Escuela
Politécnica
Superior

Deep Learning-based Cognitive Impairment Diseases Prediction and Assistance using Multimodal Data

Máster Universitario en Automática y Robótica

Trabajo Fin de Máster

Author:
David Ortiz Pérez
Supervisors:
David Tomás Díaz
José García Rodríguez

Universitat d'Alacant
Universidad de Alicante

Julio 2023

Deep Learning-based Cognitive Impairment Diseases Prediction and Assistance using Multimodal Data

Author
David Ortiz Pérez

Supervisors
David Tomás Díaz
Departamento de Lenguajes y Sistemas Informáticos
José García Rodríguez
Departamento de Tecnología Informática y Computación

Máster Universitario en Automática y Robótica

Escuela
Politécnica
Superior

Universitat d'Alacant
Universidad de Alicante

ALICANTE, Julio 2023

Abstract

In this project, we propose a mobile robot-based system capable of analyzing data from elderly people and patients with cognitive impairment diseases, such as aphasia or dementia. The project entails the deployment of two primary tasks that will be performed by the robot. The first task is the detection of these diseases in their early stages to initiate professional treatment, thereby improving the patient's quality of life. A second task focuses on automatic emotion detection, particularly during interactions with other people, in this case, clinicians. Additionally, the project aims to examine how the combination of different modalities, such as audio or text, can influence the model's results. Extensive research has been conducted on various dementia and aphasia datasets, as well as the implemented tasks. For this purpose, we utilized the DementiaBank and AphasiaBank datasets, which contain multimodal data in different formats, including video, audio, and audio transcriptions. We employed diverse models for the prediction task, including Convolutional Neural Networks for audio classification, Transformers for text classification, and a multimodal model combining both approaches. These models underwent testing on a separate test set, and the best results were achieved using the text modality, achieving a 90.36% accuracy in detecting dementia. Additionally, we conducted a detailed analysis of the available data to explain the obtained results and the model's explainability. The pipeline for automatic emotion recognition was evaluated by manually reviewing initial frames of one hundred randomly selected video samples from the dataset. This pipeline was also employed to recognize emotions in both healthy patients, and those with aphasia. The study revealed that individuals with aphasia express different emotional moods than healthy ones when listening to someone's speech, primarily due to their difficulties in understanding and expressing speech. Due to this, it negatively impacts their mood. Analyzing their emotional state can facilitate improved interactions by avoiding conversations that may have a negative impact on their mood, thus providing better assistance.

Resumen

En este proyecto proponemos un sistema integrado sobre un robot móvil capaz de analizar datos de personas mayores y pacientes con enfermedades de deterioro cognitivo, como afasia o demencia. El proyecto implica el despliegue de dos tareas principales que realizará el robot. La primera tarea es la detección de estas enfermedades en sus fases iniciales para iniciar un tratamiento profesional, mejorando así la calidad de vida del paciente. La otra tarea se centra en la detección automática de emociones, especialmente durante las interacciones con otros individuos, en este caso, los médicos. Además, el proyecto pretende examinar cómo la combinación de distintas modalidades, como audio o texto, puede influir en los resultados del modelo. Se ha llevado a cabo una amplia investigación sobre diversos conjuntos de datos de demencia y afasia, así como sobre las tareas implementadas. Para ello, utilizamos los datasets DementiaBank y AphasiaBank, que contienen datos multimodales en distintos formatos, como vídeo, audio y transcripciones de audio. Empleamos diversos modelos para la tarea de predicción de demencia, incluidas redes neuronales convolucionales para la clasificación de audio, transformers para la clasificación de texto y un modelo multimodal que combina ambos enfoques. Estos modelos se sometieron a pruebas en un conjunto de test separado, y los mejores resultados se obtuvieron utilizando la modalidad de texto, alcanzando una precisión del 90,36% en la detección de demencia. Además, realizamos un análisis detallado de los datos disponibles para explicar los resultados obtenidos y la explicabilidad del modelo. El pipeline para el reconocimiento automático de emociones se evaluó revisando manualmente los primeros fotogramas de cien muestras de vídeo seleccionadas aleatoriamente del dataset. Este pipeline también se empleó para reconocer emociones tanto en individuos sanos, como en individuos con afasia. El estudio reveló que los pacientes con afasia expresan estados emocionales diferentes a los sanos cuando están escuchando a alguien, debido principalmente a sus dificultades para comprender y expresar el habla. Debido a ello, su estado de ánimo se ve afectado negativamente. Analizar su estado emocional puede facilitar la mejora de las interacciones al evitar conversaciones que puedan tener un impacto negativo en su estado de ánimo, proporcionando así una mejor asistencia.

Agradecimientos

En primer lugar, querría agradecer el apoyo que he tenido por parte de mi familia y de mis amigos durante esta etapa académica. También me gustaría agradecer a mis tutores de este Trabajo de Fin de Máster, José García Rodríguez y David Tomás Díaz, por todos los conocimientos y consejos proporcionados durante este proyecto. Al igual que a todos los integrantes del 3D perception lab, los cuales siempre han estado para ayudar y aconsejar en este trabajo.

For once, I didn't look back.

Rick Riordan

Contents

List of Figures

List of Tables

1 Introduction

This first chapter aims to introduce the main topic of this thesis. The organization of this chapter is as follows: In Section 1.1, we provide an overview of the work carried out in this thesis. Section 1.2 explains the motivation behind this project. Our proposal and goals for this thesis are detailed in Section 1.3. Furthermore, Finally, in Section 1.4, we outline the structure of the remaining document.

1.1 Overview

The primary goal of this project is to implement an assistance system on a mobile robot to help elderly individuals and individuals with cognitive impairment diseases, such as dementia and aphasia. This system is designed to provide support to patients within the comfort of their own homes. The data collected and analyzed by the robot will be transmitted to a clinician, enabling them to monitor the patient's progress.

In this thesis, we have researched several tasks related to the prediction and analysis of cognitive impairment diseases with multimodal data. There has been research about the possible datasets where we could obtain data for this task, data about patients who are healthy or who suffer from these diseases, in this case, we have dementia and aphasia. One of our aims is to classify healthy and dementia patients, especially in the early stages of this disease, where the signs may not be as significant as in the later stages. Furthermore, we conducted a comprehensive analysis of the corpus data. Another aim has been to develop a system able to analyze the emotional mood of people with aphasia while interacting with other people. With this analysis, the robot will be able to identify situations where the patient may feel uncomfortable and attempt to improve their mood accordingly.

This work has been done from a multimodality point of view, looking for the combination of more than one modality, as can be done using text, audio, and video. For this purpose, the most suitable deep-learning techniques for these tasks have been revised, used, and tested. On the other hand, this work is a continuation from my previous Bachelor Thesis Ortiz Pérez (2022-06-30).

1.2 Motivation

The main motivation of this project is to provide a personalized assistance for elderly or dependent people with cognitive impairment diseases, improving their quality of life and having constant attention on their daily life to monitor any undesired or possible harmful behavior.

Nowadays, around 55 million people worldwide are diagnosed dementia, which is more commonly seen in older people but can also affect younger individuals. Dementia is a syndrome that affects the normal cognitive function of those who suffer from it. The most common form of dementia is Alzheimer's disease, which represents 60-70% of the cases [1]. This syndrome can affect each patient in a different way and has three different stages: early stage, middle stage, and late stage. Each stage can have different symptoms, such as losing track of time in the early stage, forgetting recent events or becoming confused at home in the middle stage, and finally, having difficulties recognizing relatives or friends in the late stage, among others. The late stage of dementia limits the individual's autonomy, requiring the support of a relative or a professional caregiver.

As mentioned before, dementia is commonly associated with older people. In recent years, the percentage of older individuals in society has been increasing. A clear sign of this aging trend is that in 2018, the number of people aged 65 years or older surpassed the number of children under five years for the first time in history [2]. Additionally, the percentage of the population over 65 years old is expected to nearly double by 2050. Consequently, the number of patients with dementia is expected to grow in the coming years. This is a problem that will become increasingly prevalent in our society.

The other cognitive impairment disease treated in this thesis is aphasia, which is a neurological disorder that occurs due to damage to certain regions of the brain involved in speech and language. This condition can cause significant communication difficulties for patients, making it challenging for them to express themselves clearly. In the United States, approximately one million people are affected by aphasia, and it is commonly associated with middle-aged and older individuals, although it can occur at any age.

The symptoms of aphasia primarily involve difficulties with language. There are various types of aphasia, determined by the specific location and extent of brain damage. The most prevalent types are Wernicke, Broca, and global aphasia. Wernicke aphasia is characterized by the use of nonsensical, long sentences and the invention of new words, making it challenging for patients to comprehend others' speech. In contrast, Broca aphasia results in patients using minimal words and constructing short, direct sentences, frequently omitting common words such as ”the,” ”and,” or ”is.” Global aphasia involves extensive brain damage and is associated with severe communication difficulties that limit patients' ability to both speak and comprehend others' speech. Other less common types of aphasia affect patients' communication abilities differently. Aphasia can result from various conditions such as strokes, brain tumors, or progressive neurological diseases like Alzheimer's disease, which is often linked to dementia National Institute of Mental Health (n.d.) Mayo Clinic (2022) Johns Hopkins Medicine (n.d.).

Another core motivation for exploring this topic in this thesis has been the collaboration with the Departamento de Tecnología Informática y Computación (DTIC) (DTIC in Spanish). In particular, the collaboration with the 3D Perception Lab group, whose main focus is on deep learning, GPU computing, and 3D computer vision. This group is actively involved in the Monitoring and Detection of human behaviors for personalized assistance and early disease detection (MoDeAsS) national project, led by Professor José García Rodríguez and Miguel Angel Cazorla-Quevedo. The MoDeAsS project aims to provide personalized assistance to dependent individuals, including those in the late stages of dementia. To achieve this, a model capable of identifying changes in their normal behavior has been proposed. This model can be used to detect anomalous behaviors or to identify diseases in the early

[1] https://www.who.int/news-room/fact-sheets/detail/dementia
[2] https://www.un.org/en/global-issues/ageing

stages, even in individuals who may appear healthy.

1.3 Proposal and goals

The main goal of this work is to develop a system capable of analyzing multimodal data from patients with cognitive impairment diseases. This system is intended to be integrated into a mobile assistance robot, which will be deployed in the patients' homes for monitoring and interaction purposes. The analysis aims to predict and identify patterns in these diseases for early treatment by professionals. In addition to prediction, understanding the emotional state of the patient is another objective of this thesis, as it enables tailored interaction with the patient based on their mood. This adaptive interaction approach can contribute to improving the patient's well-being when they are not feeling comfortable. The two main cognitive impairment diseases of focus in this work are dementia and aphasia.

Although there is no cure for dementia, treatments such as medication and therapy can help manage the symptoms. Therefore, early detection of this syndrome is crucial as it can significantly improve the quality of life for both patients and their relatives and friends (Fernández Montenegro et al., 2020)Gomez-Donoso et al. (2017)Revuelta et al. (2002). To achieve this goal, extensive research has been conducted on relevant dementia datasets, with priority given to multimodal datasets. The emphasis on multimodality is driven by the project's objective to investigate how combining multiple modalities can influence the performance of the models. The research in this project extends to other areas, including existing multimodal models and methods for text and audio classification. Following the research phase, several models have been proposed and will be tested to assess their performance. The experimental results will provide valuable insights and conclusions.

This study also aims to develop a pipeline capable of transcribing and distinguishing between patient and clinician recordings for further analysis of patients' facial expressions while listening to clinicians. The primary objective of this task is to analyze patients' emotions and identify patterns in aphasia disease, particularly examining how patients feel while listening to others, such as clinicians. Patients with aphasia may experience different moods due to difficulties in comprehending language. Analyzing their reactions and emotions can help improve communication with them, ultimately enhancing their comfort levels.

1.4 Outline

The structure of the remaining paper is organized as follows: Chapter 2 provides an introduction to the state of the art concerning the detection and assistance of cognitive impairment diseases. Chapter 3 explains the various materials and methods utilized in this project. The different approaches developed and tested will be discussed in Chapter 4. Chapter 5 presents the results obtained from testing these different approaches. Finally, in Chapter 6, we will discuss our final conclusions regarding this project.

2 State of the art

In this second chapter, we will introduce the state of the art in both of the proposed tasks that will be performed by the robot. We have previously discussed how our society tends to age. Due to this, nowadays we can find more and more works related to helping elderly and dependent people. This chapter is organized as follows: section 2.1 will explain the different datasets related to dementia and aphasia; section 2.2 will revise the different recent works over the DementiaBank dataset; section 2.3 will explain the different recent works over the DementiaBank dataset; section 2.4 will show the most interesting and novel multimodal models; section 2.5 will detail the field of natural language processing; section 2.6 will present recent approaches for the audio classification task; finally, in section 2.7 we will explain the most recent methods for face emotion recognition.

2.1 Cognitive impairment available datasets

As this is a highly sensitive topic, there are not too many options available for public use. There are some other options with other kinds of information, such as medical data, including blood test results or magnetic resonance image scans. However, we have focused on the ones which provide data that can be recorded at any time, like video, audio or text that can simply be a transcription of the audio. In the next subsections, the different examined datasets will be detailed.

2.1.1 Dementia datasets

In this subsection we will focus on the available datasets which contain information regarding people who suffer from dementia.

2.1.1.1 DementiaBank

DementiaBank (Becker et al., 1994) is a multimodal dataset which includes different corpora in different languages. Among the most used languages we can find English, Spanish, Mandarin, German and Taiwanese, each one including their different corpora with different data. The most interesting one and the one which we will be focusing on is the Pitt Corpus, which is in English. This corpus has been quite used in different works over the years, as we will see in the next section 2.2. Furthermore, this corpus has a smaller dataset, that has been balanced in terms of age and gender called the ADReSS challenge (Luz et al., 2020) that also have many recent approaches, but we will focus on the Pitt Corpus one. contains the audio of the recording as well as a transcription of the dialogue between the interviewer and the patients. The range of age of the patients goes from 46 to 90 years, including patients from both genders. The statistics about the number of healthy and dementia affected patients including the number of samples of each one can be seen in the Table 2.1.

Subjects were asked to describe an image shown to them. Specifically, the image used was the Cookie-Theft picture shown in Figure 2.1.

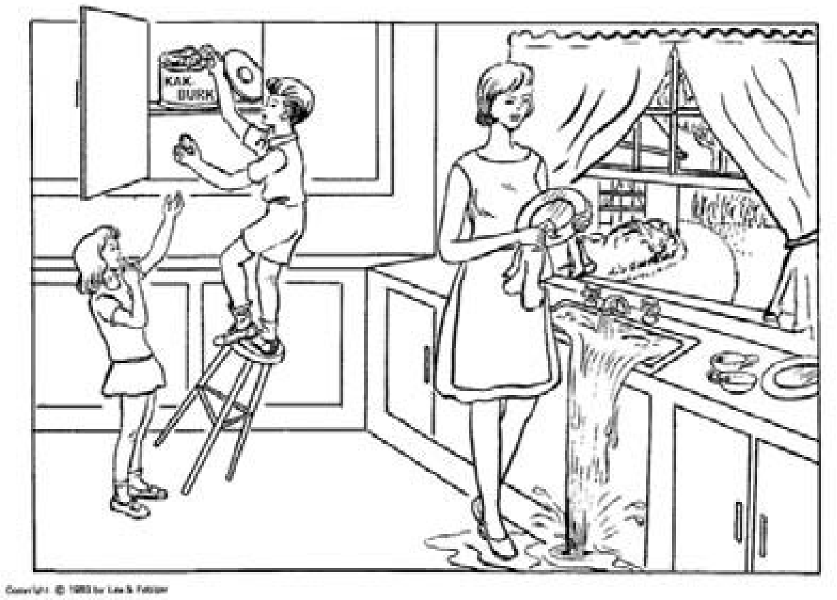


Figure 2.1: The Cookie Theft Picture, from Kokkinakis et al. (2017)

	Dementia	Control
Number of patients	194	99
Number of samples	309	243

Table 2.1: Pitt Corpus Statistics

This image has been used in clinical and experimental research, specifically in the field of mental and cognitive impairments. This experiment was designed to detect some of the signs of dementia, such as having difficulties choosing the right words, choosing wrong ones, using related or substitute words or even not finding a word at all. Other signs shown include using words with no meaning or not related to the conversation (*Dementia and language*, 2022).

2.1.1.2 DemCare

DemCare (Karakostas et al., 2017) is another multimodal dataset that contains information including: video, audio and physiological sensors data. There are different actors, who are aged over 65 years and are healthy or suffer from conditions from Mild Cognitive Impairment to mild dementia, there are some cases of full-blown dementia. These actors are recorded in the Greek Alzheimer's Association for Dementia and Related Disorders and their own houses performing different actions in their daily life. Among these daily life activities, we can find reading an article, watering plants, preparing a drug box, preparing a drink, turning on the radio or talking on the phone. The data has been recorded with several devices, where we can find static RGB cameras on the top of the rooms to record the whole room and try to identify the activity they are performing, a wearable camera that will give information about object detection to improve the recognition of the activity and an accelerometer for psychological evaluation. In addition, while some actions include a more vocal part, a wearable microphone is included to obtain information to make correlations between vocal characteristics and cognitive state. This dataset is not only focused on the detection of the disease, it tries to deeply understand how the disease affects their normal daily life to try to help them in their self-independence. The recording environment as well as the different cameras and sensors can be seen in the figure 2.2.

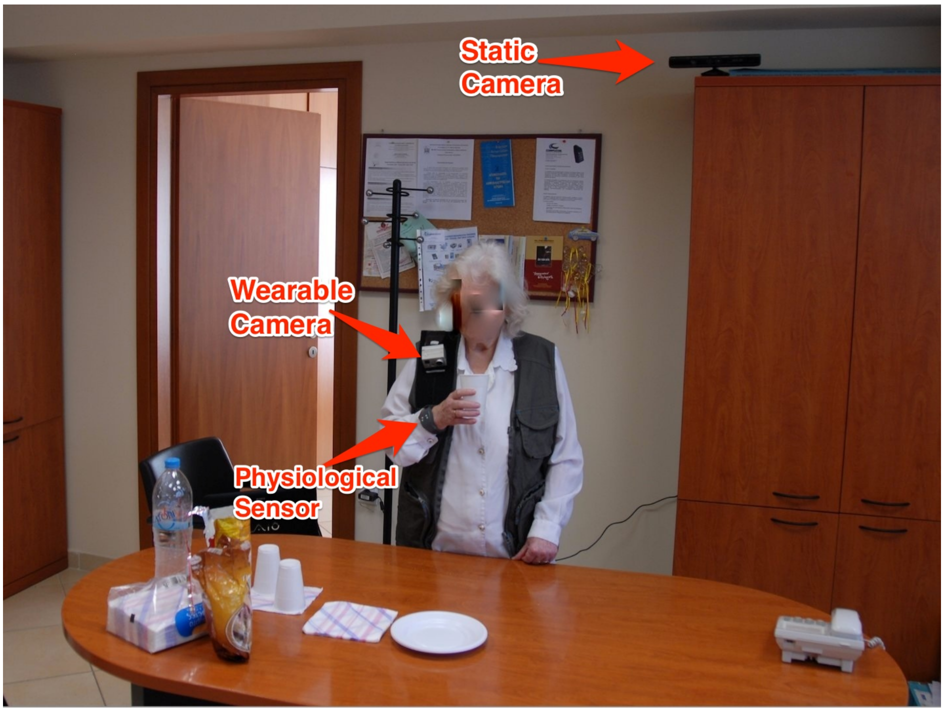


Figure 2.2: Demcare recording environment, from Karakostas et al. (2017)

2.1.1.3 PRAXIS Gesture

PRAXIS Gesture dataset (Negin et al., 2018) contains data in video with no audio. There are 64 elderly actors who are healthy or suffer from different types of dementia, with actors with vascular dementia, mixed dementia and Alzheimer among others. These actors had to perform 29 different types of simple gestures, repeating them until they did them correctly. Within the different gestures, we can identify gestures like taking the left hand to the ear, asking for silence, drinking a glass of water or hammering a nail among others. These actions can be appreciated in the figure 2.3 as well as other recorded gestures from this dataset and its corresponding names with index in figure 2.4. The data has been collected with an RGB-D camera and recorded the upper body of the patients while performing the gestures. For the recording of this dataset, the different gestures have been asked to do to 64 people, 4 clinicians and 60 elderly patients. It is more interesting to focus on the elderly patients since there will be notable differences between elderly people performing gestures and more young clinicians. In this group of 60 patients, we can find 29 patients had normal cognitive functioning, which means that they do not have dementia, and the other 31 patients had some type of dementia.

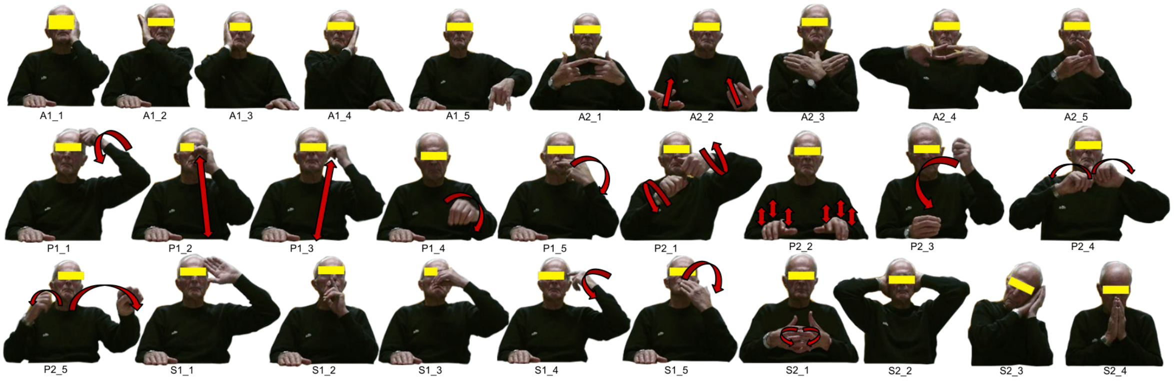


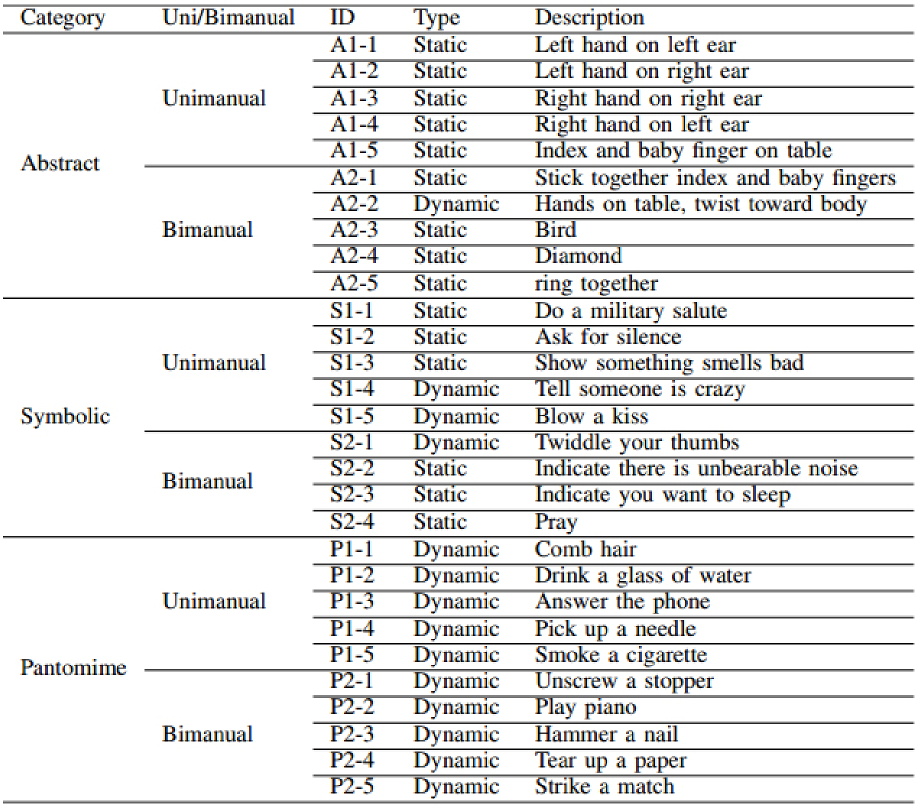
Figure 2.3: Different performed gestures in the dataset, from Negin et al. (2018)

2.1.1.4 Overview of dementia datasets

Once every available dementia dataset has been analyzed and studied, the next step has been to choose the one that better fits our project. As can be seen in the table 2.2 where the main features of each dataset are exposed, the PRAXIS Gesture dataset is the only one that does not include multimodal data, the DementiaBank dataset provides us data in audio and its transcription and finally, the DemCare dataset includes data in video, audio and psychological data. Every analyzed dataset includes patients who suffer from dementia and healthy patients, but they perform different tasks in each dataset, from simpler ones like PRAXIS Gesture to activities of their daily living or describing an image.

Dataset	Modalities	Activity	**Patients**
DementiaBank Pitt Corpus	Audio & Text	Describe image	Control, dementia, unknown
DemCare	Video & Audio & physiological	Activities of daily living	Control, MCI to mild dementia & full blown dementia
PRAXIS GESTURE	Video	Basic gestures	Control, several types of MCI and dementia

Table 2.2: Different dementia datasets



Category	Uni/Bimanual	ID	Type	Description
Abstract	Unimanual	A1-1	Static	Left hand on left ear
		A1-2	Static	Left hand on right ear
		A1-3	Static	Right hand on right ear
		A1-4	Static	Right hand on left ear
		A1-5	Static	Index and baby finger on table
	Bimanual	A2-1	Static	Stick together index and baby fingers
		A2-2	Dynamic	Hands on table, twist toward body
		A2-3	Static	Bird
		A2-4	Static	Diamond
		A2-5	Static	ring together
Symbolic	Unimanual	S1-1	Static	Do a military salute
		S1-2	Static	Ask for silence
		S1-3	Static	Show something smells bad
		S1-4	Dynamic	Tell someone is crazy
		S1-5	Dynamic	Blow a kiss
	Bimanual	S2-1	Dynamic	Twiddle your thumbs
		S2-2	Static	Indicate there is unbearable noise
		S2-3	Static	Indicate you want to sleep
		S2-4	Static	Pray
Pantomime	Unimanual	P1-1	Dynamic	Comb hair
		P1-2	Dynamic	Drink a glass of water
		P1-3	Dynamic	Answer the phone
		P1-4	Dynamic	Pick up a needle
		P1-5	Dynamic	Smoke a cigarette
	Bimanual	P2-1	Dynamic	Unscrew a stopper
		P2-2	Dynamic	Play piano
		P2-3	Dynamic	Hammer a nail
		P2-4	Dynamic	Tear up a paper
		P2-5	Dynamic	Strike a match

Figure 2.4: Details of the performed actions in figure 2.3, from Negin et al. (2018)

After analysing these datasets, PRAXIS Gesture was discarded for this work due to the lack of multimodal features, since the analysis of multimodality is one of the main goals of this research, exploring how different modalities can work separately and complement each other to improve the performance on dementia prediction.

The DementiaBank Pitt Corpus dataset was chosen for the present work due to its speech modality, which is an important feature of the dataset, as it can provide clear clues about the presence of dementia symptoms. In contrast, DemCare does not focus on this feature and instead focuses more on a visual daily tasks feature. From this visual information, dementia symptoms such as confusion, disorientation, or difficulties with coordination and motor functions can be distinguished.

On the other hand, the speech modality in the DementiaBank dataset allows observing other kinds of symptoms of dementia, such as difficulty with communication, finding words, reasoning, visual and spatial abilities, or planning. All of these abilities and difficulties can be identified by performing a task, such as describing an image with many details, as in the case of the DementiaBank dataset.

Another reason for choosing this dataset is that these difficulties can be observed not only in the text when constructing sentences to describe an image, but also in the analysis of the recorded audio. Difficulties in speech can be indicated by pauses, hesitation, doubts, and onomatopoeias. This is the main reason behind the idea of exploring different modalities and how they correlate and complement each other to improve the final performance of the system.

Therefore, the approach proposed in this work will deal with both textual and audio modalities to properly process the DementiaBank Pitt Corpus dataset.

2.1.2 Aphasia datasets

A research study has been conducted to select the most suitable dataset for the other task of automatic emotion recognition. The only dataset that provides information regarding aphasia disease is AphasiaBank Forbes et al. (2012), which will be explained in detail in section 2.3 and is provided by TalkBank. TalkBank is primarily dedicated to the research of human communication and offers other similar datasets that have been considered for our project. One such dataset is TBIBank Elbourn et al. (2019), which contains information on patients with traumatic brain injuries. This dataset is similar to the AphasiaBank corpus, as aphasia is often the result of brain damage in certain areas. Another comparable dataset is RHDBank Minga et al. (2021), which contains information on patients with right hemisphere damage.

The main factor that drove the selection of the AphasiaBank dataset over the others was the availability of video recordings and a larger number of samples. While DementiaBank only contains audio recordings and TBIBank do not provide a video modality for every sample, both AphasiaBank and RHDBank provide video recordings for each sample. Moreover, AphasiaBank offers a significantly larger number of samples for our study. In this research, the video modality is essential for emotion recognition, as it is easier to predict when the facial expressions of a person are visible.

This dataset is of particular interest due to its inclusion of video recordings, where patients were recorded during a conversation with a clinician. The videos capture the upper half of the body, including the face, which makes the facial expressions of the patients the most crucial aspect for our analysis of emotional recognition.

The dataset includes video recordings of both healthy and aphasic patients, although there are considerably more samples from the latter group. The dataset contains a total of 440 video samples from aphasic patients and 220 samples from healthy patients. The primary focus of the recordings is on the speech behavior of the patient, with the conversation and discourse tasks designed to provide data on how they express themselves. Since this database has different corpora, it is important to note that the tasks vary depending on the corpus of the dataset, and some tasks are more varied than others. A corpus is a set of data from the dataset, in this case, a set of video recordings. The main task involves initiating a conversation by inquiring about the patient's perception of their speech, while other tasks include the description of various images.

The dataset also includes CHAT transcriptions Macwhinney (2000) of the conversations, which is in line with other similar datasets, such as DementiaBank. In this sense, the information represented in the form of text that captures the speech that has been performed can be highly valuable for semantic and lexical analysis, as demonstrated in previous studies.

2.2 DementiaBank recent works

Once the dataset to use has been chosen, the next step has been to research the recent works over this dataset, DementiaBank, and more specifically in the Pitt corpus. Among the recent projects working over this dataset, we can mention the work of Warnita et al. (2018) which was released in 2018. In this work, they used only the audio data of the Pitt corpus, and the model used was a gated convolutional network, with this model a 73.6% of accuracy was achieved. Another work that uses just audio data is the one presented by Chakraborty et al. (2020) in 2020. In their project, they proposed a model that analyses the audio clips in order to obtain audio biomarkers for the detection of dementia.

There are also works working just over the text modality, as can be seen in the work of Karlekar et al. (2018) released in 2018. In their work, the best results obtained were by the use of Convolutional Neural Network (CNN)s combined with Recurrent Neural Network (RNN)s and the POS-tagging transcriptions of the utterances. The best results were obtained in this work, achieving an accuracy of 91.1%, the data used was down-sampled because not every utterance had a POS-tagging transcription accompanying.

These works were though using just one modality from the two modalities offered by this dataset, in recent years, some approaches have used both or even more data. One example of this can be seen in the work of Mittal et al. (2020) in 2021. For this work, they used both modalities, using two different models and weighting their probability of dementia. For the audio model, a Mel Spectrogram combined with an audio-based was used, and for the text model, different combinations for segment transcriptions and the full transcription were used. By using this model, they obtained an accuracy of 85.3%. All these works can be summarized in the table 2.3.

As commented before, this dataset has a smaller subset which has been balanced in terms of age and gender, called ADReSS challenge, and also different approaches with different methods (Martinc & Pollak, 2020) (Haulcy & Glass, 2021) (Mahajan & Baths, 2021).

Related works	Year	Approach	Accuracy
Amish Mittal et al.	2021	Multimodal	85.3%
Sweta Karlekar et al.	2018	Text data	91.1%
Rupayan Chakraborty et al.	2020	Audio biomarkers	81.9%
Tifani Warnita et al.	2018	Audio data	73.6%

Table 2.3: DementiaBank approaches

2.3 AphasiaBank related works

Regarding the existing work carried out on the selected dataset, AphasiaBank, there are tasks such as automatic speech recognition of aphasic individuals, as well as numerous lexical and semantic analyses, as this dataset includes transcriptions of recordings.

The task of automatic speech recognition, which involves transcribing an audio recording, has shown significant advancements in recent years, particularly with transformer-based architectures such as Whisper Radford et al. (2022) or Wav2Vec2 Baevski et al. (2020). The significance of this area lies in the added complexity of the task due to the communication difficulties faced by aphasic patients who may produce incomprehensible speech or sentences during a conversation. Additionally, there is a significant disparity in the availability of transcription data for healthy patients compared to those with the disease. In this regard, we highly appreciate the work done by Iván G. Torres et al. Torre et al. (2021), who used the AphasiaBank dataset as well.

Regarding other works focused on the semantic and lexical analysis of transcriptions from this dataset, several studies can be found. One such example is the work by Yu-Er Jiang et al. Jiang et al. (2023), which analyzed the main verbs and nouns used by patients with anomic aphasia and healthy controls. The study compared individuals of similar age and education levels to ensure a more accurate and balanced analysis. Results showed that individuals with anomic aphasia tend to use fewer core verbs and nouns than healthy individuals. Another study in this area that utilized the same dataset was conducted by Ouden Dirk-Bart et al. Ouden et al. (2015), which analyzed the use of verbs. Results showed that individuals with Broca's aphasia tend to use verbs in less complex and diverse ways than healthy individuals.

Emotional expressions and understanding are crucial in human communication. Diseases such as Aphasia and Dementia can negatively impact interactions and conversations with others. Patients with dementia may find it difficult to identify others' emotions and empathize with them. Thus, investigating the emotions of these patients is an interesting area of study. With advancements in artificial intelligence, tasks such as emotion recognition can be automated. Although there are currently no studies on how aphasia affects patients' emotions, diseases like dementia have been explored in automating emotion recognition for further analysis.

Karmele Lopez-de-Ipiña et al. López-de Ipiña et al. (2013) conducted an emotion response analysis aimed at detecting dementia by analyzing audio recordings and using audio features to determine emotions. Parkinson's disease is another illness that can affect patients' emotions, with deficits in emotional speech production. Shunan Zhao et al. Zhao et al. (2014) have performed a more complex analysis using automatic emotion recognition to investigate this disease.

In this context, there can be numerous emotions, with subtle differences between them. Psychologist Paul Ekman differentiates between six basic emotions: anger, disgust, happiness, fear, surprise, and sadness. Ekman proposed this distinction based on an analysis of eye, head, and facial muscle movements.

2.4 Multimodal models

In our daily life we perceive the world with more than one single sense, we can see objects or hear sounds for example. This is the basic idea for multimodal models, this type of model works over multimodal data, data in different modalities. An example of this data could be an image and a text describing it, it has an image as one modality and text as the other one. For each modality, the model will have different sub-networks to analyze the data. In order to get a final output, there are different approaches, such as giving a weighting for each sub-network output or concatenating the different outputs to get a final output. This example of how a multimodal model works concatenating the different outputs can be seen in the figure 2.5.

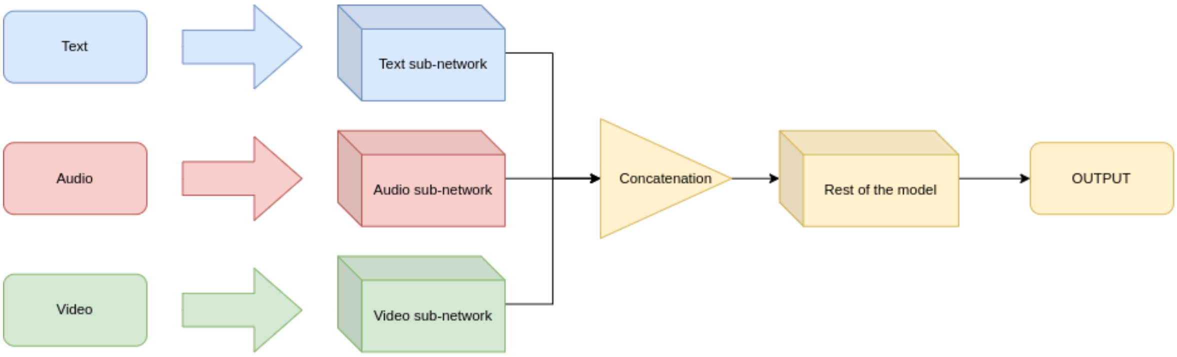
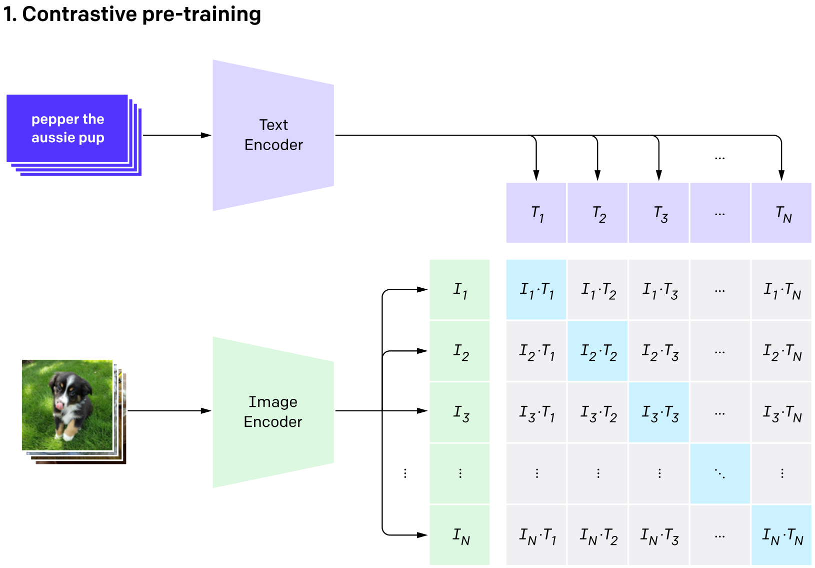


Figure 2.5: Multimodal model architecture

There are some important implementations of these types of models, for instance, MMF (Singh et al., 2020), CLIP (OpenAI, 2021) or VATT (Akbari et al., 2021). MMF is a modular framework developed by FacebookAI, it is used for vision and language multimodular research. This framework contains implementations of state-of-the-art vision and language models such as VisualBERT or VilBERT as well as different datasets to work on like Visual Question Answering (VQA). Contrastive Language–Image Pre-training (CLIP) is another multimodal model which relates a whole sentence (text) with an image. CLIP pre-trains a set of images and a set of sentences to come with relationships between them, and its architecture can be seen in the figure 2.6.



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Figure 2.6: CLIP model architecture, from OpenAI (2021)

Finally, the VATT model is an approach whose main aim is to analyze video, audio and text at the same time. Each one will have its own transformer encoder in order to process the input and finally a projection head to get the similarity between all those modalities using contrastive losses. This approach can be used later for downstream tasks in a variety of fields, such as video action recognition or audio event classification.

2.5 Natural Language Processing

Natural Language Processing (NLP) is a field of artificial intelligence that is focused on understanding how we communicate with other humans, using natural language. The understanding of this natural language is a really difficult task for computers, since our language has lots of ambiguities, such as homonyms, homophones, sarcasm and metaphors among others. Sometimes it is even hard for us to understand and recognize the sarcasm of people that we barely know. Imagine how difficult it would be a machine to completely understand it. This is a wide field, where we can find several type of different tasks related to our language, as can be seen here:

- **Speech recognition:** This task consists in the conversion of voice data to text data. This task has been present in our lives for many years, for example with the Google option to transcribe audio for the Google search engine.

- **Part of speech tagging:** This task aims to determine the part of speech of the introduced sentence. One example of this part of the speech is identifying the word “playing” with a verb, this would be applied to every single word of the sentence.

- **Word sense disambiguation:** This is the process that determines the exact meaning of a word when the word has several meanings. One example of this would be the word “left”, that can be the side, opposite of the right, like in the sentence “I will do it with my **left** hand”. The other meaning of the word is when it is related to a person who goes away from a certain place or situation, like in the sentence “Tom left the party early”.

- **Co-reference resolution:** This task tries to identify when two words in a sentence refer to the same entity. One example of this would be in the sentence “Mary was hungry, so she ate the whole pizza”, in this sentence, the words “Mary” and “she” refer to the same entity, which is Mary.

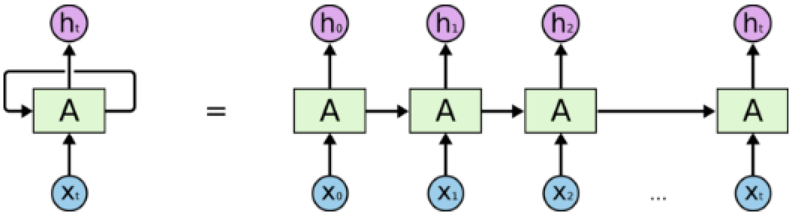
- **Sentiment analysis:** In this task, the main purpose is to identify the subject opinion of the writer from the text given. One example would be classifying reviews of a product, having as categories, “good product” and “bad product”.

- **Natural language generation** This task aims to generate or complete text from a given input text. One example of this task, would be receiving the beginning part of a newspaper article and finishing it as if a human would be.

In the last few years, there has been a huge improvement in this field with a new type of architecture, the Transformers (Vaswani et al., 2017). All those multimodal models introduced above are based on transformers. This type of architecture achieves the majority of the state-of-the-art approaches in NLP tasks. Before transformers, the majority of state-of-the-art NLP models were based on recurrent neural networks. In this section, we will explain the different architectures that have been used in recent years for NLP problems.

2.5.1 Recurrent Neural Networks

RNN (Schmidt, 2019) models sequentially process the text input word by word, using the output of one layer and as input for the next one, keeping the time dependency. This way of processing the data sequentially makes relations between one element and the following one, like in natural language, where one word is related to the following in order to construct a whole sentence. The basic architecture of this type of neural network can be seen in the figure 2.7, where the outputs from the first input X0 will be used as one input for the next layer, as well as the second input X1. In the case of NLP tasks, these inputs will be words or different kinds of representations of words. A bad thing about this architecture is that it does not perform well in long sequences, since it loses data from the initial layers.



An unrolled recurrent neural network.

Figure 2.7: RNN architecture, from Mittal (2021)

2.5.2 LSTM

Long Short-Term Memory (LSTM)s (Hochreiter & Schmidhuber, 1997) are a special type of recurrent neural networks, it differs from the normal recurrent neural networks by adding functions in order to keep the information from the past that is important. In addition, it also adds a cell state, which is like a long-term memory. This addition of more functions and states, supposes that the architecture, that can be seen in the figure 2.8, is more complex in this type of network than in conventional RNNs. On the other hand, this increase in the complexity of the network also supposes an increase in the computational cost of the network. This architecture consists of three gates. The first one, the forget gate, will be used in order to forget or keep the previous timestamp information. The input gate will be in charge of calculating the importance of the new input received. Finally, the output gate will update the hidden state and the new added cell state information of the network.

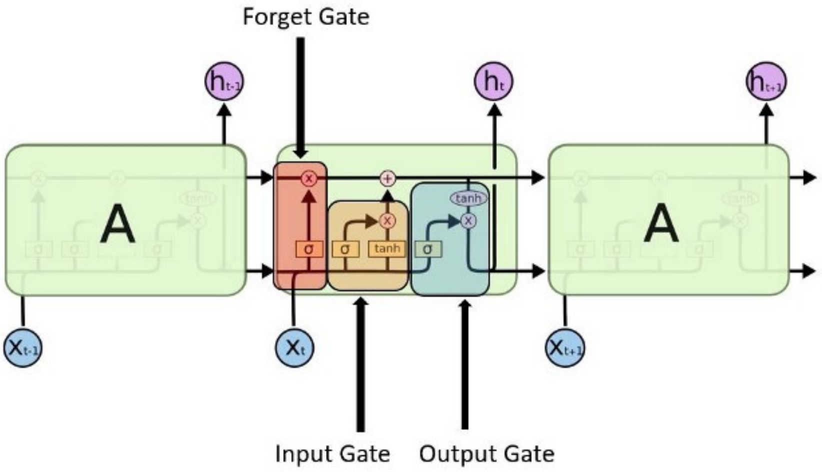


Figure 2.8: LSTM architecture, from Mittal (2021)

2.5.3 Transformers

Transformers (Vaswani et al., 2017) are a new type of network architecture, focused on the NLP tasks. They have recently shown great results in this field, achieving state-of-the-art results on many tasks. The main idea of transformers has been the attention component, with a Multi-Head Attention layer which will relate words between each other. For example in the sentence “Tim did not come to practice yesterday, he was too tired”, this attention layer will relate “Tim” and “He”. With this characteristic transformers will not require sequential processing as RNNs do. This attention function, that can be seen in the figure 2.9, consists of mapping a query (Q) and a set of key (K) values (V) pairs to an output. All these components, query, key, values and outputs are vectors and the output will simply be a weighted sum of the different values. Each value will be assigned with a weight, this weight will be calculated with a compatibility function of the query and the corresponding key.

In contrast to the models that were achieving the best results previously, which were more and more complex recurrent neural networks, the transformer was looking for simplicity, by just focusing on this attention part and not in the other components. This architecture can be appreciated in the figure 2.10.

This is the basic architecture of a transformer, consisting mainly of an encoder, with N number of encoding layers and a decoder, with N number of decoding layers. Each encoding and decoding layer has attention and feed-forward layers. The encoding layers aim to extract features from the text and the decoder will be in charge of using those features to generate a new output, such as a new sentence if the task is language translation. For this reason, the architecture of a transformer will vary with the task that will perform, using just encoders if the task is sentence classification or on the other hand using both, encoders and decoders if we want to translate between different languages.

Bi-directional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is

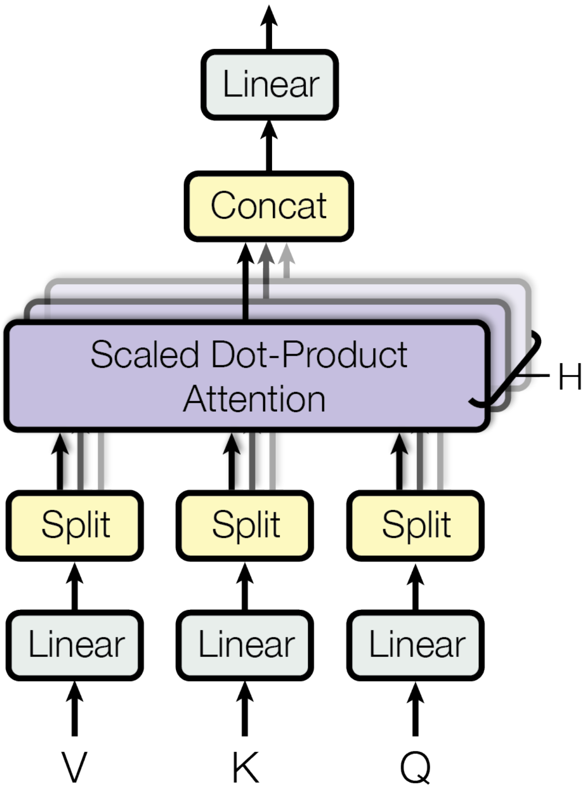


Figure 2.9: Attention function, from Vaswani et al. (2017)

an NLP model architecture based on transformers. This model can be used for many tasks such as sequence classification, question answering or natural language inference. There are two different sizes for the BERT model, the base one and the large one. This architecture , which can be seen in the figure 2.11, consists of a stack of several encoders. The base size has 12 stacked encoders and the large size has 24.

This is the main architecture of a BERT model, it consists of stacked encoders that will get as an input a tokenized sequence and will return as output its embeddings. To process the input, it will need a tokenized numerical sequence. The tokenizer will add a ['CLS'] token at the start of every new sentence and a ['SEP'] one at the end, as well as a ['PAD'] one if padding needs to be used, after this, will encode these tokens as numbers to use as inputs for the BERT model. As an output of the BERT model we obtain an array of embeddings, one embedding for each input token. These embeddings are a representation of a token, they are represented in a vector of size 768 in the case of BERT. There is a special embedding that is a representation of the whole sentence, which is the one related to the ['CLS'] token, introduced at the beginning of each sentence.

This model developed by Google, has required a lot of data in order to train the model to achieve these results. For this purpose, Google has used unlabeled data, plain text corpora from the English Wikipedia. Moreover, this model is being used by Google in their search engine, since it aims to fully understand the language, if used correctly, it can improve the searches. With these types of applications, BERT recibes more data and can learn and train

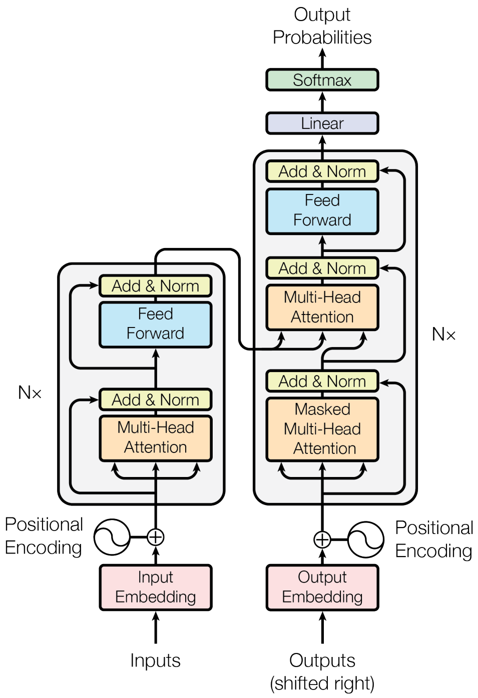


Figure 2.10: Transformer architecture, from Vaswani et al. (2017)

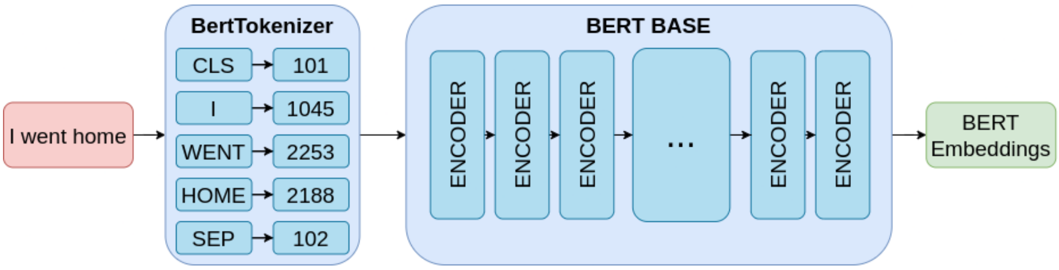


Figure 2.11: BERT architecture

over time.

Due to the good results of the understanding of the human language, with the ambiguity it has, the BERT model has expanded to other fields, like human speech. Furthermore, many different types of pretrained BERT models have been released for specific fields of knowledge, such as SciBERT (Beltagy et al., 2019), specialized in scientific papers or patentBERT (Lee & Hsiang, 2019), for the classification of patents.

Another interesting transformer-based architecture is Whisper Radford et al. (2022), whose architecture can be seen in Figure 2.12. This model is used for speech recognition and can perform multilingual speech recognition, speech translation, and language identification. This model has been pre-trained with over 680,000 hours of multilingual speech. This extensive amount of data has made the model robust against background noise, different accents depending on the region, and technical language. Additionally, since it has been pre-trained with multilingual data, it can recognize and transcribe different languages, as well as perform translations into English.

The architecture of Whisper works as follows: firstly, the audio clips are split into segments of thirty seconds, and these segments are converted into a log-Mel spectrogram. The resulting spectrograms are then fed into the encoders of the model. The output of the encoders is passed through the decoder blocks to obtain the tokens, which are subsequently translated into actual words. The utilization of log-Mel spectrograms is common in audio classification, as discussed in Section 2.6.

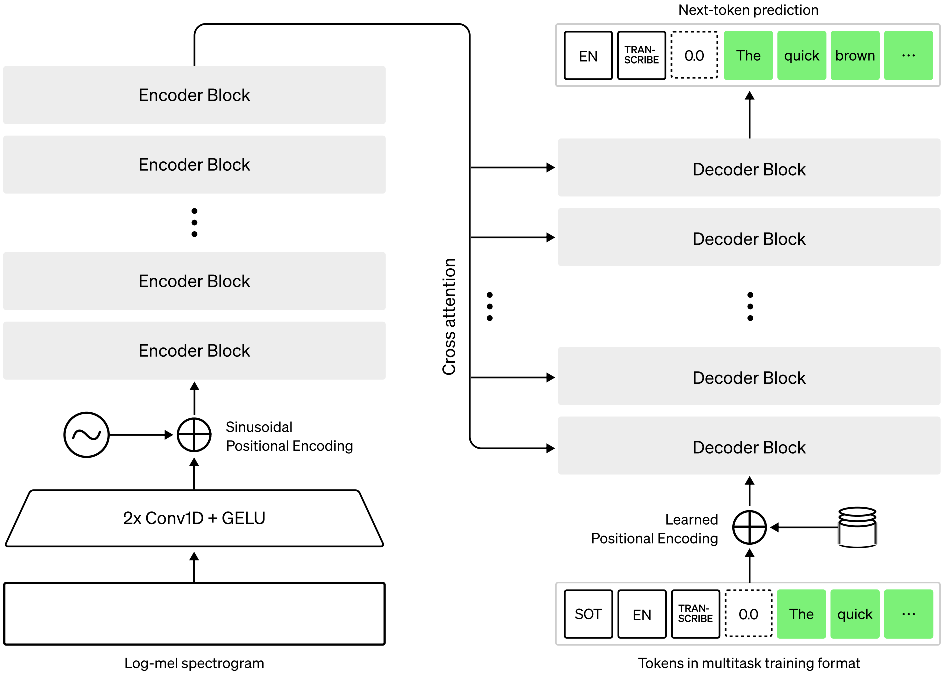


Figure 2.12: Whisper architecture, from Radford et al. (2022)

2.6 Audio classification

Recent studies have shown how we can use CNNs models for the task of audio classification and the results have achieved state-of-the-art results on various tasks (Palanisamy et al., 2020) (Hershey et al., 2016). CNNs work over images, extracting different features from them. These features are obtained by applying filters/kernels to every part of an image. With all these features extracted, it has a classification part that will be in charge of obtaining the final outputs from the previously computed features. But we are dealing with audio files, not images, a transformation needs to be done in order to work with these types of networks. When we are dealing with audio, we can obtain the waveform of that audio. This waveform is a graphical representation of the sound waves of the sound over time, as can be seen in the figure 2.13.

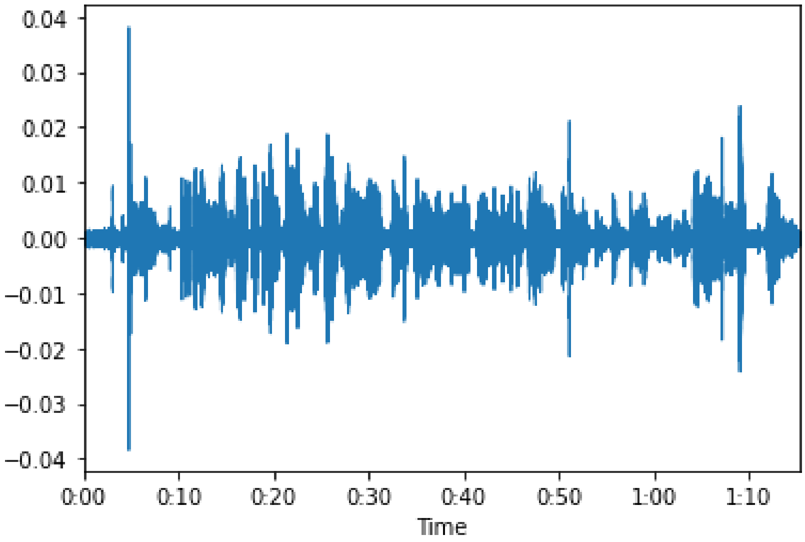


Figure 2.13: Waveform obtained from DementiaBank dataset

The human ear has shown to be better at perceiving low frequencies sounds, differentiating more easily between two low frequencies sounds than two larger ones. This is the main idea of the Mel scale, creating a scale where equally perceived sounds by a human are equally distanced as well. With this scale, we can build a Mel spectrogram (Roberts, 2020), a spectrogram in a Mel scale, with the different frequencies of sounds over the time as a human would perceive. One example of this Mel Spectrogram can be appreciated in the figure 2.14.

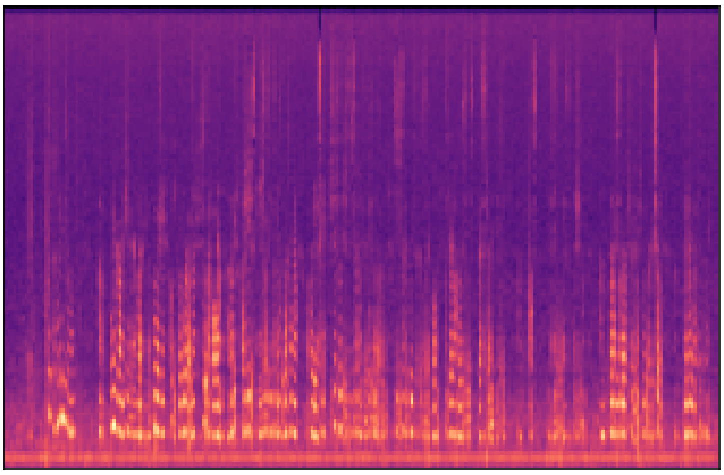


Figure 2.14: Mel Spectrogram obtained from DementiaBank dataset

These Mel Spectrograms are the inputs of our CNNs, which are treated as an image, having a height and a width that can be resized in the process of obtaining these images. In these works, the models used for analyzing these spectrograms were pretrained CNNs such as

MobileNet (Howard et al., 2017), DenseNet (Huang et al., 2016) or ResNet(He et al., 2015).

These CNNs are a great option and result in good results on the image classification task, it is due to the capability of extracting features from the images, in this case the Mel Spectrograms.

2.7 Facial emotion recognition

Facial emotion recognition is a technique used to identify and predict the emotional state of individuals based on their facial expressions. It can be applied to videos by analyzing frames or to photos. This process typically involves three steps, starting with face detection. The primary objective in this initial step is to locate and identify human faces within the frame or photo. Various methods can be employed to achieve this goal, including the following:

- **Feature-based**: This method starts by searching for the human eyes, which are the easiest feature to find on a face. After identifying the eye region, the algorithm tries to locate other important landmarks on the face, such as eyebrows, mouth, or nose. This method can be negatively influenced by noise and lighting conditions. An example of this method can be seen in Figure 2.15.

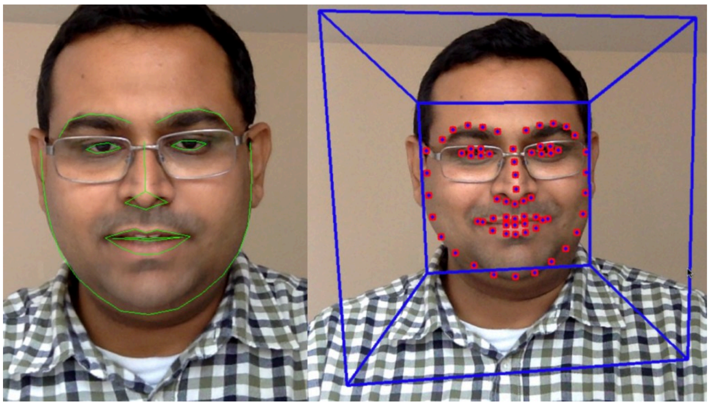


Figure 2.15: Landmark feature extraction from a face, image obtained from Mallick (2021)

- **Template matching**: This method is based on comparing images with previously stored face patterns and correlating them to detect a face. However, this method faces difficulties when there are variations in pose, scale, and shape.
- **Convolutional Neural Networks**: We have previously explained how these models work. In this case, there is a more complex architecture called R-CNN, as shown in Figure 2.16. This architecture is composed of a pipeline that starts with the initial image as input. Based on this input, the image is segmented into different regions that may contain objects. These regions are then processed using convolutional networks to extract features from each region. Finally, classification layers are added to determine the label of each region, in this case, whether it contains a face or not, among other possible labels. An improvement to this architecture is Fast R-CNN, as shown in Figure

2.17, which includes a Region of Interest (ROI) layer that adjusts the size of the region, allowing the processing of regions with different sizes. This new architecture improves both accuracy and processing time. Another architecture in this category is Faster R-CNN, as depicted in Figure 2.18. It eliminates the need for selective search and adds a Region Proposal Network (RPN) layer, which proposes regions where objects are more likely to be present for classification. This architecture achieves significant improvements in processing time.

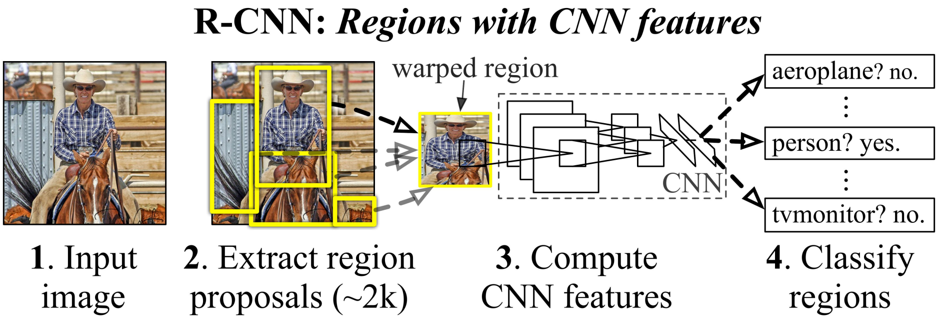


Figure 2.16: Architecture of R-CNN, image obtained from Girshick et al. (2014)

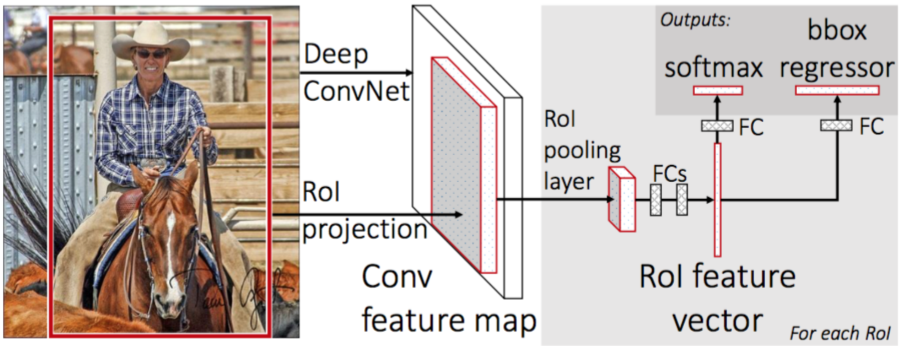


Figure 2.17: Architecture of Fast R-CNN, image obtained from GeeksforGeeks (2020)

- **Single shot detector (SSD)**: This architecture is different from the previous ones, using a single network for the location and classification of objects. This architecture extracts the characteristics of the image, from it generates different maps, and in each cell of each map predicts the class to which it belongs and its coordinates, with the combination of these maps a final result is obtained. This architecture can be seen in Figure 2.19.

The second step involves applying normalization to the detected and bounded facial expression in the frame or photo. This normalization process aims to enhance the accuracy of our models and is particularly useful for addressing illumination changes, reducing noise, or performing image smoothing, among other benefits.

In the final step is where the emotion is predicted, in this step we receive a normalized facial

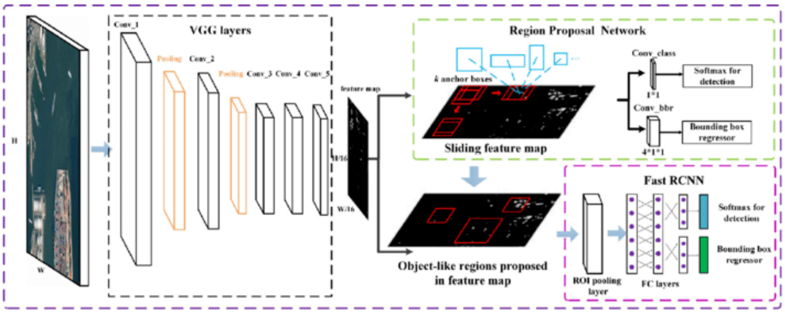


Figure 2.18: Architecture of Faster R-CNN, image obtained from Deng et al. (2018)

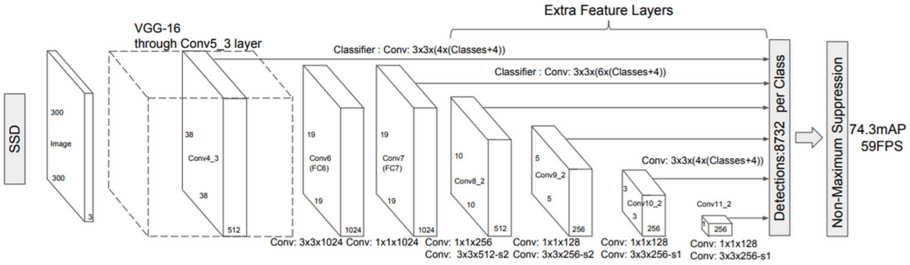


Figure 2.19: Architecture of SSD, image obtained from Khandelwal (2019)

expression. For this purpose, we extract facial features, as could be the facial landmarks and the distance, once we obtained facial features, we use classifier to predict the emotion. As classifiers, we could use classic artificial intelligence methods, like Support Vector Machine (SVM) or other methods and models like the previously seen CNNs. In this case, the best results for this task are normally achieved by the use of CNNs. In this particular case, in the first step, with the methods that used CNNs, we could add more labels to the architecture, in order to obtain the facial emotion as well.

In the final step, we predict the emotion based on the normalized facial expression. To accomplish this, we extract facial features such as Action Units (AU), the fundamental actions of muscles, facial landmarks or the distance between them. Once these facial features are obtained, we employ a classifier to predict the emotion. For this purpose, classic artificial intelligence methods such as Support Vector Machines (SVM) can be utilized, along with other methods and models like Convolutional Neural Networks (CNNs). Generally, CNNs tend to yield the best results for this task. In this particular case, when using CNNs in the first step, we can augment the architecture by adding more labels to obtain the facial emotion as well as the face detection.

3 Materials and methods

The development of artificial intelligence projects requires certain materials and methods to construct proper models. First of all, it is necessary to have appropriate hardware, which will be in charge of running the different training and tests of our approaches. Then, the use of a tool like Docker(Merkel, 2014), a platform that allows us to quickly generate different types of environments for the construction of software separate from our main infrastructure. Finally, once the environment has been set up, we can start implementing our approaches, but for this step, we need the use of different deep learning frameworks. This chapter is structured as follows: in section 3.1 the specifications of the hardware used will be described; in section 3.2 the tool docker will be introduced; in section 3.3 the main deep learning frameworks will be explained; finally, in section 3.4 the Hugging Face Transformer library will be introduced.

3.1 Hardware

Artificial intelligence projects usually require lots of data in order to train a model for its later use. This huge amount of data results in a high computing processing cost, which can result in a delay in the defined timeline due to the time it requires to get processed or even not being able to process if we do not have appropriate hardware specifications. For this project, the collaboration with the 3D perception lab has provided us with access to their servers, Asimov and Clarke. We will be working over the Asimov server.

The most important components of this server are the graphic cards, which will be in charge of the main computing process in projects related to artificial intelligence and more specifically to deep learning. In this case, Asimov is equipped with two graphic cards dedicated to research in artificial intelligence. Both graphic cards are provided by Nvidia, which are two Nvidia GeForce GTX Titan X. The rest of the components can be seen in the table 3.1. Talking about the operating system, this server runs the version 16.04 LTS of Ubuntu. Additionally, this server is connected to a Network Attached Storage (NAS) which provides 14TB more of storage with a Redundant Array of Independent Disks (RAID)5 system.

	Asimov
Motherboard	ASUS X99-A Intel X99 chipset
CPU	Intel(R) Core(TM) i7-5820K 3.30GHz 6 cores 12 threads
RAM	32 Gigabytes DDR4
GPU 1	Nvidia GeForce GTX TITAN X GP102 3840 CUDA cores 12 GB GDDR5
GPU 2	Nvidia GeForce GTX Titan X GM200 3840 CUDA cores 12 GB GDDR5
Storage (OS)	Samsung SSD 850
Storage (data)	3TB HDD (RAID1)

Table 3.1: Asimov's hardware specifications

3.2 Docker

We will work on Asimov's server, but we will not be alone working on it. Each person will need a specific environment in order to work properly, for example having different libraries installed or versions of certain libraries. For this reason, everyone will need their own and independent environment for the deployment of the software. In order to achieve this, we have some options, where we can find independent virtual machines to work, which will run over the server or using a platform like Docker (Merkel, 2014), which allows us to generate independent containers with different environments for each one. Even though they may seem similar, they have some differences. The architectures of both can be seen in the figure 3.1.

The main difference between them is that virtual machines emulate a whole operating system and Docker does not, all the containers share the kernel with the host operating system, in this case the Asimov's server. One consequence of this is that every process run in a container will be visible in the host using commands like "ps". But in the case of virtual machines, we could only see the process of the virtual machine, not any process inside. Other consequences of not emulating a whole operating system result in that containers are so much more lightweight than virtual machines, but it is not the only advantage of docker against virtual machines. Other main advantages is the start time: a docker container can be run and started in seconds, virtual machines can even last for minutes. In addition, docker containers do not require to consume as many resources as virtual machines do, they are emulating a whole operating system and that is the reason why they require lots of resources. Finally, as last advantage of using docker against virtual machines is the reusage. Docker containers are built from docker images, that can be constructed using the command "dockerbuild". These images result in a very simple way of using and sharing. In addition there is a huge repository

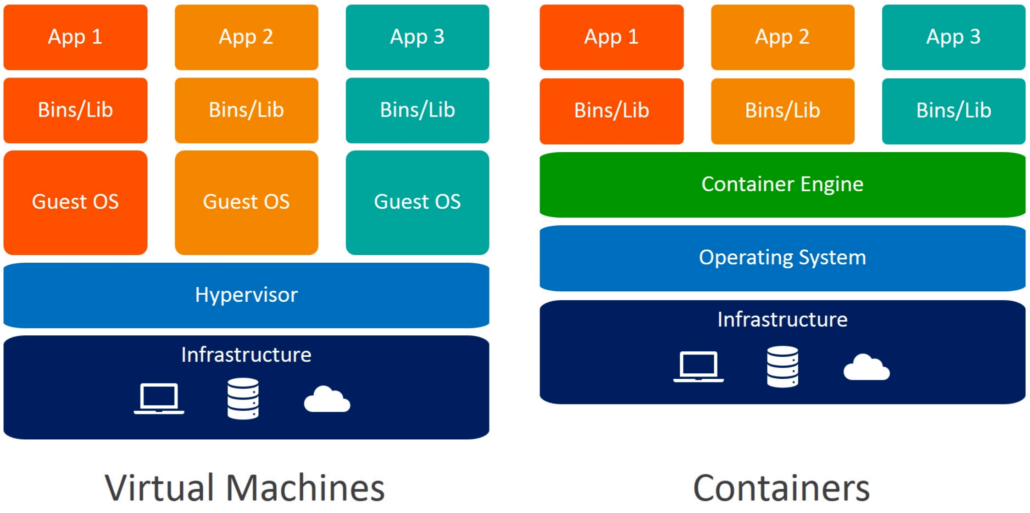


Figure 3.1: Comparison between Virtual Machine and Docker architectures, from Clancy (2021)

[1] with thousands of prebuilt docker images by other users.

Not everything from using Docker over virtual machines are advantages. On the other hand we have that are less secure than virtual machines since they share the kernel with the host. Even with this factor, we choose to use Docker because of all the advantages that present. It better suits for this type of works, since the emulation of whole new operating system is no needed and will require in more computational costs.

3.3 Frameworks

Deep learning algorithms are a highly difficult task to implement from scratch, due to its big complexity and size. Writing a model like CNNs previously explained would take days or even weeks to develop one that works. Due to this reason, nowadays, and each time there are more, we can use deep learning frameworks that will help us develop models more quickly and easily. These frameworks provide us with a clear and concise way for defining our models using pre-built different components. Over this section the different actual frameworks for deep learning will be described.

3.3.1 Tensorflow

As the first framework in this section, we can find Tensorflow (Abadi et al., 2015), an end-to-end open-source framework for machine learning developed by Google. This framework provides an abstraction to the developers in order to make them not have to deal with the underlying of the different algorithms used. Tensorflow is written in C++, Compute Unified Device Architecture (CUDA) and python and it is available to use not only in Python but also can be used in C++ and R. The mathematical operations behind this framework are performed using the code written in C++. This allows the framework to be faster,

[1] https://pytorch.org/hub/

due to the efficiency and quickness of C++ over python. This framework can be run on almost any target, a local machine, the cloud, iOS and Android devices, Central Processing Unit (CPU)s or Graphics Processing Unit (GPU)s and Tensor Processing Unit (TPU)s. TPUs are Google's custom-developed Application-Specific Integrated Circuits (ASIC)s used to accelerate machine learning when using Google's own cloud.

3.3.2 Keras

Keras (Chollet et al., 2015) is a python high-level framework that is built on the top of Tensorflow. It can also be built on The Microsoft Cognitive Toolkit (CNTK) or Theano. This framework aims at fast experimentation, if you need quick results, Keras will deal with all the core tasks and generate outputs. This framework can be also run on CPUs as well as GPUs. Keras allows the construction of two types of models. The first one a sequential model, which will sequentially execute its component. The other one is the Keras functional API, which allows us to construct more complex models, like multi-output ones or models which share layers. Keras also allows us to implement many types of layers to construct our model, such as fully connected, convolutional, pooling, recurrent and embeddings among others. It finally also offers the option of using pre-trained complex models.

3.3.3 Pytorch

Pytorch (Paszke et al., 2019) is an open-source framework for machine learning deployment developed by Facebook AI's Research Lab. Pytorch, as well as Tensorflow is written in C++, Python and CUDA and it can be used in C++, Python and Java. This framework can be run as well as the others on CPUs and GPUs and its main benefit is the flexibility that it allows over the other two previously mentioned frameworks. Pytorch has some features, the first one is that its tensors are very similar to the NumPy ndarrays, but they can be accelerated by the use of GPUs. Other key factors of the deep learning frameworks are the graphs that represent the different computations that are applied, every deep learning framework is based on this concept. Tensorflow makes this graphs computation as a static object, but Pytorch is based on dynamic computation graphs. This dynamic computation consists of being built and rebuilt at runtime, allowing us to modify the layers at each epoch for example. Finally, it counts with a Hub, where we can find pretrained models developed by users as well as share ours. If we require other pretrained models, the different libraries like torchvision offer pretrained complex models that will allow us to quickly deploy models.

3.3.4 Conclusion

Once we have researched among the actual available deep learning methods, we have to choose one in order to implement our proposal. We have chosen to use Pytorch as the deep learning framework due to the highly flexibility that it offers as well as its efficiency in the memory usage. Pytorch also allows to use python debugging tools, in contrast to Tensorflow which makes a debugging a hard task.

3.4 HuggingFace Transformer library

The HuggingFace Transformer library (Wolf et al., 2020) is an open-source python library that can be used with the deep learning frameworks Pytorch, Tensorflow and JAX. This library focuses on the usage of different types of transformers for the realization of tasks such as text classification, question answering, image classification, object detection and speech recognition among others. The best feature of this library is that it provides a really easy way to implement transformers, allowing fine-tuning them in order to adapt the models to our purposes. It not only provides transformers models, but it also provides different datasets that can save us time if it matches our purpose. It is a user-friendly library due to its simplicity and the many examples of code with different tasks and models that it provides. These examples, along with the available courses make new users not have big difficulties at the usage of this library. The other big feature offered is the possibility to use and try models that other users have previously trained and uploaded to the web.

4 Multimodal Deep-Architectures to predict and assists cognitive impairment diseases

We have chosen the datasets that better suit this project, revising the different recent approaches to process these datasets and considering the current state-of-the-art in the two proposed tasks. After this phase of research, it is time to start developing our system to predict and assist cognitive impairment diseases. For this, we have implemented different proposals for each task. All the implemented code is available in the GitHub repository [1]. This chapter is structured as follows, section 4.1 will introduce the proposed work for dementia prediction. Section 4.2 will describe the proposed pipeline for automatic emotion detection in people with Aphasia.

4.1 Dementia prediction

In this section, the proposal for the first task, which is dementia prediction will be introduced. This section is structured as follows: subsection 4.1.1 will describe the dataset used; subsection 4.1.2 will introduce the model that we have used for the audio modality; subsection 4.1.3 will present the model used for the text modality; subsection 4.1.4 will describe the model used for both modalities and finally, subsection 4.1.5 will summarize different other approaches that have been tested.

4.1.1 Dataset preprocess

As commented in previous chapters, the chosen dataset for this task has been DementiaBank. This dataset has two different types of files: an audio file and a transcription of that audio file. The audio model is going to use the Mel Spectrograms of the audio files in order to perform its classification task. The bad thing about this computation is that it takes too long to compute. It is generating images of a size 128x250 pixels, this for every audio file, with a total of 552 files. The computation of this transformation takes around thirty minutes in the servers, so we can not compute this in every epoch of the training, the training would take too long. As the image was stored in a NumPy array, the dataset has been preprocessed to store all these computed arrays in the disk. With this preprocess for the audio files, we reduced the time taken to load the audio files from thirty minutes to practically instantly.

This preprocess has been used for the audio files, but another preprocess task has had to be applied to the transcriptions files of that audios. These transcriptions are written in a CHAT format (Macwhinney, 2000), a special format used by TalkBank in their corpora, like in this case the Pitt Corpus of DementiaBank. This CHAT format implies that is available not only for the transcription of the subjects. Moreover, the interviewer transcription was

[1] https://github.com/davidorp/tfm

annotated as well as personal information from the patients and special flags representing pauses or mistaken words among others. For this reason, the original transcription files have been pre-processed in order to obtain a clean text transcription for the text model. This format also brings us certain benefits. For example, some sentences that do not have relevant information are masked with a "[EXC]" label, which tells us that we have to exclude the sentence.

These are the two preprocesses that have been applied to the dataset. Finally, we have split the dataset into two different sets, a training set and a test set, to test how well it performs. The training set is the 85% of the whole dataset and includes 469 samples and the test set is the 15% of the whole dataset and includes 83 samples.

4.1.2 Audio model

The first approach analyzed the audio files. In figure 4.1 the architecture of the implemented model is presented.

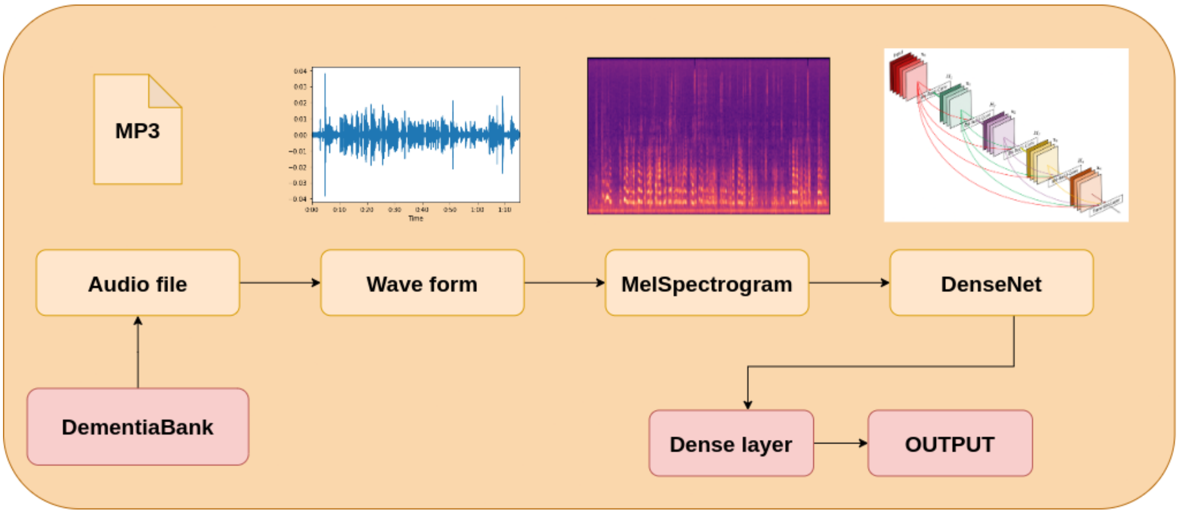


Figure 4.1: Architecture of the audio model

Each audio file has been converted to its waveform, a graphical representation of the signals over time, and then converted to a Mel Spectrogram, a Spectrogram in a Mel Scale (Roberts, 2020). This Mel Scale is inspired by the way humans perceive sounds, differentiating the low-frequency sounds rather than the high-frequency ones. In this scale, two equally distanced sounds in the pitch sound equally distanced to a listener. After this conversion, we used a Convolutional Neural Network to process this spectrogram. This model will handle the spectrogram as an image. This is because these networks are used basically for image tasks, such as image classification. In this model, different pre-trained CNNs have been tested, such as MobileNet (Howard et al., 2017), DenseNet (Huang et al., 2016) and ResNet(He et al., 2015). The best results were obtained with the DenseNet model. In chapter 5 the results of the audio model will be referencing the ones obtained by using this DenseNet. The final step of this model is a dense layer with the outputs of this CNN in order to get a final output, the prediction of the model (dementia (1) or control (0)).

4.1.3 Text model

Figure 4.2 shows the architecture of the proposed model for the text analysis.

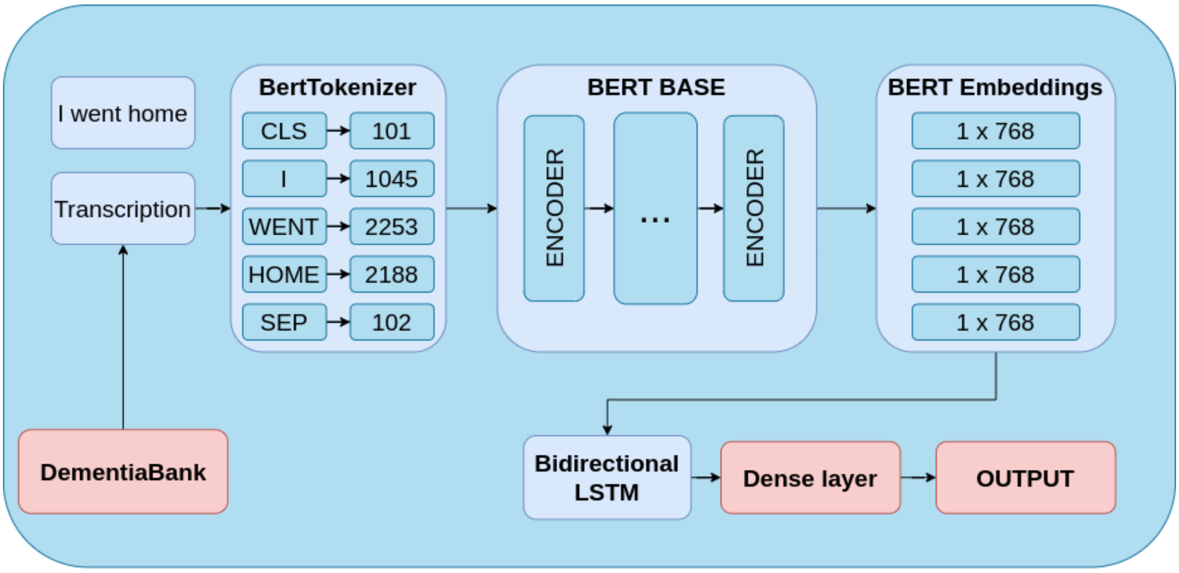


Figure 4.2: Architecture of the textual model

As it can be seen in the architecture, the famous BERT (Devlin et al., 2018) model has been used, a model which has achieved the state-of-the-art in many natural language processing tasks. This BERT model is based on Transformers (Vaswani et al., 2017), stacking different transformer encoders that will extract features from text. The most interesting aspect of these transformers is the use of attention to establishing relations between the different words in the sentence. In order to use the BERT model for this task, we fine-tune a pre-trained BERT model. The way used for fine-tuning is explained in the following paragraphs. This BERT implementation has been done using the HuggingFace's Transformers python library, which provides a more simple way and quicker way of implementing this model for our work.

The first step in order to use the BERT model has been to tokenize the input sentences. This tokenization consists of converting words that we could be using in our daily lives, to numbers, which will be the input for the BERT model. In this tokenization, several new tokens (numeric representations of the words) are added, such as the "[CLS]" token at the beginning of each sentence. After this text tokenization process, the BERT model receives as input those encoded words and returns its embeddings, a different embedding for each word received as input, having a size of 768 each embedding. These embeddings are a way of representing a word in the natural language processing models, allowing it to establish similar representations to similar meaning words.

One way of fine-tuning this BERT model is to use the embedding of the ['CLS'] token (the first one) to a final dense layer for classification. This ['CLS'] embedding is the representation of the whole text. This is the reason why this embedding is used for fine-tuning. This was not the only method tested for the text representation in our experiment with the test set. The other model tested uses these embeddings to fit a bidirectional LSTM. This type of network has proved good results for tasks with sequences such as text, and before the introduction of transformers, they were the previous state-of-the-art in many NLP tasks. LSTMs work using the output of one input (words or embeddings in this case) as the input of the following one, keeping information from the previous data. These models also have a mechanism to forget the irrelevant data from the previous segments and keep the important ones. Finally, the output of the LSTM is used as an input for a final dense layer to obtain the final output of the model. The comparison of the results of these two similar models will be seen in the chapter 5.

4.1.4 Multimodal model

After evaluating features separately, the text and audio were combined into a multimodal model to test if better results could be obtained. As previously mentioned, the main idea of this multimodal approach is to complement both modalities. For instance, adding hesitation from audio to semantic information from text provides valuable information that cannot be obtained by analyzing only one modality.

To combine both modalities, the previously defined unimodal models where used removing the final classification layers. These classification layers are simple dense layers that receive feature vectors to classify them into dementia or healthy categories. The feature vectors are the result of processing the raw data, which, in our case, is the text transcription and audio files. These models provide two feature vectors that are then combined into a single vector, adding classification layers to obtain a final multimodal prediction. This way, the prediction takes into account information regarding both modalities.

This type of combination is called late-stage fusion since the data is processed, feature vectors are obtained, and then they are combined. If the data were more similar, other methods could have been chosen, such as early-stage fusion, which involves combining the data before processing it.

The basic architecture of this model can be appreciated in figure 4.3.

4.1.5 Other proposed models

As we mentioned before, the best results on this dataset were achieved by combining text features with the POS-tagging of the text (Karlekar et al., 2018). This is the reason why we have decided to try using the POS-tagging of the text as well to see what results we obtain. After obtaining the POS-tagging of the text by using the Python library spaCy [2], the features have been introduced in an embedding and in a bidirectional LSTM. After this LSTM, a dense layer for a final classification has been used.

Since the CHAT format of the transcriptions has a lot of information and not only the plain text, there are special flags that represent, for example, a pause in the patient's response or a mistaken word. As some symptoms of dementia are having difficulties finding certain words, this can lead to pauses to think, or using mistaken words, these special flags may give us some relevant information for this task. These special tokens have been counted and compared between control and dementia patients, the figure 4.4 shows how these flags are present in the different transcriptions of healthy people and people with dementia.

[2] `https://spacy.io/`

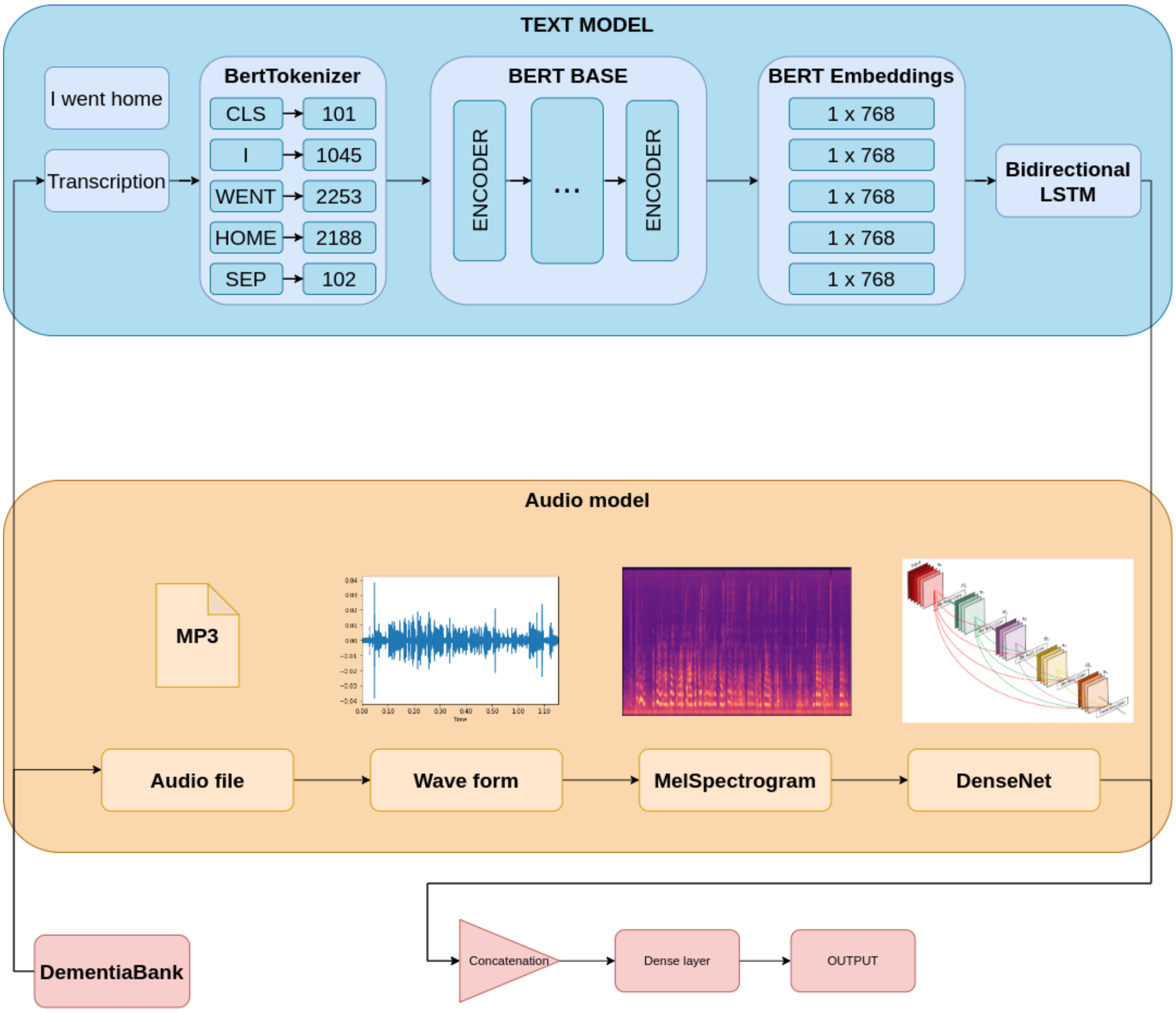


Figure 4.3: Architecture of the multimodal model

Among other flags, the ones that have shown differences between control and dementia patients are: *repetitions*, *retracing*, *pauses* and *unintelligible*. Other flags such as doubts have not shown a big difference between both.

4.2 Automatic Emotion Detection in Aphasia patients

For this task, we developed a pipeline for automatic speech recognition and speaker differentiation of the video recordings in the AphasiaBank dataset. The pipeline consists of several stages and has been applied to each sample of the dataset. First, we extract the audio information from the video and store it. Using the Whisper model developed by OpenAI, we transcribe the recording, resulting in two files: a plain transcription file and a file with transcription and time-lapse of the transcripted sentences. The latter is used for further processing.

Next, we use the speaker-diarization Bredin et al. (2020) Bredin & Laurent (2021) model

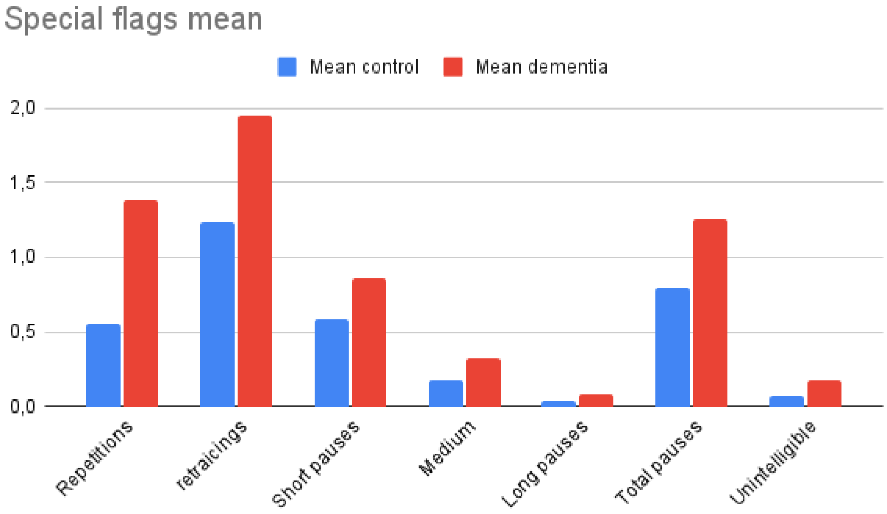


Figure 4.4: Special flags mean

provided by the HuggingFace library to differentiate between the patient and the clinician. This model enables us to obtain a time-lapse of when each speaker is talking. Both models are transformer-based, which has significantly improved the accuracy of the pipeline. Using the output from both models, we obtain a final transcription of what each speaker says. In order to distinguish between the patient and the clinician, we propose to identify the patient as the person who speaks for a longer duration in the recordings. This approach is based on the fact that the recordings are primarily focused on the speech of the patients, who are expected to speak more than the clinicians. In this scenario, the role of the clinicians is to facilitate the conversation and provide assistance to the patients when necessary.

For emotion recognition, we extract the time-lapse where the patient is listening to the clinician, and only keep the video frames during this period. The pipeline architecture is shown in the figure 4.5. Overall, our pipeline provides an efficient and accurate method for processing audio recordings and extracting important information for further analysis.

Once we have identified the video frames where the patient is listening to the clinician, we utilize the DeepFace Serengil & Ozpinar (2020) Serengil & Ozpinar (2021) Serengil & Ozpinar (2023) library's model to extract relevant information from facial expressions, as shown in the figure 4.6. While this model can provide information about age, sex, and race, our focus is solely on the emotions conveyed through facial expressions. The model identifies emotions such as anger, disgust, fear, happiness, sadness, surprise, and neutral. Those emotions are the previously mentioned in section 2.3. We will use this information to develop a method for analyzing the emotions conveyed in each sample. The other relevant information obtained through transcription and speaker differentiation with time lapses will not be used in this project. However, we will keep this information for future works.

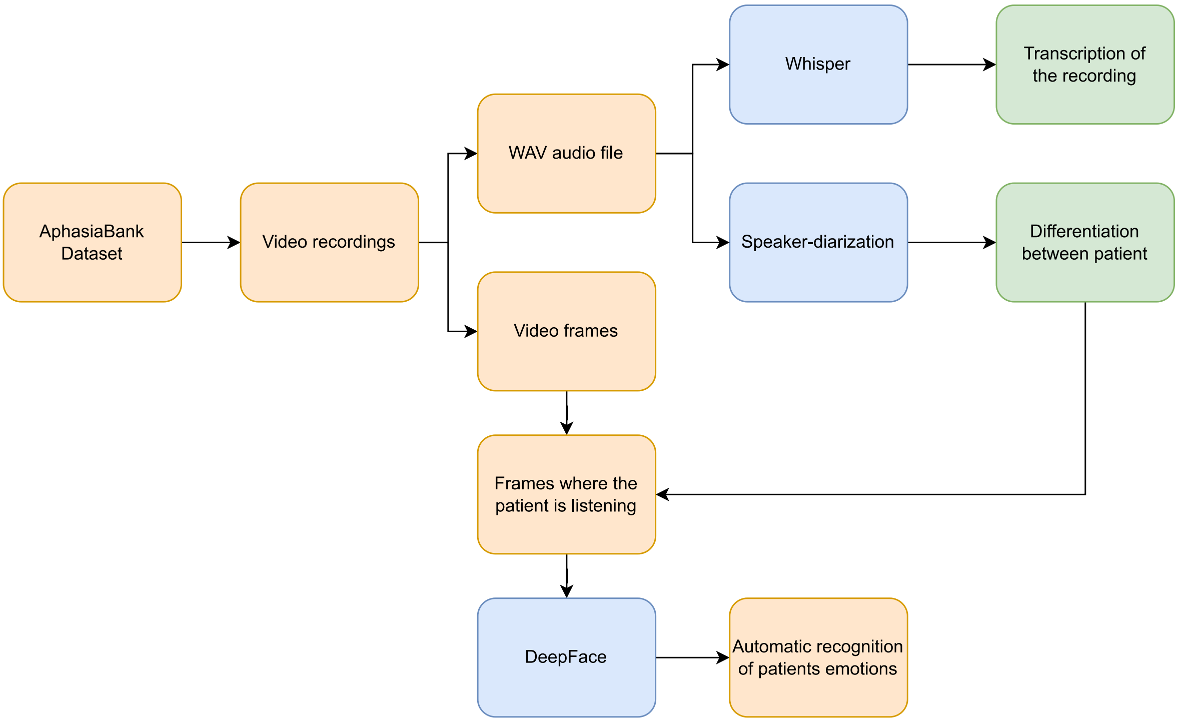
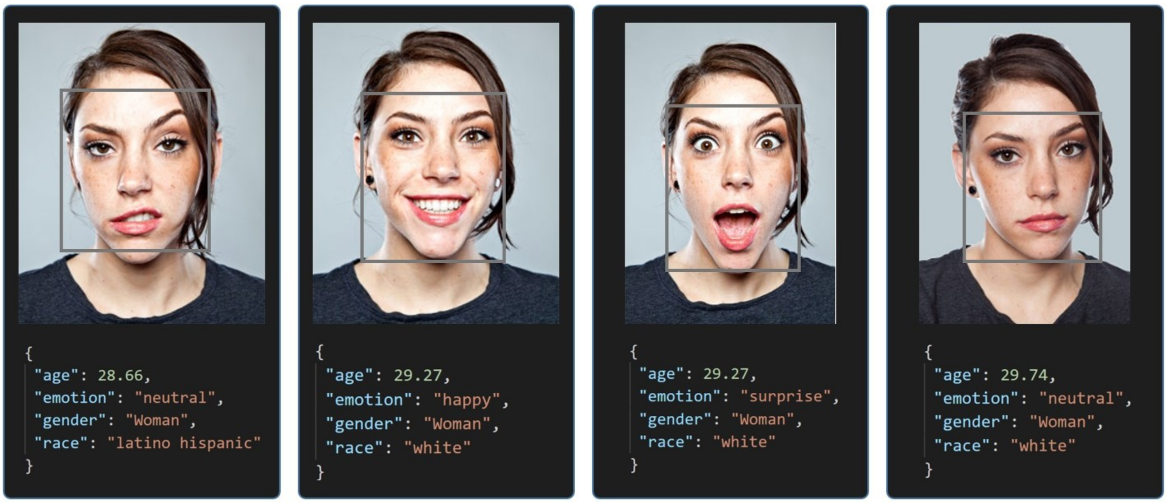

AphasiaBank Dataset
Video recordings
WAV audio file
Video frames
Whisper
Speaker-diarization
Transcription of the recording
Differentiation between patient
Frames where the patient is listening
DeepFace
Automatic recognition of patients emotions

Figure 4.5: Architecture of the proposed pipeline



```
{
  "age": 28.66,
  "emotion": "neutral",
  "gender": "Woman",
  "race": "latino hispanic"
}
```

```
{
  "age": 29.27,
  "emotion": "happy",
  "gender": "Woman",
  "race": "white"
}
```

```
{
  "age": 29.27,
  "emotion": "surprise",
  "gender": "Woman",
  "race": "white"
}
```

```
{
  "age": 29.74,
  "emotion": "neutral",
  "gender": "Woman",
  "race": "white"
}
```

Figure 4.6: Example of use of DeepFace library, from Serengil & Ozpinar (2020) Serengil & Ozpinar (2021) Serengil & Ozpinar (2023)

5 Experimentation

After the proposal of all the different models that were planned to implement, in this next step, our aim is to actually implement them and test the results in order to obtain final conclusions. This chapter is structured as follows: in section 5.1, the metrics obtained in the dementia prediction task will be discussed; in section 5.2 we will analyze in depth the DementiaBank corpus; section 5.3 will introduce the explainability of the best prediction model over the DementiaBank dataset; finally, in section 5.4 the results from the automatic emotion recognition of aphasia patients will be shown.

5.1 Obtained metrics in dementia prediction

In order to test the efficacy of each proposed model, we have focused on the accuracy obtained working over the defined test set. This accuracy measures how many times the model makes a correct prediction on if the patient has dementia or not, the main purpose of this project.

During the implementation and further testing of the proposed models, we have reached to early conclusion in the models proposed in section 4.1.5. These models were planned to analyze the POS-Tagging features of the model as well as the number of special flags such as *repetitions*, *retracing*, *pauses* and *unintelligible*. Even though there were promising results obtained by other recent works with the use of these POS-Tagging features, in our own tests, we did not achieve any good results. In this case we could not even achieve any good result neither in the training set nor in the test set. The results obtained by this model were around a 55% of accuracy in both sets. This fact of being too close to half of the accuracy and not even learning from the training set, leads us to think that the model was making random guesses on its input. For this reason, we have early discarded this model for future attempts to combine with the rest of the models.

The other discarded model has been the model which focused on the special flags previously mentioned. In this case, the statistics after analyzing the whole dataset showed us that there was a difference between healthy and dementia patients in the mean of the flags. This difference represents that dementia patients tend to have more flags like *repetitions*, *retracing*, *pauses* and *unintelligible* than healthy people. Even though these results were obtained, after deploying the model and testing it, it got similar results than the POS-Tagging model, not learning neither from the training set nor the test set. Due to this reason, we have also discarded this model for the future combination with other working models.

The next tested model has been the audio model. One special remark of this test has been the usage of different pretrained CNNs models, where we can find DenseNet, ResNet and MobileNet. All these implementations were done using the torchvision library, where the models are available in order to quickly implement them. After the implementation and testing of the three of them, we have concluded that the best results were obtained by the DenseNet model, achieving a 73.49% of accuracy.

Once we have tested the audio model, our next test step is to implement the text models and see how they perform. As commented before, there have been proposed two different models whose base is the BERT model. The first one was by using the representation of the whole sentence, the "[CLS]" token, as an input for a dense layer. This model also achieved good results, obtaining an accuracy of 84.3%. The other proposed model was by using the tokens of the whole sentence, which are the representation of the words, as an input to a bidirectional LSTM model. After this LSTM model, the output will be used as input for a final dense layer. This model achieved better results, obtaining a 90.36% of accuracy.

Finally, and as one of the objectives of this project, we wanted to test how combining different modalities into one single model performed and see if it can improve the previously obtained results. In this case, only the models that worked were proposed for this combination. In those models we include the audio and the two text models. The others have been discarded because they do not give us good results individually, so it will worsen the other ones. As we have one audio model and two text models, the multimodal approach has been combining the audio model with the two different text models. To achieve this multimodal factor, the outputs of each model have been concatenated before a final dense layer, other methods of multimodal usages, such as a weighted prediction could have been used as well. All the results obtained from these tests can be seen in the table 5.1.

Model	Description	Accuracy
Audio	Mel Spectrogram + CNN (DenseNet)	73.49%
Text 1	BERT embeddings + dense layer	84.33%
Text 2	BERT embeddings + bidirectional LSTM + dense layer	**90.36%**
Multimodal 1	Audio + Text 1	84.33%
Multimodal 2	Audio + Text 2	86.65%

Table 5.1: Comparison of the accuracy obtained in the test set of the different implemented models

As can be seen from the table, the usage of this multimodal factor has not improved our previously obtained results from the text model, even getting worse results than the ones obtained with the text models. In this case, as we achieve very good results from one modality, it is a hard task to improve those results. The best results have been obtained from the text model where all the BERT embeddings have been used, improving all the other models. One remarkable detail that we can obtain from these experiments is that we have obtained significantly better results from the text part of the dataset than from the audio one.

Finally, and having a more detailed view of the best results, in the figure 5.1 we can see how well it performed over dementia and healthy cases separately.

As can be seen from the confusion matrix obtained from the testing of the text model,

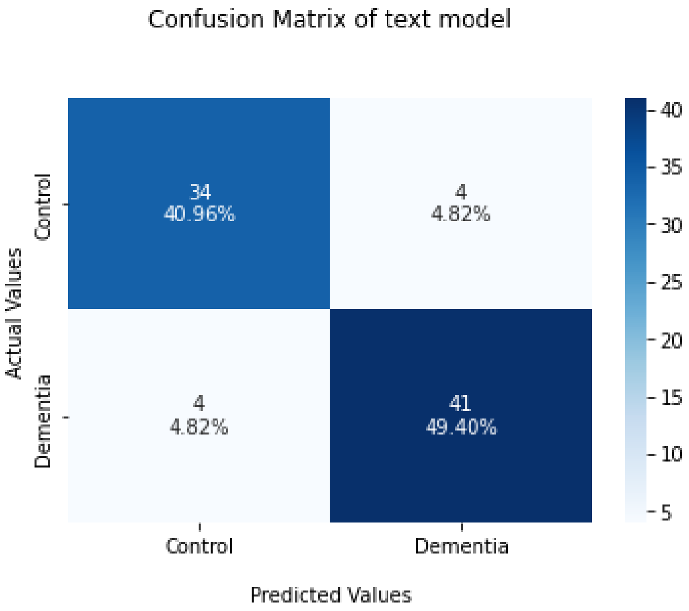


Figure 5.1: Confusion matrix of the text model results, where the percentage value represents the proportion of the square in the test set

the test set has 38 samples of control patients and 45 samples of dementia patients. The text model has been able to correctly predict 34 cases of healthy people but has missed 4 cases predicting dementia where the patient is healthy. In the case of the dementia cases, it obtains similar results, correctly predicting 41 cases of dementia and makes a mistake in 4 cases where the model predicts the patient is healthy where they are not. As the dementia cases are more common in the whole dataset and also in this subset, the percentage of correct guesses is bigger in the cases of dementia, even though it misses the same amount in both, 4 cases. For this reason, we could conclude that the text model gets slightly better results while analyzing cases of dementia rather than health.

From this confusion matrix, we can obtain several statistics about the results obtained using this model. For these statistics, we need to define the four squares of the matrix.

- **True Positive - TP:** In this case, the expected value is dementia, and we also obtain dementia from the model. In other words is correctly guessing the dementia cases, which in this case is the square with the 41 value.

- **True Negative - TN:** This is a very similar case, but this represents where it correctly guesses the healthy patients, which in this case is the square with the 34 value.

- **False Positive - FP:** This case is different from the others, in this case, the model will predict dementia, but the real value is healthy, and it will miss. In this case, is the square with the 4 value which is in the top-right position.

- **False Negative - FN:** This is the opposite case, where the model predicts to be healthy but actually the patient has dementia. In this case, is the square with the 4 value which is in the lower-left position.

After having these values into account, we can obtain statistics apart from the previously calculated accuracy such as precision, recall, F1 and accuracy. The first statistic that will be performed is precision. This metric is used in order to obtain the quality of our model when it predicts dementia:

$$precision = \frac{TP}{TP + FP} = \frac{41}{45} = 0.9111 \tag{5.1}$$

The precision in this case obtained is 0.911, which represents that 91.11% of our dementia detections will be correct. On the other hand, we have recall, which is a metric to obtain the quality of our model when it has to predict the real dementia patients. In other words, from the total dementia patients samples, how many patients is able to correctly predict as dementia the model. The equation that has to be computed is the following:

$$recall = \frac{TP}{TP + FN} = \frac{41}{45} = 0.9111 \tag{5.2}$$

The recall obtained in this case, as well as in the previous, is 0.9111, which represents that 91.11% of our detections when we have to deal with a dementia case are correct. Finally, the F1 metric is a metric that combines both recall and precision, in our case, it is not really significant since both metrics give us the same results. The equation is the following:

$$F1 = 2 * \frac{recall * precision}{recall + precision} = 0.9111 \tag{5.3}$$

As both precision and recall give us 0.9111, the F1 score also gives us the same result. In the figure 5.2 a comparison of the different metrics can be seen. As previously mentioned, the model performs slightly better in the prediction of dementia cases rather than healthy. This fact can be seen in this comparison since the recall, precision and F1 score gets an improvement over the accuracy. But this improvement is very low, it will become in a not significant difference in the real world situations.

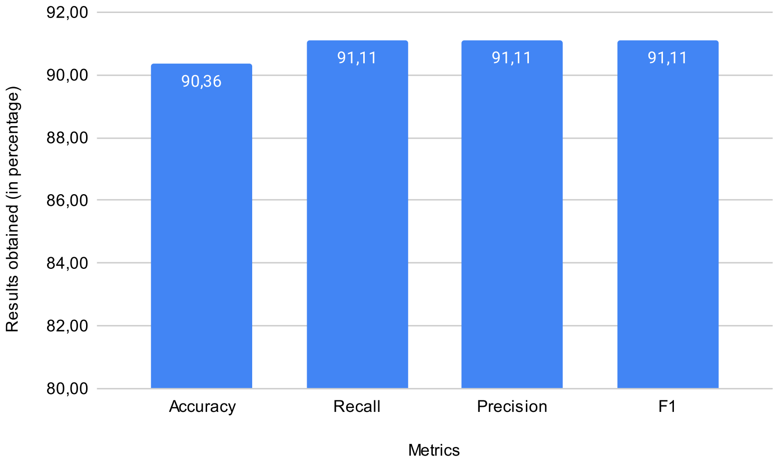


Figure 5.2: Comparison of the different metrics

5.2 DementiaBank corpus analysis

Taking into account the good performance achieved by the textual model (BERT), this section presents an analysis of the textual part of the dataset used in the experiments to provide a better understanding of their nature. The goal is to identify clues that makes text-only models to have such a good performance in this multimodal dataset.

First of all, the length (number of words) of the texts provided by healthy patients and patients with dementia were analysed. The result of this analysis is shown in Table 5.2.

Measure	Dementia	Control
Mean	450.94	503.90
Standard deviation	242.15	280.28
Min	92	109
Max	1654	2421
Median	404	432
25th percentile	279	326
75th percentile	556	613

Table 5.2: Central tendency measures for the length (number of words) of the conversations of dementia and control patients.

The table shows that, on average, the conversations uttered by dementia patients are 10% shorter than those uttered by healthy patients (450.94 words and 503.90 words, respectively). There is also more homogeneity in terms of the number of words used in patients with dementia, as they show a lower standard deviation. Looking at the median, which is less sensitive to the presence of outliers, it also indicates the presence of shorter texts in the case of patients with dementia (404 words compared to 432 words of healthy patients). Figure 5.3 and Figure 5.4 show the distribution of the length of the conversations in both dementia and control patients in more detail.

5.2.1 N-gram frequency count

Before performing additional analysis it was necessary to carry out a series of preprocessing tasks to clean the data. To this end, the NLTK[1] library was used to lowercase all the texts, remove punctuation marks, and remove *stopwords*, that is, commonly used words in English language that do not provide useful information (e.g. “the”, “a”, “and”).

The first task consisted of extracting n-grams of words from both dementia and healthy datasets. Specifically, unigrams and bigrams were identified for further analysis. Then, a straightforward count of the frequencies of n-grams in conversations was carried out. Table 5.3 shows the 50 most frequent unigrams and bigrams for dementia and healthy individuals.

Taking a closer look at these lists, many words are common between the two different classes, but there are differences in some terms. As analyzed in Section 5.3, one token that can be observed in both classes but is much more common in the dementia class is the word `well`. Another remarkable token to distinguish between the classes is the word `something`, which is present in the list of unigrams of dementia patients but not in the control group.

[1] https://www.nltk.org/

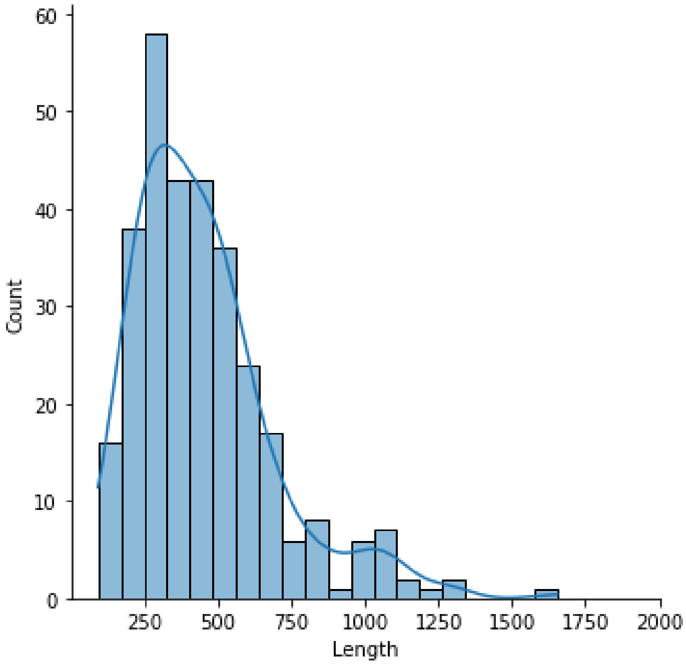


Figure 5.3: Histogram of the length (number of words) of conversations by dementia patients.

This may arise from the difficulty some dementia patients experience in properly recognizing items in the image, causing them to use a generic word instead.

The list of bigrams also reveals interesting patterns. For instance, patients with dementia tend to use the verbs `reaching` and `getting` with the noun `cookies` with almost equal frequency. In contrast, healthy patients tend to use the verb `reaching` more frequently when describing that part of the image. Another notable example is the word `gonna`, which only appears in the bigrams list of dementia patients, used in combination with the words `he's` and `fall` to describe the little boy standing up on the stool.

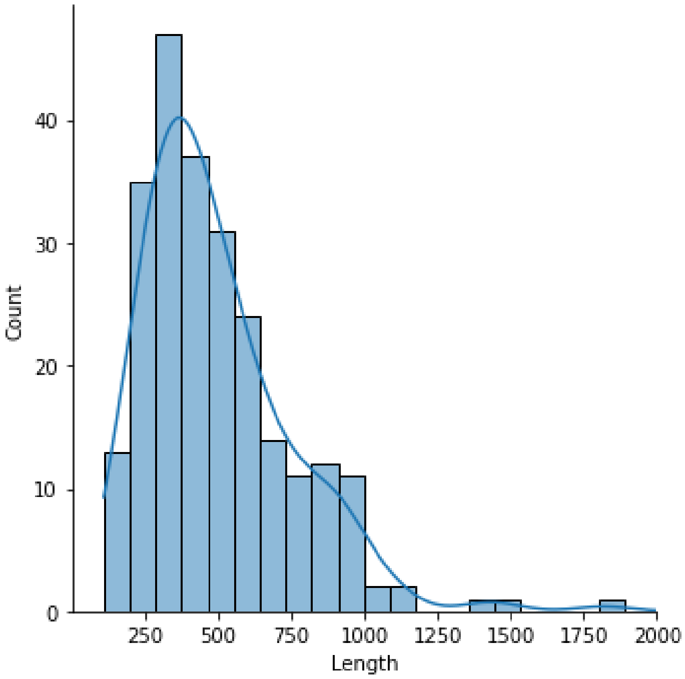


Figure 5.4: Histogram of the length (number of words) of conversations by control (healthy) patients.

	Dementia
Unigrams	uh, cookie, dishes, jar, he's, water, little, sink, stool, boy, cookies, girl, floor, well, there's, drying, mother, laughs, running, falling, washing, +, gonna, fall, get, see, getting, window, water's, one, like, going, reaching, hand, standing, um, looks, boy's, sister, got, trying, out_of, something, that's, oh, two, mother's, looking, overflowing, dish
Bigrams	cookie jar, drying dishes, washing dishes, little girl, little boy, looks like, gonna fall, he's gonna, uh uh, reaching cookie, trying get, getting cookies, dishes uh, jar he's, water running, water's running, cookies out_of, onto floor, stool he's, getting cookie, little boy's, get cookies, running sink, uh +, two cups, running floor, dishes sink, looking window, mother washing, jar uh, uh stool, sink running, get cookie, cookies cookie, sink uh, sink overflowing, he's falling, uh sink, stool falling, out_of cookie, boy getting, water run, dishes water, uh mother, jar little, boy uh, water floor, falling stool, taking cookies, uh there's
	Control
Unigrams	uh, cookie, sink, dishes, stool, water, jar, boy, little, mother, he's, girl, drying, window, um, cookies, running, reaching, hand, open, there's, floor, standing, falling, out_of, overflowing, one, getting, like, looks, water's, see, washing, mother's, fall, curtains, outside, two, well, sister, going, dish, plate, kitchen, looking, cups, onto, get, cupboard, counter
Bigrams	cookie jar, drying dishes, little girl, reaching cookie, washing dishes, looks like, little boy, water running, sink overflowing, out_of cookie, onto floor, mother drying, two cups, cookies out_of, girl reaching, dishes water, window open, dishes sink, falling stool, water's running, getting cookie, getting cookies, looking window, standing water, drying dish, sink running, jar he's, mother washing, out_of sink, taking cookies, window's open, running sink, fall stool, stool falling, standing stool, get cookie, stool he's, door open, mother's drying, stool tipping, dishes water's, left hand, uh mother, overflowing sink, cookies cookie, water overflowing, hand cookie, uh uh, stealing cookies, stool uh

Table 5.3: List of 50 most frequent n-grams for dementia and control (healthy) patients.

5.2.2 Polarized Weirdness Index

In addition to the n-gram frequency count, an analysis was carried out by computing the Polarized Weirdness Index (PWI) (Poletto et al., 2021) of the unigrams and bigrams in both dementia and healthy texts in order to extract the most characteristic words of each one. The PWI is a variant of the Weirdness Index (WI) (Ahmad et al., 1999), which is a metric to retrieve words characteristics of a special language with respect to their common use in general language. The intuition behind WI is that a word is highly weird in a specific corpus if it occurs significantly more often in that context than in a general language corpus. Given a specialist and a general corpus, the metric can be described as the ratio of its relative frequencies in the respective corpora. In the case of PWI, the metric compares the relative frequencies of a word as it occurs in the subset of a labeled corpus by one value of the label against its complement. In the present work, the PWI is used to compare the prevalence of words in dementia and healthy utterances.

”Table 5.4 displays the top 20 unigrams and bigrams extracted from the samples of dementia and healthy participants based on the PWI metric. As mentioned in the previous subsection, the bigrams of dementia patients contain the word `something`, while it is absent in the healthy patients’ list.

Dementia		Control	
Unigrams	Bigrams	Unigram	Bigrams
spilled	water run	nose	mother know
whatever	let water	daydreaming	open there's
g	got cookie	who's	blowing curtains
fell	boy's cookie	sort	um boy
way	girl wants	process	wind blowing
j	um stool	growing	getting feet
different	uh well	believe	grass growing
come	sink well	shirt	kitchen cabinets
begging	laughs he's	blowing	children getting
thing	lady washing	wind	out_of faucet
yeah	going uh	action	open curtains
wa	floor laughs	wearing	another one
wash	he's cookie	beside	cookie girl
picture	run sink	overflow	plate two
hurt	get hurt	high	mother standing
yet	jar mother's	somewhere	kitchen mother
head	uh something	brother's	okay boy
mop	there's something	raising	water looking
legs	dishes let	sort_of	sink boy
spigot's	dishes laughs	presume	standing sink

Table 5.4: List of 20 most relevant unigrams based on PWI for the dementia and control (healthy) samples in the dataset.

5.2.3 Feature selection

In addition to frequency count, a feature selection procedure using χ^2 (Pearson, 1992) was applied to identify what unigrams were considered as most relevant in order to differentiate between dementia and healthy texts. Before applying χ^2 it is necessary to transform every post into a numerical vector. The TF-IDF weighting schema was used to obtain a number representing the frequency of the token in the post (TF) and its prevalence in the dataset (IDF). The number of dimensions of each post vector is equal to the length of the vocabulary of the corpus, i.e., each dimension corresponds to one token. The value of the dimension is the TF-IDF weight if the token exists in the post or 0 otherwise. Texts were preprocessed in advance as in the previous analysis.

Table 5.5 shows the 50 best unigrams in order to differentiate dementia from healthy texts according to χ^2. This list shows some tokens that were appreciated with the previously obtained n-grams and also additional ones that are not as commonly used in the corpus but result in a good key to differentiate dementia.

`here`, `is`, `blowing`, `open`, `this`, `overflowing`, `laughs`, `down`,
`window`, `reaching`, `wind`, `out_of`, `quiet`, `finger`, `action`, `while`,
`moving`, `mouth`, `who`, `spilled`, `stepping`, `gonna`, `run`, `mother`,
`the`, `um`, `curtains`, `nose`, `be`, `something`, `her`, `faucet`, `thing`,
`breeze`, `about`, `they`, `growing`, `are`, `counter`, `well`, `get`, `hm`,
`hand`, `yeah`, `standing`, `fell`, `good`, `whatever`, `wa`, `oh`

Table 5.5: List of 50 most relevant unigrams according to χ^2.

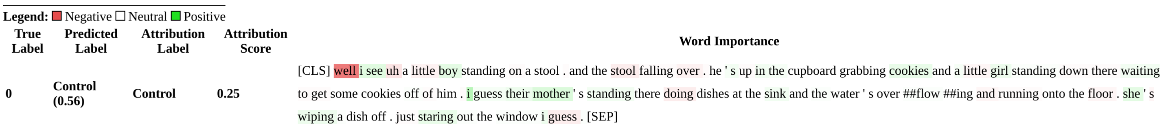
5.3 Dementia prediction model explainability

Following the premise of the previous section, with the aim of identifying why textual models work so well in the multimodal corpus, the *Transformer Interpreter* software (Pierse, 2021) was used to obtain more information about how BERT makes its predictions.

This tool provides more insights and information about how Transformer models make their decisions based on a given input. A specific weight is obtained for each token, which represents how that token influences the final decision of the model in the classification tasks. Since this model works on textual modality, each token is a word. As this classification task has only two possible outcomes (dementia or healthy), whenever a token influences positively the decision for one, it will influence negatively the decision for the other.

After analysing the tokens influence on the test set, the most significant tokens identified were those used when the patient starts to describe the image. The token `Well` at the beginning of a sentence influences positively when the model predicts dementia. This in turn means that the use of that token influences negatively when the model predicts a healthy patient. In Figure 5.5 and Figure 5.6, there are two examples of how the word `Well` influences both decisions made by the model. In that representation, the words highlighted in red will influence negatively the decision and those highlighted in green will influence it positively. The more intense the color, the more important the word has in the final decision.

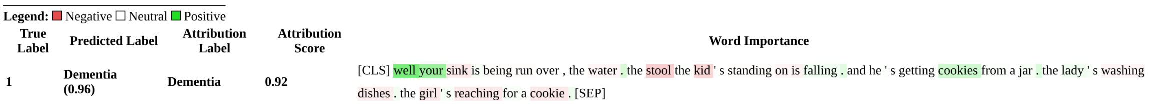
Similarly, the use of other tokens at the beginning of a sentence positively influences the



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	Control (0.56)	Control	0.25	[CLS] well i see uh a little boy standing on a stool . and the stool falling over . he ' s up in the cupboard grabbing cookies and a little girl standing down there waiting to get some cookies off of him . i guess their mother ' s standing there doing dishes at the sink and the water ' s over ##flow ##ing and running onto the floor . she ' s wiping a dish off . just staring out the window i guess . [SEP]

Figure 5.5: Influence of the word `Well` in the prediction of a healthy patient.

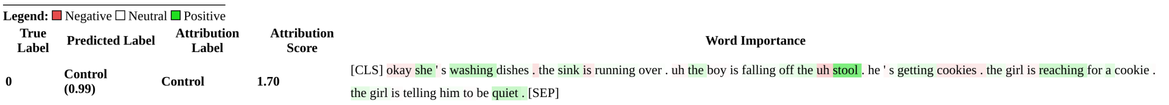


Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	Dementia (0.96)	Dementia	0.92	[CLS] well your sink is being run over , the water . the stool the kid ' s standing on is falling . and he ' s getting cookies from a jar . the lady ' s washing dishes . the girl ' s reaching for a cookie . [SEP]

Figure 5.6: Influence of the word `Well` in the prediction of a dementia patient.

prediction of dementia, as shown in the case of `Okay`, which can be observed in Figure 5.7 and Figure 5.8. These tokens are used to introduce the sentence before describing the image, and the model gives great importance to them for the final decision of predicting dementia or not. This behavior can be observed in several other words used at the beginning of a sentence, such as `So`.



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	Control (0.99)	Control	1.70	[CLS] okay she ' s washing dishes . the sink is running over . uh the boy is falling off the uh stool . he ' s getting cookies . the girl is reaching for a cookie . the girl is telling him to be quiet . [SEP]

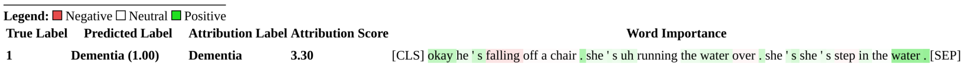
Figure 5.7: Influence of the word `Okay` in the prediction of a healthy patient.

This is related to some symptoms of dementia, such as having difficulties finding the right words to use or expressing themselves properly. These difficulties can lead to the use of auxiliary words like those explained before, with the intention of gaining more time while finding the right words to use. Other tokens used similarly that the model takes into account are expressions used to generate pauses, such as `um`, `oh` and `uh`, among others.

Another behavior observed by this model is the influence of expressing uncertainty. One example of this is by using the verb `Guess` or the adverb `Apparently`, both of which have a positive influence on predicting dementia.

The trend of hesitation and uncertainty in speech can be reflected by analyzing the length of the audio samples in the datasets. Recordings of people suffering from dementia usually are longer than healthy ones, having around 20% more duration in the files. This can be visually appreciated in Figure 5.9, where a histogram of the lengths is displayed comparing both control and dementia classes. In the figure, even though there are fewer samples of healthy patients, there are more samples in the leftmost part, representing less time of audio. And in contrast, in the rightmost part, the dementia classes predominate over the other.

One remarkable point is that the same word used in different contexts can have different influences. This is an important factor because not only are the words important, but also the way they are used in each situation. Additionally, the influence of a token can also vary from one sample to another, resulting in different scores between patients. Another remarkable



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	Dementia (1.00)	Dementia	3.30	[CLS] okay he ' s falling off a chair . she ' s uh running the water over . she ' s she ' s step in the water . [SEP]

Figure 5.8: Influence of the word `Okay` in the prediction of a dementia patient.

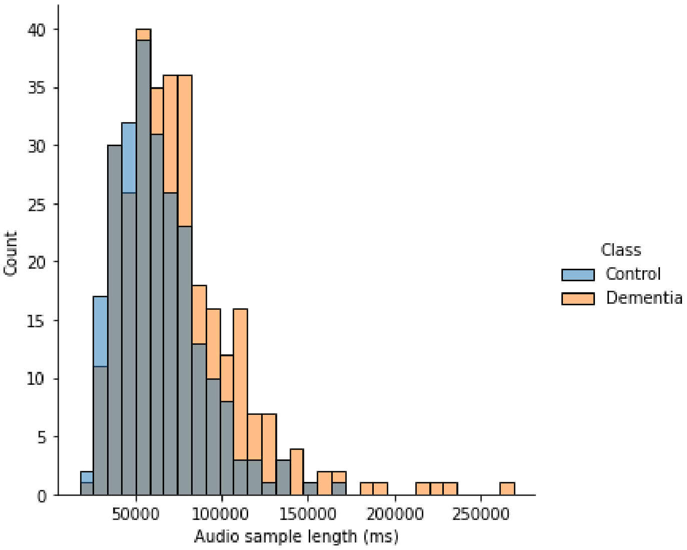


Figure 5.9: Comparison of lengths of healthy and dementia patients records.

point is that although there may be cases where a word has a very negative influence, the model can still predict the other class. This is exemplified by the use of `Well` for a healthy patient prediction.

5.4 Automatic emotion recognition in aphasia patients experimentation

In order to evaluate this task, the diarisation component was tested as the first step. One hundred random samples were selected from the dataset, with patients and clinicians properly differentiated, and were manually reviewed. The pipeline was tested by analyzing the initial frames of the selected samples along with the diarisation, as the entire files were not analyzed due to some samples being up to an hour long. The pipeline correctly distinguished 94 out of the one hundred samples, resulting in a 94% accuracy rate in distinguishing between patients and clinicians. The samples that were incorrectly distinguished were those where the clinician had to speak extensively to maintain the conversation and assist the patients who were not able to communicate fluently. In such cases, the pipeline identified the clinician as a patient since it considers the person who speaks more as the patient. Additionally, the low recording quality was another reason for incorrect labeling. Nonetheless, the pipeline generally distinguishes the majority of cases correctly. On the other hand, no evaluation has been done over the transcription text, since it has not finally used in this work.

The other evaluation metric involved comparing the results obtained from both aphasic and

Table 5.6: Average emotion detection in the different corpora of the dataset

Corpus		Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Control	**Wright**	0.153	0.021	0.238	0.041	0.409	0.044	0.094
	Capilouto	0.053	0.003	0.121	0.001	0.816	0.000	0.006
	Kempler	0.019	0.005	0.262	0.224	0.459	0.008	0.022
	Richardson	0.000	0.000	0.062	0.655	0.130	0.000	0.153
	MSU	0.083	0.000	0.088	0.139	0.475	0.001	0.214
	Total	**0.122**	**0.003**	**0.221**	**0.126**	**0.329**	**0.025**	**0.172**
Aphasia	**Wright**	0.160	0.000	0.137	0.343	0.127	0.001	0.232
	Thompson	0.113	0.000	0.105	0.347	0.169	0.040	0.226
	Adler	0.088	0.000	0.532	0.041	0.105	0.046	0.187
	UNH	0.344	0.001	0.343	0.045	0.235	0.001	0.031
	STAR	0.554	0.008	0.074	0.021	0.333	0.000	0.011
	TAP	0.359	0.019	0.267	0.032	0.265	0.002	0.056
	Garrett	0.051	0.000	0.220	0.212	0.068	0.000	0.449
	Whiteside	0.350	0.001	0.153	0.143	0.221	0.031	0.100
	Tucson	0.114	0.000	0.066	0.101	0.481	0.001	0.238
	Fridriksson	0.040	0.000	0.109	0.050	0.450	0.007	0.343
	UCL	0.296	0.000	0.139	0.024	0.198	0.029	0.313
	TCU	0.137	0.001	0.073	0.025	0.742	0.000	0.022
	Elman	0.256	0.000	0.089	0.085	0.549	0.000	0.021
	CMU	0.231	0.000	0.132	0.477	0.059	0.002	0.098
	Kurland	0.572	0.001	0.035	0.092	0.250	0.000	0.050
	TCU-bi	0.079	0.000	0.062	0.339	0.316	0.000	0.204
	Kempler	0.189	0.006	0.067	0.388	0.298	0.004	0.048
	Kansas	0.132	0.000	0.254	0.113	0.373	0.000	0.127
	SCALE	0.296	0.004	0.109	0.112	0.261	0.016	0.201
	ACWT	0.119	0.004	0.072	0.341	0.100	0.017	0.347
	Wozniak	0.266	0.002	0.088	0.039	0.322	0.041	0.242
	MSU	0.064	0.004	0.014	0.073	0.484	0.000	0.361
	Williamson	0.025	0.001	0.001	0.019	0.226	0.000	0.728
	BU	0.509	0.002	0.152	0.053	0.127	0.024	0.134
	Total	**0.201**	**0.006**	**0.150**	**0.109**	**0.32**	**0.013**	**0.198**

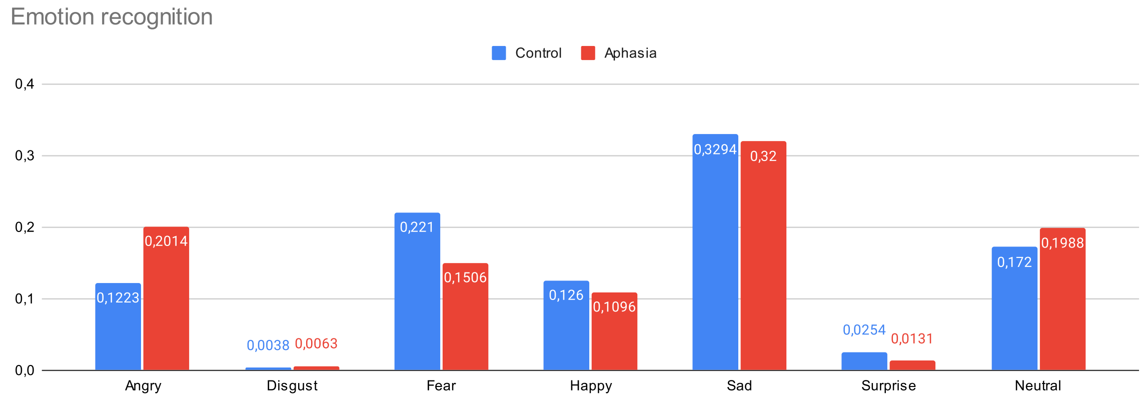


Figure 5.10: Mean of emotions represented in the analysis over patients while listening to clinicians' speech

healthy patients in terms of emotion recognition. The results are shown in the figure 5.10 and in more detail in table 5.6. The most notable difference was observed in the mean value of the ”angriness” emotion. This finding was not surprising, as patients may experience frustration and anger due to difficulties in understanding the speech of the clinician. Similarly, although it represents a small proportion of the mean of the emotions, the aphasic patients showed double the proportion of ”disgust” emotion compared to the healthy patients. Other significant differences were observed in the proportions of ”fear,” ”surprise,” and ”neutrality” emotions. The lower proportion of ”fear” and ”surprise” and the higher proportion of ”neutral” emotion may be due to the difficulty in understanding the speech. In the case of not understanding the clinician's speech, patients may not show fear or surprise as healthy patients would when they fully comprehend a sentence and are surprised by its content. Additionally, the higher proportion of ”neutral” emotion may result from the lack of expression due to poor speech recognition.

6 Conclusion

This final chapter summarizes the conclusions obtained from the work of this project. It is structured as follows, firstly, the conclusions obtained from this work will be described 6.1; then, in section 6.2 will briefly introduce the projects that can be built in future works. Finally, we mention publications obtained from this work results in section 6.3.

6.1 Conclusions

This project aimed to develop a system integrated into a mobile robot that analyzes data in patients' houses, specifically targeting elderly individuals and those with cognitive diseases such as dementia or aphasia. The analysis of this data had two main objectives.

Firstly, it aimed to enable early diagnosis of these diseases, which has the potential to significantly enhance the quality of life by initiating timely professional treatment. The increasing elderly population has resulted in a rise in the number of dependent individuals, leading to a surge in healthcare-related research endeavors. Our experiments and findings from recent studies confirm that deep learning algorithms can provide substantial help in the healthcare domain, not only for dementia but also for various other diseases. This help can be derived from early disease diagnosis as proposed in this thesis.

The second objective of the robot was to automatically recognize emotions, thereby providing enhanced support to patients during challenging situations. This capability has the potential to improve their emotional well-being and promote positive interactions with others.

During our research, the availability of dementia datasets was limited due to the sensitivity of the disease. Hence, we utilized the DementiaBank dataset, which contains audio files of patient interviews describing an image commonly used in cognitive research projects—the cookie theft picture. Additionally, we leveraged the transcriptions of these audio files, incorporating data from two modalities: audio and text. We proposed several models for dementia detection, including the use of MelSpectrograms and CNNs for audio classification, Transformers (specifically the BERT model) for text classification, and a multimodal model combining both modalities. These models were thoroughly tested on a test set of the dataset to assess their performance across different modalities. Through this experimentation, we concluded that the text modality yielded superior results, achieving an accuracy of up to 90.36% in dementia detection. To explain this result, we conducted an analysis of the textual part of the dataset, employing an explainability approach to determine the influence of specific words in identifying the nature of the patient (dementia or healthy).

Regarding the selection of the aphasia dataset, we encountered a similar limitation in terms of available options. Ultimately, we chose the AphasiaBank dataset, which comprises video recordings of patients with aphasia. The effectiveness of the pipeline was evaluated in the distinction between both speakers, in this case, the patient and the clinician. This metric has been evaluated to properly analyze the frames where the patient is listening to the clinician's speech, with the aim of recognizing the patients' emotional mood at those moments. The evaluation has been done by manually reviewing the initial frames of one hundred randomly selected video samples from the dataset, checking whether the pipeline correctly distinguishes when the clinician is speaking to the patient. The pipeline was also employed to recognize emotions in both healthy individuals and those with aphasia. Emotion detection from patients' facial expressions was accomplished using the DeepFace library. The study revealed that individuals with aphasia express emotions differently than healthy individuals when listening to speech, primarily due to their difficulties in understanding and expressing speech, which adversely affects their mood. Analyzing their emotional state can assist in improving interactions by avoiding conversations that may have a negative impact on their mood.

In conclusion, the developed system showcased the potential of integrating a mobile robot for analyzing data in patients' houses, with applications in early disease diagnosis and emotion recognition. The findings from this project emphasize the value of deep learning algorithms in the healthcare domain, shedding light on their effectiveness in dementia detection. Moreover, the system's ability to recognize and respond to emotions contributes to improving the emotional well-being and overall quality of life for elderly individuals and patients with cognitive disorders.

6.2 Future work

For future work in this field, there are several areas we plan to explore. One direction is to utilize the trained model with a different type of dataset focusing on other diseases that share similarities with dementia, such as Traumatic Brain Injuries. This investigation aims to understand the effects of these diseases on patients and further expand our knowledge in this domain.

Another aspect of future work involves proposing and implementing a more sophisticated system for analyzing patients' facial expressions. This enhanced system would encompass additional features beyond emotions, enabling the identification of various facial expression patterns between healthy individuals and those with Aphasia. Additionally, a deeper analysis in the transcriptions of the collected samples holds promise for identifying patterns in patients' expressions and the content they listen to, which could lead to intriguing avenues for further research. By delving into the transcription analysis, a more profound emotional analysis can be conducted to identify specific types of speech that have a negative impact on patients' mood. It is important to note that facial expressions are not the sole aspect to be analyzed, as observing the evolution of a patient's pose and movements during tasks or conversations can also provide valuable insights. These movements can be affected by cognitive diseases, and leveraging this information can contribute to a more comprehensive understanding of the data.

6.3 Publications

As a result of this thesis and project, we have published our proposal and results in the following journals:

- Neurocomputing, 5.779 impact factor, entitled *A Deep Learning-Based Multimodal Architecture to predict Signs of Dementia*, doi: `https://doi.org/10.1016/j.neucom.2023.126413` Ortiz-Perez et al. (2023).
- 18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023), entitled *Deep Learning-based emotion detection in Aphasia patients*.

Acknowledgment

We would like to thank “A way of making Europe” European Regional Development Fund (ERDF) and MCIN/AEI/10.13039/501100011033 for supporting this work under the MoDeaAS project (grant PID2019-104818RB-I00) and AICARE project (grant SPID202200X139779IV0). Furthermore, we would like to thank Nvidia for their generous hardware donation that made these experiments possible.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from `https://www.tensorflow.org/` (Software available from tensorflow.org)

Ahmad, K., Gillam, L., & Tostevin, L. (1999). University of surrey participation in TREC8: weirdness indexing for logical document extrapolation and retrieval (WILDER). In E. M. Voorhees & D. K. Harman (Eds.), *Proceedings of the eighth text retrieval conference, TREC* (Vol. 500-246, pp. 1–8). Gaithersburg, Maryland, USA: National Institute of Standards and Technology (NIST). Retrieved from `http://trec.nist.gov/pubs/trec8/papers/surrey2.pdf`

Akbari, H., Yuan, L., Qian, R., Chuang, W., Chang, S., Cui, Y., & Gong, B. (2021). VATT: transformers for multimodal self-supervised learning from raw video, audio and text. *CoRR, abs/2104.11178.* Retrieved from `https://arxiv.org/abs/2104.11178`

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A framework for self-supervised learning of speech representations.*

Barney, N., & Bernstein, C. (2023, Apr). *What is face detection and how does it work?* TechTarget. Retrieved from `https://www.techtarget.com/searchenterpriseai/definition/face-detection`

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994, 06). The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *Archives of Neurology*, *51*(6), 585-594.

Beltagy, I., Cohan, A., & Lo, K. (2019). Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, *abs/1903.10676.* Retrieved from `http://arxiv.org/abs/1903.10676`

Bredin, H., & Laurent, A. (2021, August). End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. interspeech 2021.* Brno, Czech Republic.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., … Gill, M.-P. (2020, May). pyannote.audio: neural building blocks for speaker diarization. In *Icassp 2020, ieee international conference on acoustics, speech, and signal processing.* Barcelona, Spain.

Chakraborty, R., Pandharipande, M., Bhat, C., & Kopparapu, S. K. (2020). *Identification of dementia using audio biomarkers.* arXiv. Retrieved from `https://arxiv.org/abs/2002.12788` doi: 10.48550/ARXIV.2002.12788

Chollet, F., et al. (2015). *Keras.* `https://keras.io`.

Clancy, M. (2021, Oct). *A practical guide to choosing between docker containers and vms.* Retrieved from https://www.weave.works/blog/a-practical-guide-to-choosing-between-docker-containers-and-vms

Dementia and language. (2022, Mar). Retrieved from https://www.alzheimers.org.uk/about-dementia/symptoms-and-diagnosis/symptoms/dementia-and-language

Deng, Z., Sun, H., Zhou, S., Zhao, J., Lei, L., & Zou, H. (2018, 05). Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, *145*. doi: 10.1016/j.isprsjprs.2018.04.003

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. Retrieved from http://arxiv.org/abs/1810.04805

Dougherty, A. (2020, Nov). *Magic of the sobel operator.* Towards Data Science. Retrieved from https://towardsdatascience.com/magic-of-the-sobel-operator-bbbcb15af20d

Elbourn, E., Kenny, B., Power, E., & Togher, L. (2019, 09). Psychosocial outcomes of severe traumatic brain injury in relation to discourse recovery: A longitudinal study up to 1 year post-injury. *American Journal of Speech-Language Pathology*, *28*, 1-16. doi: 10.1044/2019_AJSLP-18-0204

Fernández Montenegro, J. M., Villarini, B., Angelopoulou, A., Kapetanios, E., Garcia-Rodriguez, J., & Argyriou, V. (2020). A survey of alzheimer's disease early diagnosis methods for cognitive assessment. *Sensors*, *20*(24). Retrieved from https://www.mdpi.com/1424-8220/20/24/7292 doi: 10.3390/s20247292

Forbes, M., Fromm, D., & Macwhinney, B. (2012, 08). Aphasiabank: A resource for clinicians. *Seminars in speech and language*, *33*, 217-22. doi: 10.1055/s-0032-1320041

GeeksforGeeks. (2020, Mar). *Fast r-cnn: Ml.* Author. Retrieved from https://www.geeksforgeeks.org/fast-r-cnn-ml/

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 ieee conference on computer vision and pattern recognition* (p. 580-587). doi: 10.1109/CVPR.2014.81

Gomez-Donoso, F., Orts-Escolano, S., Garcia-Garcia, A., Garcia-Rodriguez, J., Castro-Vargas, J. A., Ovidiu-Oprea, S., & Cazorla, M. (2017). A robotic platform for customized and interactive rehabilitation of persons with disabilities. *Pattern Recognition Letters*, *99*, 105-113. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167865517301903 (User Profiling and Behavior Adaptation for Human-Robot Interaction) doi: https://doi.org/10.1016/j.patrec.2017.05.027

Guan, D. (2020, Jul). *Classical architectures in cnn.* Retrieved from https://guandi1995.github.io/Classical-CNN-architecture/

Haulcy, R., & Glass, J. (2021). Classifying alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology*, *11*. Retrieved from `https://www.frontiersin.org/article/10.3389/fpsyg.2020.624137` doi: 10.3389/fpsyg.2020.624137

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, *abs/1512.03385*. Retrieved from `http://arxiv.org/abs/1512.03385`

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., … Wilson, K. W. (2016). CNN architectures for large-scale audio classification. *CoRR*, *abs/1609.09430*. Retrieved from `http://arxiv.org/abs/1609.09430`

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, *abs/1704.04861*. Retrieved from `http://arxiv.org/abs/1704.04861`

Huang, G., Liu, Z., & Weinberger, K. Q. (2016). Densely connected convolutional networks. *CoRR*, *abs/1608.06993*. Retrieved from `http://arxiv.org/abs/1608.06993`

Jiang, Y.-E., Liao, X.-Y., & Liu, N. (2023, 03). Applying core lexicon analysis in patients with anomic aphasia: Based on mandarin aphasiabank. *International Journal of Language & Communication Disorders*, *n/a*(n/a). Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/1460-6984.12864` doi: https://doi.org/10.1111/1460-6984.12864

Johns Hopkins Medicine. (n.d.). *Aphasia.* Retrieved from `https://www.hopkinsmedicine.org/health/conditions-and-diseases/aphasia`

Kanopoulos, N., Vasanthavada, N., & Baker, R. L. (1988). Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, *23*(2), 358–367.

Karakostas, A., Briassouli, A., Avgerinakis, K., Kompatsiaris, I., & Tsolaki, M. (2017). The dem@care experiments and datasets: a technical report. *CoRR*, *abs/1701.01142*. Retrieved from `http://arxiv.org/abs/1701.01142`

Karlekar, S., Niu, T., & Bansal, M. (2018). Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. *CoRR*, *abs/1804.06440*. Retrieved from `http://arxiv.org/abs/1804.06440`

Khandelwal, R. (2019, Nov). *Ssd;: Single shot detector for object detection using multibox.* Towards Data Science. Retrieved from `https://towardsdatascience.com/ssd-single-shot-detector-for-object-detection-using-multibox-1818603644ca`

Kokkinakis, D., Lundholm Fors, K., Björkner, E., & Nordlund, A. (2017, 05). Data collection from persons with mild forms of cognitive impairment and healthy controls-infrastructure for classification and prediction of dementia..

Lee, J., & Hsiang, J. (2019). Patentbert: Patent classification with fine-tuning a pre-trained BERT model. *CoRR*, *abs/1906.02124*. Retrieved from http://arxiv.org/abs/1906.02124

Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge. In *Proceedings of interspeech 2020.* Shanghai, China. Retrieved from https://arxiv.org/abs/2004.06833

López-de Ipiña, K., Alonso, J.-B., Travieso, C. M., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., … Lizardui, U. M. d. (2013). On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis. *Sensors*, *13*(5), 6730–6745. Retrieved from https://www.mdpi.com/1424-8220/13/5/6730 doi: 10.3390/s130506730

Macwhinney, B. (2000, 01). The childes project: tools for analyzing talk. *Child Language Teaching and Therapy*, *8*. doi: 10.1177/026565909200800211

Mahajan, P., & Baths, V. (2021, 02). Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech. *Frontiers in Aging Neuroscience*, *13*. doi: 10.3389/fnagi.2021.623607

Mallick, S. (2021, May). *Facial landmark detection: Learnopencv.* LearnOpenCV. Retrieved from https://learnopencv.com/facial-landmark-detection/

Martinc, M., & Pollak, S. (2020, 11). Tackling the adress challenge: A multimodal approach to the automated recognition of alzheimer's dementia.. doi: 10.21437/Interspeech.2020-2202

Mayo Clinic. (2022, Jun). *Aphasia.* Mayo Foundation for Medical Education and Research. Retrieved from https://www.mayoclinic.org/diseases-conditions/aphasia/symptoms-causes/syc-20369518

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, *2014*(239), 2.

Minga, J., Johnson, M., Blake, M., Fromm, D., & Macwhinney, B. (2021, 01). Making sense of right hemisphere discourse using rhdbank. *Topics in Language Disorders*, *41*, 99-122. doi: 10.1097/TLD.0000000000000244

Mittal, A. (2021, Aug). *Understanding rnn and lstm.* Medium. Retrieved from https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e

Mittal, A., Sahoo, S., Datar, A., Kadiwala, J., Shalu, H., & Mathew, J. (2020). Multi-modal detection of alzheimer's disease from speech and text. *CoRR*, *abs/2012.00096*. Retrieved from https://arxiv.org/abs/2012.00096

National Institute of Mental Health. (n.d.). *What is aphasia? - types, causes and treatment.* U.S. Department of Health and Human Services. Retrieved from https://www.nidcd.nih.gov/health/aphasia

Negin, F., Rodriguez, P., Koperski, M., Kerboua, A., Gonzàlez, J., Bourgeois, J., … Bremond, F. (2018). Praxis: Towards automatic cognitive assessment using gesture recognition. *Expert Systems with Applications.*

OpenAI. (2021, Jun). *CLIP: Connecting Text and Images.* Author. Retrieved from `https://openai.com/blog/clip/`

Ortiz-Perez, D., Ruiz-Ponce, P., Tomás, D., Garcia-Rodriguez, J., Vizcaya-Moreno, M. F., & Leo, M. (2023). A deep learning-based multimodal architecture to predict signs of dementia. *Neurocomputing*, *548*, 126413. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0925231223005362` doi: https://doi.org/10.1016/j.neucom.2023.126413

Ortiz Pérez, D. (2022-06-30). *Deep learning-based dementia prediction using multimodal data.*

Ouden, D.-B., Malyutina, S., & Richardson, J. (2015, 04). Verb argument structure in narrative speech: Mining the aphasiabank. *Frontiers in Psychology*, *6*. doi: 10.3389/conf.fpsyg.2015.65.00085

Palanisamy, K., Singhania, D., & Yao, A. (2020). Rethinking CNN models for audio classification. *CoRR*, *abs/2007.11154*. Retrieved from `https://arxiv.org/abs/2007.11154`

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`

Pearson, K. (1992). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Methodology and distribution* (pp. 11–28). New York, NY: Springer New York. doi: 10.1007/978-1-4612-4380-9_2

Pierse, C. (2021, 2). *Transformers Interpret.* Retrieved from `https://github.com/cdpierse/transformers-interpret`

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, *55*(2), 477–523. Retrieved from `https://doi.org/10.1007/s10579-020-09502-8` doi: 10.1007/s10579-020-09502-8

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision.*

Revuelta, F. F., Chamizo, J. M. G., Garcia-Rodrguez, J., & Sáez, A. H. (2002). Representation of 2d objects with a topology preserving network. In J. M. I. Quereda & L. Micó

(Eds.), *Pattern recognition in information systems, proceedings of the 2nd international workshop on pattern recognition in information systems, PRIS 2002, in conjunction with ICEIS 2002, ciudad real, spain, april 2002* (pp. 267–276). ICEIS Press.

Roberts, L. (2020, Mar). *Understanding the mel spectrogram.* Analytics Vidhya. Retrieved from https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

Schmidt, R. M. (2019). Recurrent neural networks (rnns): A gentle introduction and overview. *CoRR*, *abs/1912.05911*. Retrieved from http://arxiv.org/abs/1912.05911

Serengil, S. I., & Ozpinar, A. (2020). Lightface: A hybrid deep face recognition framework. In *2020 innovations in intelligent systems and applications conference (asyu)* (p. 23-27). Retrieved from https://doi.org/10.1109/ASYU50717.2020.9259802 doi: 10.1109/ASYU50717.2020.9259802

Serengil, S. I., & Ozpinar, A. (2021). Hyperextended lightface: A facial attribute analysis framework. In *2021 international conference on engineering and emerging technologies (iceet)* (p. 1-4). Retrieved from https://doi.org/10.1109/ICEET53442.2021.9659697 doi: 10.1109/ICEET53442.2021.9659697

Serengil, S. I., & Ozpinar, A. (2023). *An evaluation of sql and nosql databases for facial recognition pipelines.* https://www.cambridge.org/engage/coe/article-details/63f3e5541d2d184063d4f569. Cambridge Open Engage. Retrieved from https://doi.org/10.33774/coe-2023-18rcn (Preprint) doi: 10.33774/coe-2023-18rcn

Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., … Parikh, D. (2020). *Mmf: A multimodal framework for vision and language research.* https://github.com/facebookresearch/mmf.

Torre, I. G., Romero, M., & Álvarez, A. (2021). Improving aphasic speech recognition by using novel semi-supervised learning methods on aphasiabank for english and spanish. *Applied Sciences*, *11*(19). Retrieved from https://www.mdpi.com/2076-3417/11/19/8872 doi: 10.3390/app11198872

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). *Attention is all you need.*

Warnita, T., Inoue, N., & Shinoda, K. (2018, 09). Detecting alzheimer's disease using gated convolutional neural network from audio data. In (p. 1706-1710). doi: 10.21437/Interspeech.2018-1713

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-demos.6

Yusculearning, P. (2020, Aug). *Redes convolucionales.* Retrieved from https://demachinelearning.com/redes-convolucionales/

Zhao, S., Rudzicz, F., Carvalho, L. G., Marquez-Chin, C., & Livingstone, S. (2014). Automatic detection of expressed emotion in parkinson's disease. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 4813-4817). doi: 10.1109/ICASSP.2014.6854516

List of acronyms

ASIC	Application-Specific Integrated Circuits.
BERT	Bi-directional Encoder Representations from Transformers.
CLIP	Contrastive Language–Image Pre-training.
CNN	Convolutional Neural Network.
CNTK	The Microsoft Cognitive Toolkit.
CPU	Central Processing Unit.
CUDA	Compute Unified Device Architecture.
DTIC	Departamento de Tecnología Informática y Computación.
GPU	Graphics Processing Unit.
LSTM	Long Short-Term Memory.
MoDeAsS	Monitoring and Detection of human behaviors for personalized assistance and early disease detection.
NAS	Network Attached Storage.
NLP	Natural Language Processing.
RAID	Redundant Array of Independent Disks.
RNN	Recurrent Neural Network.
TPU	Tensor Processing Unit.
VQA	Visual Question Answering.