ML model in real-life healthcare settings.

5.3.1 Bias

Another question that arises when gathering the data, concerns who or rather what kind of group the data should represent. Overall it would be desired to have a model performing well on subjects from all genders, ages and demographic backgrounds. Therefore, it is important to ensure the diversity between the individuals in the training dataset. Failure to do so can result in models that perform well only on certain demographic groups but poorly on others, escalating health inequities.

Overall, AI models, including the deep learning models used in this study, are only as good as the data they are trained on.

5.3.2 Data importance and privacy

There are relatively few datasets available on patient speech, the ones existing are small and only consist of a few speakers and a few different settings (recording environment and specific recording device) which is not universal and therefore generalizable. There is a need for bigger datasets representing a greater amount of human populations with better quality of recordings. Otherwise there could be chosen to diagnose only with the use of constant specific conditions. ML models could then be fine-tuned on specific regions or age categories, instead of being universal, but this is all sub-optimal.

The high confidential status of medical data stands in the way of these desired datasets. Solutions should be searched in anonymizing the patient data. Still, patients should be exstensively informed about how their data is being used for transparency. The advancements in healthcare through the use of ML models would be enormous when all patient data could be used.

5.4 Future Work

The results of this study open up several avenues for future research. Firstly, the performance of the models can be further improved by incorporating larger and more diverse datasets. The creation of these datasets could play a big role in the process of improving AI in healthcare. This would not only improve the generalizability of the models but also their ability to handle a wider range of voice disorders. In addition, future studies could explore the use of other types of features or different deep learning architectures to see if they offer improved performance. This field is very new and is improving at a vast rate. Also, the application of these models to other voice disorders and their potential for predicting disease progression or response to treatment should be explored. Lastly, future work should also focus on translating these models into practical tools that can be used in clinical settings or for example apps. This will involve addressing challenges related to user-interface design, data security, and integration with existing healthcare systems. Then the pros and

cons of deploying such an application should be considered. An example of this is the mole checker which causes many false positives to visit a doctor, increasing the pressure on healthcare.

6 Conclusion

In conclusion, this thesis investigated the efficiency of pre-trained self-supervised deep learning models, specifically wav2vec 2.0, for recognizing diseases from speech. The performance has been compared with the more traditional ML method of extracting MFCCs and classifying accordingly with a 1D CNN and a MLP, all using the TORGO database on dysarthria. The study was driven by the primary research question, "How effective are pre-trained self-supervised deep learning models in recognizing diseases or disorders, comparing to traditional ML methods on the TORGO database on dysarthria?"

Through experimentation and evaluation, the results indicated that traditional methods, especially when using Mel Frequency Cepstral Coefficients (MFCCs) as input features to a 1D Convolutional Neural Network (CNN), performed best across the majority of evaluation metrics. These metrics included accuracy, precision, recall, and F1-score. Specifically, the MFCC + 1D CNN model achieved high scores: an accuracy of 90.10%; a precision of 85.09%; a recall of 97.49%; and an F1-Score of 90.87%. When considering the AUC-ROC metric, which is particularly important when dealing with imbalanced classes, so not essentially for the used TORGO database, the MFCC + MLP model performed best with a score of 93.79%.

Contrarily, the wav2vec 2.0 model, despite its status as a state-of-the-art solution in the domain of audio processing, performed less in this particular task. This outcome underscores the importance of problem-specific model selection and highlights that more complex, advanced models do not always outperform simpler, established techniques. It also raises a question on the extent of fine-tuning required for these pre-trained models to adapt to specialized tasks such as disease recognition from speech.

Furthermore, this thesis extensively elaborated on relevant points of discussion, differences between results in papers, possible directions for future work and ethical considerations. The ethical considerations of deploying ML models in a healthcare setting, underlining the significance of data privacy, potential biases, and the need for transparency and interpretability.

The findings of this thesis offer valuable insights into the domain of disease recognition from speech using machine learning models. However, it also reveals that further research is needed to release the power of self-supervised deep learning models for these specialized tasks. Potential future improvements might involve a more extensive fine-tuning process, or integrating the strengths of traditional methods, such as MFCC feature extraction, with these advanced models in a multi-model way.

In essence, while pre-trained self-supervised models promise significant advancements in the field of speech-based disease recognition, their advantage in disease recognition compared to traditional methods still has to be shown. However, these models, backed by the big AI companies behind them, are improving them at a vast rate and, as stated, a lot of individuals are working on these models in a collaborative manner. These facts show a extremely promising future of AI in healthcare.

References

[Ahmed et al., 2013] Ahmed, S., Haigh, A.-M. F., de Jager, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. *Brain*, 136(12):3727–3737.

[Albes et al., 2020] Albes, M., Ren, Z., Schuller, B., and Cummins, N. (2020). Squeeze for sneeze: Compact neural networks for cold and flu recognition. *INTERSPEECH 2020*.

[Baevski et al., 2020] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

[Baird et al., 2020] Baird, A., Cummins, N., Schnieder, S., Krajewski, J., and Schuller, B. (2020). An evaluation of the effect of anxiety on speech-computational prediction of anxiety from sustained vowels. *INTERSPEECH 2020*.

[Balagopalan and Novikova, 2021] Balagopalan, A. and Novikova, J. (2021). Comparing acoustic-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2106.01555*.

[Braun et al., 2022] Braun, F., Erzigkeit, A., Lehfeld, H., Hillemacher, T., Riedhammer, K., and Bayerl, S. P. (2022). Going beyond the cookie theft picture test: Detecting cognitive impairments using acoustic features. In *Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings*, pages 437–448. Springer.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[BT et al., 2020] BT, B., Hee, H. I., Teoh, O., Lee, K., Kapoor, S., Herremans, D., and Chen, J.-M. (2020). Asthmatic versus healthy child classification based on cough and vocalised/ɑ:/sounds. *The Journal of the Acoustical Society of America*, 148(3):EL253–EL259.

[Cesarini et al., 2021] Cesarini, V., Casiddu, N., Porfirione, C., Massazza, G., Saggio, G., and Costantini, G. (2021). A machine learning-based voice analysis for the detection of dysphagia biomarkers. In *2021 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4. 0&IoT)*, pages 407–411. IEEE.

[Connaghan et al., 2021] Connaghan, K. P., Wertheim, C., Laures-Gore, J. S., Russell, S., and Patel, R. (2021). An exploratory study of student, speech–language pathologist and emergency worker impressions of speakers with dysarthria. *International Journal of Speech-Language Pathology*, 23(3):265–274.

[Cooper and Greene, 2005] Cooper, S. and Greene, J. (2005). The clinical assessment of the patient with early dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 5):v15–v24.

[Costantini et al., 2023] Costantini, G., Cesarini, V., Di Leo, P., Amato, F., Suppa, A., Asci, F., Pisani, A., Calculli, A., and Saggio, G. (2023). Artificial intelligence-based voice assessment of patients with parkinson's disease off and on treatment: Machine vs. deep-learning comparison. *Sensors*, 23(4):2293.

[Costantini et al., 2021] Costantini, G., Di Leo, P., Asci, F., Zarezadeh, Z., Marsili, L., Errico, V., Suppa, A., and Saggio, G. (2021). Machine learning based voice analysis in spasmodic dysphonia: An investigation of most relevant features from specific vocal tasks. In *BIOSIGNALS*, pages 103–113.

[Costantini et al., 2022] Costantini, G., Robotti, C., Benazzo, M., Pietrantonio, F., Di Girolamo, S., Pisani, A., Canzi, P., Mauramati, S., Bertino, G., Cassaniti, I., et al. (2022). Deep learning and machine learning-based voice analysis for the detection of covid-19: A proposal and comparison of architectures. *Knowledge-Based Systems*, 253:109539.

[Cummins et al., 2020] Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V. S., Christensen, H., Blackburn, D., Schuller, B. W., Magimai-Doss, M., Strik, H., et al. (2020). A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition. In *Interspeech 2020*, pages 2182–2186. ISCA-International Speech Communication Association.

[Darley et al., 1969] Darley, F. L., Aronson, A. E., and Brown, J. R. (1969). Differential diagnostic patterns of dysarthria. *Journal of speech and hearing research*, 12(2):246–269.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Enderby, 2013] Enderby, P. (2013). Disorders of communication: dysarthria. *Handbook of clinical neurology*, 110:273–281.

[Feldman and Woodward, 2005] Feldman, H. and Woodward, M. (2005). The staging and assessment of moderate to severe alzheimer disease. *Neurology*, 65(6 suppl 3):S10–S17.

[Gómez-García et al., 2019] Gómez-García, J. A., Moro-Velázquez, L., and Godino-Llorente, J. I. (2019). On the design of automatic voice condition analysis systems. part i: Review of concepts and an insight to the state of the art. *Biomedical Signal Processing and Control*, 51:181–199.

[Han et al., 2020] Han, J., Qian, K., Song, M., Yang, Z., Ren, Z., Liu, S., Liu, J., Zheng, H., Ji, W., Koike, T., et al. (2020). An early study on intelligent analysis of speech under covid-19: Severity, sleep quality, fatigue, and anxiety. *arXiv preprint arXiv:2005.00096*.

[Jeancolas et al., 2019] Jeancolas, L., Mangone, G., Corvol, J.-C., Vidailhet, M., Lehéricy, S., Benkelfat, B.-E., Benali, H., and Petrovska-Delacrétaz, D. (2019). Comparison of telephone recordings and professional microphone recordings for early detection of parkinson's disease, using mel-frequency cepstral coefficients with gaussian mixture models. In *INTERSPEECH 2019: 20th annual conference of the International Speech Communication Association*, pages 3033–3037. International Speech Communication Association (ISCA).

[Joshy and Rajan, 2022] Joshy, A. A. and Rajan, R. (2022). Automated dysarthria severity classification: A study on acoustic features and deep learning techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1147–1157.

[Kent, 1996] Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology*, 5(3):7–23.

[Kent, 2000] Kent, R. D. (2000). Research on speech motor control and its disorders: A review and prospective. *Journal of Communication disorders*, 33(5):391–428.

[Kodrasi et al., 2020] Kodrasi, I., Pernon, M., Laganaro, M., and Bourlard, H. (2020). Automatic discrimination of apraxia of speech and dysarthria using a minimalistic set of handcrafted features. In *INTERSPEECH*, pages 4991–4995.

[Kotsiantis et al., 2007] Kotsiantis, S. B., Zaharakis, I., Pintelas, P., et al. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24.

[López-Barroso et al., 2013] López-Barroso, D., Catani, M., Ripollés, P., Dell'Acqua, F., Rodríguez-Fornells, A., and de Diego-Balaguer, R. (2013). Word learning is mediated by the left arcuate fasciculus. *Proceedings of the National Academy of Sciences*, 110(32):13168–13173.

[Ma et al., 2020] Ma, A., Lau, K. K., and Thyagarajan, D. (2020). Voice changes in parkinson's disease: What are they telling us? *Journal of Clinical Neuroscience*, 72:1–7.

[McKhann et al., 2011] McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., et al. (2011). The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):263–269.

[Millet and Zeghidour, 2019] Millet, J. and Zeghidour, N. (2019). Learning to detect dysarthria from raw speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5831–5835. IEEE.

[Milling et al., 2022] Milling, M., Pokorny, F. B., Bartl-Pokorny, K. D., and Schuller, B. W. (2022). Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell. *Frontiers*.

[Narendra et al., 2021] Narendra, N., Schuller, B., and Alku, P. (2021). The detection of parkinson's disease from speech using voice source information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1925–1936.

[Pokorny et al., 2017] Pokorny, F. B., Schuller, B., Marschik, P. B., Brueckner, R., Nyström, P., Cummins, N., Bölte, S., Einspieler, C., and Falck-Ytter, T. (2017). Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach. *INTERSPEECH 2017*.

[Ramoo, 2021] Ramoo, D. (2021). *Psychology of Language*. BCcampus.

[Rejaibi et al., 2022] Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., and Othmani, A. (2022). Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71:103107.

[Ren et al., 2019] Ren, Z., Han, J., Cummins, N., Kong, Q., Plumbley, M. D., and Schuller, B. W. (2019). Multi-instance learning for bipolar disorder diagnosis using weakly labelled speech data. In *Proceedings of the 9th International Conference on Digital Public Health*, pages 79–83.

[Rudzicz et al., 2012] Rudzicz, F., Namasivayam, A. K., and Wolff, T. (2012). The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46:523–541.

[Schu et al., 2023] Schu, G., Janbakhshi, P., and Kodrasi, I. (2023). On using the ua-speech and torgo databases to validate automatic dysarthric speech classification approaches. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[Sonnenburg et al., 2007] Sonnenburg, S., Braun, M. L., Ong, C. S., Bengio, S., Bottou, L., Holmes, G., LeCunn, Y., Muller, K.-R., Pereira, F., Rasmussen, C. E., et al. (2007). The need for open source software in machine learning. *Journal of Machine Learning Research 8 (2007)*.

[Szatloczki et al., 2015] Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., and Pakaski, M. (2015). Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in aging neuroscience*, 7:195.

[Teichert et al., 2020] Teichert, G., Natarajan, A., Ven, A., and Garikipati, K. (2020). Scale bridging materials physics: Active learning workflows and integrable deep neural networks for free energy function representations in alloys. *npj Computational Materials*.

[Tiwari, 2010] Tiwari, V. (2010). Mfcc and its applications in speaker recognition. *International journal on emerging technologies*, 1(1):19–22.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Winslow, 2020] Winslow, T. (2020). Larynx anatomy, child.

[Yunusova et al., 2008] Yunusova, Y., Weismer, G., Westbury, J. R., and Lindstrom, M. J. (2008). Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research (JSLHR*.