



Detecting Alzheimer's Disease using NLP Methods

by

Anastasios Sarafidis

Submitted

in partial fulfilment of the requirements for the degree of

Master of Artificial Intelligence

at the

UNIVERSITY OF PIRAEUS

February 2023

Author

II-MSc “Artificial Intelligence”

February 21, 2023

Certified by.....

Rentoumi
Vassiliki
Researcher
Thesis
Supervisor

Certified by.....

Petasis Georgios
Researcher
Member of
Examination
Committee

Certified by.....

Giannakopoulos
Georgios
Researcher
Member of
Examination
Committee

Detecting Alzheimer’s Disease using NLP Methods

By

Anastasios Sarafidis

Submitted to the II-MSc “Artificial Intelligence” on February 28, 2023,
in partial fulfillment of the
requirements for the MSc degree

Abstract

Alzheimer's disease (AD) is a progressive brain disease that cannot be treated, but only be slowed down or stopped with medical treatment. Language changes may indicate that a patient's cognitive functions have been compromised, potentially leading to an earlier diagnosis. The challenging manual diagnosis of such neurodegenerative disorders could be assisted by the use of Machine Learning algorithms able to automatically detect those disorders using linguistic features. For this purpose, we use the ADReSS Challenge dataset and we develop NLP methods to classify and analyze the linguistic characteristics of Alzheimer's disease patients. To distinguish between language samples from probable AD and control patients, we propose the implementation of an XGBoost classification model, which has not been present in similar cases literature, along with three other models that are most often used. XGBoost’s final scores indicate that this classification model, with the right adjustments in terms of data and features, is able to compete with or even surpass in performance the other models.

Περίληψη

Η νόσος του Αλτσχάιμερ είναι μία προοδευτική ασθένεια του εγκεφάλου, η οποία δεν μπορεί να θεραπευτεί, αλλά μόνο να επιβραδυνθεί λαμβάνοντας φαρμακευτική αγωγή. Πιθανές γλωσσικές διαφοροποιήσεις μπορεί να υποδηλώνουν ότι οι γνωστικές λειτουργίες του ασθενούς έχουν υποβαθμιστεί, οδηγώντας σε πρόωμη διάγνωση. Δεδομένης της δυσκολίας στην στον παραδοσιακό τρόπο διάγνωσης τέτοιων νευρολογικών διαταραχών, η χρήση αλγορίθμων μηχανικής μάθησης που

είναι σε θέση να ανιχνεύουν αυτόματα τις εν λόγω διαταραχές χρησιμοποιώντας γλωσσικά χαρακτηριστικά θα μπορούσε να φανεί πολύ βοηθητική. Για το λόγο αυτό, αποφασίσαμε να χρησιμοποιήσουμε το σύνολο δεδομένων του ADReSS Challenge και να αναπτύξουμε μεθόδους Επεξεργασίας Φυσικής Γλώσσας (NLP) για την ανάλυση και ταξινόμηση των γλωσσικών χαρακτηριστικών των ασθενών με νόσο Αλτσχάιμερ. Για την κατηγοριοποίηση των γλωσσικών δειγμάτων που έχουμε συλλέξει από πιθανούς ασθενείς της νόσου και από υγιείς συμμετέχοντες, προτείνουμε την εφαρμογή του XGBoost μοντέλου, το οποίο δεν χρησιμοποιείται συχνά σε αντίστοιχες περιπτώσεις, παράλληλα με την υλοποίηση τριών επιπλέον μοντέλων. Τα τελικά αποτελέσματα του XGBoost υποδεικνύουν πως το συγκεκριμένο μοντέλο, υπό τις κατάλληλες ρυθμίσεις, είναι σε θέση να αποφέρει παρόμοια ή και καλύτερα αποτελέσματα από τα υπόλοιπα μοντέλα.

Επιβλέπουσα: Βασιλική Ρεντούμη

Ακαδημαϊκή Θέση: Ερευνήτρια

Thesis Supervisor: Vassiliki Rentoumi

Academic Title: Researcher

Acknowledgments

At this point I would like to thank the people who helped make this thesis possible. First of all, I would like to express my deepest appreciation to my supervisor, Mrs Vassiliki Rentoumi, who trusted me and gave me the opportunity to be engaged in the impressive field of Natural Language Processing. I would also like to thank Mr Nikiforos Pittaras, Postdoctoral Researcher at NCSR Demokritos, whose crucial comments and recommendations helped me accomplish a thorough thesis. I am also grateful to each one of the MSc professors. Every lecture, assignment and conversation between us has been a true guidance and inspiration for me. A major acknowledgment also goes to Mr Brian MacWhinney and the TalkBank project that granted me the necessary dataset in order to proceed.

Special thanks to my girlfriend, Dimitra, who supported me from the beginning and always does. Her belief in me has played a major part in the accomplishment of this project. Last but not least, I owe a big thank you to my parents, Dimitris and Erifili, who are always by my side supporting me and giving me the necessary tools to achieve my goals.

Any opinions, findings, conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of University of Piraeus and Inst. of Informatics and Telecom. of NCSR “Demokritos”.

Table of Contents

List of Figures	8
1 Introduction	9
2 State of the Art.....	11
2.1 Dementia.....	11
2.2 Alzheimer’s Disease (AD)	12
2.3 Studies.....	13
2.4 Datasets	13
2.5 Features	14
2.6 Methods Used	15
2.7 Goals and Contributions	15
3. AI & NLP	17
3.1 Intelligence.....	17
3.2 Artificial Intelligence	18
3.2.1 History of AI	19
3.2.2 AI Fields	20
3.2.3 AI in Healthcare.....	21
3.3 Machine Learning.....	22
3.4 Natural Language Processing.....	23
3.3.1 NLP History.....	23
3.3.2 NLP Techniques.....	26
3.3.3 NLP Applications	27
4 NLP for Neurodegenerative Disease Detection	29
4.1 AI Techniques in Healthcare	29
4.2 NLP in Healthcare.....	31
4.3 NLP in Detecting Dementia Diseases	33
4.4 NLP in Detecting Alzheimer’s Disease.....	35
5 Methodology and Implementation.....	37
5.1 Dataset	39
5.1.1 The Cookie Theft Picture.....	40
5.1.2 The ADReSS Challenge Dataset.....	44
5.2 Data Preprocessing	44
5.3 Feature Extraction.....	55
6 Model Training and Evaluation	62

6.1 Model Implementation	63
6.1.1 XGBoost Classifier	63
6.1.2 SVM Classifier	70
6.1.3 Decision Tree Classifier	74
6.1.4 Random Forest Classifier	78
6.2 Classifier Comparison.....	82
6.3 Literature Comparison	84
7. Conclusion and Future Work	89
7.1 Conclusion.....	89
7.2 Future Work.....	90
References	92

List of Figures

Figure 1: Cookie Theft Picture.....	41
Figure 2: Example of initial transcription format	46
Figure 3: Possible AD patient’s spontaneous speech after cleaning the transcript ..	47
Figure 4: Healthy participant’s spontaneous speech after cleaning the transcript ..	48
Figure 5: Part of the processed_data.csv file.....	49
Figure 6: Dataset class distribution.....	50
Figure 7: Tokenized text.....	51
Figure 8: Lemmatization of a possible AD patient's tokens.....	52
Figure 9: POS Tagging of a possible AD patient's transcript	54
Figure 10: Total amount of POS Tags in descending order	54
Figure 11: Confusion matrix.....	65
Figure 12: XGBoost Confusion Matrix	66
Figure 13: XGBoost ROC Curve	67
Figure 14: XGBoost Stratified 10-fold CV ROC Curves & Mean ROC Curve	70
Figure 15: SVM Confusion Matrix	72
Figure 16: SVM ROC Curve	72
Figure 17: SVM Stratified 10-fold CV ROC Curves & Mean ROC Curve.....	74
Figure 18: Decision Tree Confusion Matrix	75
Figure 19: Decision Tree ROC Curve	76
Figure 20: DT Stratified 10-fold CV ROC Curves & Mean ROC Curve.....	78
Figure 21: Random Forest Confusion Matrix.....	79
Figure 22: Random Forest ROV Curve.....	80
Figure 23: RF Stratified 10-fold CV ROC Curves & Mean ROC Curve	82
Figure 24: Precision - Recall Curves after Stratified 10-fold CV.....	83

1 Introduction

It has been estimated that Alzheimer's disease, the leading cause of dementia, affects over 20 million people worldwide. Alzheimer's disease is an incurable progressive brain disease that can only be slowed down with medical treatment. It is characterized by a growing cognitive decline which begins with memory and linguistic impairment. Detecting such linguistic deficits may lead us to an earlier dementia disease diagnosis, such as Alzheimer's disease. Machine Learning algorithms have started being developed in order to automatically diagnose neurological disorders using linguistic features. This offers an important boost to physicians who all these years analyze symptoms and medical examinations manually in order to formulate a diagnosis. With the automatic diagnosis of dementia diseases, the healthcare systems could save a great amount of time, while also reducing the required costs effectively.

In the present thesis, we develop Natural Language Processing techniques in an attempt to analyze linguistic features of several experimental participants and later classify them in one of two categories: possible Alzheimer's patients and healthy controls. To achieve that, we use the ADReSS Challenge dataset, taken from the DementiaBank database. After reviewing several studies on this subject, we have identified that the XGBoost classifier is not that commonly used in comparison with other classification algorithms on such cases. For this reason, we propose the development of XGBoost, an algorithm built with model performance and computational speed in mind, as well as three other algorithms: SVM, Decision Tree and Random Forest. The goal is to run experiments on every model with the same dataset and features in order to evaluate them, compare their performances and establish XGBoost's place regarding the detection of Alzheimer's disease. Results show that XGBoost's performance is not far from other algorithms' performance and that with the right optimizations in terms of data, features and techniques it could produce even higher classification scores.

In the following chapters we report every theoretical and practical step that we studied, followed or implemented in order to present a scientifically comprehensive thesis. Chapter 2 presents the current state of dementia and Alzheimer's disease

scientific research, as well as the current datasets, features, Machine Learning and NLP techniques used in detecting neurological diseases. In Chapter 3, we make a brief review of Artificial Intelligence, its corresponding fields and their applications. Chapter 4 briefly reports the use of NLP in healthcare system and focuses on its role in detecting dementia and neurological diseases. Furthermore, in Chapter 5 we report the methodology that we followed in order to obtain the required dataset, clean and preprocess it and extract important features for the model training. We proceed with Chapter 6, where we present our implementation of each classifying model that we have built. We also present their individual classification scores and their respective mean metrics after a stratified 10-fold cross validation. We finish our thesis with Chapter 7, in which we present our final conclusions and the work that can be possibly done in the future for improving the outcomes.

2 State of the Art

In this chapter we present a systematic literature review on the state of the art research work regarding neurological diseases and the implementation of NLP methods for detecting such diseases.

2.1 Dementia

According to Centers for Disease Control and Prevention (CDC), there are more than 5 million Americans over 65 years-old who suffer from a disorder called dementia. Dementia is not a single disease, but it covers a wide range of specific diseases, like Alzheimer's disease, Parkinson's disease Dementia, Vascular Dementia or Lewy Body Dementia. It is a general term used to describe a group of symptoms that affect or damage the ability to remember, think or make decisions concerning everyday activities. Disorders that are grouped under this general term are caused by abnormal brain changes that trigger a decline in cognitive abilities. These brain changes can affect severely the patient by interfering with their behavior, feelings or even relationships. In early stages, dementia may just affect the patient's functioning, but in most severe stages, patients are completely dependent on others for basic activities or even living.

Some falsely believe that dementia is part of normal aging and that, as a person grows old, he shows signs of memory related problems, like forgetting names or struggling to find words. The above may be true to some extent, as every adult loses some healthy neurons or nerve cells while growing up. However, people with dementia tend to experience a far greater loss of these neurons making them susceptible to various symptoms of dementia. Those symptoms can be distinguished into two main categories; cognitive and psychological changes. By cognitive changes, we mean mental difficulties that affect the patient's everyday skills, such as memory loss or finding difficulty with simple tasks like communicating with others, finding the right words, reasoning or problem-solving, planning, organizing or being confused and disoriented. Furthermore, a person suffering from dementia will most likely develop psychological disorders. Personality changes, depression, anxiety,

inappropriate behavior, hallucinations are the most common psychological symptoms of dementia.

As mentioned earlier, dementia is not a single disease, but a wider term that covers several other diseases that cause dementia symptoms. The most common one is the Alzheimer's disease (AD) which accounts for 60 to 80% of dementia cases and it is caused by specific changes in the brain. Another well-known disease that causes dementia is the Vascular Dementia. This kind of dementia is caused by damage to the vessels that help the blood flow towards the brain. Once this damage occurs, the patient suffers more from difficulties with problem-solving, slower thinking or loss of focus and less from memory loss. Furthermore, Lewy Body Dementia is one of the most common types of progressive dementia and it is often diagnosed on Alzheimer's and Parkinson's patients. More serious symptoms are hallucinations, problems with focus and attention.

2.2 Alzheimer's Disease (AD)

As already mentioned, the most common disease that causes dementia is the Alzheimer's disease, accounting for over 60% of dementia cases. Alzheimer's disease is a progressive neurodegenerative disease that causes problems with memory, thinking and general behavior. According to the National Institute on Aging (NIH), even though the first symptoms of Alzheimer's disease vary from person to person, problems with memory are typically the most common ones.

Except for the memory loss, Alzheimer's disease at its earliest stages may develop light symptoms regarding word-finding problems, impaired reasoning or judgment. Even though a person suffering from mild Alzheimer's disease may seem to be healthy, there are several symptoms that are gradually becoming more severe making that person's daily life more and more difficult. Such symptoms may be mild memory loss, repeating questions, losing things, wandering or getting lost.

As the disease progresses, symptoms progress as well, making the patient's life even more difficult. A patient diagnosed with moderate Alzheimer's disease usually suffers from increased memory loss, difficulty with language, problems recognizing family and friends and many more symptoms. From this stage on, the patient is often incapable of completing tasks independently, making intensive supervision and care

more necessary. Finally, patients at the latest stages of Alzheimer's disease become completely dependent on others, as they are incapable of communicating or perform any other action.

2.3 Studies

From the above, we understand that early detection, diagnosis and cure of the Alzheimer's disease are of utmost importance for the well-being of Alzheimer's disease patients as well as for their family and friends. However, current means of diagnosis are time-consuming and severely expensive. This is the main reason why more and more studies aim to find quick and secure ways to automatically distinguish between a healthy person and an Alzheimer's disease patient.

Many scientists are focusing on the language difficulties that Alzheimer's disease causes in order to identify early dementia signs. The reason why they focus on language features lies on the fact that Alzheimer's disease affects considerably the content and acoustics of speech. By exploiting text, linguistic and acoustic features, scientists are able to compare healthy and non-healthy samples and automatically predict Alzheimer's disease in its early stages using Natural Language Processing and Machine Learning technologies. Extracting and analyzing linguistic characteristics offer the chance to detect several speech impairments that a subject may be suffering from. As Klimova B. & Kuca K. state (2016), in early stages of Alzheimer's disease, patients basically present a mild impairment concerning lexical-semantic aspects of the language, such as naming various things or finding the right words in a sentence. Suffering from these early symptoms of lexical disorders does not mean that patients' communication is affected. On the contrary, communication is still completely fluent. However, in the more serious cases of Alzheimer's disease, communication begins lacking fluency and ends up being completely impossible in the most severe cases. According to the same research (Klimova B. & Kuca K., 2016), the key language impairments detected on Alzheimer's disease patients are finding the right word for objects, naming objects and comprehending words.

2.4 Datasets

After having certain knowledge of the key symptoms and impairments, scientists must find an effective way to collect enough data from patients in order to examine

them. In order to extract the most important features that will lead to observations and results, scientists tend to experiment on subjects by numerous ways. As subjects' spontaneous speech and how they behave in casual conversations are essential ways to evaluate a person's linguistic and acoustic characteristics, scientists perform experiments where the subjects' objective is to freely talk about their own lives or several situations. Another usual way of deriving linguistic features from the subjects is based on image descriptions. On this kind of experiments, participants are required to describe an image that they are shown. This way, scientists are able to conduct linguistic analysis on subjects' answers. By collecting and processing those data, we are able to extract certain important features that we can use in order to detect the necessary symptoms that will lead to a positive or negative Alzheimer's disease diagnosis.

Scientists have already done an important work on collecting essential data so that we would be able to process and utilize them in order to conduct experiments and produce certain results and conclusions. Therefore, most studies have used already available datasets such as Iflytek's Dataset [2], ADNI Dataset [12, 15] or DementiaBank the largest open source dataset for Alzheimer's Disease Classification, which also contains the ADrESS Challenge Dataset, the most commonly used one [3, 4, 5, 6, 7, 8, 11, 20]. In this thesis, we will also use the ADrESS Challenge Dataset for the purpose of detecting Alzheimer's disease using Machine Learning methods.

2.5 Features

Having gathered the necessary data, the next step is to determine which features are more important than others and select them in order to produce the optimal classification by using a Machine Learning method. The majority of the literature that we studied use feature extraction algorithms to extract acoustic, semantic or linguistic features and conduct the needed researches. These features are extracted from the subject's spontaneous speech, interview or any other method that is being used for gathering the data and they can be observed in various forms that usually represent the symptoms of Alzheimer's disease. Semantic and lexical errors, repetition, long pauses or coherence dysfunction are some examples of available features that can be extracted straight from the subjects' recording which can be used for an early detection of Alzheimer's disease symptoms.

Meanwhile, other researchers decided to build their own features in an attempt to examine deeper the diagnosis of Alzheimer's disease, better distinguish patients with different levels of severity and to possibly end up to better overall performance of Alzheimer's disease detection. Hand-crafted features are built by using already available data and features in a way that can produce additional features which can shed brighter light on the whole procedure of detecting Alzheimer's disease. For example, Sarawgi U. et al. ([4], 2020) built the Disfluency feature which consists of 11 distinct features taken from the transcripts, like word rate and pause rate.

2.6 Methods Used

With the necessary features available, the classification process takes place. Subjects can be classified using several different methods, thus scientists must decide which one is the best in order to produce desirable results in each case.

Among the studies that we examined, the most commonly used learning methods were the SVM [2, 3, 9, 11, 13, 16], Logistic Regression [2, 13, 16, 17], Random Forest [2, 3] and Neural Networks [3, 6, 7, 9, 11, 15, 16, 17, 18]. Undoubtedly, there are several other studies that use different methods learning methods, however, the above shows that more and more scientists tend to choose those specific algorithms for performing their experiments. Each of these studies managed to come to remarkable conclusions as well as classification results that helped lead automatic detection of Alzheimer's disease on step forward.

2.7 Goals and Contributions

It is clear that the combination of linguistics and NLP techniques can produce new tools that contribute to achieving a faster and cost-effective diagnosis of AD and other cognitive impairments. By utilizing the ideal features extracted from the available data and using a proper classification algorithm, we have the ability to establish a quick and objective evaluation of an individual's cognitive ability and ultimately diagnose the Alzheimer's disease.

As stated earlier, by studying several researches we have observed that most studies focus on the same most common classification algorithms for their tasks, such as SVM or Neural Networks. However, the lack of use of XGBoost, an algorithm famous

for its execution speed and performance, caught our attention and lead us to perform our classification experiments using this specific ML method.

The goal of this thesis is to present an innovative way of predicting AD which exploits the ADReSS Challenge Dataset, extracts the necessary features and proceeds with the subjects' classification by using the XGBoost algorithm. After completing the classification task, we will be able to compare our method's performance to other methods implemented using the same dataset and evaluate its efficiency.

3. AI & NLP

This chapter introduces the scientific field that is discussed in this thesis and its applications.

3.1 Intelligence

Ever since the industrial revolution, every small or bigger step in the field of computer science has guided us to the great era of Artificial Intelligence. Through the years, researchers focused on inventing new technology that could deal with hard manual work and assist or even replace humans. One of those great inventions was Artificial Intelligence which currently stands as one of the most important scientific fields. But what exactly is Artificial Intelligence? Why do scientists try to build systems that follow human characteristics, as well as human behavior? To achieve this goal, scientists first had to understand the deep meaning of human intelligence in order to be able to build theories and models based on human behavior.

In order to understand and describe the term “Intelligence”, all we have to do is think of some general attributes that may be found among the humanity. Below we present some of the main human attributes that researchers tend to look for and reproduce in order to build an “intelligent” system [1]:

Perception: comprehension, manipulation and interpretation of any information that may be sensed by the surroundings

Action: use of any possible means in order to accomplish a given task

Adaptation and Learning: the ability of handling and adapting in any changes that may occur in an environment, as well as learning to observe, discover, explain and exploit those changes

Communication: communicate with other subjects that belong to the same environment, using signals, symbols, language, pictures or any other communication method

Autonomy: decide, plan and execute the most suitable action or actions in order to attain a certain goal, adapt to any unanticipated new circumstances

Of course, the presence alone of those features in an artificial system does not provide that this system is intelligent. However, displaying these features on a broader environment and while being under a range of constraints may be a sign of Artificial Intelligence.

3.2 Artificial Intelligence

Artificial Intelligence was first introduced to the world as an imaginary creation by philosophers and writers. Philosophers invented the concept of artificial intelligence in an attempt to explain the great importance of being a human being [2]. Rene Descartes built a metaphor of a “mechanical man” in order to present the possible existence of intelligence machines. Except for the philosophers, writers also played a vital role in presenting early signs of artificial intelligence to the humanity. Science fiction writers like Jules Verne or Isaac Asimov also adopted the theoretical existence of artificial intelligence so that they could add some more fantasy to their novels. In 1907, it was Frank Baum who wrote about an “extra-responsive, thought-creating, perfect-talking mechanical man, who thinks, speaks acts and does everything, but live” in his children novel “The Wonderful Wizard of Oz”. In 1942, another writer, Isaac Asimov, wrote the story “Runaround” in which the main character was a robot developed by engineers Gregory Powell and Mike Donovan. It was those works that inspired many researchers to start their own researches in the field of Computer Science, Robotics and Artificial Intelligence.

The most influential work towards AI was done around the same time by the English scientist Alan Turing. After inventing the machine that managed to decipher Enigma code, a code used by German army in the World War 2 in order to encode their messages, he wrote about the creation and testing of intelligence machines, known as “the Turing test” [3]. The Turing test was designed to provide an early definition of Artificial Intelligence. A computer can be considered as an intelligent system if it can mimic human behavior under certain circumstances. Therefore, if a human is not in a position to distinguish between another human and a machine by interacting with them, then this machine passes the test and it can be considered as intelligent. However, for a computer to be able to pass such a test, there are six disciplines that it must follow [4]:

- *Natural Language Processing*: so that it can communicate in English

- *Knowledge Representation*: in order to store any information it can gather from its environment
- *Automated Reasoning*: to use the gathered information to confront questions and obstacles
- *Machine Learning*: to adapt in new conditions
- *Computer Vision*: to notice any present objects
- *Robotics*: to manipulate objects

As of today, the above six disciplines compose the biggest part of Artificial Intelligence and have dominated the focus of researchers.

3.2.1 History of AI

After Alan Turing's groundbreaking work, researches on AI were greatly intensified. In 1956, mathematician John McCarthy convinced several researchers from various scientific fields to host a two-month workshop which would focus on the study of artificial intelligence. As the official proposal stated, "An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves." [4]. Even though there were no new breakthroughs at this workshop, it certainly made it clear to the scientific society that artificial intelligence had to become an individual field. According to Russel S. & Norvig P. [4], "AI from the start embraced the idea of duplicating human faculties such as creativity, self-improvement and language use. None of the other fields were addressing these issues."

What followed this workshop can be considered as the history of AI. At first, given the fact that computers seemed to be simple machines that just computed simple arithmetic actions, any new and even a bit more clever action was hailed by everyone. As time passed, more and more scientists managed to successfully design programs that imitated human behavior and protocols. In 1961, Newell and Simon designed the General Problem Solver (GPS), a program that managed to handle a number of puzzles in a way that was very similar to the way that humans approached the same puzzles [5]. In 1952, Arthur Samuel began writing a series of programs for playing checkers, which eventually reached a stage in which they were able to play checkers at a strong amateur level. This program quickly managed to play a game at a better level than its creator.

Even though researches on AI were significantly increased, access to tools and resources was limited, either because of the lack of those, or because of the expensive ways to obtain the necessary resources. It was John McCarthy who, in 1958, gave a solution to those problems by presenting the programming language LISP, a language which came to be the default programming language for artificial intelligence for the next 30 years.

Soon, the first difficulties emerged. At first, most programs succeeded simply thanks to syntactic manipulations made by their developers. This meant that the programs were not able to autonomously deal with complicated circumstances, but they had to be specifically designed to confront any obstacle. This problem was clearly encountered by American researchers who tried to translate Russian papers using an artificial intelligence program. The program did not have any deeper syntactic or linguistic knowledge of both Russian and English languages, resulting in faulty translations, like the translation of the sentence “the spirit is willing but the flesh is weak” as “vodka is good but the meat is rotten” [4]. Furthermore, despite the fact that AI problems were initially easier to solve for the programs developed due to the limited number of objects and possible actions to use, later on problems became much more complicated making their solutions more difficult, as the need for more resources and more computational complex systems had began increasing.

After several ups and downs and as technology kept growing, researchers managed to build systems that could face computational complexity and finally establish AI as one of the most important parts for computer science, as well as for plenty of other different fields like healthcare, education, finance etc. and make AI, often in cooperation with Machine Learning or some other field, really helpful for humanity.

3.2.2 AI Fields

One major field in which AI has been proved necessary is social media platforms. AI is widely used in social media platforms for various reasons, like showing personalized advertisements to users, detecting fake news or deleting offensive comments (hate, racist, discrimination, etc). Moreover, digital assistants are becoming more and more used in society, as the biggest technological companies have already developed their own AI assistant (Siri by Apple, Alexa by Amazon, Cortana by Microsoft etc). While digital assistants began their “journey” providing

limited features to their owner, now they are able to accomplish tasks like finding various information (weather, traffic, news etc), organizing schedules, listen and answer to questions, perform actions in the house (control devices, lights etc), make and take phone calls, translate foreign languages and many more actions that facilitate our every-day routines.

AI has also changed business industry and how people do their businesses, as it provides tools for marketing and customer service purposes that help companies target the right customers through advertisement, data collection and analysis. Furthermore, AI has not changed solely small businesses, but also bigger companies. For example, car companies are now able to build self-parking or even self-driving cars that, thanks to AI, have the ability to think, learn and do various driving actions.

3.2.3 AI in Healthcare

Of course, one of the most important fields in which AI has provided tons of help is the healthcare system. AI technologies are now able to analyze health data and with the use of computer vision, speech recognition and natural language processing, as well as with the help of advanced software and hardware, the use of AI systems has become easier and more common in order to extract important features and insights from health data [6]. Most common data used for medical purposes are, among others, clinical, behavioral, patient or drug data.

Clinical data usually consist of medical images, health records and physiological signals. The analysis of medical images is based on feature engineering. By conducting feature engineering on medical images, scientists have the opportunity to extract features and descriptors that will be later used on learning methods for classification methods. According to reports, a doctor may review over 200,000 images of clinical data over decades of work, while an AI program would need only some days to analyze the same images [6]. This ability of examining a large collection of clinical data in a small amount of time is what makes AI an important tool for the medicine society.

Behavioral data are also linked with a human's health status. Since the way such data are collected is different than the way scientists collect clinical data, several research groups are assigned to gather behavioral data and examine their relationship with health in general. There are plenty ways of collecting those data. Now that social

media has become a vital part of our lives, they are a somewhat effortless way to analyze health status by examining posts on Facebook, Twitter, Instagram or LinkedIn. A research by Sinnenberg et al. [7] has shown an association between Twitter posts and cardiovascular diseases. The research claims that users with cardiovascular disease can be characterized by the tone, style and perspective of their tweets and some basic demographics. There are also several research groups that, after examining social media analytics and mental health, have identified indicators in social media associated with psychotic symptoms, schizophrenia, risk of suicidal thoughts and depression [6].

Another way to collect and examine behavioral data is by video or conversation methods. Use of video or conversation data is now widely used in an attempt to detect and diagnose mild cognitive disease, Parkinson's disease, Alzheimer's disease or dementia. By examining a patient's behavior through a video or a conversation, we can identify any known symptoms of a specific disease. Especially for a dementia-linked disease (like Alzheimer's disease which we examine in this thesis), gathering linguistic data by interviewing subjects and analyzing them using Natural Language Processing methods has shown high numbers of successful diagnosis.

3.3 Machine Learning

Machine learning (ML) is an Artificial Intelligence branch that allows computers to "self-learn" from training data and improve over time without being explicitly programmed to. Machine Learning algorithms are used to discover valuable patterns in the available data that would otherwise be difficult to find. These patterns can later be used to predict future events or make complex decisions.

The advantage of Machine Learning is that it can process massive amounts of data and perform much more accurately than humans. For this reason, we can detect machine learning everywhere, from every-day tasks such as automated translation, image recognition or voice search, to more advanced and groundbreaking cases, like diseases' detection or self-driving cars.

3.4 Natural Language Processing

Natural Language Processing belongs in the field of Artificial Intelligence. It helps computers “understand”, interpret and manipulate human language. NLP technology is quickly advancing, thanks to the increased tendency and interest in communication between humans and machines. As most modern disciplines, the origin of NLP is mixed, as it is strongly influenced by different groups whose backgrounds are influenced by each other. The most notable contributors to NLP use and behavior are Linguistics, that focus on formal models of language and the discovery of language universals, Computer Science which focuses on developing representations of data and Cognitive Psychology that examines language usage as a window into cognitive processes [12]

3.3.1 NLP History

The field of Natural Language Processing (NLP) began in the 1940s, when governments looked for ways to automatically translate Russian language to English during the Cold War. Researchers’ goal was to build machines capable of translating different languages in order to monitor the enemy’s correspondence and activity. Machine translation (MT), being one of the first non-numeric computer operations, was considered a crucially important aid for human to human communication.

Active research on NLP really began in the 1950’s when several experiments, exhibitions, publications and international conferences started taking place. A first automatic translation from Russian to English was exhibited in 1954, while international conferences had already began since 1952. In 1961, the Teddington International Conference on Machine Translation of Languages and Applied Language Analysis was possibly the peak of this first period of NLP. In this conference, there were several reports from different countries that presented works done on many aspects of NLP, like morphology, syntax and semantics in interpretation and generation [8]. The first phase of NLP was met with great enthusiasm and optimism, as scientists had a new challenge: the use of computers for processing non-numerical data in an era when there were no modern computing resources or high level programming languages.

Even though those engaged in NLP researches were scientists with great backgrounds in linguistic and language study, there were three issues that lead to the

end of the first NLP phase. The first issue was the emphasis that had to be shown on syntax and semantics. Another issue was whether generalization was needed or ad hoc particularization would be adequate in order to obtain expected results, while the final issue was concentrated on the actual value of those results. Because of those issues, Machine Translation research was almost canceled in 1966 by the ALPAC Report, which stated that this scientific field was not able to achieve positive results and for that, it should be cut from government funding. Nevertheless, the first steps were made. The importance of computational language processing was established, in relation to syntactic analysis. Several aspects of language were examined, scientists began to develop new tools and formalisms, while new state-of-the-art ideas started appearing.

A solution was attempted to be found in 1963, with the creation of Backus-Naur Form (BNF) notation, which is considered a “context-free grammar” (CFG). Context-free grammars are more complex than regular languages and they are commonly used by developers of programming languages to specify syntax rules for their language. A language’s BNF main characteristic is its set of production rules that validate a code syntactically. In 1970s, lexical-analyzer and parser generators came to the surface, utilizing grammars and BNF specifications. A lexical-analyzer’s task was to transform a given text into tokens, while a parser’s task was to validate the said token. Their importance lies in the fact that they can greatly simplify the implementation of a programming language by taking regular-expressions and BNF features as input and produce codes and look-up tables capable of determining lexing decisions.

Because of human language’s size and ambiguities, scientists encountered two serious problems when tried to follow parsing approaches that relied purely on symbolic, hand-crafted rules. The first problem lied in the fact that grammars that specify relationships between parts of speech (nouns, verbs etc) address syntax primarily. Having that in mind, scientists could extend grammars to address natural language semantics by establishing more rules and constraints. However, those rules would eventually become great in numbers, resulting in unpredictable “behavior” when a word would possess more than one interpretation. The second problem that NLP scientists had to face was that the rules which they had established handled in a

poor way possible spoken speech that did not follow grammatical rules, even though humans could easily comprehend such speech.

Because of those limitations, NLP needed a fundamental reorientation. In 1980s, four changes were established that resulted in the birth of statistical NLP [9]:

- Simple robust approximations replaced deep analysis
- Evaluation became more precise
- Machine-learning methods using probabilities became more prominent
- Large annotated textbooks were employed in order to train machine learning algorithms and provide standards for the evaluation.

Thanks to those changes, new rules are now established through machine-learning algorithms that were trained on annotated corpora, rules that are now fewer in numbers and broader than the old ones.

By the end of 1980s, system builders could benefit from relatively well-understood forms of grammar and parsing algorithms, as well as from actual grammars. At that moment, NLP applications had the ability to accomplish several tasks, including message processing and translation. In the same period, researchers started showing interest in text generation through NLP. The ultimate goal of the time was the development of the lexicon. Because of its crucial role in grammatico-logical approaches, lexicon's growth was of utmost importance for developing dictionaries in forms that could be read by machines. Those dictionaries would eventually lead to the exploitation of text sources in order to validate, enhance or customize lexical data. [8]

As statistical language processing became more and more popular, corpus data showed a remarkable raise in numbers, thus we observe a constant increment in the amount of resources available for research. Massive databases have been developed, such as WordNet, a lexical database that not only includes nouns, verbs, adjectives and adverbs, but it also groups them in cognitive synonyms based on their meanings. This rapid growth does not only supply researchers with new sources of data, but it has also guided us to new ways of researching and using NLP in a broader area of text processing.

3.3.2 NLP Techniques

As stated earlier, NLP's goal is to "fill the gap" between human communication and computer understanding. To do so, it draws from several disciplines, including computational linguistics and computer science. Because of the wide variety of text and voice data forms, NLP consists of many different techniques in order to interpret human language, ranging from statistical and machine learning methods to rules-based and algorithmic approaches [10]. Regardless of the technique used, there are several NLP tasks that are considered as the basis of any approach that is going to be developed. Those tasks are common tasks that we have already met in school. They are greatly important for NLP, in order to prepare text, words or documents for further processing.

A fundamental step in NLP methods is called tokenization. It is a way of separating a text piece into smaller units (tokens). These tokens can be words, characters or sub-words. Tokens are essentially the building blocks of a natural language, making them an important tool while modeling text data. After having tokenized the text, scientists are able to use tokens in order to build a vocabulary, which is basically a set of unique tokens and it later is used as feature or as input in NLP approaches.

Another important step in NLP approaches is lemmatization. The method of lemmatization focuses on reducing a word to its canonical or dictionary form. The produced root word is called a "lemma" and it is used to map different words to their cognitive set (e.g. "was", "are", "is" are reduced to "be"). Lemmatization has mainly uses in situations where it is necessary to get valid words, such as tagging, search engines or indexing. A process similar to lemmatization is stemming. While stemming reduces words to their root forms, just like lemmatization, it is possible that the resulting root (stem) is not a valid word. This happens because stem is the part of the word to which you add inflectional affixes, such as "-ed", "-ize" etc. Thus, stemming a word with an "-ing" affix (e.g. troubling) will produce as a result the stem "troubl", which is not a valid word. In contrast with lemmatization applications in which a valid word is needed to obtain the desired results, stemming is useful in applications that need simplicity and speed, such as information retrieval or search queries.

One final but equally important process is the part-of-speech tagging. POS tagging focuses on categorizing words according to a particular part of speech, based on those words definition and context. Thanks to POS tagging, we manage to get a large amount of information about a word or a sentence. This process is usually used in tasks such as information retrieval and extraction, parsing or text-to-speech applications.

3.3.3 NLP Applications

Due to the huge advancement of data, algorithms and powerful technologies, NLP is a rapidly progressing science in the field of AI. The foundations of NLP lie in several disciplines, information sciences, linguistics, mathematics, artificial intelligence etc [11]. NLP supplies both theory and implementation for several applications, some of which are the following:

- *Information Extraction (IE)*: it focuses on recognizing, tagging and extracting key elements of information in a structured representation (e.g. people, locations, organizations from a large collections of text) Extractions can later be used for a number of applications, such as question-answering, visualization and data mining.
- *Summarization*: NLP models that can reduce a large text into a shorter one without missing any critical information
- *Machine Translation (MT)*: As stated in previous chapters, MT is the first NLP application that was implemented and it focuses on automatic translation between different languages.
- *Speech Recognition*: A technology that allows the computer to convert voice input machine readable format. There are multiple fields where speech recognition is used, such as virtual assistants, automatic translating speech, sending e-mails etc.
- *Voice Assistants and Chatbots*: Software that uses NLP and speech recognition in order to understand a user's voice commands and act accordingly. Likewise, chatbots are programs designed to assist and respond to any query that the user might have.
- *Sentiment Analysis*: Even though human sentiments or expressions like sarcasm, hate or admiration are very hard to be recognized by a computer,

NLP is able to identify different sentiments. With this implementation, we are able to analyze customer reactions, social media discussions etc.

- *Social Media Analytics*: Using both NLP and sentiment analysis we are able to gather and analyze an enormous amount of data through social media.

Of course, an industry in which NLP applications are of major importance is the healthcare industry. By gathering medical or patient data and developing NLP models and implementations, not only do we assist doctors in their important work, but we also give them the opportunity to invest more time in patients. With the help of NLP, healthcare industries can make the best use of their gathered data, extract patterns from these data, or facilitate the managerial duties. Clinical Documentation is a helpful application for clinicians, as they are “freed” from EHR’s tasks and they eventually have more time to invest on their patients and their treatment. NLP technologies can also help doctors by pointing out relevant and important data with the help of speech recognition equipment.

This thesis focuses on the early detection of Alzheimer’s disease, by using AI implementations, such as data gathering and processing or feature extraction, as well as NLP techniques, like tokenization and lemmatization. Our purpose is to examine several patients’ interviews, detect important data concerning Alzheimer’s disease symptoms, analyze them and find out if the candidates are possible future Alzheimer’s disease patients, with the use of Machine Learning algorithms.

4 NLP for Neurodegenerative Disease Detection

As stated in the previous chapter, AI and especially NLP applications play a vital role in the healthcare industry. It is widely known that the healthcare system is an industry that consists of vast and growing amount of information obtained from discharge summaries, physicians' case notes, pathologists or radiologists' reports etc. This information is usually gathered arbitrary without a standard form in healthcare systems, which make it difficult for systems to understand the contained information. The challenge that rises from this unstructured way of storing information is getting access to valuable and meaningful healthcare information for decision making purposes. Moreover, with enormous amounts of data present in a patient record, manual retrieval of important information for a specific clinical task is often challenging and may cause cognitive overload and inefficiency for physicians.

Before 2010, healthcare technology companies were focusing on medical products that provided historic and evidence-based care, while after 2010, they turned on real-time medical platforms and outcome-based care. However, since 2020, technology has moved towards medical solutions that are achieved by using robotics, virtual and augmented reality and Artificial Intelligence [13]. With the help of AI and its sub-domains, we can now have intelligent solutions focused in collaborative and preventative care.

NLP techniques are being used in order to structure narrative information in healthcare [14]. This happens thanks to the capability of NLP techniques for capturing unstructured information, analyzing it grammatically, determining its meaning and translating it so that it can be understood by the electronic healthcare systems. Additionally, NLP techniques help reduce the costs, as well as improve healthcare quality levels.

4.1 AI Techniques in Healthcare

The ongoing advance of Information Technology and its modern tools and applications have offered, and still offer, a great help to our personal and

professional lives. Of course, the field of healthcare could not be left unaffected, as it was highly benefited by modern technology and tools such as Electronic Medical Records (EMR) or Electronic Health Records (EHR). These tools managed to improve healthcare processes as they assisted in providing timely access to important information, reducing healthcare cost and errors or ensuring confidentiality and security of healthcare information. Moreover, an effective method was established in order to store large amounts of information relating to diagnosis, medication, test results or imaging data that were, as stated earlier, highly unstructured. We can classify AI applications in healthcare in various domains, such as surgery, nursing assistant, medical consultation, machine vision, automatic and preliminary diagnosis, health monitoring and clinical trials. The above applications can be actualized by the following AI methods:

- *Machine Learning (ML)*: It can be categorized in three subareas. First subarea is Supervised Learning algorithms, which build models using data that contain both the inputs and the desired outputs. Then we have the Unsupervised Learning algorithms, which take data containing only inputs and their goal is to discover useful structure by which to generate informative outputs. Finally, there are the Reinforcement Learning algorithms which do not assume any prior knowledge of a model or an environment and they focus on autonomously discovering strategies that result in large rewards, given an objective function.
- *Natural Language Processing (NLP)*: It consists of techniques and tools that allow computers to read, understand, manipulate and extract a meaning from human language.
- *Neural Networks (NN)*: They receive digitized inputs, such as image or speech, which then are processed through several hidden layers of artificial neurons which in their turn detect features and provide outputs.
- *Deep Learning (DL)*: Based on artificial NNs, DL algorithms use multiple layers to extract higher level features from raw input.

Thanks to those methods, healthcare industry has been greatly advanced in a wide range of fields. Multiple fields of surgery such as cardiac, thoracic, orthopedic or general surgery have taken advantage of robotic-assisted surgical systems (RASS) and computer-assisted surgery (CAS) since 1985 [13]. With the help of those systems,

surgeons can now use cameras and tools to perform procedures with a speed and precision that we could only imagine in the past. Physicians are also able to collect data for further analysis, enhance the surgery process, use computers for surgical planning and improve surgical efficiency and postoperative efficacy. According to several researches [13], RASS and CAS have shown multiple benefits in surgeries, such as reduction of length of hospital stay, reduction of complications, reduction of operation times and reduction of costs.

Another field which was majorly benefited is the medication management and medication error reduction (MMMER). MMMER services can provide healthcare cost reduction and minimize unnecessary injuries and deaths. With the help of AI, we have observed an improvement in medical safety by detecting potential medication errors or issues, a reduction in time and expenses, as well as high rates in predicting health risks and outcomes.

A field that was also assisted by AI is the field of preliminary diagnosis and precision (PDP), which is also the field that we will examine in this thesis. Latest advances in AI researches have shown that AI methods are currently outperforming physicians in terms of speed and accuracy concerning medical diagnosis and prediction. For many years, specialists have been using health history and diagnosis data in order to obtain a more accurate patient diagnosis. However, using AI methods has been proved a much accurate and safe way to predict several diseases, like diabetes, cancer and psychiatric or neurological diseases. With ML and NLP algorithms, we can now detect symptoms and predict possible diseases with really high accuracy levels in considerably less time compared to manual operations.

4.2 NLP in Healthcare

In biomedical and health areas, natural language is the primary means of storing and circulating knowledge and data concerning matters that vary from scientific articles to patient information and reports. Because of the enormous size of information and data circulated between scientists and healthcare facilities, it has become difficult for physicians to keep up with every detail, thus they need help to find, manage and analyze knowledge and data needed for any occasion. A way to facilitate physicians' tasks is NLP, which gives them the ability to automate methods with high reliability and validity. Mainly, NLP techniques aid us to transform data in a computable

format in a way that allows us to use natural language, while, at the same time, enabling important applications for processing data effectively [15].

NLP consists of the following levels of analysis, which take advantage of any lexico-syntactic attributes of a language and help ML methods to confront several health-related problems and challenges: [14]

- *Phonological Analysis*: it is related to speech sounds of a language and their interpretation within words.
- *Morphological Analysis*: it is defined as the scientific study of the structure of words in a language and the relationship between each other. Given the fact that medical terminology has a very rich vocabulary and morphological structure for chemicals and procedures, the morphological analysis of all these aspects with NLP methods can produce important findings concerning the cognitive ability of a patient.
- *Lexical Analysis*: the process of converting a sentence or a sequence of characters into a sequence of tokens, also known as Tokenization.
- *Syntactic Analysis*: it is related to the construction of each sentence in a language and the relationship between the words in this sentence.
- *Semantic Analysis*: it focuses on the meaning of words, phrases and sentences in a language and the possible interactions that may exist among word meanings in the sentence. This analysis is of great importance, as many words in a natural language have diverse meanings, making it challenging for a system, or a person, to distinguish the proper meaning of a word or a sentence.
- *Pragmatic Analysis*: it focuses on the way different sentences combine in order to form a paragraph, document or dialogue. It also concerns the interpretation of each individual sentence in its context.

NLP methods can be applied in various healthcare fields which exploit human natural language, such as clinical notes, patient records or medical forms. According to Attrey R. & Levit A. [16], NLP applications in healthcare can be focused on four possible categories:

- *Patients*: The use of NLP-enabled chatbots, which provide spoken or written dialogue, could assist nurses and physicians by providing basic disease

information or instructions for healthcare to patients. Since NLP techniques have already been established as successful tools for analyzing clinical text, chatbots are able to handle healthcare-related text and help reduce operational costs, provide services in a patient's native language or even detect cognitive impairment symptoms through language patterns.

- *Physicians*: Electronic health records play a vital role in clinical decision-making by physicians. However, these records have a large amount of unstructured text data making them difficult for health professionals to manage. NLP algorithms are able to analyze all these unstructured data and help clinical decision support systems (CDSS). NLP gives the ability to CDSS to observe more concerning the crucial data and alert physicians for possible suspicious symptoms, or even the complete detection of a disease.
- *Researchers*: NLP offers a great chance for qualitative studies, such as an analysis of patient interviews which can offer more data and new insights.
- *Healthcare Management*: NLP can gather patients' feedback in an unstructured written or spoken form and produce summaries that would help healthcare management implement quality improvement actions.

In this thesis we focus more on the category of Physicians. By using NLP techniques and developing NLP and ML algorithms, we focus on processing gathered unstructured data, extracting important features and detecting dementia-related diseases, especially Alzheimer's disease.

4.3 NLP in Detecting Dementia Diseases

As mentioned in the state-of-the-art chapter, Alzheimer's disease and dementia-related diseases in general, represent an advancing health crisis worldwide. This health crisis has led to high demands of diagnostic services related to defects in memory and cognitive performance. Physicians can assess a dementia disease biologically by examining specific biomarkers through positron emission tomography (PET), or through magnetic resonance imaging (MRI). However, these techniques require a well-resourced medical and professional environment, while at the same time they do not determine cognitive decline with certainty, as there are cases in which brain pathology does not translate into clinical expression. For that reason, a neuropsychological exam is now conducted to possible patients of dementia diseases

in the form of in-person interview. This is currently the primary method which, in conjunction with clinical exams, is able to detect cognitive impairment in early stages [17]. Given that cognitive decline can take years to evolve to more severe stages, automatic detection of early stages, also known as Mild Cognitive Impairment (MCI), is crucial in order for physicians to intervene as earlier as possible.

Several studies have shown that spoken language is a rich source of information and plays a vital role in evaluating a person's cognitive status. Patients with dementia usually begin with exhibiting naming deficits and as the disease advances, all aspects of language are affected. Patients present word-finding problems, sentence comprehension deficits and lack of cohesion in discourse. Even though symptoms may vary between specific dementia diseases, (Alzheimer's disease (AD), semantic dementia (SD) or Parkinson's disease (PD)), data gathering, their processing and analysis are achieved in the same way regardless of the disease examined.

Data gathering is achieved by submitting possible patients to various tests, like in-person interviews, picture description or word-picture matching. One common test is the Mini-Mental State Examination (MMSE), a 30-point 11-question measure that is used to systematically and thoroughly assess mental state [19]. It tests five areas of cognitive ability: orientation, registration, attention and calculation, recall and language. The questionnaire takes 5-10 minutes and it includes simple questions and problems, like mentioning the time and place of the test or repeating certain words. The MMSE is an effective tool for detecting and separating patients with cognitive disabilities from those without it, as well as for identifying possible changes in cognitive status after a physician's intervention. However, given the fact that MMSE relies on verbal response, writing and reading, patients with hearing or visually impairments, with low specific language literacy or with other communication problems could perform poorly on this test, even if they are cognitively intact.

Another commonly used test for cognitive evaluation is the Clock Drawing Test (CDT). The test is a nonverbal screening tool that begins by asking the patient to draw a simple clock. The next step for the patient is to place the numbers around the clock and then draw the hands in the clock to indicate a certain hour. Placement of the numbers requires visual and spatial abilities, as well as numerical sequencing abilities. The CDT overall is also able to assess patient's long-term attention,

memory, motor programming and frustration tolerance. It is considered a better test than MMSE, given the fact that by its nature it does not include any language knowledge or cultural biases.

A final test that physicians use is the Cookie Theft Picture Description (CTP), designed by the Boston Diagnostic Aphasia Examination. During this test, patients are shown a familiar picture of domestic scene with basic vocabulary terms, distinct characters, time and place. Patients are then instructed to tell everything that is going on in the picture. A basic advantage of this test is the ability to retrieve appropriate lexical items, while it can also quantify the patients' discourse by measuring the frequency of syllables or the total number of complete words. Another basic advantage of CTP test is the vast range of semantic categories that the words it contains belong to. Choosing the correct word or phrase to describe all the entities, help clinicians to examine patients' semantic level. Furthermore, CTP allows us to examine the subject's ability to detect causal and temporal relations between entities in the picture. Finally, CTP is an ideal tool to assess a subject's structural language skills and motor speech production. Subjects' picture descriptions may present impairments related to phonology, syntax or semantics or they may even contain pauses, as subjects look for specific words. If the search for the right word is unsuccessful, non-specific words like "someone" or "stuff" will replace them.

To be able to gather the ideal dataset, scientists need their patients to freely talk about a specific topic without any intervention from outside factor. This is why tests like describing a picture or answering to a specific questionnaire are ideal for these cases, as they give patients the chance to talk in their own pace and rhythm which can produce more liable and independent results. Through all the above tests, scientists are able to obtain the necessary dataset and process it in a way that it enables them to identify semantic, lexicosyntactic or any other important feature for each dementia-related disease, using NLP techniques. From there, ML algorithms are deployed in order to be trained and finally detect a possible dementia-related disease.

4.4 NLP in Detecting Alzheimer's Disease

Patients with AD have trouble remembering names and other words. To deal with this problem, AD patients often choose to substitute the "missing" word with

pronouns (e.g., “he”, “it”), use different words that are conceptually similar to the “missing” one (e.g., “dog” instead of “horse”), or even pause when they cannot find the right word in a conversation. Moreover, extralinguistic deficits could produce anomia in spontaneous speech or picture-naming tasks. Those extralinguistic deficits can also include inattention to a task, forgetting the needed word or being easily distracted by other responses. Furthermore, AD patients may suffer from semantic impairments resulting in a loss of semantic memories, lexical production difficulties or difficulty in any other task related to comprehension, knowledge of category relationships and attributes. [18]

Studies through word-picture matching done by several researchers have shown that the semantic system of an AD patient is able to distinguish semantically related words between each other. Other tasks have also shown that AD patients perform similar to healthy controls in definition and word similarity tasks. Even though early signs show that the loss of semantic features is minimal and it does not interfere with semantic tasks, the overall picture derived from several studies is that as the disease advances, errors change in quality (e.g. “apple” in early stages is substituted by “pear”, which later on is substituted by “fruit” or “I don’t know”), comprehension is declined and difficulties in other tasks become more apparent. At this point, errors can be associated to extralinguistic and semantic deficits.

With all of the above in mind, scientists have began building ML diagnostic models using vocal and lexical features extracted from voice recordings or text transcripts. Those models usually take either an audio file or a text transcript as input, and produce a prediction of possible AD positive subject, using ML and NLP techniques. By preprocessing the available data and extracting any important feature based on the known symptoms of AD, scientists are able to train the algorithms in order to evaluate and detect a possible AD patient.

In the following chapters we are going to present all the steps that we followed in this thesis in order to obtain our dataset, complete all the necessary NLP processes for data manipulation and feature extraction and build the ML models for detecting Alzheimer’s disease.

5 Methodology and Implementation

As we have already stated, the purpose of this thesis is to build an NLP system able to detect AD in its early stages. Detecting neurological diseases as early as possible is of utmost importance for any patient, as well as for medical staff. With the help of NLP methods, clinicians are able to enhance their examinations and their diagnoses in order to identify possible patients in a quicker way. An early AD diagnosis allows patients to access treatment in order to reduce cognitive and functional decline or even lessen symptoms such as memory loss and confusion. Patients diagnosed with early AD have also the chance to prioritize and change their lifestyle or even participate in clinical trials. Of course, another great benefit of early AD detection is the cost saving for both the patients' families and the government. According to Alzheimer's Disease Facts and Figures [35], an annual report released by the Alzheimer's Association, among all citizens in the USA, if those with Alzheimer's disease were diagnosed when they had mild cognitive impairment, before dementia, it would save approximately \$7 trillion in health and long-term care costs.

To achieve our goals, there is a number of “steps” that we need to follow, each with its own great importance for the final result. First of all, there is the gathering of the necessary data on which our model will be trained and finally evaluated. Finding the right data is both a challenging and a vital task for a project, as they are essentially the core of the system. If data is falsely gathered, few in numbers or if there is any other issue, final outcome may not be as valid as expected. Data is considered as the most important piece of Machine Learning, as without data, ML models would not be able to learn, grow and make necessary decisions. Data is usually gathered through surveys, questionnaires, interviews etc. If the above techniques are not possible, then data augmentation takes place, a technique that produces fake data which can later be used in machine learning.

The next step is the data preprocessing part, which can often have a great impact on the performance of a machine learning algorithm [20]. After having collected the necessary data that we are going to use in order to train and evaluate our model, we

have to be sure for certain aspects. Firstly, data have to be enough in numbers for the model to reach our ultimate goals. Having a very small amount of data often leads to a poor predictive performance that comes from two main disadvantages, over-fitting and under-fitting. In the first, the model is trained by reading too much of the very few available data. The above means that the model eventually memorizes the data patterns and produces low training errors and high test errors. In the case of under-fitting, because of the dataset's small size, the model is unable to detect relationships in the data and it ends up being a really simple model. In any machine learning or data analytics task, the models' success highly depends on the relevance and comprehensiveness of the training data, making it clear that the higher the quality of the data, the better the performance of our models. Some of the top finest characteristics of quality data are:

- *Accuracy*: Indicates how accurate the data is. Outdated, redundant or data with typos are considered as data of poor quality.
- *Consistency*: Having consistent data ensures that the results will be the same for each experiment ran.
- *Validity*: Indicates if the data format is the right type or not. Invalid data makes the whole process more difficult.
- *Completeness*: Having complete data leads to having entire information for the model training. Incomplete information or missing values result to poor accuracy of the results.

From the above we can understand the high importance of data preprocessing and how it affects any AI development and its respective accuracy and results.

Moreover, a common problem scientists have to confront when dealing with newly collected data, is the possible presence of noise instances or the absence of several data parts. To cope with these problems, we follow the process of data cleaning in order to fix possible incorrect, incomplete or duplicate data. Data cleaning helps improve the quality of the dataset and provide more accurate and reliable information that will later be used for training. This is also an important process as we use it to solve several problems and correct typos or invalid data, adjust data format so that all data are consistent with each other, identify and remove or merge duplicate data etc.

Even though real-world data usually is consisted of too many features, we may only need a few of them for a project. We may also detect several correlated features, so that it is not necessary to use all of them for our model, thus we could exclude the redundant ones. Also, we must detect possible dependencies between two or more features of the dataset which carry important information and it would be wrong if we included one of those features on its own. To address the above challenges, we have to follow the process known as “feature selection”. Feature selection allows us to identify and remove irrelevant and unnecessary information. By doing so, we manage to reduce the dimensionality of the data and allow our models to operate faster and more effectively [20]. For this thesis, we have mainly used NLP methods in order to extract the ideal features.

At last, after turning the data in the optimal format and extracting the necessary features, we are ready to build the model to be trained and evaluated. Depending on the task and the available dataset, ML model training can be supervised or unsupervised, both containing several methods based on the project purpose. In our case, since we have a collection of labeled data and our goal is to predict if a subject suffers from AD or not, we are using supervised machine learning classification methods.

In the next sections we are going to describe each step that we followed for our AD detection model.

5.1 Dataset

Given that this thesis focuses on NLP methods for detecting neurological diseases (specifically Alzheimer's Disease), the ideal dataset would be a collection of subjects' natural language samples derived from interviews, conversations, essays or any other possible way of collecting such data. By using such a dataset and not, for example, a collection of MRI images we have the opportunity to investigate subjects' communication characteristics and their capability or incapability of using natural language properly.

A significant work towards building a database with data regarding cognitive disorders is done by Brian MacWhinney, Professor of Psychology and Modern Languages at Carnegie Mellon University. Through his project, TalkBank [36], and

with the support and cooperation of several members and contributors, he has set a goal to promote research done in the studies of human communication with emphasis on spoken communication. According to their website, www.talkbank.org, TalkBank currently provides repositories in 14 research areas and data in the project have been contributed by hundreds of researchers in over 34 languages all over the world. Among the research categories available in this project, there are conversation banks, child language banks, multilingualism banks, clinical banks etc. The project has been highly utilized for spoken language analysis and clinical research.

For this thesis, we have decided to use the DementiaBank, a database with clinical multimedia interactions built for the study of communication in dementia¹. DementiaBank contains sound clips and their transcripts of 156 real interviews that took place between clinical investigators and participants. Both sound clips and transcripts are provided, while subjects' final diagnosis is also given in the format of "Control" if subject is healthy and "AD" if subject is a possible AD patient. For those interviews, picture description tasks have gained a great attention and they are being used extensively by clinicians in order to examine people with language disorders. These tasks may also be used to assist the assessment of formal language or to extract sentence production during therapy.

5.1.1 The Cookie Theft Picture

During the interviews for the gathering of the dataset that we used, participants were given the cookie-theft picture (Figure 1), a picture from the Boston Diagnostic Aphasia Examination that has managed to dominate in similar clinical processes more than any other task [21]. This picture depicts a domestic scene, in which a mother and her two children are shown in the kitchen. The mother is drying dishes next to the sink and, as she is not paying attention, she has left the tap on making the

¹ To gain access to this dataset, one has to become a member of DementiaBank by getting in touch with Brian MacWhinney. For students, though, a supervisor professor must join as a member and then give access to the students. For that, I have to thank my supervisor, Researcher Vassiliki Rentoumi, who gained access to DementiaBank and provided me with the necessary dataset that was used in this thesis.

water overflow from the sink. Meanwhile, the two children are trying to reach a cookie jar from a cupboard when their mother is not looking. In order to get up to the cupboard, the boy has climbed on a stool which is moving dangerously. The girl is standing next to the stool with her hand stretched so that she can take the cookies. The interviewer asks the participants to describe everything they see and everything that is going on in the picture without being guided by the examiner. The examiner may point out some neglected elements of the drawing or ask for more details if the subject's answer is less than might be expected.

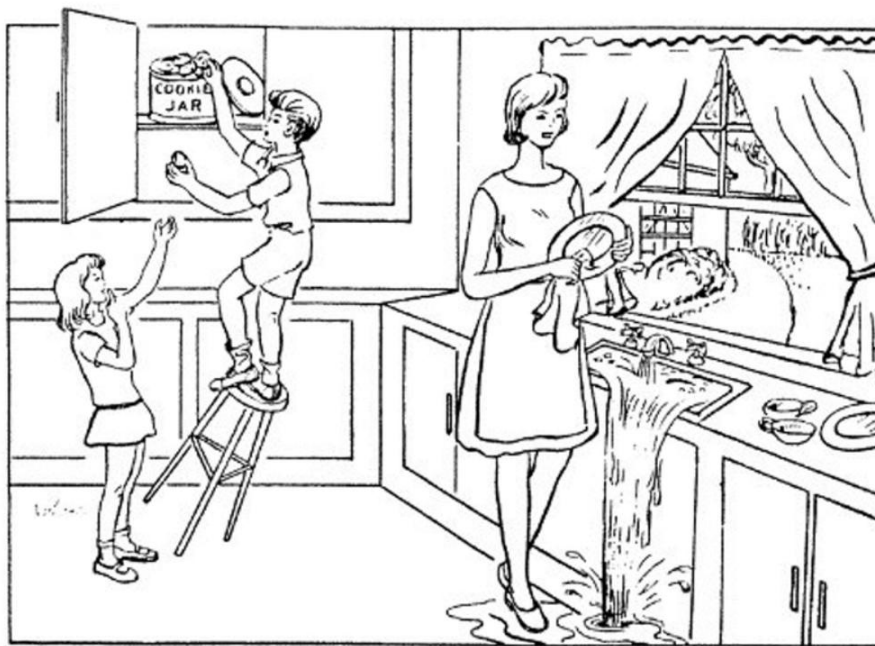


Figure 1: Cookie Theft Picture

The Cookie Theft picture is being used by examiners to detect language deficits for several clinical conditions, such as adults with corticobasal syndrome and progressive aphasia, adults who have experienced right and left-hemisphere strokes or adults with mild cognitive impairment and Alzheimer's disease [21]. The Cookie Theft picture's large diversity of uses shows its greatness and importance in the attempts to detect cognitive-linguistic impairments. According to Professor Louise Cummings, there are seven important features in this task which can assist the examination of the above impairments:

1. *Salience of information*: The Cookie Theft picture includes several degrees of salience. The three characters in the scene and their actions belong to the

highest level of salience, while less important information is shown in the background such as the items shown in the kitchen, the plants and trees shown in the garden through the window or the clothes that the characters are wearing. The ability or not to present the information ordered from the most important subjects to the less important ones is a way to acknowledge an examinee's neurological impairment.

2. *Semantic categories*: The task contains objects, people and actions which can be described by using a words from a wide range of semantic categories. The subjects must have the ability to use words for living entities (mother, boy) or for inanimate objects (table, window). They must also be able to express concepts (the boy is falling) or more abstract ones (the mother is daydreaming). Furthermore, most elements in the picture can be described either by using general terms (woman, child, plant, doing the dishes) or by using more specific terms (mother, son, tree, drying the plate). These different levels of generality or specificity allow the clinicians to inspect the subjects' ability regarding terms interrelationships within a semantic network.
3. *Referential cohesion*: The Cookie Theft picture task allows examiners to investigate the subject's ability to introduce people and objects and refer to them later with the correct use of pronouns. Because of the existence of more than one similar characters (two female characters, two children etc.), it is really important for the subjects to have the ability to use anaphoric references in order to achieve referential cohesion. If the examinee uses just the personal pronoun "she", the interviewer is not able to understand which of the two female characters is referenced. The same confusion can arise during the description of inanimate objects, such as the stool, the table, the cookie jar which are all referred to by using the pronoun "it". Studies have shown that referential cohesion is a really difficult task to succeed for Alzheimer's disease patients when they are asked to describe a picture.
4. *Causal and temporal relations*: Even though the Cookie Theft picture is a static scene, an examinee must be able to capture and describe several events that are linked through causal and temporal relations in order to have a

complete description of the picture. Describing events with such causal and temporal relations is a task that Alzheimer's disease patients find difficult to accomplish.

5. *Mental state language*: Thanks to the rich context provided by the Cookie Theft picture, clinicians are able to study the subjects' mental state language and their cognitive ability to associate mental states to their own or the others' minds in order to explain something or predict a behavior. In the Cookie Theft image, the subject must be able to understand that the boy, for example, is climbing the stool because he wants to grab the cookie jar. An examinee, who is able to present such a language during an image description, is a person with intact mental state skills.
6. *Structural language and speech*: Like any other image description task, the Cookie Theft picture task allows clinicians to assess the examinee's structural language skills. Phonology, syntax and semantics impairments are usually obvious during a picture description. Subjects with dementia may present considerable difficulties in finding the right words. They may also use many filled and unfilled pauses while looking for a particular word and if they are unable to find the right word, they may use non-specific vocabulary like "someone" or "stuff" in their place. Among the structural language problems, we may also detect limited syntax, usage of simple subject-verb-object word order, or avoidance of passive voice.
7. *General cognition and perception*: A description of the Cookie Theft picture can be considered as successful, if the subject demonstrates the ability to address and perceive every aspect of the picture. Patients with neurological impairments have a decline in general cognitive and perceptual skills that are required for the description task.

From the above, we can understand that the Cookie Theft picture description task gives us the ability to investigate and detect a great range of cognitive-linguistic skills and deficits in patients with dementia.

5.1.2 The ADReSS Challenge Dataset

The ADReSS challenge [22] is a public challenge created by the team of DementiaBank and its main objective is to make available a benchmark dataset of spontaneous speech on which different approaches will be built focusing on the automated recognition of Alzheimer’s dementia. It consists of two tasks:

1. A classification task, where participants are required to implement a model that predicts the healthy or non healthy label for a subject.
2. An MMSE score regression task, where participants have to create a model deduce the subject’s Mini Mental Status Examination (MMSE) score.

The dataset we use for this thesis is the same dataset that members gain access to in order to participate in the challenge. The dataset consists of the following files:

- *Full wave enhanced audio folder*: contains the audio recordings after noise removal
- *Normalized audio-chunks folder*: contains the .wav files which are extracted from the audio recordings after applying voice activity detection
- *Transcription folder*: contains the transcription of audio recordings along with meta-data, such as age, gender and MMSE score
- *Two text files*: contain the meta-data (age, gender, MMSE score) for each healthy and non-healthy subject

For our thesis, we use the audio recordings transcriptions of 156 participants. Transcriptions are saved in .cha files and they contain both the interview that took place between the examiner and each subject, and some other technical details regarding the interview, such as character encoding, language or participants’ details. Since we already have gathered the necessary dataset thanks to the DementiaBank database, we now have to prepare the data in order to use them for the training and the evaluation of our models.

5.2 Data Preprocessing

Our first engagement with the dataset is the data preprocessing part. As already mentioned, real-world data are usually incomplete, noisy or inconsistent. Thus, we have to make sure that the dataset that is going to be used for the model training will

be of the best possible quality. By completing the data preprocessing part and eliminating inconsistencies, duplicates, missing values or any other form of errors, we will eventually have a dataset easier for the computer to interpret and use.

The first step is to clean the transcription files and give them a format that will later enable us to use them in order to extract necessary features. By cleaning the available data, we are able to offer complete and accurate samples to our machine learning models. An example of our transcripts' form is shown in Figure 2. According to the TalkBank manuals [23], a .chat transcript begins with a series of headers which begin with a "@" and consist of lines of text that contain information about the participants and the setting. Some important headers appearing in Figure 2 are described as follows:

- *@UTF*: This is a hidden header and it is used to mark the fact that the file is encoded in UTF8.
- *@PID*: A hidden header that declares identification for the file.
- *@Begin*: This is always the first visible header and it is always placed in the beginning of the file.
- *@Languages*: The second visible header. It declares the language in which the dialogues are performed.
- *@Participants*: It lists all of the speakers that appear in the file.
- **INV, *PAR*: They denote which participant talks at the specific line of text.
- *%mor*: Lines that begin with this header contain the morphological analysis of the text that is written in the above line.
- *%gra*: These lines represent the grammatical dependency analysis of the above line.

```

1 @UTF8
2 @PID: 11312/t-00002184-1
3 @Begin
4 @Languages: eng
5 @Participants: PAR Participant, INV Investigator
6 @ID: eng|Pitt|PAR|74;|male|Control||Participant||
7 @ID: eng|Pitt|INV||||Investigator||
8 @Media: S001, audio
9 *INV: tell me everything that you see going on in that picture . NAK0_2360NAK
10 %mor: v|tell pro:obj|me pro:indef|everything pro:rel|that pro:per|you
11 v|see n:gerund|go-PRESP adv|on prep|in det:dem|that n|picture .
12 %gra: 1|0|ROOT 2|1|OBJ2 3|1|OBJ 4|6|LINK 5|6|SUBJ 6|3|CMOD 7|6|OBJ 8|6|JCT
13 9|6|JCT 10|11|DET 11|9|POBJ 12|1|PUNCT
14 *INV: everything that you see happening . NAK2360_4266NAK
15 %mor: pro:indef|everything pro:rel|that pro:per|you v|see
16 n:gerund|happen-PRESP .
17 %gra: 1|0|INCR00T 2|4|LINK 3|4|SUBJ 4|1|CMOD 5|4|OBJ 6|1|PUNCT
18 *PAR: well there's a mother standing there &uh &uh washing the dishes
19 an(d) the sink is overflowing [: overflowing] [* s:r] . NAK4266_13310NAK
20 %mor: co|well pro:exist|there~cop|be&3S det:art|a n|mother
21 part|stand-PRESP adv|there part|wash-PRESP det:art|the n|dish-PL
22 coord|and det:art|the n|sink aux|be&3S over#part|flow-PRESP .
23 %gra: 1|3|COM 2|3|SUBJ 3|0|ROOT 4|5|DET 5|3|PRED 6|5|XMOD 7|6|JCT 8|6|XJCT
24 9|10|DET 10|8|OBJ 11|10|CONJ 12|13|DET 13|15|SUBJ 14|15|AUX 15|11|COORD
25 16|3|PUNCT
26 *PAR: an(d) &uh the window's open . NAK13310_20608NAK
27 %mor: coord|and det:art|the adj|window&dn-POSS adj|open .
28 %gra: 1|4|LINK 2|4|DET 3|4|MOD 4|0|INCR00T 5|4|PUNCT
29 *PAR: and outside the window there's a <walk with a> [//] &c curved walk
30 with a garden . NAK20608_27071NAK
31 %mor: coord|and prep|outside det:art|the n|window
32 pro:exist|there~cop|be&3S det:art|a part|curve-PASTP n|walk
33 prep|with det:art|a n|garden .
34 %gra: 1|6|LINK 2|6|JCT 3|4|DET 4|2|POBJ 5|6|SUBJ 6|0|ROOT 7|9|DET 8|9|MOD
35 9|6|PRED 10|9|NJCT 11|12|DET 12|10|POBJ 13|6|PUNCT
36 *PAR: and you can see another &uh &uh building there . NAK27284_32813NAK
37 %mor: coord|and pro:per|you mod|can v|see pro:indef|another
38 part|build-PRESP adv|there .
39 %gra: 1|4|LINK 2|4|SUBJ 3|4|AUX 4|0|ROOT 5|4|OBJ 6|5|XMOD 7|6|JCT 8|4|PUNCT
40 *PAR: looks like a garage or something with curtains and the grass in the
41 garden . NAK32813_37356NAK
42 %mor: v|look-3S conj|like det:art|a n|garage coord|or pro:indef|something
43 prep|with n|curtain-PL coord|and det:art|the n|grass prep|in
44 det:art|the n|garden .

```

Figure 2: Example of initial transcription format

After understanding the transcription’s structure, we have to remove any unnecessary information that exists in it. As we stated earlier, transcriptions begin with general information, such as encoding (@UTF8), language (@Languages), participants (@Participants) etc. These lines offer no important information for our future model so they have to be removed. Moreover, since we only need to investigate subjects’ spontaneous speech, we have to remove any part that belongs to the examiner which is declared by the identifier “*INV” at the beginning of the specific lines. Finally, we must also remove any other line that does not belong to the subjects’ speech, the parts that begin with “%mor” or “%gra”.

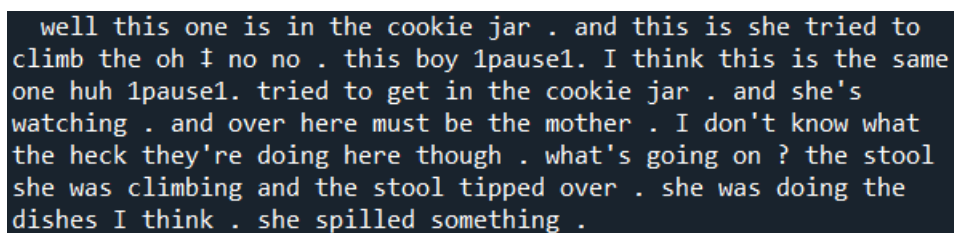
Eventually, we only need to keep participant’s speech which is denoted by “*PAR” in the file, as well as participant’s personal information which are written in the 6th line of the transcription and include subject’s age, gender and AD category (*Control or ProbableAD*). However, getting only the subjects’ parts of the interview is not enough. We also have to identify and remove any extra content like throat clearings

('=&clears throat'), tags and fillers ('&=', '['*'' etc.) or interviewer's responses ('mhm', 'okay', 'oh' etc.).

After managing to get only the necessary parts of the file (Figures 3, 4), we have to distinguish any possible important elements in the subjects' transcripts which may be later used as features for the training of our models. Based on bibliography and any minor or major known symptoms of Alzheimer's disease, we are able to identify and extract those aspects from the transcripts. Aspects that we are looking for in the transcript are the following:

- *Trailings*: Subject's voice becomes softer and softer as he/she speaks until it completely stops.
- *Pauses*: A rest, a hesitation or a temporary stop in subject's speech.
- *Unintelligible words*: Subject's use of words that is very difficult or completely impossible to be understood by the examiner.
- *Repetitions*: Subjects use single words or whole sentences multiple times in a short period of time.

All of the above show signs of cognitive and linguistic difficulties that are in direct correlation with Alzheimer's and other neurological diseases. The more often these signs appear in a subject's interview, the more possible it is for the subject to suffer from Alzheimer's disease.



```
well this one is in the cookie jar . and this is she tried to
climb the oh ‡ no no . this boy 1pause1. I think this is the same
one huh 1pause1. tried to get in the cookie jar . and she's
watching . and over here must be the mother . I don't know what
the heck they're doing here though . what's going on ? the stool
she was climbing and the stool tipped over . she was doing the
dishes I think . she spilled something .
```

Figure 3: Possible AD patient's spontaneous speech after cleaning the transcript

```
many or the mother's washing the dishes and the sink is overflowing . she has some of them dried already on the side as she's looking out the window while the little boy is falling off the stool because he's getting into the cookie jar to give to his little sister who's reaching up to get the cookie also . there's water all over the floor . there's the garden is outside and the mother's not paying any attention to what they do . the stool is tipping . the cookie jar the door is open . there may be a little breeze coming in because the this window is open . 1pause1. the little girl is saying has her finger to her mouth shh we won't tell mother while you give me the cookie . 1pause1. it's in the kitchen ofcourse and the cups two cups and a dish already have been dried . and the mother's stepping in the water and she's probably so engrossed in what she's doing outside she neither knows what the children are doing nor is she paying any attention that the water's overflowing .
```

Figure 4: Healthy participant's spontaneous speech after cleaning the transcript

The data preprocessing steps are implemented in the files “*data_processing.py*” and “*feature_extraction.py*”. The *data_processing* file produces a *.csv* file (*processed_data.csv*) which contains comma-separated values divided in the following twelve columns:

- *filepath*: The original *.cha* file name
- *age*: Subject's age
- *gender*: Subject's gender
- *mmse*: Subject's mini mental state examination (MMSE) score
- *pause1*: Unfilled pause that takes up a “normal” duration and is marked only by silence. In the transcript, it is represented by “(.)”.
- *pause2*: Longer pauses between words. They are represented by “(.)”.
- *pause3*: Very long pauses. Represented by “(..)”.
- *unintelligible_words*: Words with unclear phonetic shape. They are represented by “xxx”.
- *trailings*: Incomplete, but not interrupted, utterance. It is symbolized with “+...” and it occurs when the speakers shift their attention away from what they are saying. It is usually followed by a pause in the conversation.
- *repetitions*: The speaker says the same word or group of words several times in a row without any changes.
- *category*: The clinical category in which the speaker belongs, which can be either “Control” or “ProbableAD”. If the speaker belongs to the first category

(Control) then he/she is healthy. Otherwise, if the speaker belongs to the “ProbableAD” category, then he/she is a possible Alzheimer’s disease patient.

- *data*: The subject’s spontaneous speech as it was documented in the transcript and after it was cleaned by our preprocessing techniques.

An example of this exported .csv file can be seen in Figure 5.

	A	B	C	D	E	F	G	H	I	J	K	L
1	filepath	age	gender	mmse	pause1	pause2	pause3	unintelligible_words	trailings	repetitions	category	data
2	S001.cha	74	male		0	0	0	0	0	1	Control	well there's a mother standing there washing the dishes and the sink is overflowing . and the window
3	S002.cha	62	female	30	0	0	0	0	0	0	Control	somebody's getting cookies outof the cookie jar , standing on a stool . the stool's gonna tip over . ar
4	S003.cha	69	female	29	0	1	0	0	1	2	Control	there's a little boy and he's getting he's standing on a stool that's upsetting . and he's getting a cook
5	S004.cha	71	female	30	1	0	0	0	0	4	Control	are you ready ? well the sink is overflowing . mother is standing in the water like a jerk . she's wiping
6	S005.cha	74	female	30	2	0	0	0	0	0	Control	many or the mother's washing the dishes and the sink is overflowing . she has some of them dried
7	S006.cha	67	female	29	3	1	0	0	0	0	Control	1pause1. mother is drying the dishes but the water is going out over the sink onto the floor . it's a p
8	S007.cha	71	male	28	0	0	1	1	0	0	Control	boy taking cookies outof a cookie jar . the stool is falling . the little girl is reaching . water is running
9	S009.cha	67	male	30	0	0	1	0	0	2	Control	a boy is taking cookies from the cookie jar giving one to his sister . he's also falling off the stool he i
10	S011.cha	70	female	30	1	1	0	0	0	0	Control	a girl and a boy and a stool . cookies . cookie jar . open closet . curtains . oh the little boy's reaching
11	S012.cha	77	male	29	1	0	0	0	0	0	Control	the mother is wiping a dish at the sink . the water is overflowing from the sink . a youngster's about
12	S013.cha	57	male	30	1	0	0	1	0	2	Control	fellow falling off a stool . also taking cookies from the cupboard . sister standing . outstretched har
13	S015.cha	70	female	29	0	0	0	2	0	0	Control	all of the action you see going on . this is in the kitchen . and the little boy is climbing up to the cool
14	S016.cha	63	female	30	1	0	0	0	0	1	Control	well the girl is watching the boy go into the cookie jar . he has a cookie in his hand . he's on the stoc
15	S017.cha	65	female	28	0	0	0	0	1	0	Control	well I see the mother doing the dishes . the sink the water running over in the sink . the boy's taking
16	S018.cha	72	male	29	2	0	0	0	0	1	Control	all of the action . just go ahead and tell you ? the mother is drying a plate . and the water's sink is cl
17	S019.cha	57	female	27	0	0	0	0	0	0	Control	well the mother is drying the dishes . the sink is overflowing . the little girl's reaching for a cookie . i
18	S020.cha	58	male	27	3	0	0	1	0	0	Control	1pause1. boy's falling off a stool . the lid is falling off a cookie jar . he's grabbing a cookie in his han
19	S021.cha	64	female	30	1	0	0	0	0	2	Control	the boy is taking cookies . the girl is is saying quiet . he's he's the boy is falling off the stool . the mot
20	S024.cha	73	female	30	0	0	0	0	0	2	Control	the mother is standing at the kitchen sink . the water is overflowing the sink and she's paying no att
21	S025.cha	67	female	28	3	2	0	0	0	1	Control	water running outof the sink . lady drying a plate . and a child getting cookies outof the cookie jar s
22	S027.cha	78	male	29	0	0	0	1	0	0	Control	look at the picture ? oh one of the the boy is on the stool getting cookies from the cookie jar and g
23	S028.cha	59	male	29	0	0	0	0	0	1	Control	washing washing dishes or wiping dishes . the water's running over the sink . the kid is stealing the ci
24	S029.cha	72	female	29	0	0	0	0	0	0	Control	action . a lady's drying dishes . the boy is was standing on a stool but the action is that the stool has
25	S030.cha	70	female	30	1	0	0	0	0	5	Control	okay the little boy is on a stool about to fall . the stool's about to upset . and he has a cookie in eac
26	S032.cha	57	female	28	1	0	0	0	0	0	Control	well the little girl is saying to be quiet to her brother . and her brother's in the cookie jar and he's fal
27	S033.cha	61	female	30	0	0	0	0	0	0	Control	well the water's running over on the floor . the chair is tilting . the boy is into the cookie jar . and hi
28	S034.cha	65	female	29	0	0	0	0	1	0	Control	well this little boy is up on the stool taking cookies handing them down to his sister and she's telling
29	S035.cha	66	female	30	0	0	0	0	0	0	Control	. touching lip . raising arm . is that what you mean ? reaching for cookie . handing cookie down . slip
30	S036.cha	73	male	28	0	0	0	0	0	2	Control	well for one thing this boy's on the stool getting cookies . and his stool's about to fall . and and the l
31	S038.cha	57	male	30	1	0	0	0	0	0	Control	everything that I see going on . well a little boy is stealing a cookie from the cookie jar . he's also ha

Figure 5: Part of the processed_data.csv file

We can also confirm that our dataset is equally distributed in the two available classes, by looking at Figure 6. Having a balanced dataset is an essential characteristic for our classification models to be as more accurate as possible.

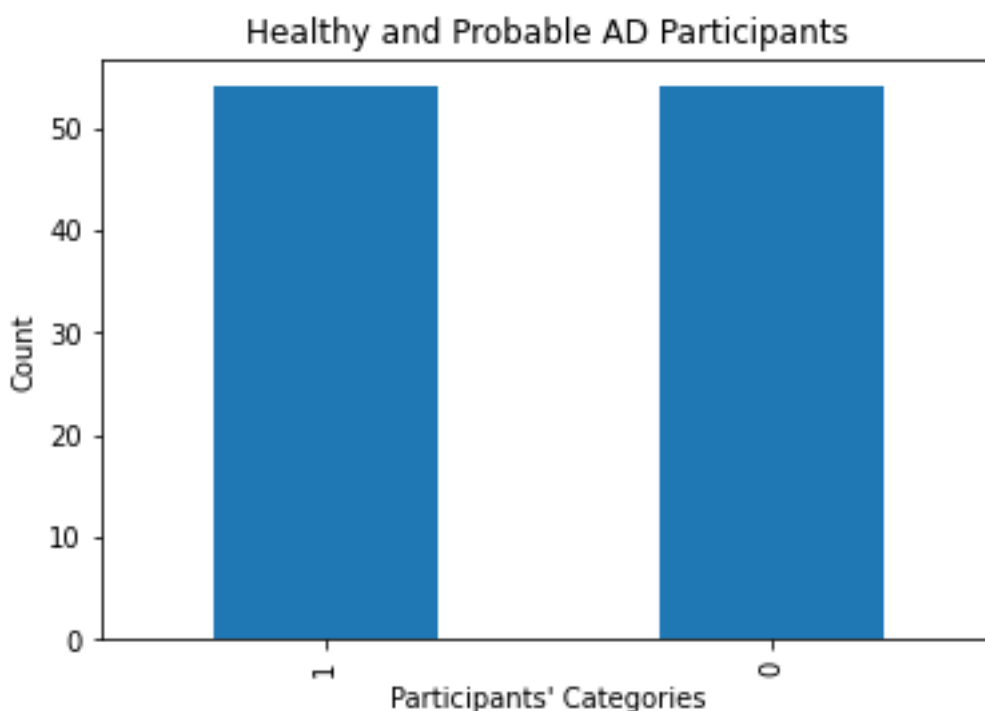


Figure 6: Dataset class distribution. 1 is for probable patients and 0 is for healthy ones

Before beginning to identify and extracting the most important features, we have also implemented several NLP preprocessing techniques by using the NLTK library for Python language [37]. NLTK is a standard Python library that provides a variety of NLP algorithms. It is one of the most popular NLP and Computational Linguistics libraries. With the help of NLTK, we have managed to accomplish several methods vital for feature extraction and the future training of our model.

First, there is the tokenization method, a crucial step in NLP methods, which enables us to separate a piece of text into smaller units known as tokens. Since tokens are the building blocks of Natural Language, the most common method of processing raw text occurs at the token level. Tokenization can be performed on word, character, or sub-word level and the produced tokens are then used to create a vocabulary that refers to the corpus's set of unique tokens. In this thesis, we used the word tokenization method, which splits the text into individual words. By splitting the available text in separate words, it becomes easier to understand the context or to develop the NLP model. Eventually, with tokenization, we convert the unstructured text document into a numerical data structure suitable for machine learning. In Figure 7 we can see a possible Alzheimer's disease patient's text (from fig. 3) after the tokenization method is completed.

```
['well', 'this', 'one', 'is', 'in', 'the', 'cookie', 'jar', '.',  
'and', 'this', 'is', 'she', 'tried', 'to', 'climb', 'the', 'oh',  
'I', 'no', 'no', '.', 'this', 'boy', '1pause1', '.', 'I',  
'think', 'this', 'is', 'the', 'same', 'one', 'huh', '1pause1',  
'.', 'tried', 'to', 'get', 'in', 'the', 'cookie', 'jar', '.',  
'and', 'she', "'s", 'watching', '.', 'and', 'over', 'here',  
'must', 'be', 'the', 'mother', '.', 'I', 'do', "n't", 'know',  
'what', 'the', 'heck', 'they', "'re", 'doing', 'here', 'though',  
'.', 'what', "'s", 'going', 'on', '?', 'the', 'stool', 'she',  
'was', 'climbing', 'and', 'the', 'stool', 'tipped', 'over', '.',  
'she', 'was', 'doing', 'the', 'dishes', 'I', 'think', '.', 'she',  
'spilled', 'something', '.']
```

Figure 7: Tokenized text

After having completed the tokenization of our text document, we begin the lemmatization method. Lemmatization is used to switch any kind of word to its base root word, by using vocabulary and morphological analysis. Essentially, lemmatization is a linguistic term that refers to the act of grouping together words that have the same root or lemma but different accents or meaning derivatives so that they can be analyzed as one item. The process of lemmatization tries to remove suffixes and prefixes in order to reveal the basic dictionary form of the word. For example, the word “tried” that appears in our transcripts is changed to the word “try”. Similarly, the word “trying” is also changed to “try”. Thus, both of these words can be analyzed as one.

Lemmatization is a really important method for our project since it is highly accurate in understanding the meaning of a dialogue or an interview, which is a crucial task for examining a patient’s transcript and detecting a neurological disease such as Alzheimer’s. An example of our lemmatization process is presented in Figure 8, where we observe that words like “tried”, “watching” or “dishes” from Figure 7 have been changed to “try”, “watch” and “dish” accordingly.

```
[ 'well', 'this', 'one', 'is', 'in', 'the', 'cookie', 'jar', '.',
  'and', 'this', 'is', 'she', 'try', 'to', 'climb', 'the', 'oh',
  'i', 'no', 'no', '.', 'this', 'boy', '1pause1', '.', 'I',
  'think', 'this', 'is', 'the', 'same', 'one', 'huh', '1pause1',
  '.', 'try', 'to', 'get', 'in', 'the', 'cookie', 'jar', '.',
  'and', 'she', "'s", 'watch', '.', 'and', 'over', 'here', 'must',
  'be', 'the', 'mother', '.', 'I', 'do', "n't", 'know', 'what',
  'the', 'heck', 'they', "'re", 'do', 'here', 'though', '.',
  'what', "'s", 'go', 'on', '?', 'the', 'stool', 'she', 'be',
  'climb', 'and', 'the', 'stool', 'tip', 'over', '.', 'she', 'be',
  'do', 'the', 'dish', 'I', 'think', '.', 'she', 'spill',
  'something', '.']
```

Figure 8: Lemmatization of a possible AD patient's tokens

The final method of the data preprocessing is the Parts of Speech tagging, or simply POS tagging. POS tagging refers to classifying words in a text in correspondence with a specific part of speech, based on the definition of the word and its context. In other words, it is used to identify nouns, verbs, adjectives, adverbs, etc. However, writing the full terms of parts of speech each time we perform text analysis can become a tiring and time-consuming process. Therefore, we use the following shorten “tags” to represent each lexical category:

- *CC*: coordinating conjunction (*for, and, but*)
- *CD*: cardinal digit (*1, 2, 3*)
- *DT*: determiner (*the, a*)
- *EX*: existential there (*there is, there exists*)
- *FW*: foreign word (*ciao*)
- *IN*: preposition / subordinating conjunction (*since, after, because*)
- *JJ*: adjective (*large*)
- *JJR*: adjective, comparative (*larger*)
- *JJS*: adjective, superlative (*largest*)
- *LS*: list market
- *MD*: modal (*could, should, might*)
- *NN*: noun, singular (*cookie, tree*)
- *NNS*: noun, plural (*cookies, trees*)
- *NNP*: proper noun, singular (*Sarah*)
- *NNPS*: proper noun, plural (*Americans*)
- *PDT*: predeterminer (*all, both*)
- *POS*: possessive ending (*'s*)
- *PRP*: personal pronoun (*I, he, she*)

- *PRP\$*: possessive pronoun (*my, mine, his, her*)
- *RB*: adverb (*great, swiftly, easily*)
- *RBR*: adverb, comparative (*greater*)
- *RBS*: adverb, superlative (*greatest*)
- *RP*: particle (*about*)
- *TO*: infinite marker (*to*)
- *UH*: interjection (*uh, oh*)
- *VB*: verb, base form (*take*)
- *VBG*: verb, gerund/present participle (*asking*)
- *VBD*: verb, past tense (*took*)
- *VBN*: verb, past participle (*taken*)
- *VBP*: verb, present tense not 3rd person, singular (*wrap*)
- *VBZ*: verb, present tense with 3rd person, singular (*bases*)
- *WDT*: wh-determiner (*that, what*)
- *WP*: wh-pronoun (*who*)
- *WRB*: wh-adverb (*how*)

These tags convey information about a word, including its morphological structure and inflectional paradigm, as well as its potential grammatical role in a clause [27]. By POS tagging each participant's document, we have the chance to examine their incompetence in using the appropriate parts of speech during the interview. Figure 9 depicts the POS tagging for the same patient's text document as in previous figures, while Figure 10 presents all POS tags of the same document in descending order.

well	RB	1pause1	CD	.	.	going	VBG
this	DT	.	.	and	CC	on	IN
one	NN	I	PRP	over	IN	?	.
is	VBZ	think	VBP	here	RB	the	DT
in	IN	this	DT	must	MD	stool	NN
the	DT	is	VBZ	be	VB	she	PRP
cookie	NN	the	DT	the	DT	was	VBD
jar	NN	same	JJ	mother	NN	climbing	VBG
.	.	one	CD	.	.	and	CC
and	CC	huh	NN	I	PRP	the	DT
this	DT	1pause1	CD	do	VBP	stool	NN
is	VBZ	.	.	n't	RB	tipped	VBD
she	PRP	tried	VBH	know	VB	over	IN
tried	VBD	to	TO	what	WP	.	.
to	TO	get	VB	the	DT	she	PRP
climb	VB	in	IN	heck	NN	was	VBD
the	DT	the	DT	they	PRP	doing	VBG
oh	JJ	cookie	NN	're	VBP	the	DT
‡	NNP	jar	NN	doing	VBG	dishes	NNS
no	DT	.	.	here	RB	I	PRP
no	DT	and	CC	though	IN	think	VBP
.	.	she	PRP
this	DT	's	VBZ	what	WP	she	PRP
boy	NN	watching	VBG	's	VBZ	spilled	VBD
						something	NN
						.	.

Figure 9: POS Tagging of a possible AD patient's transcript

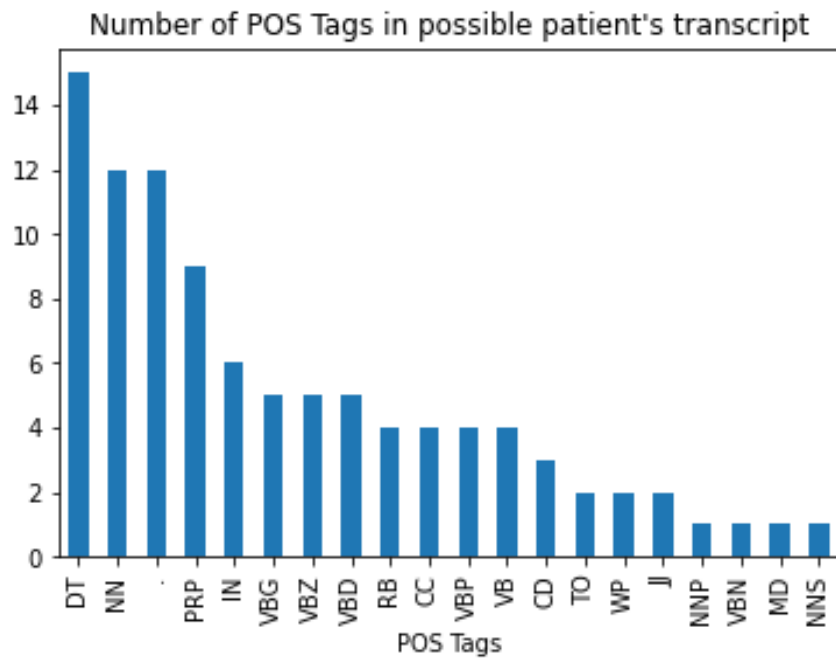


Figure 10: Total amount of POS Tags in descending order

Now that we have gathered together the most important pieces of our dataset and have transformed them in the most appropriate way for our project, we can proceed with identifying and extracting the necessary features that will allow us to train our model in the best possible way.

5.3 Feature Extraction

Feature extraction is a step in the dimensionality reduction process that divides and reduces an initial set of raw data to more manageable subgroups. As a result, processing becomes simpler. The most important feature of these large data sets is the large number of variables. These variables need a significant amount of computing power to process. As a result, feature extraction aids in obtaining the best feature from large data sets by selecting and combining variables into features, effectively reducing the amount of data. These features are simple to process, as well as accurate in describing the initial data set.

The importance of feature extraction lies in its ability to reduce the number of resources without losing any important or relevant information. When we have a large dataset and we need to reduce its numbers, the feature extraction technique comes in handy, as it assists in reducing the amount of redundant data in a data set. Eventually, data reduction allows the model to be built with less machine effort while increasing the speed of the learning in the machine learning process.

In NLP, there are two main feature extraction techniques:

- *Bag of Words* [40]: one of the most basic methods for converting tokens into a set of features. The bag-of-words model is an NLP simplifying representation. A text is represented in this model as a bag of its words, ignoring grammar and even word order but retaining multiplicity. The model is only concerned with whether known words appear in the document, not with where they appear.
- *TF-IDF* [41]: TF-IDF stands for Term Frequency-Inverse Document Frequency. It is intended to indicate the importance of a word to a document in a collection or corpus. The problem with scoring word

frequency is that highly frequent words begin to dominate the document (e.g., higher score), but may not provide as much "informational content" to the model as rarer but possibly domain specific words. One approach is to rescale the frequency of words based on how frequently they appear in a document, so that frequent words like "the" that appear frequently in any document are penalized. The TF-IDF value rises in proportion to the number of times a word appears in the document and is offset by the frequency of the word across all documents, which helps to account for the fact that some words appear more frequently than others.

For this thesis, our implementation was based on manual feature engineering methods and it is present in the `feature_extraction.py` file.

First, we determine which features from our cleaned dataset we consider as the most important in a subject's speech, in order to detect cognitive and Alzheimer's disease symptoms. To discover the most important features, we performed a wide research on several studies and scientific papers related to this specific subject. From this research, we have deduced that linguistic, acoustic and semantic features are of the greatest importance in detecting Alzheimer's Disease symptoms. Therefore, characteristics such as pauses, trailings, parts of speech or statistics related to the number of words and concepts mentioned are considered essential for this thesis' goals.

After defining the most important features, we proceed with manually selecting and saving them to the `feature_extraction.csv` file. According to the studies and papers that we have reviewed, we have decided to select the following features for the training of our classifying models:

- *Category*

As stated earlier, a participant can belong in one of two clinical categories: "Control" or "ProbableAD". If the speaker belongs to the first category (Control) then he/she is healthy. Otherwise, if the speaker belongs to the "ProbableAD" category, then he/she is a possible Alzheimer's disease patient. For the feature extraction process, we label each possible output as follows:

- 0, if category is “Control”
- 1, if category is “ProbableAD”

The categories for each participant are already given by the initial dataset, so we will be using them in order to train and later evaluate our model.

- *Age*

The most important risk factor for Alzheimer's disease is age. It primarily affects people over the age of 65. Above this age, the risk of developing Alzheimer's disease roughly doubles every five years.

- *TTR*

The type-token ratio (TTR) is a measure of the variation in vocabulary within a written text or a person's speech. It is the total number of unique words (types) divided by the total number of words (tokens) in the whole interview transcript. TTR is demonstrated to be a useful measure of lexical variety within a text. It can be used to track changes in children and adults who have trouble with their vocabulary.

- *num_concepts_mentioned*

The number of basic concepts present in the Cookie Theft picture and are mentioned by the participants during their interview. We have defined the following lists of the basic concepts that appear in the Cookie Theft picture:

- | | |
|-----------|------------|
| ○ cookie | ○ lady |
| ○ jar | ○ girl |
| ○ stool | ○ daughter |
| ○ steal | ○ sister |
| ○ sink | ○ boy |
| ○ kitchen | ○ son |
| ○ window | ○ child |
| ○ curtain | ○ kid |
| ○ fall | ○ brother |
| ○ mother | ○ dish |
| ○ woman | ○ plate |

- | | |
|------------|-----------|
| ○ cup | ○ wash |
| ○ overflow | ○ faucet |
| ○ spill | ○ counter |
| ○ running | ○ cabinet |
| ○ dry | ○ water |

All of the above concepts are considered essential for the description of the Cookie Theft task, thus, their appearance, or not, in a subject's transcript can play a vital role in the overall diagnosis.

- *num_utterances*

The total number of utterances in a participant's interview. An utterance is the smallest unit of speech in spoken language analysis. It is a continuous speech that begins and ends with a clear pause. It is generally, but not always, bounded by silence in the case of oral languages.

- *num_unique_words*

The total number of unique words used by the participant during his/her interview. This feature displays the total number of words, without counting their respective repetitions.

- *prp_count*

The total number of pronouns used by the participant. Unlike the limited use of pronouns in normal speech, the use of pronouns by Alzheimer's patients is common and frequently inappropriate. When the intended phonological information (the target word) is not activated during sentence production, AD patients substitute a higher frequency, easily retrievable, and grammatically valid replacement for the target word that they cannot find. Pronouns are successfully and relatively easily activated precisely because they are frequent and allow patients to produce fluent and grammatical sentences [24].

- *prp_noun_ratio*

The number of pronouns in proportion to the nouns existing in the transcript. We need not only to count the pronouns, but also their ratio related to the total number of nouns used in the same transcript.

- *Gerund_count*

The number of gerund types used by the participant. Reduced sentences are spoken subordinated sentences that are abbreviated, lack a conjunction, and have nominal verb forms that are either infinitive or gerund in English grammar. Thus, analyzing the gerund uses in a participant's interview can give us probable signs of Alzheimer's disease.

- *NP_count*

The number of noun phrases used by the participant. A noun phrase is a group of words that functions like a noun, usually consisting of a noun and a modifier (such as an adjective, adverb, or article). A noun phrase can be the subject, object, subject complement, or object complement in the sentence it appears in. Noun phrases usage can show the participant's ability of understanding which subject in the picture does a certain task that the participant tries to describe.

- *VP_count*

The number of verb phrases. A verb phrase is made up of a verb and another word that emphasizes the verb tense, action, and tone. The dependents, which can be adverbs, prepositional phrases, helping verbs, or other modifiers, are the other word or words tied to a verb in a verb phrase. This group of words defines the intention and timing of the verb's action. Similar to noun phrases, the presence of verb phrases in a participant's transcript denotes his/her ability to recognize any action and the responsible subject of that action.

- *num_of_sentences*

The number of sentences used in a subject's transcript. Sentences can be calculated by counting the number of periods (".") and question marks ("?"), since one of these punctuation marks define the end of a sentence. Participant's ability to comprehend and create sentences is a health sign, since even in the early stages of Alzheimer's disease sentence comprehension is impaired [25].

- *word_sentence_ratio*

The words to sentences ratio during a participant's interview. With this feature we can measure the average number of words in each sentence. The bigger the ratio, the more words are used in the participants interview.

- *MLU*

Mean Length Utterance represents the ratio of total number of words to total number of utterances. MLU has been initially used to assess grammar development in children with Specific Language Impairment (SLI) and later it was used to determine language disorder in Alzheimer's disease and related dementias. According to Orimaye S.O. et al [26], MLU, among other features, significantly distinguish the disease group from the healthy elderly group.

- *count_pauses*

The total number of all kinds of pauses present in each transcript. As mentioned earlier, we have three kinds of pauses: pauses that take up a "normal" duration and are marked only by silence, longer pauses between words and very long pauses. Pauses in speech production are usually considered as a sign of a patient's lexical-semantic decline, making them an important feature for early Alzheimer's disease detection.

- *unintelligible_words*

Total number of unintelligible words. As stated in the previous chapter, unintelligible words are words with unclear phonetic shape. Their presence in a transcript declares that the participant is having language impairments and should be considered as a possible AD patient.

- *trailings*

Total number of incomplete not interrupted utterances. They occur when a speaker trails off before completing an utterance or a sentence. Again, trailing off is a possible symptom of Alzheimer's disease, since dementia patients have difficulties articulating their thoughts in writing or in conversation, resulting to trailing off mid-sentence and not knowing how to continue.

- *repetitions*

Total number of repetitions in a transcript. A person with Alzheimer's disease may do or say something repeatedly, such as repeating a word, question, or activity, or undo something that has just been completed. In the case of repetition, the person may not remember that she or he has just asked a question or completed a task.

- *repetitions_ratio*

The ratio of repeated words to total words. It gives us percentage of present repetitions in the participant's transcript. The bigger the percentage, the more likely it is for the subject to be a possible Alzheimer's disease patient.

- *unintelligible_ratio*

The ratio of unintelligible words to total words. Similar to the repetitions ratio, unintelligible ratio produces the percentage of unintelligible words that exist in each participant's transcript. Again, if the percentage is high, the possibility for the participant to be Alzheimer's disease patient is also higher.

- *trailing_ratio*

The percentage of trailings related to the total words in a transcript. Its importance is similar to the other ratios.

- *pauses_ratio*

The number of pauses in proportion to the total number of words. Again, its importance is similar to the above ratios.

Identifying and extracting the most important features is the final step of the entire preprocessing task which eventually allows us to reduce the amount of the data that will be "fed" to our models in order to be trained and later evaluated.

6 Model Training and Evaluation

Since we have managed to modify the initial dataset and extract any important feature, we have to build the machine learning models that will try to examine and make an Alzheimer's disease diagnosis as more accurate as possible. As the result is expected to be one of two choices, healthy or non-healthy, the diagnosis of a disease belongs in the category of classification problems.

Classification is a task that challenges machine learning algorithms to learn how to assign class labels to problem domain examples. It refers to any type of problem in which a specific type of class label must be predicted from a given field of data input. Any classification model needs a training dataset with many examples of inputs and outputs in order to train itself. Moreover, the training data must include all possible problem scenarios and have enough data for each label for the model to be correctly trained.

There are three main types of classification tasks that you may encounter in your daily challenges:

- *Binary Classification*: It refers to tasks that produce one of two class labels as the output.
- *Multi-Label Classification*: These classification problems do not have a fixed pair of labels but can have one or more numbers of labels.
- *Multi-Class Classification*: It refers to classification tasks in which we must assign three or more specific class labels that can be predicted for each example.

We must also notice the *Imbalanced Classification*, which refers to tasks in which the number of examples in each class is unequally distributed.

Apparently, the task of identifying a disease is a binary classification problem, thus we are going to build binary classification models able to detect a possible Alzheimer's disease patient.

6.1 Model Implementation

Since there is no established theory for the best model [38], we have to experiment on several models and determine which algorithm is the best for any classification task. There are several machine learning algorithms that are suitable for classification tasks, such as Support Vector Machines or Decision Tree. After reviewing several studies regarding NLP and neurological diseases, we have detected the lack of a specific machine learning algorithm, the XGBoost algorithm, in such cases [28, 29, 30, 31, 32, 33]. With that in mind, we have decided to build an XGBoost classifier along with three other classifiers: SVM, Decision Tree and Random Forest.

Once our models are built, it is necessary to evaluate them and compare them with each other in order to determine which one has the best overall performance. Several metrics exist that can be used to assess a classification algorithm, such as the confusion matrix, the AUC-ROC curves, accuracy, precision, f1-score etc. By examining all of the appropriate metrics, we will be in place to resolve the most efficient model.

For each of the models that we built, there are some standard initial steps that we follow. First we divide the dataset in train and test sets, which account for 70% and 30% of the total dataset respectively. Furthermore, we remove the “Category” column from the train set and we keep this column’s values in a different array which will be used for testing each algorithm’s results. Then, we are ready to run the experiments for each classifier.

6.1.1 XGBoost Classifier

XGBoost stands for eXtreme Gradient Boosting and is a gradient boosting machine implementation that pushes the limits of computing power for boosted trees algorithms, as it was built and developed with model performance and computational speed in mind. It is especially popular because it was the winning algorithm in several open data science competitions for prediction or any other kind of task. According to Chen T. & Guestrin C. [34], on a single machine the system runs more than ten times faster than existing popular solutions and scales to billions of examples in distributed or memory-limited settings. XGBoost's scalability is due to several

important system and algorithmic optimizations. They continue by stating that XGBoost uses out-of-core computation to allow data scientists to process hundreds of millions of examples on a single desktop. Finally, it is even more interesting to combine these techniques to create an end-to-end system that scales to even larger data sets, while using the fewest cluster resources.

In our implementation we used the *xgboost* library for Python² which, according to its documentation, is “an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable”. After experimenting and fitting our model with several different combinations of parameters, we have observed that the best performance is achieved when we use the *gblinear* booster parameter, while for the evaluation metric parameter we have defined the binary classification error rate, as it is the best fit for a classification task.

After having trained our XGBoost model, we can have an early evaluation of our classification model by reporting the four metrics presented in Tables 1a and 1b for both train and test sets.

Table 1a: XGBoost Classification report with a 70-30 train-test split – Train Set

Class	Precision	Recall	F1-score	Accuracy
Control (0)	92%	88%	90%	89.33%
ProbableAD (1)	86%	91%	89%	

Table 1b: XGBoost Classification report with a 70-30 train-test split - Test Set

Class	Precision	Recall	F1-score	Accuracy
Control (0)	80%	84%	82%	78%
ProbableAD (1)	77%	71%	74%	

² <https://xgboost.readthedocs.io/en/stable/python/index.html>

In addition, we extract the confusion matrix, as another performance measure (Figure 11). Confusion matrix is basically a 2x2 table with four different combinations of predicted and actual values, which are described below:

- *True Positive (TP)*: Model prediction is positive and that is true.
- *True Negative (TN)*: Model predicted negative and it is true.
- *False Positive (FP)*: Prediction is positive, but it is false.
- *False Negative (FN)*: Prediction is negative, but it is false.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 11: Confusion matrix

By examining the confusion matrix of any classification model, we are able to deduce important outcomes regarding the model's performance. The confusion matrix displays the various ways in which a classification model is confused when making predictions. It provides information not only about the errors made by the classifier, but also about the types of errors made.

For the XGBoost model, the produced confusion matrices for both the train and test set are shown in Figures 12a and 12b.

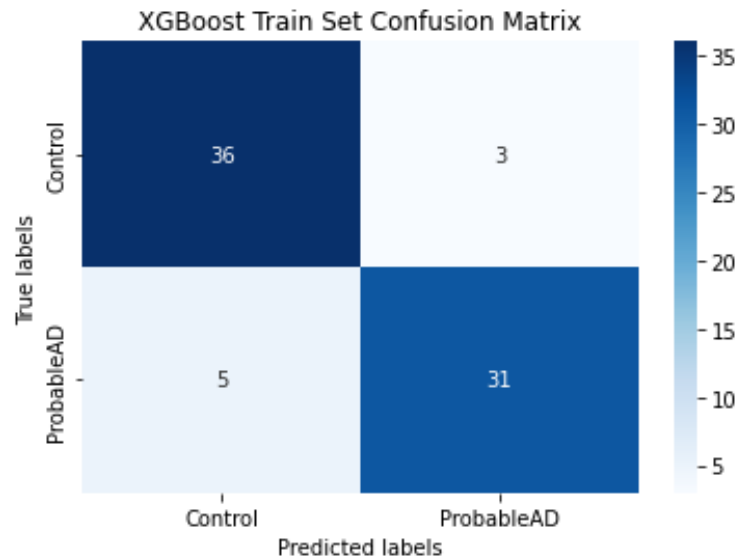


Figure 12a: XGBoost Train Set Confusion Matrix

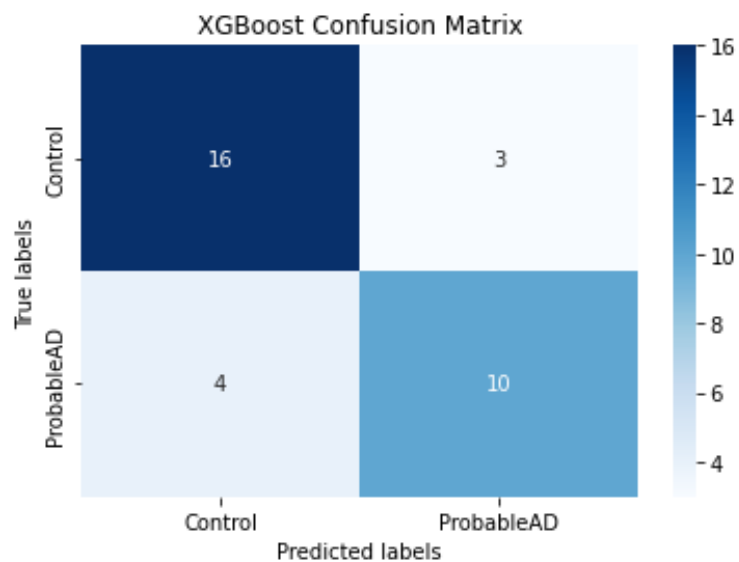


Figure 12b: XGBoost Test Set Confusion Matrix

We can also evaluate our model by examining the AUC-ROC curve (Figure 13). The AUC-ROC curve is a performance metric for classification problems at various threshold levels. When we get a TP prediction, we make a step upward on the y-axis, and when we have a FP prediction, we get a step rightward on the x-axis. Step sizes are inversely proportional to the number of actual positives (in the y-direction) or negatives (in the x-direction), thus the path always ends at coordinates (1, 1). The result is a plot of true positive rate

(TPR, or sensitivity) against false positive rate (FPR, or $1 - \text{specificity}$), which form the ROC curve. One way to summarize it in a single value is to compute the area under the curve (AUC). This indicates how well the model can distinguish between classes. The greater the AUC, the better the model, thus, the higher the AUC, the better the model distinguishes between patients with and without Alzheimer's disease.

In the same figure we observe that the area under the curve (AUC) is equal to 0.86, a high enough value that shows that the classifier is able to detect more numbers of True Positives and True Negatives than False Negatives and False Positives.

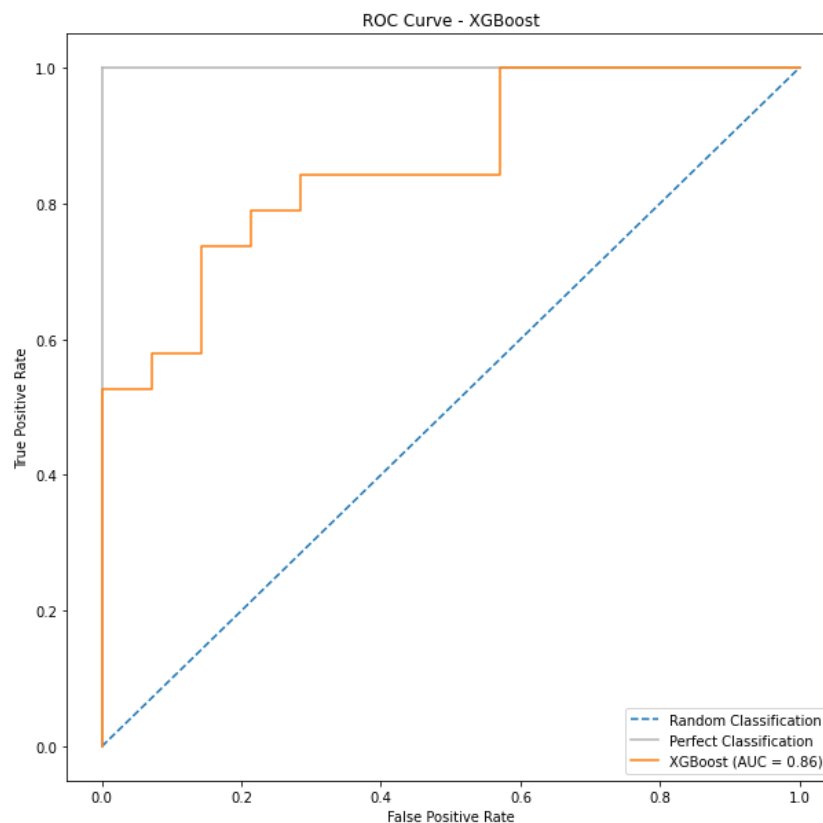


Figure 13: XGBoost ROC Curve

Although the above metrics present an overall positive performance for an XGBoost algorithm, if we want to get a more precise evaluation of its performance we have to proceed with a validation technique. Machine learning validation techniques are used to obtain the error rate of the model, which is close to the true error rate of the population. If the data size is large enough to be representative of the population, validation techniques may not

be required. In real-world scenarios, however, we work with data samples that may not be a true representation of the population. This is where validation methods become important.

K-Fold cross-validation is the most commonly used validation technique, which involves splitting the training dataset into k folds. The first k-1 folds are used for training, and the remaining fold is held for testing. This process is repeated for K folds. K folds are fit and evaluated in total, and the mean metrics for all of these folds are returned. This process can produce positive results for balanced classification tasks but it usually fails for imbalanced datasets. This is due to cross-validation, which also randomly splits the data without accounting for class imbalance. The solution to this issue is to stratify the data rather than split it randomly. The stratified k fold cross-validation technique is a classification problem extension of the cross-validation technique. It keeps the same class ratio as the original dataset throughout the K folds. For this thesis, we have used a stratified 10-fold cross-validation which has produced the following average metrics for the test and the train set (Table 2a and 2b):

Table 2a: XGBoost Train Set Mean Results after Stratified 10fold Cross Validation

Mean F1 Score	73.80%	SD F1 Score	18.61%
Mean Recall Score	74.17%	SD Recall	25.67%
Mean Precision Score	78.17%	SD Precision	16.49%
Mean Accuracy Score	77.50%	SD Accuracy	13.22%

Table 2b: XGBoost Test Set Mean Results after Stratified 10fold Cross Validation

Mean F1 Score	78.23%	SD F1 Score	19.52%
Mean Recall Score	77.33%	SD Recall	25.38%
Mean Precision Score	84.88%	SD Precision	15.54%
Mean Accuracy Score	81.55%	SD Accuracy	12.81%

With the stratified 10-fold cross validation completed, we are in a place to obtain a more accurate assessment of the XGBoost model that we built. From this table we notice that mean classification scores are over or around 80%, which makes our model a good Alzheimer’s disease classifier. Standard Deviation values show us that our data are more spread out.

Again, we can evaluate the stratified 10-fold validation by observing the mean ROC Curve (Figure 14). In this figure we can see that in each of the 10 folds the AUC value is very high, thus the mean value is also really high, establishing our XGBoost algorithm as an efficient Alzheimer’s disease classifier.

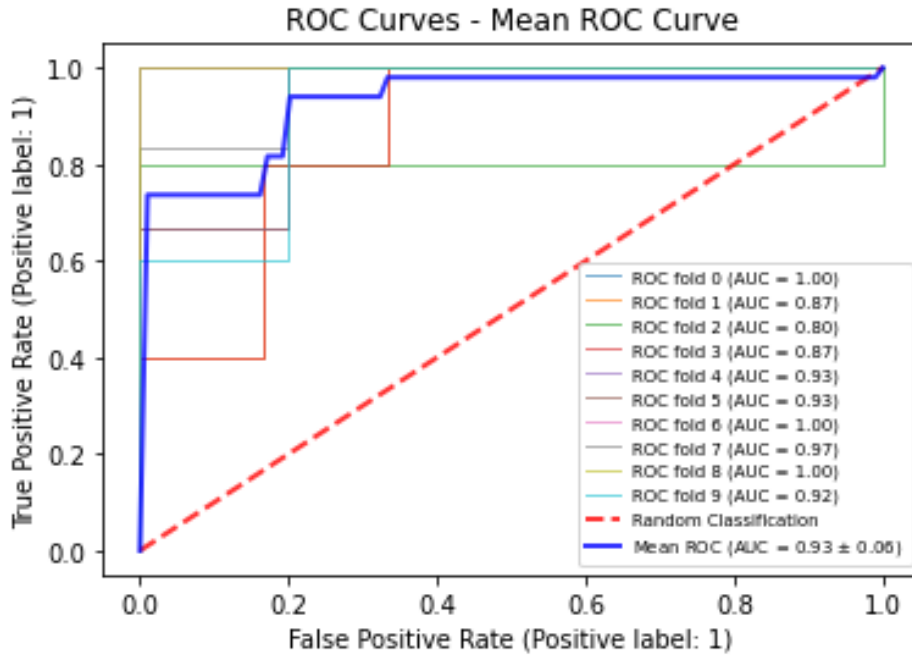


Figure 14: XGBoost Stratified 10-fold CV ROC Curves & Mean ROC Curve

6.1.2 SVM Classifier

The next classification model that we have implemented is the SVM Classifier [39]. A Support Vector Machine is an algorithm that learns to label objects by example. A support vector machine (SVM) is a supervised machine learning algorithm that can be used for classification and regression tasks. Each data item is plotted as a point in n -dimensional space (where n is the number of features) with the value of each feature being the value of a specific coordinate in the SVM algorithm. Then, we perform classification by locating the hyperplane that best distinguishes the two classes. Ideally, the hyperplane must segregate the two classes in the best possible way.

SVM does not perform well when we have a large dataset, as it requires a longer training time. It is also not so efficient when the dataset contains more noise. For that reason, Support Vector Machine is considered a fitting classifying algorithm for our project, for which we have a small amount of data and it is cleared of noises. Regarding the tuning parameters, we have used the linear kernel, since it is the most appropriate for a binary classification task such as the Alzheimer's disease detection.

The first evaluation metrics for the train and test sets are presented in Tables 3a and 3b correspondingly.

Table 3a: SVM Classification Report with a 70-30 train-test split - Train Set

Class	Precision	Recall	F1-score	Accuracy
Control (0)	90%	88%	89%	88%
ProbableAD (1)	86%	89%	87%	

Table 3b: SVM Classification Report with a 70-30 train-test split - Test Set

Class	Precision	Recall	F1-score	Accuracy
Control (0)	87%	87%	87%	87.88%
ProbableAD (1)	89%	89%	89%	

Furthermore, we present the confusion matrices of SVM train and test sets (Figures 15a, 15b).

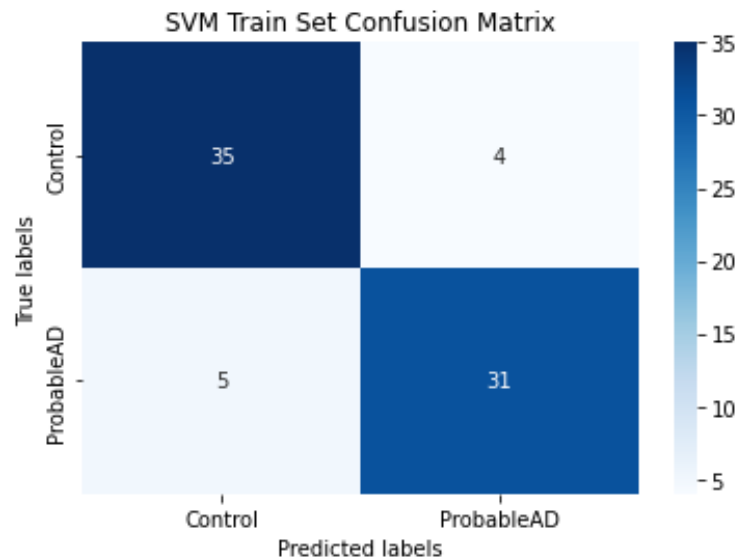


Figure 15a: SVM Train Set Confusion Matrix

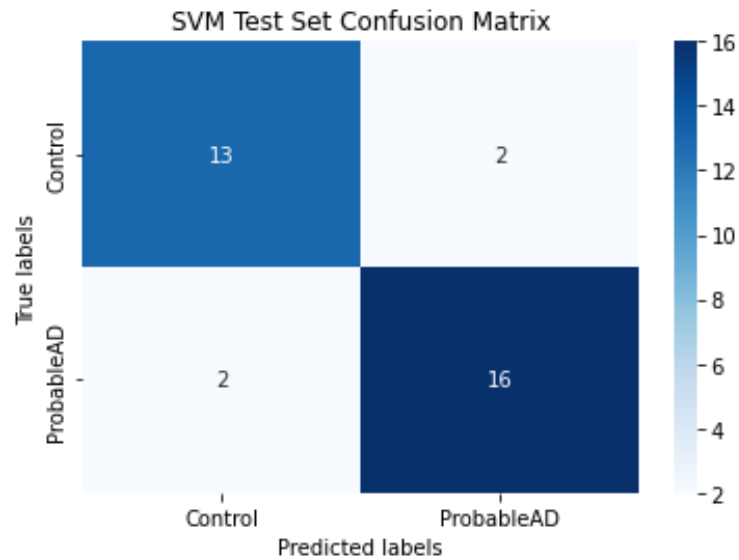


Figure 15b: SVM Test Set Confusion Matrix

We can proceed with examining the ROC curve of the SVM classification model (Figure 16). AUC value is 0.94 in this case, declaring that the model is even better in distinguishing between patients with and without Alzheimer’s disease.

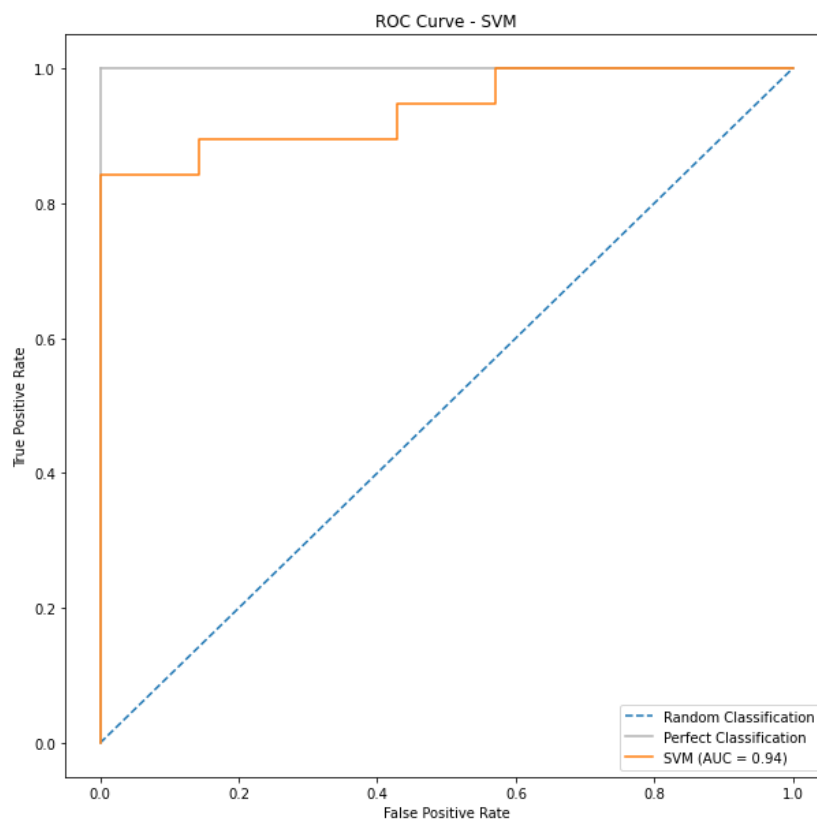


Figure 16: SVM ROC Curve

Again, we want to have a better evaluation of SVM's performance, so we proceed with the stratified 10-fold cross validation. The f1, recall, precision and accuracy scores of train and test sets are present in Tables 4a and 4b.

Table 4a: SVM Train Set Mean Results after Stratified 10fold Cross Validation

Mean F1 Score	75.07%	SD F1 Score	18.19%
Mean Recall Score	75.83%	SD Recall	24.57%
Mean Precision Score	79.17%	SD Precision	16.42%
Mean Accuracy Score	78.57%	SD Accuracy	12.58%

Table 4b: SVM Test Set Mean Results after Stratified 10fold Cross Validation

Mean F1 Score	79.75%	SD F1 Score	14.26%
Mean Recall Score	77.33%	SD Recall	20.10%
Mean Precision Score	86.79%	SD Precision	15.19%
Mean Accuracy Score	81.64%	SD Accuracy	11.51%

Finally, we present the mean ROC curve after the stratified 10fold cross validation (Figure 17). Similar to the XGBoost classifier, SVM manages to achieve high AUC values.

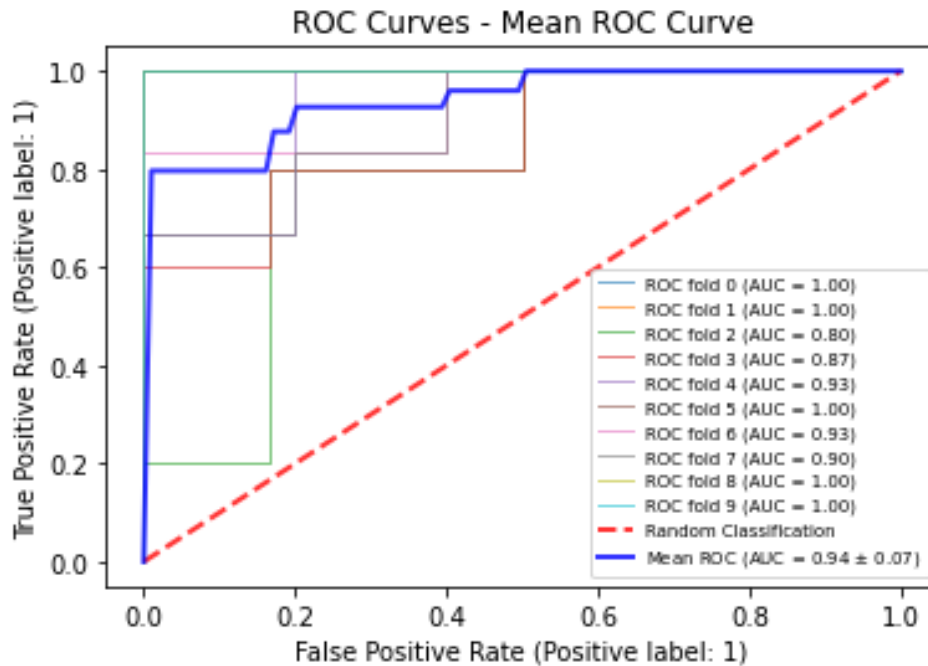


Figure 17: SVM Stratified 10-fold CV ROC Curves & Mean ROC Curve

6.1.3 Decision Tree Classifier

The Decision Tree belongs to the supervised learning algorithms and it can be used to solve regression and classification problems. It is a tree-structured classifier in which internal nodes represent dataset features, branches represent decision rules, and each leaf node represents the result. The goal of using a Decision Tree is to build a training model that can predict the class or value of a target variable by learning simple decision rules from training data. Decision trees classify examples by descending the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the example's classification. Each node in the tree represents a test case for some attribute, and each edge descending from the node represents one of the test case's possible answers. This recursive process is repeated for each sub tree rooted at the new node.

We built the Decision Tree model with the “gini” criterion, a max depth of 5 and the “best” splitter parameters. Initially we had chosen the “random” splitter, which selects a set of features, splits it randomly and it does not have large computational needs. However, the results are inconsistent showing

both very high and very low classification scores. Thus, we have decided to continue with the “best” splitter parameter. The train and test sets’ classification scores are present in Tables 5a and 5b, while the respective confusion matrices are shown in Figures 18a and 18b.

Table 5a: Decision Tree Classification Report with a 70-30 train-test split - Train Set

Class	Precision	Recall	F1-score	Accuracy
Control (0)	97%	84%	90%	89.33%
ProbableAD (1)	82%	97%	89%	

Table 5b: Decision Tree Classification Report with a 70-30 train-test split - Test Set

Class	Precision	Recall	F1-score	Accuracy
Control (0)	71%	80%	75%	75.76%
ProbableAD (1)	81%	72%	76%	

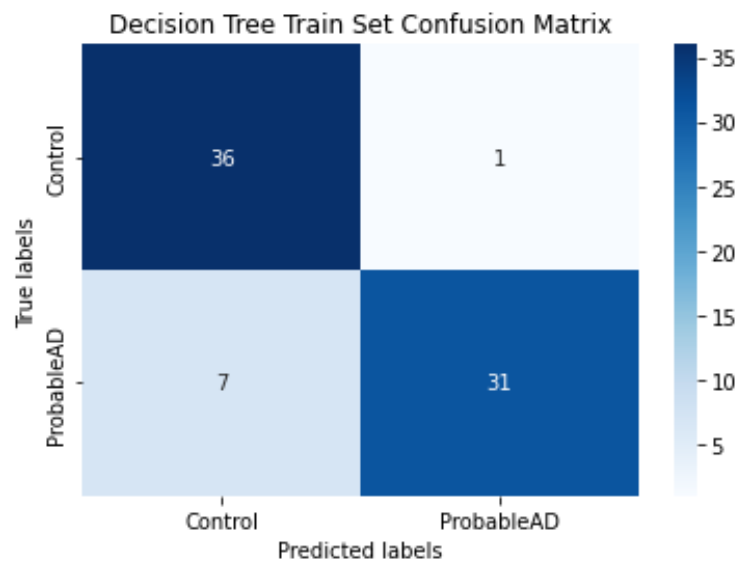


Figure 18a: Decision Tree Train Set Confusion Matrix

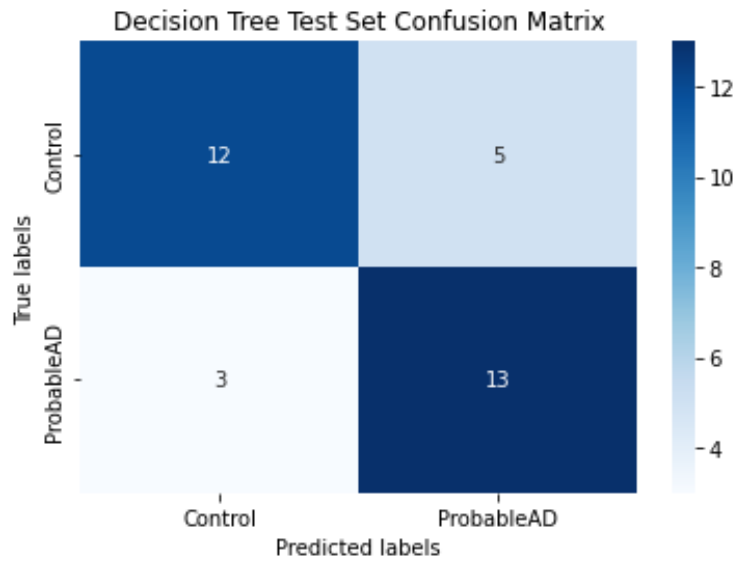


Figure 18b: Decision Tree Test Set Confusion Matrix

Finally, we present the ROC curve (Figure 19), with an AUC value of 0.75 in this experiment.

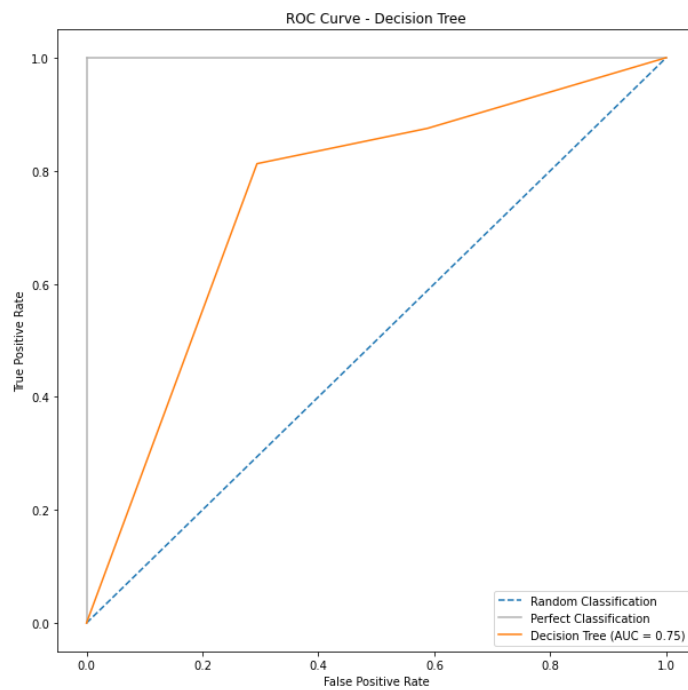


Figure 19: Decision Tree ROC Curve

The next step, similar to the previous classifiers, is the stratified 10-fold cross validation. The respective mean classification scores for train and test sets are shown in Tables 6a and 6b.

Table 6a: Decision Tree Mean Classification Scores – Train Set

Mean F1 Score	70.48%	SD F1 Score	22.72%
Mean Recall Score	72.50%	SD Recall	24.73%
Mean Precision Score	75.83%	SD Precision	22.50%
Mean Accuracy Score	76.25%	SD Accuracy	18.75%

Table 6b: Decision Tree Mean Classification Scores – Test Set

Mean F1 Score	73.87%	SD F1 Score	14.53%
Mean Recall Score	72.00%	SD Recall	19.10%
Mean Precision Score	77.90%	SD Precision	12.40%
Mean Accuracy Score	76.00%	SD Accuracy	11.64%

For the test set, we observe a mean value over 70% for every classification score. Moreover, the mean ROC curve (Figure 20) has a mean AUC value of 0.75.

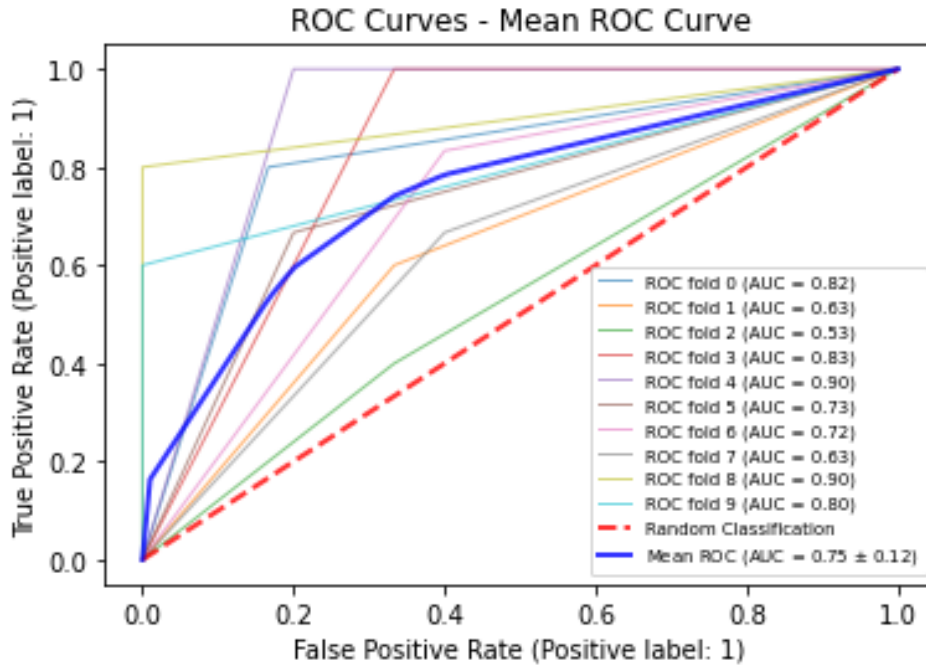


Figure 20: DT Stratified 10-fold CV ROC Curves & Mean ROC Curve

6.1.4 Random Forest Classifier

The random forest classifier is made up of a large number of individual decision trees that work together as an ensemble. Each individual tree in the random forest produces a class prediction, and the class with the most votes becomes the prediction of our model. This model works sufficiently in classification problems because a large number of uncorrelated models operate as a group and are able to outperform any of those models individually. A main advantage of this algorithm is the fact that trees can “protect” each other, meaning that while some trees may predict a label or class falsely, there are many other trees that at the same time will be right on their prediction, making the algorithm move in the correct direction towards the desired outcome.

We have defined 100 `n_estimators`, a max depth of 3 and the entropy criterion as the specified parameters for the Random Forest classifier. First, we take a look at the classification scores for train and test sets (Tables 7a, 7b).

Table7a: Random Forest Classification Report with a 70-30 train-test split - Train Set

Class	Precision	Recall	F1-score	Accuracy
Control (o)	100%	88%	94%	93.33%
ProbableAD (1)	87%	100%	93%	

Table7b: Random Forest Classification Report with a 70-30 train-test split - Test Set

Class	Precision	Recall	F1-score	Accuracy
Control (o)	82%	88%	85%	84.85%
ProbableAD (1)	88%	82%	85%	

In addition, the confusion matrices (Figures 21a, 21b) shows us that for the test set we had 28 predictions that were accurate, 14 healthy and 14 non healthy participants.

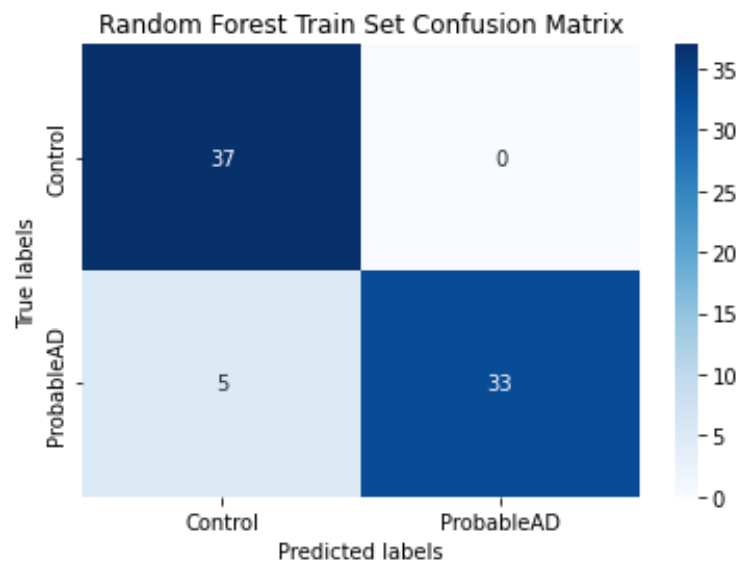


Figure 21a: Random Forest Confusion Matrix – Train Set

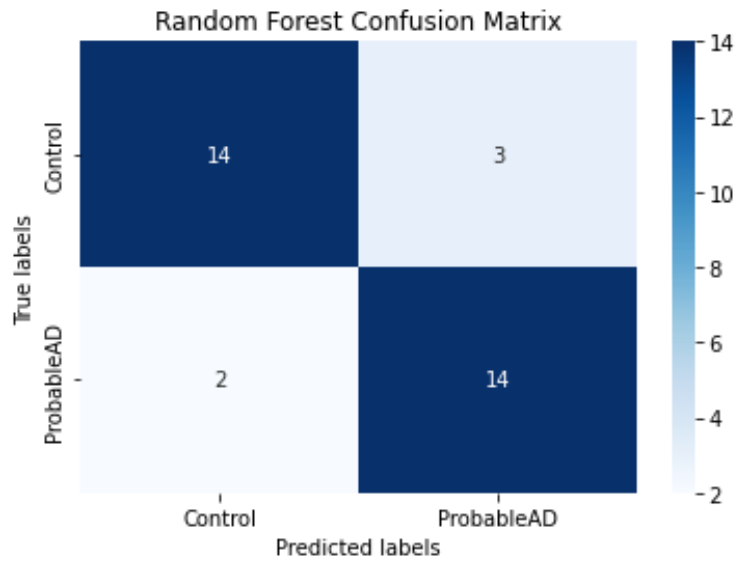


Figure 121b: Random Forest Confusion Matrix – Test Set

The ROC curve is depicted in Figure 22, which has an AUC value of 0.86.

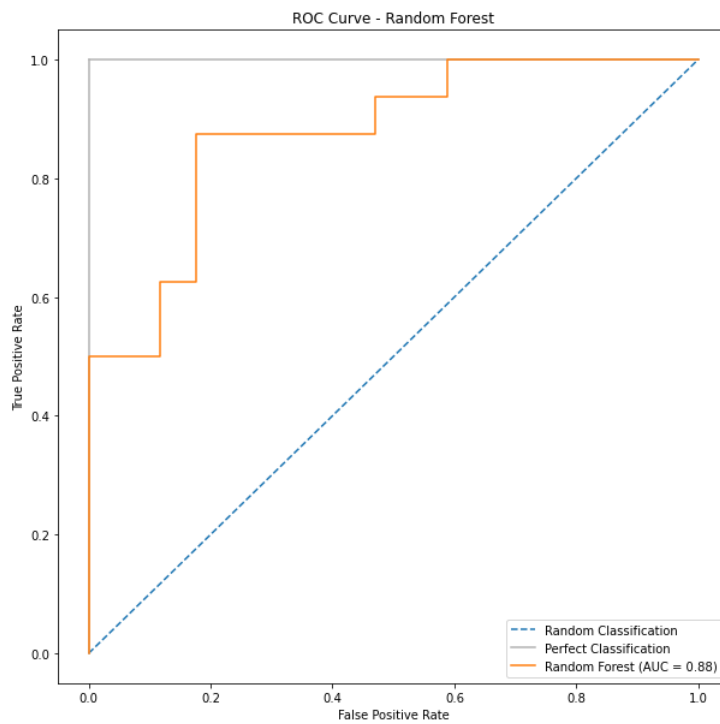


Figure 22: Random Forest ROV Curve

After running the stratified 10-fold cross validation, we get mean classification scores over 70% for both train and test sets (Tables 8a and 8b).

Table 8a: Random Forest Mean Classification Scores - Train Set

Mean F1 Score	74.17%	SD F1 Score	20.75%
Mean Recall Score	70.00%	SD Recall	25.60%
Mean Precision Score	84.67%	SD Precision	21.72%
Mean Accuracy Score	77.68%	SD Accuracy	16.56%

Table 8b: Random Forest Mean Classification Scores – Test Set

Mean F1 Score	76.24%	SD F1 Score	13.42%
Mean Recall Score	72.33%	SD Recall	19.27%
Mean Precision Score	84.50%	SD Precision	12.07%
Mean Accuracy Score	78.91%	SD Accuracy	10.76%

Finally, we extract the mean ROC Curve for all 10 folds of the Random Forest classifier (Figure 23)

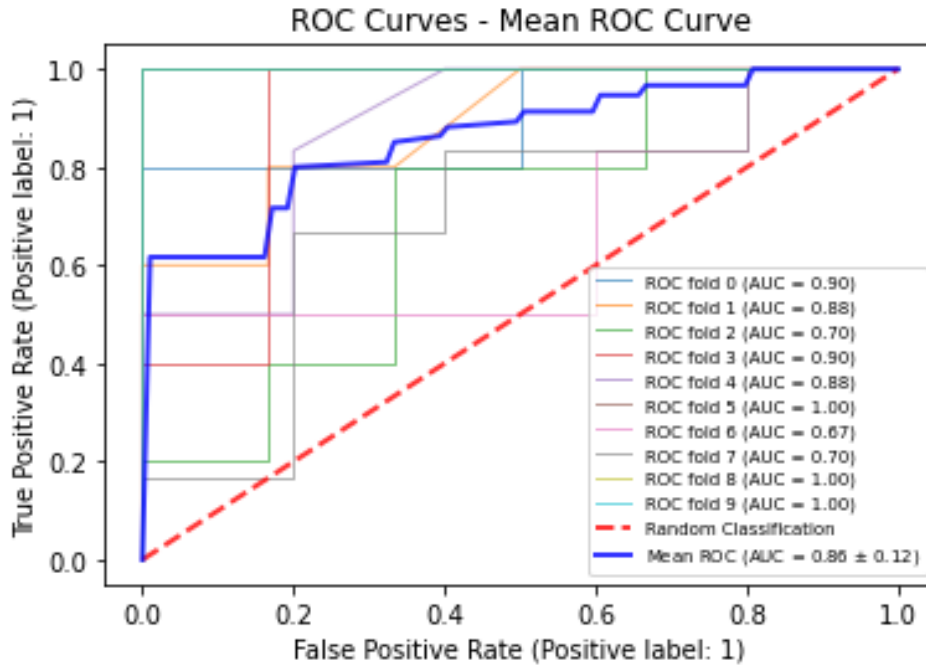


Figure 23: RF Stratified 10-fold CV ROC Curves & Mean ROC Curve

6.2 Classifier Comparison

For the eventual comparison of all the classifying models built in this thesis, we will be using only the evaluation scores and metrics derived after the stratified 10-fold cross validation that we ran in each one of the models. The evaluation of a classifier's performance is more accurate after running a k-fold or a stratified k-fold cross validation, so we are examining the means of precision, recall and F1 scores after the stratified 10-fold cross validation.

A first step for the comparison of the four implemented classifiers would be to check their Precision-Recall Curves (Figure 24). The precision-recall curve depicts the tradeoff between precision and recall for various threshold values. A high area under the curve indicates that both precision and recall are high, with high precision indicating that the model has a low false positive rate and high recall corresponding to low false negative rate. If both of those metrics are high, we can assume that the classifier is producing accurate results (high precision) and it is also producing the majority of all positive results (high recall).

From the Precision-Recall curve we can understand that the best performing model is the Support Vector Machine, which presents the highest area under the curve, thus the highest numbers in both precision and recall. The worst model is clearly the Decision Tree, which has the lowest area under the curve and also it starts reducing its precision score very early during the precision-recall tradeoff.

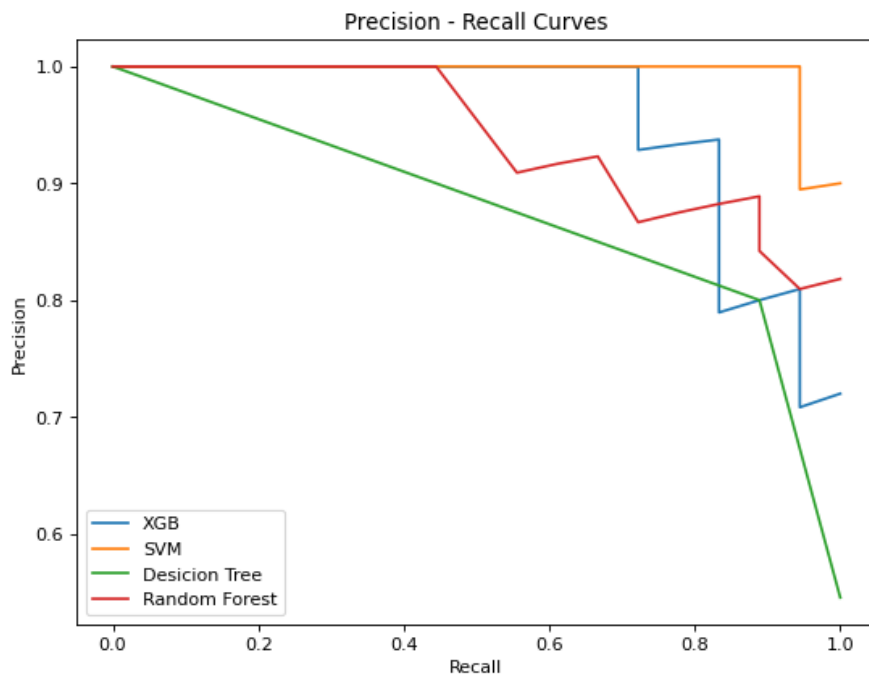


Figure 24: Precision - Recall Curves after Stratified 10-fold CV

Regarding the two remaining classifiers, XGBoost and Random Forest show a similar “path”. XGBoost seems to hold a high precision score, compared to Random Forest, while recall is increasing, but it terminates with a lower precision than the Random Forest.

To resolve this matter, we have to examine their respective mean F1 scores so that we can have a clearer image on which classifier is the most effective. The F1-score combines a classifier's precision and recall into a single metric by taking their harmonic mean so, in our case, we can use this metric in order to examine deeper and identify any differences between our classifiers. Table 9 contains mean F1 scores and their respective mean standard deviations for each one of our classifier.

Table 9: Mean F1 and Mean SD per Classifier

	XGB	SVM	DT	RF
Mean F1	78.23%	79.75%	73.58%	76.24%
Mean SD	19.52%	14.26%	14.53%	13.42%

In this table we observe that three of the classifiers have mean F1 scores above 75%, with the SVM model having the greatest one with 79.75%, backing up the Precision-Recall curve from which we suggested that the SVM is the most effective classifier. We can also confirm that the Decision Tree classifier is the worst of the four, having the lowest mean F1 score. Between the two other models, we notice that XGBoost has a higher F1 score (78.23%) than the Random Forest classifier (76.24%). However, if we examine their respective mean Standard Deviation, Random Forest has a lower mean SD. A low standard deviation represents a more reliable classifier, since the data are clustered around the mean, in contrast to a high standard deviation which means that the data are widely spread and the classifier is less reliable.

To summarize, we have deduced that the SVM classifier is the best performing model in detecting healthy and probable Alzheimer’s disease patients, while the Decision Tree is definitely the worst. XGBoost produces a better F1 score than Random Forest, but we have to acknowledge that the XGBoost model has the highest standard deviation, meaning that it could be improved in order to be more stable.

6.3 Literature Comparison

Now that we have completed the models’ implementation and evaluation, we have the opportunity to compare our results with those of other scientific papers that examine the same subject. In order to examine, comprehend and analyze Alzheimer’s Disease and ultimately build the prediction models that are present in this thesis, we have studied a large amount of literature on this subject.

Among the literature mentioned above, we find several papers that tackle the challenge of gathering the necessary dataset, extracting important features and building models able to predict an Alzheimer’s Disease patient. In some of those papers, authors have used the same dataset that we have also used for this thesis, the ADRess Challenge Dataset which is made by the team of DementiaBank. This fact gives us a great opportunity to compare our models’ performance with those of other scientists’ methods and get a better insight of this thesis’ overall efficiency.

One of the papers that we are going to compare our thesis with is written by Balagopalan A. et al [28]. In this paper, authors have gained access to the ADRess Challenge Dataset and have built SVM, Random Forest and Naïve Bayes models. They also have implemented fine-tuned pre-trained Bidirectional Encoder Representations from Transformer (BERT) classification models. Moreover, they have manually extracted lexico-syntactic, acoustic and semantic features from the available dataset in order to obtain the most important characteristics that indicate cognitive impairments. After running a 10-fold Cross Validation they have managed to produce the following results on train and test sets per model [Table 10a, Table 10b], which show that the BERT-based model is performing better than the other classifiers.

Table 10a: Balagopalan A. et al Train Set results

Model	Accuracy	Precision	Recall	F1
SVM	79.6%	81%	78%	79%
RF	73.8%	73%	76%	74%
NB	75%	76%	74%	75%
BERT	81%	84%	79%	81%

Table 10b: Balagopalan A. et al Test Set Results

Model	Class	Accuracy	Precision	Recall	F1
SVM	non-AD	79.6%	83%	79%	81%
	AD		80%	83%	82%
RF	non-AD	75%	71%	83%	77%
	AD		80%	67%	73%
NB	non-AD	72.9%	69%	83%	75%
	AD		79%	63%	70%
BERT	non-AD	83.3%	86%	79%	83%
	AD		81%	88%	84%

Another paper that we are able to compare our results to is the one written by Campbell E.L. et al [31]. In this approach, several linguistic features are constructed from the subjects' interventions. Later, they propose four speech-based systems and two text-based systems for automatically distinguishing patients with and without Alzheimer's disease. As our thesis focuses on text-based data and features, we will only investigate the text-based systems of this paper. For those systems, Campbell and his team have built a recurrent neural network (RNN) and a SVM for classification. This study's results can be seen in Table 11.

Table 11: Campbell E.L. et al Test Set Results

Model	Accuracy	Precision	Recall	F1
RNN	74.07%	69.14%	87%	77.05
SVM	68.52%	67.86%	70.37%	69.09%

The last study for comparison is written by Farrús M. & Codina-Filbà J. [33]. The two researchers extracted 28 acoustic features and 51 transcription features based on lexical and turn-taking information from the ADReSS Challenge Dataset in order to analyze and build several predictive models. For

the classification task they built three different classifiers: Random Forest (RF), Support Vector Machine (SVM) and Multilayer Perceptron (MLP). This paper’s results are shown in Tables 12a and 12b.

Table 12a: Farrús M. & Codina-Filbà J Train Set Results

Model	Accuracy
RF	78.7%
SVM	84.4%
MLP	78.7%

Table 12b: Farrús M. & Codina-Filbà J Test Set Results

Model	Class	Accuracy	Precision	Recall	F1
RF	non-AD	87.5%	82%	96%	84%
	AD		95%	79%	86%
SVM	non-AD	79.2%	85%	71%	77%
	AD		75%	87%	81%
MLP	non-AD	81.2%	86%	75%	80%
	AD		78%	87%	82%

By examining the above studies and comparing them with our thesis (Table 13), we observe that Farrús and Codina-Filbà’s paper achieve better overall performance results. Regarding our thesis, we can derive that it effectively competes with other similar papers that study and tackle the detection of Alzheimer’s Disease using the same dataset.

Table 13: Thesis comparison

Study	Accuracy	Precision	Recall	F1
Balagopalan et al	81%	84%	79%	81%
Campbell et al	74.07%	69.14%	87%	77.05%
Farrús & Codina-Filbà	87.50%	95%	96%	86%
Sarafidis	81.64%	86.79%	77.33%	79.75

A thorough understanding of the dataset's characteristics and how different features behave in the context of Alzheimer's disease is required to perform automatic prediction tasks. Farrús & Codina-Filbà have extracted 79 features that lead them to achieve the best performance for their classifier. By taking advantage of 28 acoustic features, they were able to identify pauses, speech rate, syllable duration and other prosodic and voice quality features that can reveal Alzheimer's Disease related symptoms. They also extracted 51 lexical features that mostly focus on the way subjects describe the Cookie Theft Picture and its main present concepts.

The combination of acoustic-based systems with lexical features derived from transcriptions is the main difference between this best-performing paper and our thesis. Given that acoustic-based systems have the advantage of being language-independent, as they do not rely on human transcriptions, they give us the chance to clean or improve the existing dataset and ultimately extract a more accurate and a more complete set of acoustic-dependent features.

The fact that we find the present thesis in the second place (almost a "tie" with the paper of Balagopalan et al) of this list is really encouraging for the job we have done. Every step of this project is chosen after thoroughly researching the necessary literature regarding Alzheimer's Disease and its symptoms, data manipulation, feature engineering and ML and NLP techniques.

7. Conclusion and Future Work

In this chapter we present the final conclusions that we have derived from this whole thesis procedure, as well as the proposed future work that could improve our results.

7.1 Conclusion

This thesis' purpose was to study neurological disorders and cognitive dysfunction along with dementia symptoms and ultimately propose and build Machine Learning methods based on NLP techniques in order to detect Alzheimer's disease in early stages. After reviewing the necessary literature, we obtained the dataset thanks to the supervisor of this thesis, Researcher Vassiliki Rentoumi, and the TalkBank project. Significant work has been done in order to clean the available data, preprocess them and extract all the necessary features for the training and evaluation of our algorithms. Our suggestion was to implement the XGBoost classifier, a model that, as we observed, was not present in the reviewed literature regarding the development of NLP methods for detecting Alzheimer's disease or other neurological disorders, such as Parkinson's disease or dementia in general.

Except for the XGBoost classifier, we have also implemented three more classification models, Support Vector Machine, Decision Tree and Random Forest, in order to evaluate them and compare them with XGBoost and with each other. All models were trained on the same cleaned and preprocessed data, and they were validated with a stratified 10-fold cross validation. By extracting several decisive classification metrics, such as F1 scores, ROC curves or Precision-Recall curves, we have deduced that our proposal, the XGBoost classifier, is able to detect probable Alzheimer's disease patients with above the average scores. XGBoost's classification scores are not so far away from the corresponding scores of SVM, the best performing model in our thesis. This gives us the aspiration that the XGBoost could reach the highest levels of performance after some more tuning and improvements and be

considered as a sufficient classifying algorithm for detecting Alzheimer's disease.

Furthermore, we have presented a comparison between our best performing model and the corresponding results of three other papers that use the same dataset and engage in the same classification task. The comparison results are promising, since the features that we have manually proposed and selected have helped our best performing model (SVM) achieve results comparable or even better than other published scientific researches.

Finally, given the fact that our thesis approaches a crucial worldwide health issue such as Alzheimer's disease, we believe that every classification model presented here should also be improved in an attempt to achieve better overall performance. Acquiring a better dataset or discovering more and better features is a first step towards the aspiring improvement.

7.2 Future Work

As already stated, not only XGBoost, but all classifiers built in this thesis could be improved in order to be able to produce better classification scores and eventually achieve nearly perfect results in predicting Alzheimer's disease.

With that in mind, a future task could be the implementation of more NLP methods in order to use new and more specific features regarding dementia and Alzheimer's disease. With more features available for input, any classification model could be trained with more deep information related to Alzheimer's disease and eventually it would produce better predictions. We can either create features that will help our models reach our performance goals, or try to extract new features from the available dataset that we may have omitted in this thesis.

Another possible future work would be to use a bigger volume of data, either by collecting our own new dataset or by enriching the already available DementiaBank dataset by interviewing several participants. It is crucial to have the right size of data to avoid under-fitting or over-fitting of our algorithms, so in order to achieve that, we have to experiment on various sizes of datasets and establish the one that fits the most with our case.

In any case, it is only through trials and experiments that we will be able to understand all that is required in terms of data, features or techniques in order to produce machine learning algorithms able to efficiently classify between healthy or probable Alzheimer's disease patients.

References

- [1]: Honavar, V. (2006). "Artificial intelligence: An overview." *Artificial Intelligence Research Laboratory*, 1-14.
- [2]: Ghosh, R. (2022). *Learning Outcomes of Classroom research*
- [3]: Turing, A. M. (1950). "Computing machinery and intelligence" *Mind*, 36, 433-460.
- [4]: Russel, S. & Norvig, P. (2010). *Artificial Intelligence: A modern approach*. Upper Saddle River, NJ : Prentice Hall
- [5]: Newell, A. & Simon, H.A. (1961). "GPS, a program that simulates human thought". *Science*, 134(3495), 2011-2017
- [6]: Wang, F. & Preininger, A. (2019). "AI in health: state of the art, challenges, and future directions". *Yearbook of medical informatics*, 28(01), 16-26.
- [7]: Sinnenberg, L., DiSilvestro, C.L., Mancheno, C., Dailey, K., Tufts, C., Bottenheim, A.M., Barg, F., Ungar, L., Schwartz, H., Brown, D., Asch, D.A., Merchant & R.M. (2016). "Twitter as a Potential Data Source for Cardiovascular Disease Research". *Jama Cardiology*, 1(9), 1032-1036.
- [8]: Jones, K.S. (1994). "Natural Language Processing: A Historical Review". *Current issues in computational linguistics: in honour of Don Walker*, 3-16.
- [9]: Nadkarni, P.M., Ohno-Machado L. & Chapman W.W. (2011). "Natural Language Processing: An introduction". *Journal of the American Medical Informatics Association*, 18(5), 544-551.
- [10]: *Natural Language Processing (NLP) What it is and why it matters*. sas.com from https://www.sas.com/el_gr/insights/analytics/what-is-natural-language-processing-nlp.html.
- [11]: Chowdhury, G.G. (2005). "Natural Language Processing". *Annual Review of Information Science and Technology*, 37, 51-89.
- [12]: Liddy, E.D. (2001). "Natural Language Processing". *Encyclopedia of Library and Information Science*
- [13]: Vaananen, A., Haataja, K., Vehvilainen-Julkunen K. & Toivanen P. (2021). "AI in healthcare: A narrative review". *F1000Research*, 10.6:6.

- [14]: Iroju, O.G. & Olaleke, J.O. (2015). "A Systematic Review of Natural Language Processing in Healthcare". *International Journal of Information Technology and Computer Science*, 8, 44-50.
- [15]: Friedman, C. & Elhadad, N. (2014). "Natural Language Processing in Health Care and Biomedicine". *Biomedical informatics*, 255-284.
- [16]: Attrey, R. & Levit, A. (2018). "The promise of natural language processing in healthcare". *University of Western Ontario Medical Journal*, 87(2), 21-23.
- [17]: Amini, S., Hao, B., Zhang, L., Song, M., Gupta, A., Kardaji, C., Kolachalama, V.B., Au, R. & Paschalidis, I. (2022). "Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach". *Alzheimer's & Dementia*.
- [18]: Kempler, D. & Goral, M. (2008). "Language and Dementia: Neuropsychological Aspects". *Annual review of applied linguistics*, 28, 73-90.
- [19]: Kurlowicz, L. & Wallace, M. (1999). "The Mini Mental State Examination (MMSE)". *Journal of gerontological nursing*, 25(5), 8-9.
- [20]: Kostiantis, S.B., Kanellopoulos, D. & Pintelas, P.E. (2006). "Data Preprocessing for Supervised Learning", *International journal of computer science*, 1(2), 111-117.
- [21]: Cummings, L. (2019). "Describing the Cookie Theft picture: Sources of breakdown in Alzheimer's dementia", *Pragmatics and Society*, 10(2), 153-176.
- [22]: Luz, S., Haider, F., de la Fuente, S., Fromm, D. & MacWhinney, B. (2020). "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge", *arXiv preprint arXiv:2004.06833*.
- [23]: MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing Talk*. 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates
- [24]: Almor, A., Kempler, D., MacDonald, M.C., Andersen, E.S. & Tyler L.K. (1999). "Why Do Alzheimer Patients Have Difficulty with Pronouns? Working Memory, Semantics, and Reference in Comprehension and Production in Alzheimer's Disease", *Brain and language*, 67(3), 202-227.
- [25]: Liu, X., Wang, W., Wang, H. & Sun, Y. (2019). "Sentence comprehension in patients with dementia of the Alzheimer's type", *PeerJ*, 7, e8181.
- [26]: Orimaye, S.O., Wong, J.S.M. & Golden, K.J. (2014). "Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementia using Verbal Utterances", *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality*, 78-87.

- [27]: Moeller, S., Liu, L. & Hulden, M. (2021). “To POS Tag or Not to POS Tag: The Impact of POS Tags on Morphological Learning in Low-Resource Settings”, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 966-978.
- [28]: Balagopalan, A., Eyre, B., Rudzicz, F. & Novikova, J. (2020). “To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer’s disease detection.”, *arXiv preprint arXiv:2008.01551*.
- [29]: Pappagari, R., Cho, J., Moro-Velazquez, L. & Dehak, N. (2020). “Using State of the Art Speaker Recognition and Natural Language Processing Technologies to Detect Alzheimer's Disease and Assess its Severity”, *INTERSPEECH*, 2177-2181.
- [30]: de la Fuente Garcia, S., Haider, F. & Luz, S. (2020). “Cross-corpus feature learning between spontaneous monologue and dialogue for automatic classification of alzheimer’s dementia speech”, *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 5851-5855.
- [31]: Campbell, E. L., Docío-Fernández, L., Raboso, J. J. & García-Mateo, C. (2020). “Alzheimer's Dementia Detection from Audio and Text Modalities”, *arXiv preprint arXiv:2008.04617*.
- [32]: Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z. & Church, K. (2020). “Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease”, *INTERSPEECH*, 2020, 2162-2166.
- [33]: Farrús, M. & Codina-Filbà, J. (2020). “Combining prosodic, voice quality and lexical features to automatically detect Alzheimer's disease”, *arXiv preprint arXiv:2011.09272*.
- [34]: Chen, T. & Guestrin, C. (2016). “XGBoost: A Scalable Tree Boosting System”, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785-794.
- [35]: 2022 Alzheimer’s disease facts and figures. (2022). *Alzheimer’s & dementia: the journal of the Alzheimer’s Association*, 18(4), 700-789
- [36]: Talkbank. <https://www.talkbank.org>
- [37]: Bird, S., Klein, E. & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*, O’Reilly Media
- [38]: Wolpert, D. H. & Macready, W. G. (1997). “No free lunch theorems for optimization», *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, 67-82

[39]: Cortes, C. & Vapnik, V. (1995). “Support-vector networks”, *Machine Learning*, 20(3), 273–297

[40]: Salton, G., Wong, A. & Yang, C. S. (1975). “A vector space model for automatic indexing”, *Communications of the ACM*, 18(11), 613-620

[41]: Salton, G. & Buckley, C. (1988). “Term-weighting approaches in automatic text retrieval”, *Information processing & management*, 24(5), 513-523