POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Machine learning-based analysis of spontaneous speech to detect and monitor decline of cognitive functionality in elderly people

Tesi di Laurea Magistrale in
Biomedical Engineering - Ingegneria Biomedica

Author: **Chiara Giangregorio**

Student ID: 944978

Advisor: Prof.ssa Simona Ferrante

Co-advisors: Emilia Ambrosini, Eugenio Lomurno

Academic Year: 2020-21

Abstract

In the context of the worldwide phenomenon of aging, the decline of cognitive functions has been one of the main focuses of healthcare policies due to the high costs for diagnosis and maintenance of patients. Since speech and language capacity is a well-established early indicator of cognitive deficits including dementia, speech processing methods offer great potential to automatically screen for prototypical indicators in real-time. Therefore, this work aims at analyzing cognitive and functional decline from spontaneous speech-extracted acoustic features via machine learning techniques. First, machine learning performances have been evaluated with acoustic features computed at different time scales (5-10-15s) on 3 datasets of different idioms, Italian, Spanish, and English, respectively. The objective was to find the optimal duration on which to compute the features to speed up the computational time. The binary classification between subjects with no cognitive impairment and those with a mild one was performed yielding 71%, 69%, and 67% in accuracy on the test set, respectively for each dataset. Regarding the multiclass classification task, performances were good on the 3 classes, providing a good discrimination power (66%, 55%, and 49% on test set). Finally, prediction of Mini-Mental State Examination was performed for the first time on Italian and Spanish subjects, whereas the results obtained with the English dataset were compared to those outlined in the ADReSS challenge, since the starting dataset was the same. For the first 2 datasets, MMSE prediction showed promising results, while for the last one, prediction was slightly improved with respect to the challenge. This work shows that good outcomes can be obtained even with features computed at shorter time-lengths and regardless of the language involved, suggesting that longitudinal language-independent monitoring of cognitive decline can be obtained. In this regard, it would be useful to develop a mobile application to run in background during phone calls for automatic feature extraction.

Keywords: acoustic features, dementia, MMSE, machine learning

Sommario

Il declino delle funzioni cognitive è uno dei principali focus delle politiche sanitarie per via degli elevati costi per la diagnosi e la cura dei pazienti. Poiché il parlato e la capacità di linguaggio sono degli indicatori precoci di deficit cognitivi, come la demenza, i metodi di elaborazione dei segnali sul linguaggio sono un'importante risorsa per lo screening automatico in tempo reale. Pertanto, questo lavoro mira ad analizzare il declino cognitivo e funzionale a partire da parametri acustici estratti dal linguaggio spontaneo tramite tecniche di machine learning. In primo luogo, le prestazioni dei modelli di machine learning sono state valutate su parametri acustici calcolati a diverse scale temporali (5-10-15s) su 3 set di dati di lingue diverse, rispettivamente italiano, spagnolo e inglese. Un primo obiettivo è stato trovare la durata ottimale su cui calcolare le caratteristiche per velocizzare i tempi di calcolo dei parametri. La classificazione binaria tra soggetti senza decadimento cognitivo e soggetti con decadimento lieve è stata eseguita ottenendo un'accuratezza sulla validazione rispettivamente di 71%, 69%, e 67% sul test set per ciascun set di dati. Per quanto riguarda il compito di classificazione multiclasse, le prestazioni sono state buone sulle 3 classi, fornendo un buona capacità di discriminazione (66%, 55%, and 49% sul test set). Infine, la previsione del Mini-Mental State Examination è stata eseguita per la prima volta su soggetti italiani e spagnoli, mentre i risultati ottenuti con il dataset inglese sono stati confrontati con quelli riportati nel challenge ADReSS, poiché il dataset di partenza era lo stesso. Per i primi 2 set di dati, la previsione dell'MMSE ha mostrato risultati promettenti, mentre nell'ultimo caso, la previsione è stata leggermente migliorata rispetto alla sfida. Questo lavoro mostra che è possibile ottenere buoni risultati anche con parametri calcolati su durate temporali più brevi e indipendentemente dalla lingua coinvolta, suggerendo che è possibile ottenere un monitoraggio longitudinale indipendente dalla lingua del declino cognitivo.

Parole chiave: parametri acustici, demenza, MMSE, machine learning

Contents

Abstract i

Sommario iii

Contents v

List of Figures ix

List of Tables xiii

1 Introduction 1

2 State of the Art 5

2.1 Voice as a diagnostic tool 5

2.2 Evaluation of cognitive decline 6

2.3 Voice features to assess alterations in cognitive functionality 8

2.4 Tasks for speech induction - an overview 9

2.5 Mini-Mental State Examination 10

2.5.1 The ADReSS challenge 11

2.6 Supervised Learning 12

2.7 Classification 13

2.7.1 Logistic Regression 14

2.7.2 Support Vector Machines 15

2.7.3 CatBoost Classifier 17
2.8 Regression 18
2.8.1 Ridge Regression 18
2.8.2 Support Vector Regressor 19
2.8.3 CatBoost Regressor 20
2.9 Model evaluation techinques 21
2.10 Model explainability - SHAP 23
2.11 Prediction based on acoustic features from speech 25
2.11.1 Classification of cognitive level 25
2.11.2 Prediction of MMSE 26

3 Methods 29

3.1 Optimization of Matlab algorithms for features extraction 29
3.1.1 Pauses 30
3.1.2 Spectral centroid 31
3.1.3 Mel-Frequency Cepstral Coefficients 32
3.1.4 Speech Temporal Regularity 33
3.2 Data 35
3.2.1 MoveCare datasets 35
3.2.2 Pitt corpus 37
3.3 Analysis of time scales for feature extraction 39
3.3.1 Anderson-Darling Test 40
3.4 Data Normalization 41
3.5 Classification tasks 42
3.6 Regression analysis 43

4 Experiments and results 45

4.1 Italian Dataset 45
4.1.1 Binary classification 45
4.1.2 Multiclass classification 49
4.1.3 Regression analysis 53
4.2 Spanish Dataset 56
4.2.1 Binary classification 56
4.2.2 Multiclass classification 60

4.2.3 Regression analysis 63

4.3 Pitt Corpus 67

4.3.1 Binary classification 67

4.3.2 Multiclass classification 71

4.3.3 Regression analysis 75

5 Discussion and Limitations 81

6 Conclusions and Future Developments 85

A Speech Emotion Recognition 87

A.1 Alterations of acoustic profiles in emotion expression . . 87

A.2 Acoustic features for emotion recognition 88

A.2.1 Frequency related features 88

A.2.2 Energy 89

A.2.3 Vocal tract features 89

A.2.4 Spectral features 90

A.2.5 Syllable-related features 90

A.3 Dimensional approach for emotion representation 90

A.3.1 Self-Assessment Manikin (SAM)-Likert Scale . . . 91

A.4 Classification of emotional level 92

B Mini-Mental State Examination - MMSE 95

C Geriatric Depression Scales - GDS 99

Bibliography 103

Acknowledgments 113

List of Figures

1.1 *Life Expectancy at birth in the last 60 years* 1

1.2 *Population shift from 2019 to 2070* 2

2.1 *Cookie Theft picture* 11

2.2 *Example of Logistic Regression algorithm* 15

2.3 *Maximal margin separating hyperplanes for a nonlinearly separable dataset* 16

2.4 *Gradient Boosting Schema*[26] 18

2.5 *Nested k-fold schema* 23

2.6 *Example of SHAP summary plot*[29] 25

3.1 *Workflow of the study* 30

3.2 *Example of pauses calculation* 31

3.3 *Spectral Centroid distribution over a 10s audio segment* . 32

3.4 *Mel-Frequency Cepstrum of a 10s audio segment* 33

3.5 *Distribution of MMSE in Italian dataset* 36

3.6 *Distribution of MMSE in Spanish dataset* 36

3.7 *Distribution of MMSE in Pitt Corpus* 38

4.1 *Classification - Binary settings - Confusion Matrix on test set - Italy* 47

4.2 *Classification - Binary settings - Feature ranking - Italy* . 48

4.3 *Classification - Multiclass settings - Confusion Matrix on test set - Italy* 51

4.4 *Classification - Multiclass settings - Feature ranking - Italy* 52
4.5 *Regression - MMSE prediction - Density plot of residuals on test set - Italy* 54
4.6 *Regression - MMSE prediction - Density plot of residuals per class on test set - Italy* 55
4.7 *Regression - MMSE prediction - Box-plot of residuals on test set - Italy* 55
4.8 *Regression - MMSE prediction - Feature ranking - Italy* . 56
4.9 *Classification - Binary settings - Confusion Matrix on test set - Spain* 59
4.10 *Classification - Binary settings - Feature ranking - Spain* 59
4.11 *Classification - multiclass settings - Confursion matrix of test set - Spain* 62
4.12 *Classification - multiclass settings - Feature ranking - Spain* 63
4.13 *Regression - MMSE prediction - Density plot of residuals on test set - Spain* 65
4.14 *Regression - MMSE prediction - Density plot of residuals on test set - Spain* 65
4.15 *Regression - MMSE prediction - Box-plot of residuals on test set - Spain* 66
4.16 *Regression - MMSE prediction - Feature ranking - Spain* 67
4.17 *Classification - Binary settings - Confusion Matrix on test set - Pitt Corpus* 70
4.18 *Classification - Binary settings - Feature ranking - Pitt Corpus* 71
4.19 *Classification - multiclass settings - Confursion matrix of test set - Spain* 74
4.20 *Classification - multiclass settings - Feature ranking - Pitt Corpus* 75
4.21 *Regression - MMSE prediction - Density plot of residuals on test set - Pitt Corpus* 77
4.22 *Regression - MMSE prediction - Density plot of residuals on the test set per class - Pitt Corpus* 78

4.23 *Regression - MMSE prediction - Box-plot of residuals on test set - Pitt Corpus* 78
4.24 *Regression - MMSE prediction - Feature ranking - Pitt Corpus* 79

A.1 *PAD space* 91
A.2 *Self-Assessment Manikin Likert Scale* 92

List of Tables

3.1 Summary table of extracted features 34

3.2 *Summary of the composition of the datasets - numerosity for each group* 39

3.3 *Dataset Compositions* 39

4.1 *Classification - Binary settings - Only Acoustic Features - Italy* 46

4.2 *Classification - Binary settings - Acoustic Features with Demographic Information - Italy* 46

4.3 *Classification - Binary settings - Performance on test set - Italy* 47

4.4 *Classification - Multiclass settings - Only Acoustic Features - Italy* 49

4.5 *Classification - Multiclass settings - Acoustic Features with Demographic Information - Italy* 50

4.6 *Classification - Multiclass settings - Performance on test set - Italy* 50

4.7 *Regression - MMSE prediction - Italy* 53

4.8 *Regression - MMSE prediction - Performance on test set - Italy* 54

4.9 *Classification - Binary settings - Only Acoustic Features - Spain* 57

4.10 *Classification - Binary settings - Acoustic and Demographic Features - Spain* 58
4.11 *Classification - Binary settings - Performance on test set - Spain* 58
4.12 *Classification - Multiclass settings - Only Acoustic Features - Spain* 61
4.13 *Classification - Multiclass settings - Acoustic Features with Demographic Information - Spain* 61
4.14 *Classification - Multiclass settings - Performance on test set - Spain* 62
4.15 *Regression - MMSE prediction - Spain* 64
4.16 *Regression - MMSE prediction - Performance on test set - Spain* 64
4.17 *Classification - Binary settings - Only Acoustic Features - Pitt Corpus* 68
4.18 *Classification - Binary settings - Acoustic Features with Demographic Information - Pitt Corpus* 68
4.19 *Classification - Binary settings - Performance on test set - Pitt Corpus* 69
4.20 *Classification - Multiclass settings - Only Acoustic Features - Pitt Corpus* 72
4.21 *Classification - Multiclass settings - Acoustic Features and Demographic Information - Pitt Corpus* 73
4.22 *Classification - Multiclass settings - Performance on test set - Pitt Corpus* 73
4.23 *Regression - MMSE prediction - Pitt Corpus* 76
4.24 *Regression - MMSE prediction - Performance on test set - Pitt Corpus* 76

1 | Introduction

According to *The 2021 Ageing Report* by the European Commission, life expectancy is shown in continuous trend both for both males and females over the past years [1]. Moreover, this increase is projected to continue over the period 2019-2070, causing a shift in the composition of population based on age, as it can be seen in Figure 1.2.

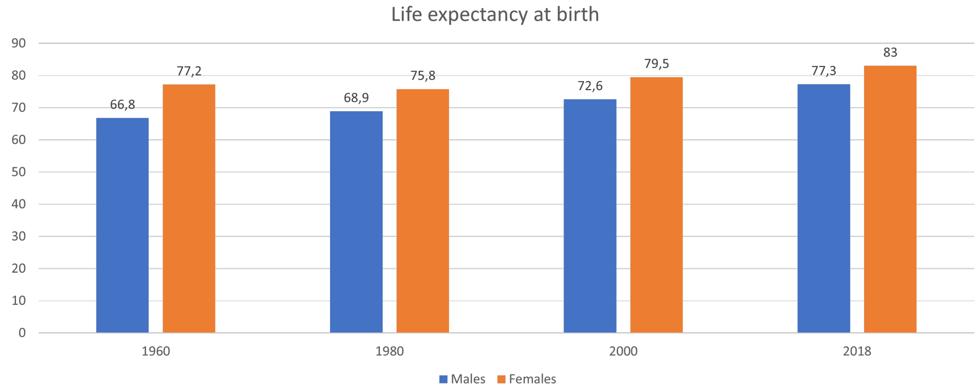

Figure 1.1: *Life Expectancy at birth in the last 60 years*

As life expectancy keeps increasing in Western countries, so does the need for care due to the outburst of chronic illnesses that cause decline of physical and cognitive functions.

Alzheimer's Disease (AD) has become one of the most expensive chronic diseases in society [2]. In comparison with other diseases, the costs entailed for people suffering from AD and other forms of dementia are more

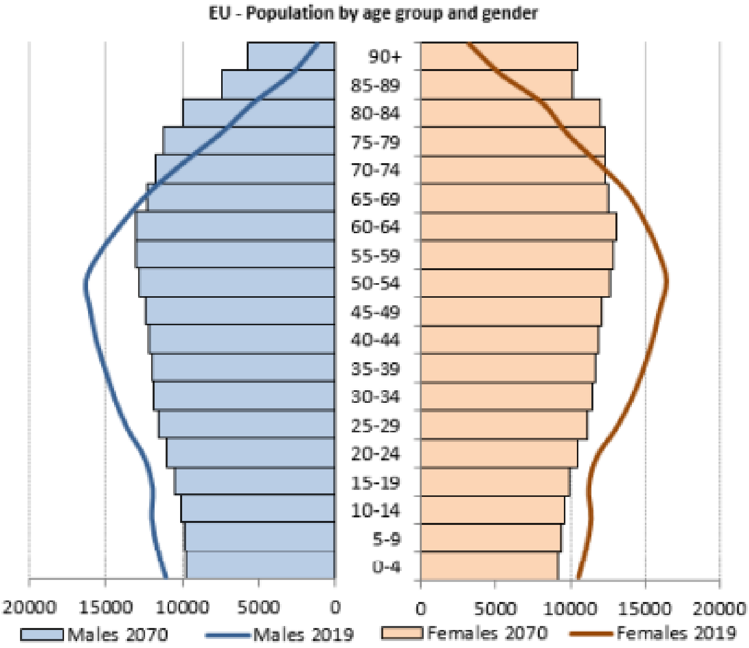

Figure 1.2: *Population shift from 2019 to 2070*

than double those of patients of the same age suffering from cancer, and 74% higher than patients with cardiovascular diseases [3]. Besides the high costs of diagnosis and pharmacological costs, costs of care for people with dementia are high and constitute the major expense, making it a burden for patients and families. According to [4] the care of a person with dementia requires, more than 25 hours per week more than for people without it.
Since there is no effective cure for these diseases, there is the need to have sensitive tools to detect very subtle changes in the pre-clinical population (65-year-old people) in order to react before damage becomes worse. In this context, analysis of acoustic features of speech can be used as a method to early detect the decline of the cognitive functions.

The current work focuses on the analysis of acoustic features of speech by means of machine learning techniques to recognize cases of the cognitive decline. The long-term goal is to develop a mobile app for large-scale remote monitoring, hence there is the need to automatically extract the acoustic features. For this reason, features must be computed on smaller time-lengths, i.e. at most on 15s, to reduce computational cost, and hence avoiding storage of recordings. Therefore, the same features have been computed in Matlab at different time scales of 5-10-15 seconds and performances obtained at different scales have been evaluated. The features have then been analyzed with SHAP explainer to see which ones contributed the most in the prediction of cognitive decline.

Moreover, the extracted features from the different time lengths were considered altogether, and their performances were compared with the results obtained from the datasets with features computed at a singular scale.

The current approach has been tested on three different datasets, two of Latin languages (Italian and Spanish) and the other one of non-Latin language (English).

2 | State of the Art

In this chapter, the previous works related to voice analysis and an overview of the main techniques employed to detect cognitive decline will be outlined.

2.1. Voice as a diagnostic tool

Voice is one of the most studied digital biomarkers in recent years. It is widely employed since it allows an ecological and rapid acquisition of measures concerning many assessments that must usually occur in presence of clinicians, thus allowing these trials to be carried out during everyday activities [5]. The use of digital biomarkers facilitates frequent testing, promising to provide richer and more detailed data, yielding more sensitive measures of symptoms and disease.

It has been demonstrated that alterations in voice and speech prosody encode information about different aspects of physiological and pathological states. Indeed, speech offers insights into cognition and function and it is affected by many psychiatric and neurodegenerative diseases. The analysis of its features allowed for example the detection of people with depression [6, 7], schizophrenia, [8] and autism spectrum disorder [9, 10].

In [6], for example, for the prediction of depressed subjects, it was found that there are indeed vocal differences in loudness, and in the Cepstral Coefficients between the groups, regardless of the emotion involved.

In particular, these features are lower in people with depression with smaller variances than in healthy people. These vocal differences indicate that the depressed voice is untoned, low-pitched, and weak. Moreover, regarding studies for Autism Spectrum Disorder (ASD), [10] a shift in pitch frequencies is noticed between ASD patients during the standardized test for Autism diagnosis (ADOS-2 test) and the control group, as well as changes in loudness. Indeed, ASD is associated with lower pitch value and a wider loudness range compared to those with typical development (TD).

Moreover since the outbreak of the COVID-19 pandemic, speech and coughs have been studied as an attempt to diagnose the disease in a rapid, inexpensive and non-invasive way [11, 12]. The breakout of the pandemic indeed reinforced the concept of telemonitoring patients in a non-invasive way, boosting the development of new toolboxes for the analysis of voice and a new outbreak in the study of its parameters.

2.2. Evaluation of cognitive decline

In the recent years, voice has been widely employed to study the onset of cognitive decline in elderly people. Indeed, the increase of life expectancy in industrialized countries is associated with a severe increase in geriatric diseases. The most common one is dementia, a chronic progressive disease accompanied by loss of autonomy in everyday life.
Dementia is a category of neurodegenerative diseases that entails a long-term and usually gradual decrease in cognitive functioning. It is characterized by a set of symptoms including memory loss, thought difficulties, poor executive functions (e.g. problem-solving, decision-making, planning), language impairment, motor problems, lack of motivation, and emotional distress. Throughout the disease, the severity of these symptoms increases, reducing the patient's autonomy and well-being, as well as their caregivers'. Those cognitive symptoms may be a consequence of the neuropathology of different diseases, such as Alzheimer's Disease (AD; 50% of dementia cases), cerebrovascular disease (25% of cases, in-

cluding those that also manifest AD), Lewy body disease (15% of cases), and other brain diseases (5%), including Parkinson's, frontotemporal dementia and stroke[13].

Current diagnostic procedures require a thorough examination by medical specialists, which are too cost- and time-consuming to be provided on a large scale. Since speech and language capacity is a well-established early indicator of cognitive deficits including dementia, speech processing methods offer great potential to automatically screen for prototypical indicators in real-time. Moreover, they allow to present fast analyses and results to the medical specialists so that they can be included as additional information sources when diagnosing cognitive deficits [14].

There is a need for cost-effective and scalable methods for the detection of dementia from its most subtle forms, such as the preclinical stage of Subjective Memory Loss (SML), to more severe conditions like Mild Cognitive Impairment (MCI) and Alzheimer's Dementia (AD) itself. These pathologies, caused by brain damage and neural functional disruption, start silently up to 20 years before there are clear and observable cognitive symptoms, and there is no effective treatment for them. Therefore, it is fundamental to find strategies to detect the problem as early as possible, to enhance therapy effectiveness and quality of life [13].

Although memory loss is often considered the main symptom of AD, language is a valuable source of clinical information, as well. Moreover, the efficiency and easiness of speech acquisition and processing has led to a large number of studies investigating speech and language features for the detection of AD [15].

It is important to stress that with the onset of AD, emotional response capacity is also affected and behavioral changes are often detected, probably due to memory loss [16]. Emotions and rationality are the main characteristics of human beings. They affect the perception and daily life, for example, the communicative process and decision-making and they are expressed through speech, facial expressions, gestures and other non-verbal clues. Differences between emotional states can also be considered as one of the most important evaluation criteria to measure the performance of cognition processes[3]. In this regard, an analysis of literature about speech emotion recognition can be found in Appendix A.

2.3. Voice features to assess alterations in cognitive functionality

Over the years, several acoustic parameters have been employed to assess cognitive functionality. The most popular ones are those described by:

- the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [17]
- the *emobase* feature set [18]
- the *ComParE* feature set [19]

The eGeMAPS comprises 25 Low-level descriptors (LLD) such as:

- frequency related parameters - pitch, jitter, first 3 formant frequencies, their bandwidth and their related energy and Harmonic-to-Noise Ratio (HNR)
- Amplitude or energy related parameters - energy, shimmer,loudness
- Spectral parameters - first 4 Mel-Frequency Cepstral Coefficients (MFCCs) and Spectral flux

For each of them statistics such as mean, standard deviation and percentile are usually computed, reaching a total of 88 features [17].

The emobase feature set is mainly used for Speech Emotion Recognition tasks, but, since many features in the former task overlap with those employed for early detection of cognitive decline, it has been widely employed and tested in this field of research. It contains the following low-level descriptors (LLD): Intensity, Loudness, 12 MFCC, Pitch (F_0), Probability of voicing, envelope, 8 LSF (Line Spectral Frequencies), Zero-Crossing Rate. Delta regression coefficients are computed from these LLD, and the following functionals are applied to the LLD and the delta coefficients: maximum and minimum value and respective relative position within input, range, arithmetic mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1–3, and 3 inter-quartile ranges, yielding 988 acoustic features for each speech utterance.

The acoustic *COMPARE* feature set is the one defined in the INTERSPEECH 2013 Computational Paralinguistics ChallengE which provides 65 LLDs and their related statistics for a total of 6 373 static features [19].

The aforementioned features sets are all extracted via the openSmile toolbox, an open-source toolkit for audio feature extraction and classification of speech and music signals. In the current work, however, a Matlab feature extractor algorithm with a lower number of features devised in [20] has been employed and it has been optimized by adding new parameters. The extracted features will be described more in detail in Section 3.1.

2.4. Tasks for speech induction - an overview

Feature extraction is usually carried out via processing of audio recordings of vocal signals obtained during the accomplishment of different tasks. The most recurring ones are:

- Interviews
- Movie recalls
- Day descriptions
- Event descriptions (positive or negative)
- Recalling of a dream

- Picture description

The Picture description task is indeed the most common one. One of the main reasons why this task is widely exploited is because of its great test-retest reliability [21], as well as high degree of agreement among raters (inter-rater reliability). Moreover, it has the advantage of focusing attention and overcoming interference from memory difficulties, thus enabling even those subjects with severe memory problems to stay on track more easily [22]. It has been employed in several studies to detect different stages of cognitive decline, such as MCI, AD, Primary Progressive Aphasia (a neurodegenerative disorder in which an alteration in speech is one of the first clinical symptoms), Parkinson's disease and depression. A variety of parameters and acoustic characteristics can be evaluated, e.g. intonation and prosody, discourse fluency and speech rate.

There exists a variety of pictures to be administered from the most complex ones such as "Chaos" to the simplest ones e.g. the Rockwell's "Easter Morning" picture and the "Cookie Theft" of the Boston Diagnostic Aphasia Examination test in Figure 2.1. The latter is the most used one as well as one of the tasks that have been employed in the current work for cognitive evaluation.

2.5. Mini-Mental State Examination

When dementia is suspected, its diagnosis is usually performed by using Mini-Mental State Examination (MMSE, presented in Appendix B), that provides a score from 0 to 30 where 0 corresponds to the major cognitive decline and 30 to no cognitive decline at all [5]. It is based on 30 questions to address short and long-term memory, attention span, concentration, language and communication skills, ability to plan and ability to understand instructions. The aim of the test alone is not to provide a diagnosis for any particular disease but to give an indication of onset of cognitive impairment. A score of 26 or higher is usually classified as normal. If the score is below 25, the result highlights a possible cognitive impairment which may be classified as follows [23]:

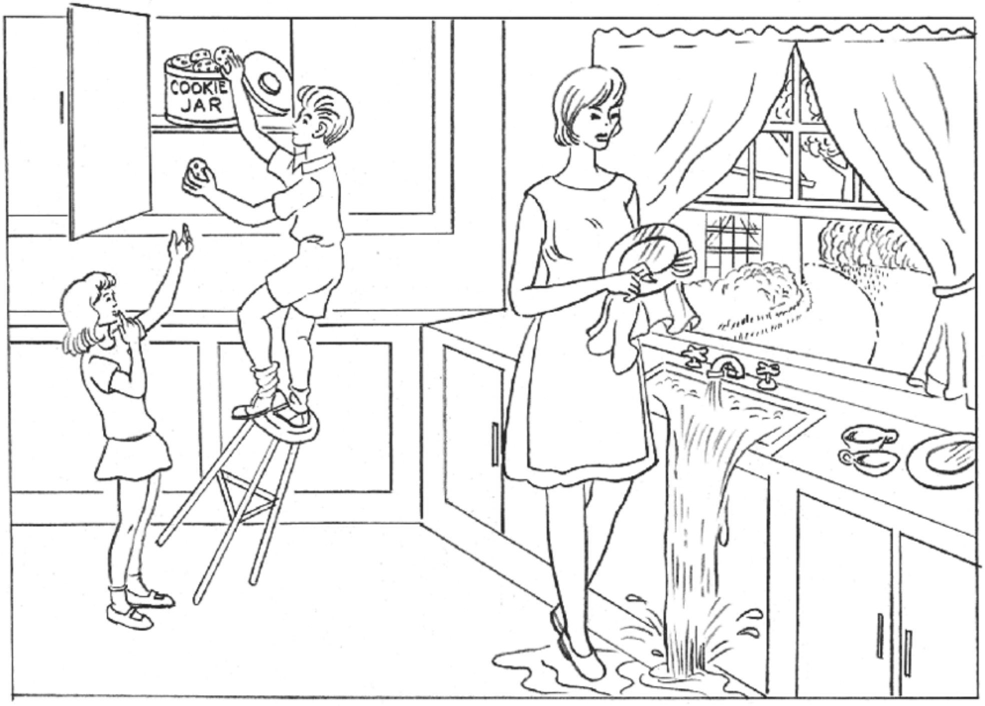

Figure 2.1: *Cookie Theft picture*

- mild — 21 ≤ MMSE score ≤ 25
- moderate — 10 ≤ MMSE score ≤ 20
- severe — MMSE score < 10

Although the test is highly employed and handy, it needs to be performed with the help of a clinician, hence slowing times for diagnosis. Therefore, there is the need to develop other faster non-invasive methods that can be used for large-scale monitoring of dementia in every-day activities.

2.5.1. The ADReSS challenge

In the context of prediction of Mini-Mental State Examination score without a thorough examination, but instead by employing acoustic features, the ADReSS challenge plays a key role.
Indeed, previous studies related to this topic have outlined several signal

processing and machine learning methods for this task, but the field still lacked a balanced and standardised datasets on which these different approaches could be systematically compared.
Therefore, the ADReSS challenge provides researchers with the very first available benchmark, acoustically pre-processed and balanced in terms of age and gender. ADReSS defines two different prediction tasks:

- the AD recognition task, which requires researchers to model participants' speech data to perform a binary classification of speech samples into AD and non-AD classes;
- the MMSE prediction task, which requires researchers to create regression models of the participants' speech in order to predict their scores in the Mini-Mental State Examination (MMSE)

The challenge has been based on audio recordings collected in the Pitt corpus which contains audio recordings of 459 English-speaking subjects whose age is between 55 and 80 years old. Each subject has been asked to describe the cookie theft picture in Figure2.1 adding as much details as possible. The audio recordings are distributed by DementiaBank which is an online shared database of multimedia interactions for the study of communication in dementia.

2.6. Supervised Learning

In the context of searching for tools that are able to detect mental disease and cognitive decline from a set of data related to patients, artificial intelligence techniques such as Supervised Learning plays a paramount role.

Supervised Learning is a subcategory of machine learning and artificial intelligence to detect patterns and relationships between the input data and the output, allowing to predict accurate labeling results when new unseen data are presented. Given a set of N training samples in the form $(x_1, y_1), \dots, (x_n, y_n)$ so that x_i is the input feature vector of the i-th sample and y_i is the output, called label, a learning algorithm seeks a function that can be defined as:

$$g : X \to Y \tag{2.1}$$

where g is the function that maps the independent variables x onto the output target variable Y. A general model works as follows: since it learns by example, it has a training phase during which the model is fed with a set of input variables and the corresponding correct labels. In this way, the algorithm can learn how the output (the label) is related to each input value. The evaluation of the best algorithm is then performed on the validation set, which is a set of samples separated from the training to avoid the overestimate of model performance. The validation set is usually employed to tune the hyperparameters of a model and for model selection: it is indeed exploited to compare the performance of different candidates and choose the best one. Finally, the trained model is presented with test data, which is a set of data that has not been seen by the algorithm either in training or in validation phase. During this phase, data has been labeled, but the labels have not been revealed to the algorithm. Therefore the testing phase aims at measuring how accurately the algorithm performs on unseen data.

In supervised learning, the two main approaches that are widely used, each one addressing a type of data analysis problem, are:

- Classification
- Regression

They will be further explained in the next sections.

2.7. Classification

Classification is the process of categorizing a set of data into classes, often referred to as label or categories. The main goal is to identify which class or category the new data will fall into, by approximating the mapping function from the input variables to discrete output variables.

In the next sections, the following models that are employed for the classification task will be further explained:

- Logistic Regression
- Support Vector Classifier
- CatBoost Classifier

2.7.1. Logistic Regression

Binary logistic model is a statistical model that estimate the probability of one event out o 2 occurring by computing the log-odds (logarithm of the odds) for the event as a linear combination of independent variables. Logistic regression estimates the parameters of the logistic model. In binary logistic regression there is a binary dependent variable, that takes the values $0, 1$. The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1").
The function that converts log-odds to probability is the *standard logistic function*, also known as *sigmoid* function.

$$p(x) = \frac{1}{e^{-(x-\mu)/s}} \tag{2.2}$$

where μ is the midpoint of the curve and s is a scale parameter. Therefore, the posterior probability $P(y|\mathbf{x})$ of the target conditioned on the vector $\mathbf{x}$ is given by:

$$P(y = 0|\mathbf{x}) = \frac{1}{1 + e^{\mathbf{w}'\mathbf{x}}} \tag{2.3}$$

$$P(y = 1|\mathbf{x}) = \frac{e^{\mathbf{w}'\mathbf{x}}}{1 + e^{\mathbf{w}'\mathbf{x}}} \tag{2.4}$$

where $\mathbf{w}$ is the vector of the slope regression coefficients. An example of how Logistic Regression works is represented in Figure 2.2.

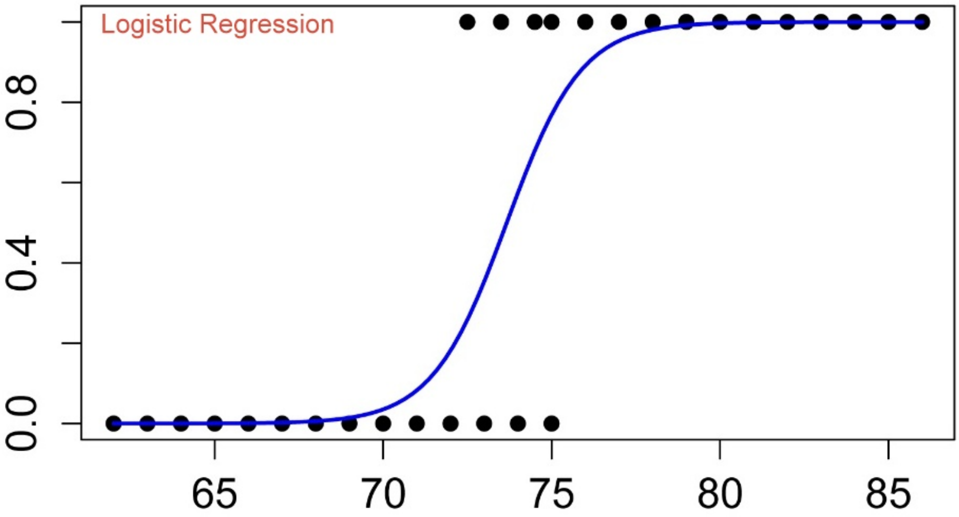

Figure 2.2: *Example of Logistic Regression algorithm*

2.7.2. Support Vector Machines

The aim of Support Vector Machines (SVM) is to find an hyperplane in an N-dimensional space, where N is the number of features, that classifies data points. An hyperplane is a decision boundary that help classify data points. There are several hyperplanes that can be selected. The objective of SVM is to find the plane that maximizes the distance between data points. Therefore, Support Vector Machines identify a set of samples, called indeed support vectors, which are the most representative observations for each target class, playing therefore a more critical role than the other samples, by defining the position and orientation of the separating hyperplane generated by the classifier in the space of features.

SVM maps training samples to points in space so as to maximize the distance between the classes. New samples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. The output of the linear function is taken and if that output is greater than 1, it is identified with one class whereas if the output is -1, with another class. Therefore, the reinforcement range of values([-1,1]) is obtained which acts as margin. A good separation is achieved when the hyperplane has the maximum distance to the separation margin, that are the nearest data points of any class, because the larger will the margin be, the higher the generalization power.

It often occurs that classes are not linearly separable in space, therefore, it is easier making the separation by mapping the original space into a higher-dimensional one. In this case, a penalty parameter is introduced, guaranteeing a trade-off between increasing the margin size and the new samples being of the correct side of the hyperplane. Therefore, in the non-separable case, the optimization problem can be formulated as follows:

$$min_{\mathbf{w},b,d}\frac{1}{2}||\mathbf{w}||^2 + \lambda\sum_{i=1}^{m} d_i \tag{2.5}$$

The objective function is composed of the weighted sum of two terms representing respectively the reciprocal of the margin of separation and the empirical error. The parameter λ is introduced in order to guarantee a trade-off between the generalization capability, represented by the reciprocal of the margin, and the accuracy on the training set, evaluated as the sum of the slack variables [24].

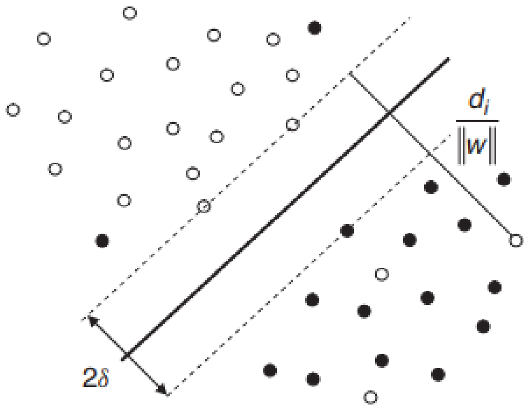

Figure 2.3: *Maximal margin separating hyperplanes for a nonlinearly separable dataset*

2.7.3. CatBoost Classifier

CatBoost (Categorical Boosting) is one of the best boosting algorithms, developed by Yandex researchers and engineers. Main advantages of this model are that it achieves high prediction performances even without parameter tuning, it improves accuracy by reducing overfitting and it works well with less data. Moreover, it works well in heterogeneous datasets, hence with data with high variability of types and formats [25].

Boosting is an ensemble learning technique in which several models (weak learners) are sequentially generated, giving more weight to the errors obtained with the previous models. Indeed, the final model will be an ensemble model with the best accuracy of each constituting model. In particular, CatBoost is based on gradient boosted decision trees. During training phase, a set of decision trees is built consecutively, modifying at each iteration tree's parameters in order to reduce loss function. Therefore, each successive tree is built with reduced loss compared to the previous ones. Training stops when loss function satisfies some specified constraints or there are no improvements on the validation set or the maximum number of trees is reached. A general schema of Gradient Boosting is showed in Figure 2.4.

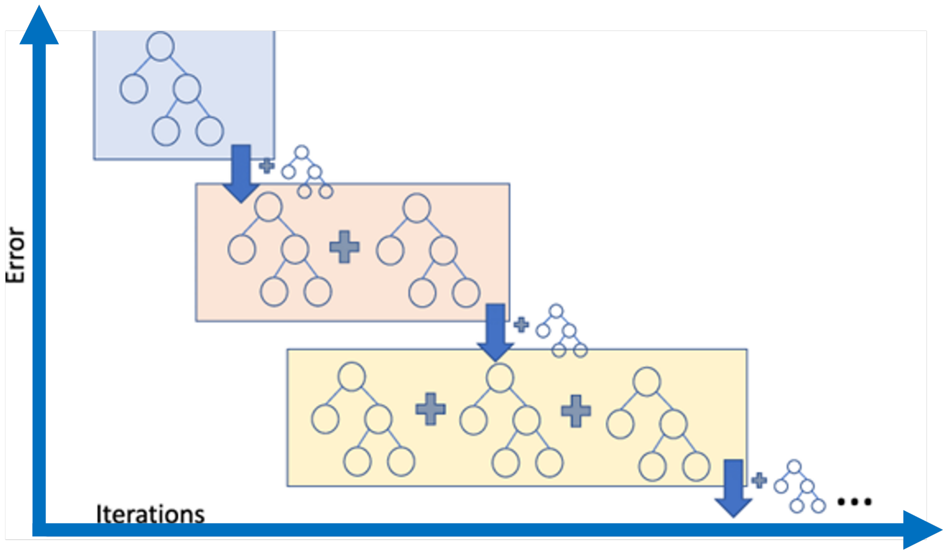

Figure 2.4: *Gradient Boosting Schema*[26]

2.8. Regression

The purpose of regression models, also known as explanatory models, is to identify a functional relationship between the target variable, which in this case is continuous, and a subset of the remaining attributes contained in the dataset. Moreover, they are used to predict the future value of the target, based upon the identified relationships.
The following regressors will be further explained in the next sections:

- Ridge Regression
- Support Vector Regressor (SVR)
- CatBoost Regressor

2.8.1. Ridge Regression

Ridge regression is a linear model, i.e. a model that assumes a linear relationship between the input variables ($\mathbf{x}$) and the single output variable (y). More specifically, as with linear regression, the target y can be calculated from a linear combination of the input variables ($\mathbf{x}$).

To estimate the values of the coefficients, the Ordinary Least Squares (OLS) is one of the most common techniques. The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data the distance from each data point to the regression line is calculated, squared, and all of the squared errors are summed together. Defining $\omega = (\omega_1, ..., \omega_p)$ as the coefficients of the linear model, X as the input data and y as the target, linear regression seeks to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation, therefore:

$$min_w ||Xw - y||_2^2 \tag{2.6}$$

The main difference with linear regression is that ridge regression addresses Ordinary Least Squares through the regularization of the weighys by imposing a penalty on the size of the coefficients, which is why ridge is also called *weight decay*. Indeed, the ridge coefficients minimize a penalized residual sum of squares:

$$min_\omega ||X\omega - y||_2^2 + \alpha ||\omega||_2^2 \tag{2.7}$$

The complexity parameter $\alpha \geq 0$ controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity.

2.8.2. Support Vector Regressor

Support Vector Regressor is more flexible with respect to Linear Regression since it allows to define how much error is acceptable and will find an appropriate line (or hyperplane in higher dimensions) to fit the data. Contrary to Ordinary Least Squares, the aim of the objective function of SVR is to minimize the coefficients and not the squared error. The error term is instead considered in the constraints, where the absolute error is set less than or equal to a defined margin, named maximum error, ϵ, that can be tuned to gain the desired accuracy of the model. The objective function is:

$$\frac{1}{2}||\mathbf{w}||^2 \tag{2.8}$$

and constraint:

$$|y_i - w_i x_i| \leq \epsilon \tag{2.9}$$

2.8.3. CatBoost Regressor

As for the classifier, CatBoost Regressor belongs to the family of gradient boosting algorithms. For regression problems, boosting is a form of "functional gradient descent". It applies a numerical optimization technique for minimizing the loss function by adding, at each step, a new tree that best reduces the loss function. The first regression tree is the one that, for the selected tree size, maximally reduces the loss function. For each following step, the focus is on the variation in the response that has not been explained by the model so far [27]. General algorithm formulation F_m for the prediction of y_i considering the input data x_i is:

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i) \tag{2.10}$$

where h_m are the weak learners, in this case decision tree regressors. The algorithm is built in a greedy fashion:

$$F_m(x) = F_{m-1}(x) + h_m(x), \tag{2.11}$$

where h_m is fitted to minimize the sum of losses L_m, therefore:

$$h_m = argmin_h L_m = argmin_h \sum_{i=i}^{n} l(y_i, F_{m-1}(x_i) + h(x_i)), \tag{2.12}$$

where $l(y_i, F(x_i))$ is the considered loss function to be minimized.

2.9. Model evaluation techinques

To have an estimation of the model performances on data, the most common validation technique is k-fold cross-validation. It consists in splitting the whole set into k parts. Iteratively, for i in the range from 1 to k, the i-th part of the k folds is kept out, while the k-1 parts are used to fit the model. Performance score is then evaluated on the i-th fold. In the end, there will be k scores that will be averaged to obtain mean performance score (e.g. accuracy for the classification tasks).

When standard Cross-Validation is applied on a dataset with a small number of samples, there is a higher risk to obtain a biased test set, hence overestimating model performances. To have a less biased and more correct estimation of the model capability, Nested k-fold Cross Validation. It is a more complex procedure compared to the one previously described to simultaneously select the best and most robust machine learning model for the dataset and to tune the best set of hyperparameters, which involves two loops, an outer and an inner one.

The procedure involves treating model hyperparameter optimization as part of the model itself and evaluating it within the broader k-fold cross-validation procedure for evaluating models for comparison and selection. The k-fold cross-validation procedure for model hyperparameter optimization is nested inside the k-fold cross-validation procedure for model selection. The use of two cross-validation loops also leads the procedure to be called *"double cross-validation"*. The algorithm is describe in Algorithm 2.5 [28].

In particular, in the outer loop, the dataset is split into k sets. Iteratively, for each k, the k-th set is chosen as test set while the other k-1 sets are involved in the training set as it happens with the traditional k-fold Cross-Validation.

In the internal loop, each training set from the outer loop is split into I sets. Again, iteratively, for each i, the i-th set is kept as the inner test set, while the other I-1 sets yield the inner training set.

Algorithm 2.1 Nested K-fold Cross-Validation

```
Define hyperparameters combination C, for current model (C is empty if there are no hyperparameters)
Divide data in K folds with approximately equal distributions of classes
for Parameter combination c in C do
    for fold $k_i$ in the K folds do
        Set the fold $k_i$ as test set
        for Fold $k_i$ in $K - 1$ folds do
            Set fold $k_j$ as validation set
            Train model on $K - 2$ folds
            Evaluate model performance on fold $k_j$
        end for
        Calculate average performance on $K - 2$ folds
    end for
    Train model on $K-1$ folds using hyperparameter combination that produced best performance in inner loop
    Evaluate model performance on fold $k_i$
end for
Calculate average performance over K folds
```

The validation set scores in the inner loop are averaged on k*I sets. The best average score of validation scores, associated with a certain set of hyperparameters is used to establish the optimal hyperparameters for that model.
In the end, the best model with the tuned hyperparameters is re-trained on the outer test set and evaluated on the outer test set, averaged on k scores.

Under this procedure, during hyperparameter search, the opportunity to overfit is highly reduced with respect to other validation methods

since it uses only to a subset of the dataset provided by the outer cross-validation procedure. In this way, the risk of overfitting is highly reduced, almost eliminated, giving an estimate of a model's performance on the dataset with a lower bias.
A downside of nested cross-validation is the dramatic increase in the number of model evaluations performed. In the case of Nested 10-fold CV, there is an increase of 10 times in computational cost. In Figure 2.5 the schema of Nested k-fold CV is shown.

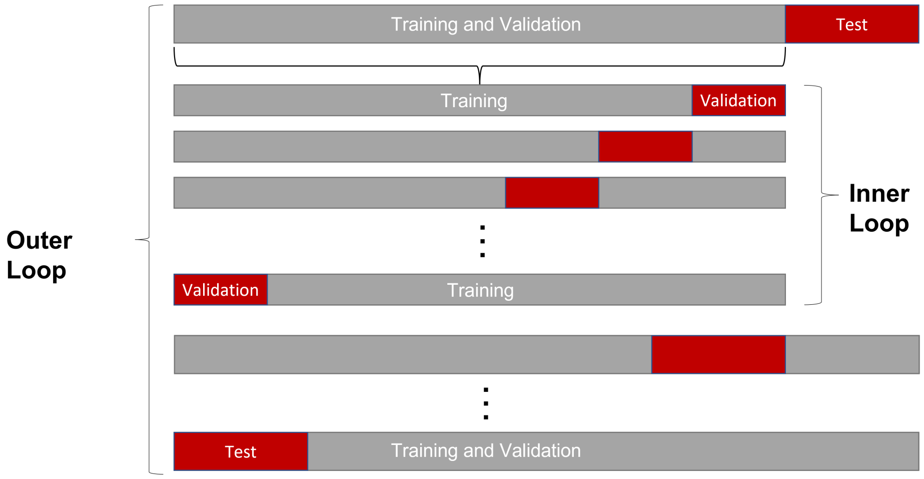

Figure 2.5: *Nested k-fold schema*

2.10. Model explainability - SHAP

SHapley **A**dditive ex**P**lanations (SHAP), is a method based on co-operative game theory, used to increase transparency and interpretability of machine learning models. It is a solution concept in cooperative game theory. Cooperative games are games in which it is possible to forge alliances between players forming "coalitions". In this case the considered players are the features of the dataset. Therefore, by analyzing features through the framework of cooperative game theory, the focus will be on predicting which coalitions will be formed, and the resulting collective payoffs. To understand the impact of the introduction of a specific feature on the final prediction, for each observation, SHAP computes the

marginal contribution of each feature on the output of the model. It starts with base values for prediction based on a-priori knowledge. For each feature all the combinations with the other features are evaluated; then the same prediction is computed without considering the specific feature, obtaining the marginal contribution of the feature as the difference between the two outcomes.
The main difference with other algorithms for feature importance evaluation is that they usually compute score for all input features, evaluated individually, whereas SHAP, having foundation in game theory evaluates contribution of features inside the coalitions.

The contribution of the features in SHAP for each instance can be explained through *summary plots* which combine feature importance with feature effects. An example of summary plot can be seen in Figure 2.6. A summary plot is a density scatter plot of Shapley values for each feature. Indeed, each point corresponds to a Shapley value for a feature and an instance, whereas the color represents the value of the feature from low (blue) to high (red). Moreover, features are ordered based on their importance on the y-axis in descending order. Finally, the position along the x-axis gives indications of the impact on the prediction.

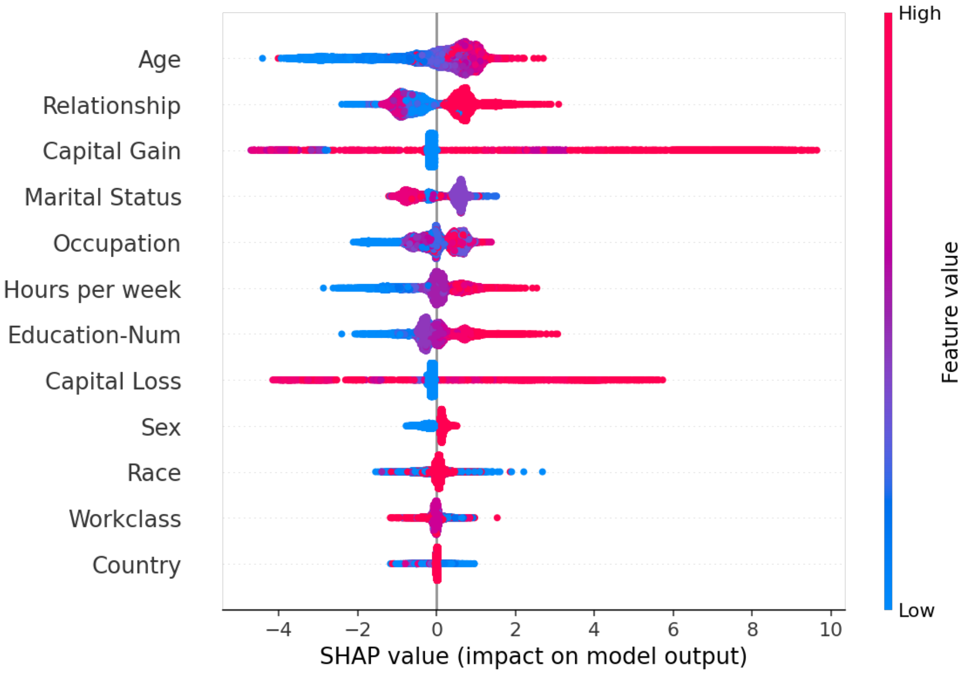

Figure 2.6: *Example of SHAP summary plot*[29]

2.11. Prediction based on acoustic features from speech

In the following sections the most relevant results in predicting cognitive decline - via classification - and MMSE - through regression - from acoustic features will be outlined.

2.11.1. Classification of cognitive level

In [30], Authors analysed the temporal parameters of reading fluency to discriminate between Spanish-speaking asymptomatic subjects and those with AD. The algorithms applied to the recordings were capable of differentiating between AD patients and controls with an accuracy of 80% (specificity 74.2%, sensitivity 77.1%) based on speech rate.
Moreover, in [31] it has been demonstrated that it is possible to differentiate between several kinds of dementia and Mild Cognitive Impairment

both in binary and multiclass scenarios, through free speech tasks with high classification accuracy. In [32], it was showed that acoustic parameters such as speech rate, hesitation ratio, number of pauses and articulation rate yield significant results in the discrimination between MCI and healthy subjects in the movie recall task, obtaining an F1-score of 78.8%. Moreover, in [33] Authors were able to discriminate between controls and MCI subjects with Random Forest and Support Vector with a high F1-score of around 75% with the nested-leave-one-subject-out cross-validation.
Finally, some studies have been carried out in a multimodal context, by employing both vocal and eye-tracking features, such as in [34], in which Fraser et al. discriminated between a cohort of 26 MCI subjects and 29 healthy, with a series of cascaded classifiers with multimodal features, obtaining an accuracy of 83%.

2.11.2. Prediction of MMSE

In literature there are few evidences of attempts at prediction of Mini-Mental State Examination score. The main contribution is the ADReSS challenge at INTERSPEECH 2020 which defines a shared task through which different approaches to the automated recognition of AD based on spontaneous speech can be compared[15]. The Authors have established a baseline RMSE of 6.14 for acoustic features. In this context, with the same dataset, Authors of [35] obtained a RMSE of 7,1 and a MAE of 6,2 with support vector machine and linear regression, with features from the ComParE set. In [36], a RMSE of 6,49 was obtained with prosody features and support vector regressor. In [37], Authors chose the mean absolute error (MAE) as their score for performance evaluation and the lowest one that they achieved was of 3.83, by computing features correlation for selecting the top 40 features. In [38], linear regression analysis showed that fusion of acoustic features, age, sex and years of education provided better results (MAE = 4.97, and R2 = 0.261) with respect to the use of acoustic features alone (MAE = 5.66 and R2 = 0.125).

With respect to the previous work, the main novelties of this thesis are:

- Evaluation of optimal duration of speech segments for feature extraction
- Optimization of Matlab feature extraction algorithm with addition of new features
- Prediction of Mini-Mental State Examination from Italian and Spanish voice segments
- Validation of model performances through Nested 10-fold cross-validation

3 | Methods

This work aims at classifying cognitive decline in elderly people as well as predicting Mini-Mental State Examination score from voice features. The feature extraction code was optimized from a previous work [20, 39] with the introduction of new features. Moreover, while in [39] features were computed on the whole length of the audio recordings, in the current work different time segments have been taken into account to analyze dependence of features with time and to find the optimal duration length for features extraction. The workflow of the main phases of this work is shown in Figure 3.1.

3.1. Optimization of Matlab algorithms for features extraction

With the new Matlab Audio Toolbox [40], the following features have been added to the original set:

- Pauses
- Spectral centroid
- Mel-Frequency Cepstral Coefficients
- Speech temporal regularity

Moreover, the mean fundamental frequency (mean F0) has been sub-

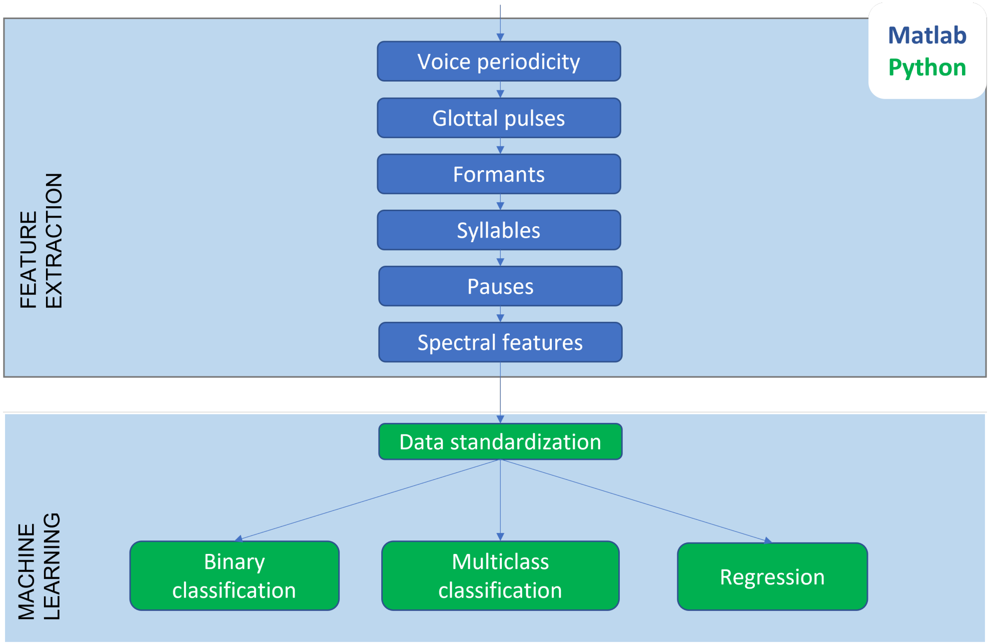

Figure 3.1: *Workflow of the study*

stituted with mean pitch and its standard deviation, computed with the same window-length and overlap of [20].

3.1.1. Pauses

The number of pauses and their mean duration inside each segment have been computed with the function *detectSpeech* from the Matlab Audio Toolbox [40]. The function uses a thresholding algorithm based on energy and spectral spread on each frame to highlight the indices corresponding to the boundaries of speech signals, as it is shown in Figure 3.2.

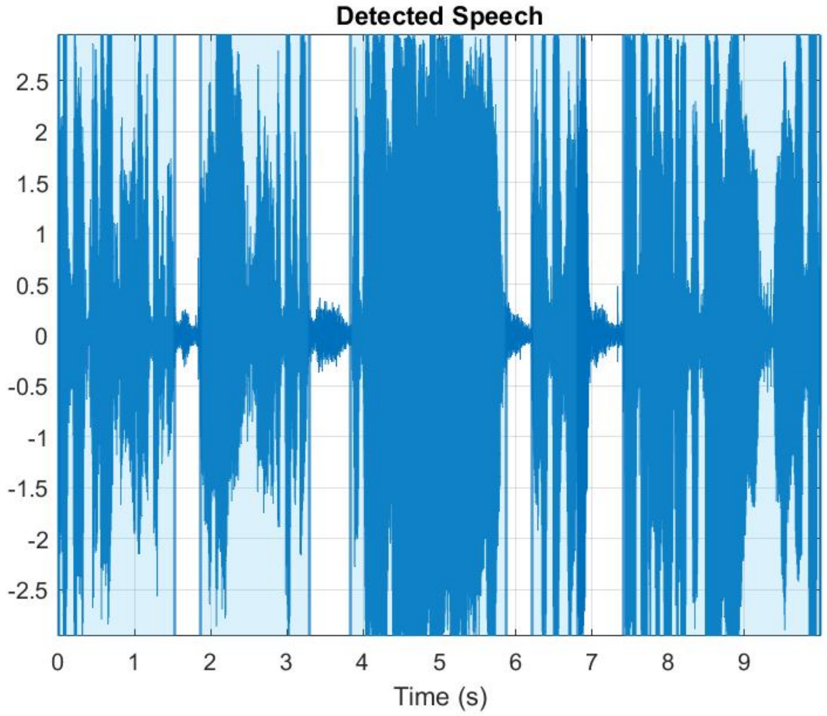

Figure 3.2: *Example of pauses calculation*

3.1.2. Spectral centroid

The spectral centroid is the center of 'gravity' of the spectrum. The value of spectral centroid, C_i, of the i^{th} audio frame is defined as:

$$C_i = \frac{\sum_{k=b_1}^{b_2} f_k s_k}{\sum_{k=b_1}^{b_2} s_k} \tag{3.1}$$

with f_k being the frequency in Hz corresponding to bin k, s_k the spectral value at bin k and b_1 and b_2 being the band edges over which to calculate the spectral centroid. The Matlab function returns a centroid for each frame as it can be seen in Figure 3.3. To compute a singular value for each time segment, the length of the frame was set equal to the length of the segment with no overlap.

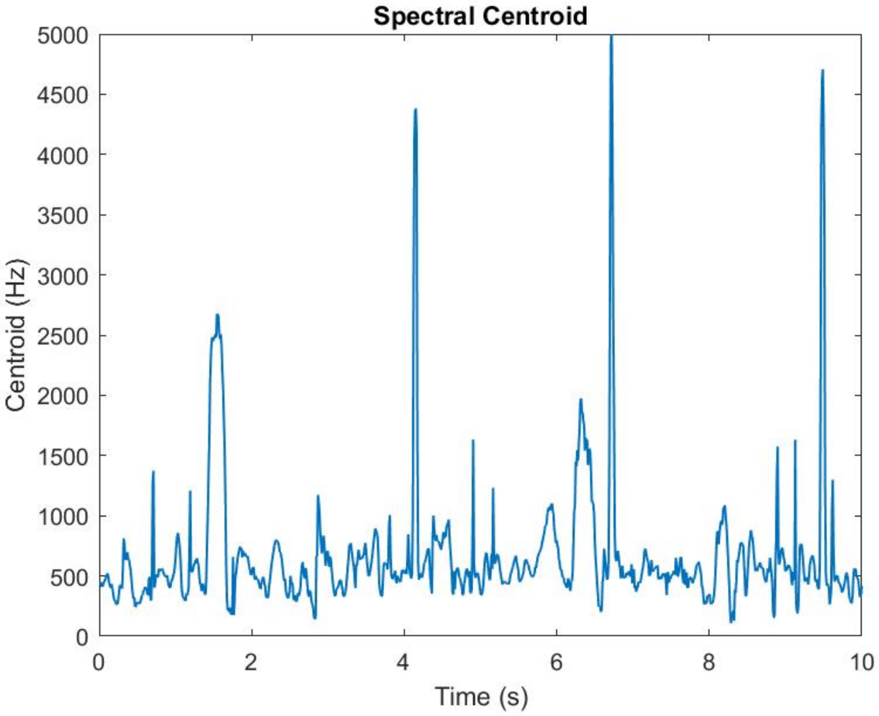

Figure 3.3: *Spectral Centroid distribution over a 10s audio segment*

3.1.3. Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) are a type of cepstral representation of the signal, where the frequency bands are distributed according to the mel-scale, instead of the linearly spaced approach. They are the coefficients making up a Mel-Frequency Cepstrum (MFC), a representation of a short-term power spectrum of a sound [41]. In Figure 3.4 the Mel-Frequency Cepstrum is shown, which is the information of rate of change in spectral bands. In Matlab, the first 16 coefficients were computed.

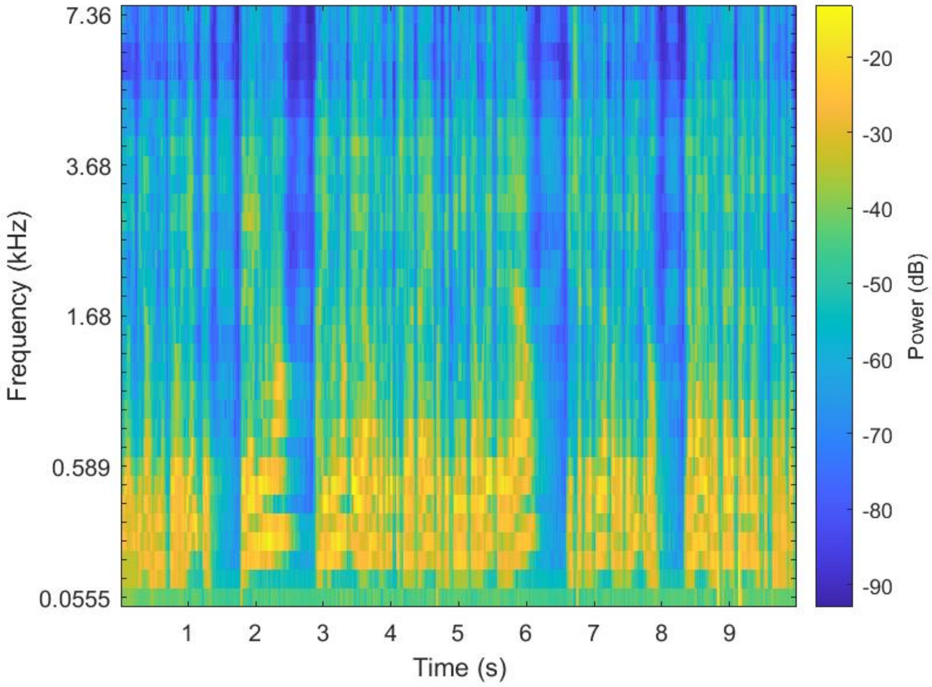

Figure 3.4: *Mel-Frequency Cepstrum of a 10s audio segment*

3.1.4. Speech Temporal Regularity

Speech temporal regularity captures the temporal structure of speech segments. It is calculated as the sum of the absolute values of the first 16 cepstrum coefficients (MFCC1 - MFCC16) which yields the measure of the temporal regularity of the sequence. In normal voices, the duration of the contiguous speech segments tends to be longer - in the order of a few seconds on average - and more “regular”, which typically results in higher values of MFCC1 - MFCC16. Conversely, for AD cases, the duration of the contiguous speech segments tends to be shorter, often tenths of a second, and less “regular”, which typically results in lower average values of MFCC1 - MFCC16 [42].

The final set of acoustic features is summarized with a brief description in Table 3.1, along with an indication of how they change in subjects with dementia with respect to healthy subjects.

Table 3.1: Summary table of extracted features

Type	**Feature**	**Description**	**Trend with dementia**
Voice periodicity	Unvoiced percentage	Percentage of aperiodic parts in the audio segment	↑
	Voiced and unvoiced parts	Mean, median, 15 and 85 percentile of the parts of the signal with and without periodic nature	↓ voiced and ↑ unvoiced
	Pitch	Contour of fundamental frequency F0	↓
	Shimmer	Random cycle-to-cycle temporal changes of the amplitude of the vocal fold vibration	↓
Glottal pulses	Total voice breaks	Percentage of distances between consecutive glottal periods	↑
Formants	Standard deviation of the 3rd formant	degree of tonal modulation of the voice	↑
Syllables	Speech Rate	Number of syllables per second	↓
	Phonation Percentage	Percentage of syllables throughout the speech signal	↓
	Articulation Rate	Number of syllables over the phonation time	↓
	Intersyllabic time	Duration between syllables	↑
	Intrasyllabic time	Duration of syllables	↑
Pauses	Pauses	Number of pauses for each audio segment	↑
	Duration of pauses	Mean duration of pauses for each audio segments	↑
Spectral features	MFCC	First 16 Mel-Frequency Cepstral Coefficients	↓
	Speech Temporal Regularity	Temporal structure of speech segments	↓
	Centroid	Location of the center of mass of the spectral signal	↑

3.2. Data

Prediction of cognitive decline was carried out on three already existing datasets, two of which of Latin-derived-language speaking subjects (Italian and Spanish) and the other one of a non-Latin language (English). The following section will describe them in terms of participants and acquisition protocol.

3.2.1. MoveCare datasets

The first two datasets come from a previous European project, namely MoveCare, which developed a multi-actor platform for the independent living of the elders at home by monitoring and promoting activities to contrast decline and social exclusion [43]. The project took place in Italy and Spain, therefore the two datasets are composed by 153 Italian and 150 Spanish-speaking subjects, respectively.
Experimental protocol and excluding criteria are described in [20]. Basic requirement for patient recruitment was an age over 65. Exclusion criteria were:

- Subjects with MMSE score ≤ 10
- Non-native speaking subjects
- Depressed subjects
- Subjects affected by hearing loss

Participants have been divided in three groups based on the Mini-Mental State Examination (MMSE) score they achieved. Therefore, the Italian recruitment groups were composed as follows:

- Group 1: 62 subjects with MMSE > 26
- Group 2: 46 subjects with $20 \leq$ MMSE ≤ 26
- Group 3: 45 subjects with a $10 <$ MMSE < 20

Whereas, the Spanish recruitment groups were:

- Group 1: 60 subjects with MMSE > 26
- Group 2: 45 subjects with $20 \leq$ MMSE ≤ 26
- Group 3: 45 subjects with $10 <$ MMSE < 20

The distribution of the scores of MMSE are shown in Figures 3.5 and 3.6, for the Italian and Spanish dataset respectively.

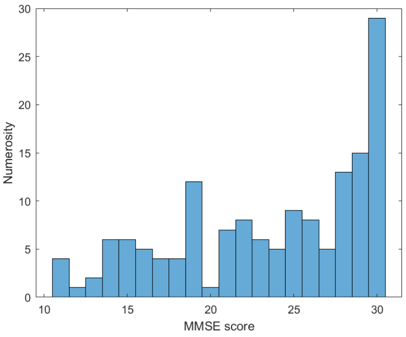

Figure 3.5: *Distribution of MMSE in Italian dataset*

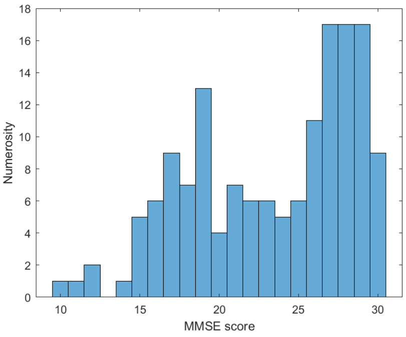

Figure 3.6: *Distribution of MMSE in Spanish dataset*

Each participant carried out the following tasks:

- 3 story-telling tasks
- Picture description task

For the first three tasks, subjects were asked to tell three short stories in an interrupted way for approximately two minutes each:

1. Positive story
2. Negative story
3. Episodic story, an event in the recent past that did not involve strong emotions, in a neutral tone

Within the picture description task, subjects were asked to describe a picture freely in an uninterrupted way, trying to add as many details about the scene as possible. The picture was the *Cookie Theft* picture of the Boston Diagnostic Aphasia Examination test, presented in Figure 2.1, considered an ecologically valid approximation to spontaneous speech [44].

To avoid having depressed participants among the healthy group, subjects from Group 1 filled out the Geriatric Depression Test (GDS, presented in Appendix C), which is a self-report assessment used to identify depression in the elderly population [45]. One point is assigned to each answer and the cumulative score is rated on a scoring grid. For the Short-Form GDS, subjects fill in a 15-item questionnaire about their satisfaction on life, their interests and social interactions. Subjects with a score over 6 are considered mildly depressed and severely as the score increases to 9. Concerning the Long-Form GDS, instead, subjects fill in a 30-item questionnaire about the same topics outlined above. In this case, subjects with a score over 9 are considered mildy depressed and a score over 20 corresponds to severe depression. As for the MMSE, the GDS questionnaire does not provide a definitive diagnosis, but it is a first indication of probable depression. In this study, the Italian subjects filled the Long-Form GDS whereas the Spanish subjects filled the Short-Form, therefore, only the subjects with score < 10 have been included for the Italian dataset and those with a score < 7 for the Spanish dataset. In this way, 43 subjects were kept from Group 1 of the Italian dataset and 44 subjects for Group 1 of the Spanish dataset, guaranteeing also more balanced classes.

3.2.2. Pitt corpus

Pitt corpus contains audio recordings of 459 English-speaking subjects whose age was between 55 and 80 years old. Each subject was asked to describe the *Cookie Theft* picture adding as much details as possible. In this case as well, to each subject a Mini-Mental State Examination score was associated after they underwent the related test.

The database is distributed by DementiaBank which is a shared database of multimedia interactions for the study of communication in

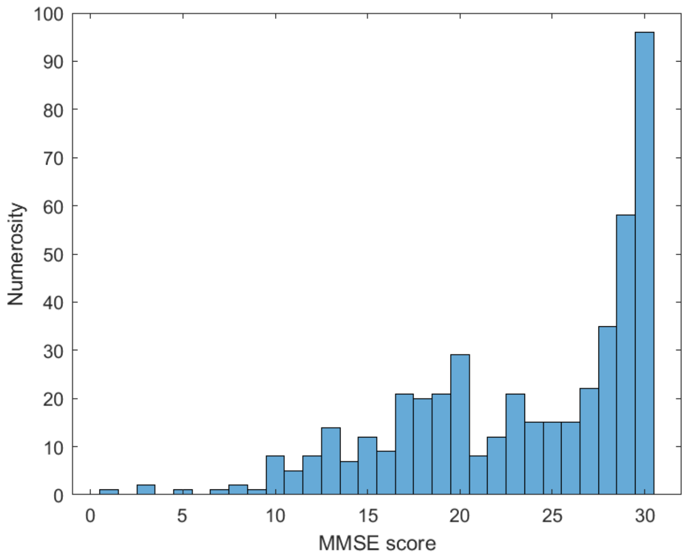

Figure 3.7: *Distribution of MMSE in Pitt Corpus*

dementia. Access to the data in DementiaBank is password protected and restricted to members of the DementiaBank consortium group.
For the Pitt corpus, subjects were divided in 3 groups according to their MMSE, as well, therefore, obtaining:

- Group 1: 211 subjects with MMSE > 26
- Group 2: 115 subjects with $20 \leq$ MMSE ≤ 26
- Group 3: 113 subjects with MMSE < 20

In this case the classes were unbalanced and no exclusion criterion was introduced.
The distribution of the MMSE score among subjects of the Pitt Corpus is shown in Figure 3.7.

The composition of the 3 final datasets after the exclusion criteria are summarized in Tables 3.2 and 3.3. In Table 3.2, each row corresponds to a specific dataset whereas on the columns, the number of subjects contained in each group of cognitive decline is specified. In Table 3.3, for each group, mean and standard deviation of age, years of education and MMSE are reported, as well as the ratio between men and women.

Table 3.2: *Summary of the composition of the datasets - numerosity for each group*

	Group 1	Group 2	Group 3	Total
ITALY	44	45	44	133
SPAIN	43	45	45	133
PITT	211	115	113	459

Table 3.3: *Dataset Compositions*

Dataset		Group 1	Group 2	Group 3
ITALY	Age(years)*	76,6 (4,9)	82,8 (4,6)	86 (5,7)
	Men/Female	6/39	11/33	7/37
	Years of education*	12,4 (3,6)	8,7 (4,3)	7,4 (4,5)
	MMSE*	29 (1)	24 (2)	16 (3)
SPAIN	Age(years)*	79,7 (7,5)	82,4 (6,9)	85,5 (6,6)
	Men/Female	22/21	9/36	17/28
	Years of education*	6 (4,2)	5 (3,2)	6,3 (3,9)
	MMSE*	28 (1)	23 (2)	7 (2)
PITT	Age(years)*	64,6 (8,3)	71,2 (8,4)	72,1 (8,4)
	Men/Female	89/122	40/75	40/73
	Years of education*	14 (2,6)	12,7 (2,9)	11,8 (2,7)
	MMSE*	29 (1)	23 (2)	15 (4)

*. Mean (standard deviation)

3.3. Analysis of time scales for feature extraction

In the previous work, features were computed considering the whole length of the audio signals, that lasted on average 2 minutes each [20]. Thanks to the outburst of telemonitoring, the monitoring and delivering of therapy at a distance in an automatic way is frequently employed and its use is constantly increasing. In this context, it would be useful, in the near future, to develop a mobile app for automatic real-time acquisition of voice features. Therefore, in order to compute the features through an app, there would be the need to reduce computational cost by computing features on smaller time scales. Hence, it was analyzed how machine learning performances change when extracting features at different time scales, to find the optimal time interval. Recordings have been divided in segments of a pre-defined length on which features were computed. The process has been done by considering the duration of segments in the range between 5 and 15s with a 5s step. As a result, 9 different starting datasets were obtained (3 datasets from the Italian dataset, 3 from the Spanish one and 3 from the Pitt Corpus). Moreover, a further analysis was carried out, by considering datasets in which the features computed at different scales were considered *altogether*.

In order to reduce noise, for each one of the features mean and standard deviation or median and interquartile range have been computed over the 4 recordings, depending on whether their distribution was normal or not. In this way, in the final dataset each subject is represented by a single entry. Therefore, a normality distribution check for each feature was carried out through Matlab by performing the Anderson-Darling test (*adtest*).

3.3.1. Anderson-Darling Test

The Anderson-Darling test is commonly used to test whether a data sample comes from a normal distribution. It measures the distance between the hypothesized distribution (i.e. normal), $F(x)$, and the empirical cumulative distribution function (cdf) $F_n(x)$ as

$$\int_{-\infty}^{\infty} (F_n(x) - F(x))^2 w(x) dF(x) \tag{3.2}$$

over the ordered sample values $x_1 < x_2 < ... < x_n$, where w(x) is a weight function and n is the number of data points in the sample. The weight function for the Anderson-Darling test is

$$w(x) = [F(x)(1-F(x))]^{-1} \tag{3.3}$$

that places higher weight on the observations in the tails of the distribution. In this way, the test is more sensitive to outliers and better at detecting deviation from normality in the tails of the distribution [46].
In *adtest*, the decision to reject or not the null hypothesis is based on comparing the p-value for the hypothesis test with the specified significance level, not on comparing the test statistic with the critical value.
Normality check was done for each dataset and for each patient and in the end the feature distribution was considered normal if in the majority of the patients it presented that specific nature.

3.4. Data Normalization

The extracted features in Matlab were used for the aforementioned three tasks (binary and multiclass classification, and regression). To reduce computational cost and improve models' performances, data were pre-processed in Python. In particular, each feature was normalized and rescaled in the range [0,1]. Considering x as a single feature, the following transformation has been applied:

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)} \tag{3.4}$$

From previous studies, it has been demonstrated that age, sex, and years of education are strong indicators of cognitive level, but in an aim of monitoring cognitive decline over time, it was chosen to study if the results obtained only with acoustic features were similar to those obtained with demographic characteristics as well. Therefore, each dataset was tested twice, considering:

- Dataset 1 – only acoustic features
- Dataset 2 – acoustic features, plus age, sex and years of education

3.5. Classification tasks

Classification was evaluated with the following 2 strategies:

- Binary classification
- Multiclass classification

Binary classification was performed to distinguish between Group 1 (no cognitive decline) and Group 2 (Mild cognitive decline), to evaluate the capability of the model to detect early stages of dementia. Multiclass classification between the three classes was implemented considering the *One vs All* technique.
To perform classification, the following standard Machine Learning classifiers were applied:

- Logistic regression (LR)
- Support Vector Machines (SVC)
- CatBoost (CATBOOST)

Each classifier was implemented in Python using the Scikit-learn libraries and catboost library and trained and validated on each one of the aforementioned datasets [47, 48].

The performances of the different classifiers were compared in terms of F1-score. Moreover, accuracy, recall, and precision were measured on the test set. Accuracy is the proportion of correct outcomes, therefore the sum of true positives (TP) and true negatives (TN), among the total number of cases examined:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

where FP = False Positive and FN = False Negative.

Recall is also defined the true positive rate or sensitivity and it represents the proportion of correctly-predicted positive instances with respect to the total positive instances:

$$Recall = \frac{TP}{TP + FN} \tag{3.6}$$

Precision instead descibes the quality of a positive prediction made by the model, thus it is the proportion of true positives with respect to the total number of positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{3.7}$$

F1 score summarizes contribution of precision and recall since it is calculated as their harmonic average:

$$F1score = \frac{2TP}{2TP + FP + FN} \tag{3.8}$$

It is important to stress that in the case of multi-class classification, the true positive condition is associated to one specific class and the true negative one to the other two.

3.6. Regression analysis

Regression was applied in order to predict the MMSE score. In this case, the target variable is not considered categorical, instead continuous. The following standard Machine Learning regressors were applied:

- Ridge Regression
- Support Vector Regressor (SVR)
- CatBoost Regressor

Regression performances were evaluated by computing mean absolute error (MAE), which compares the predicted versus the observed value by calculating the average absolute difference between the two.

$$MAE = \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{N} \tag{3.9}$$

where y_i is the i-th observation (the true value) and $\hat{y}_i$ is the i-th predicted value of the total N samples.
Moreover, for the test set, root mean squared error (RMSE) was computed. RMSE represents the square root of the second sample moment of the differences between predicted values and observed values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}} \tag{3.10}$$

Due to the relatively small number of subjects involved, to obtain a more robust testing, both classification and regression algorithm were validated with Nested 10-fold Cross-Validation procedure.

4 | Experiments and results

In this Chapter, the results of this thesis are illustrated, from the perspective of the performances of the selected subsets in the machine learning tasks. For each dataset, results of binary classification, multi-class classification, and MMSE regression are presented. For each task, two summary tables compare all the employed models. The first one is related to the dataset in which only acoustic features are considered and the latter refers to the one that considers the acoustic features, plus age, sex, and years of education of each patient.

4.1. Italian Dataset

4.1.1. Binary classification

The performance of binary classification models in dividing Italian-speaking subjects between group 1 and group 2 is shown in Tables 4.1 and 4.2. The models were applied on datasets of features computed at 5, 10, and 15s and then on the dataset with the features at the different time scales altogether. The comparison was carried out in terms of F1-score on the validation sets and when the performances were the same, the simplest model was chosen as the best one. The choice of the best model was carried out only by comparing the results obtained by the dataset with the acoustic features, in an aim to employ these algorithms for longitudinal monitoring.

Table 4.1: *Classification - Binary settings - Only Acoustic Features - Italy*

	Time segment	Accuracy	Recall	F1-score
CATBOOST	5	**0,78** ($\pm 0,02$)	0,76 ($\pm 0,05$)	0,76 ($\pm 0,03$)
	10	**0,78** ($\pm 0,02$)	**0,76** ($\pm 0,02$)	0,76 ($\pm 0,03$)
	15	0,77 ($\pm 0,03$)	0,74 ($\pm 0,05$)	**0,77** ($\pm 0,03$)
	5-10-15	0,77 ($\pm 0,03$)	0,74 ($\pm 0,04$)	0,74 ($\pm 0,03$)
SVM	5	0,73 ($\pm 0,02$)	0,70 ($\pm 0,04$)	0,71 ($\pm 0,03$)
	10	0,71 ($\pm 0,03$)	0,65 ($\pm 0,04$)	0,67 ($\pm 0,04$)
	15	0,68 ($\pm 0,03$)	0,58 ($\pm 0,04$)	0,63 ($\pm 0,05$)
	5-10-15	0,71 ($\pm 0,04$)	0,68 ($\pm 0,06$)	0,68 ($\pm 0,05$)
LR	5	0,74 ($\pm 0,03$)	0,68 ($\pm 0,04$)	0,71 ($\pm 0,04$)
	10	0,72 ($\pm 0,02$)	0,66 ($\pm 0,03$)	0,69 ($\pm 0,04$)
	15	0,72 ($\pm 0,03$)	0,63 ($\pm 0,04$)	0,68 ($\pm 0,04$)
	5-10-15	0,72 ($\pm 0,03$)	0,67 ($\pm 0,04$)	0,69 ($\pm 0,03$)

Table 4.2: *Classification - Binary settings - Acoustic Features with Demographic Information - Italy*

	Time segment	Accuracy	Recall	F1-score
CATBOOST	5	0,84 ($\pm 0,04$)	0,83 ($\pm 0,05$)	0,83 ($\pm 0,03$)
	10	**0,84** ($\pm 0,02$)	0,82 ($\pm 0,03$)	0,83 ($\pm 0,04$)
	15	**0,84** ($\pm 0,02$)	**0,83** ($\pm 0,04$)	**0,83** ($\pm 0,02$)
	5-10-15	0,82 ($\pm 0,03$)	0,78 ($\pm 0,04$)	0,80 ($\pm 0,02$)
SVM	5	0,78 ($\pm 0,01$)	0,73 ($\pm 0,02$)	0,76 ($\pm 0,03$)
	10	0,75 ($\pm 0,04$)	0,70 ($\pm 0,04$)	0,73 ($\pm 0,04$)
	15	0,78 ($\pm 0,04$)	0,73 ($\pm 0,04$)	0,76 ($\pm 0,04$)
	5-10-15	0,74 ($\pm 0,04$)	0,70 ($\pm 0,05$)	0,71 ($\pm 0,04$)
LR	5	0,79 ($\pm 0,03$)	0,74 ($\pm 0,03$)	0,76 ($\pm 0,04$)
	10	0,77 ($\pm 0,03$)	0,70 ($\pm 0,03$)	0,74 ($\pm 0,04$)
	15	0,78 ($\pm 0,03$)	0,71 ($\pm 0,04$)	0,76 ($\pm 0,04$)
	5-10-15	0,76 ($\pm 0,03$)	0,71 ($\pm 0,03$)	0,74 ($\pm 0,03$)

Performances on the validation set are similar, regardless of the time scale of the features. Overall, CatBoost achieves the best accuracy, recall, and F1-score. In particular, the best model is CatBoostClassifier with the features computed at 15s which in validation reaches 77% both in accuracy and F1-score, and on the test set, it achieves 71% and 66%, respectively in accuracy and F1-score, as it can be seen in Table 4.3. Table 4.2 confirms the results from literature that states that performances of the models indeed improve when the demographic information is added to the acoustic feature set, reaching even 84% with CatBoostClassifier.

Table 4.3: *Classification - Binary settings - Performance on test set - Italy*

Time segment	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
15	CATBOOST	0,71	0,74	0,63	0,66

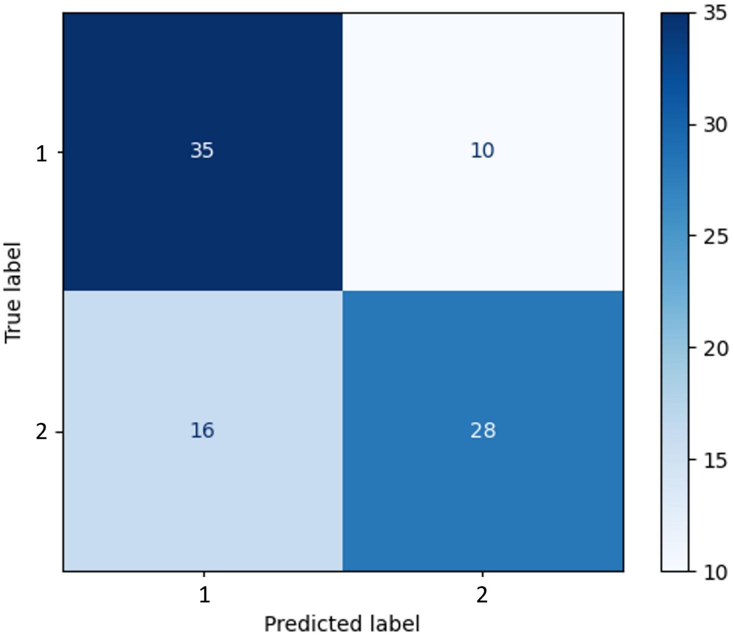

Figure 4.1: *Classification - Binary settings - Confusion Matrix on test set - Italy*
Class 1 represents Group 1 subjects (MMSE>26) and Class 2 represents Group 2 subjects (20≤MMSE≤26).

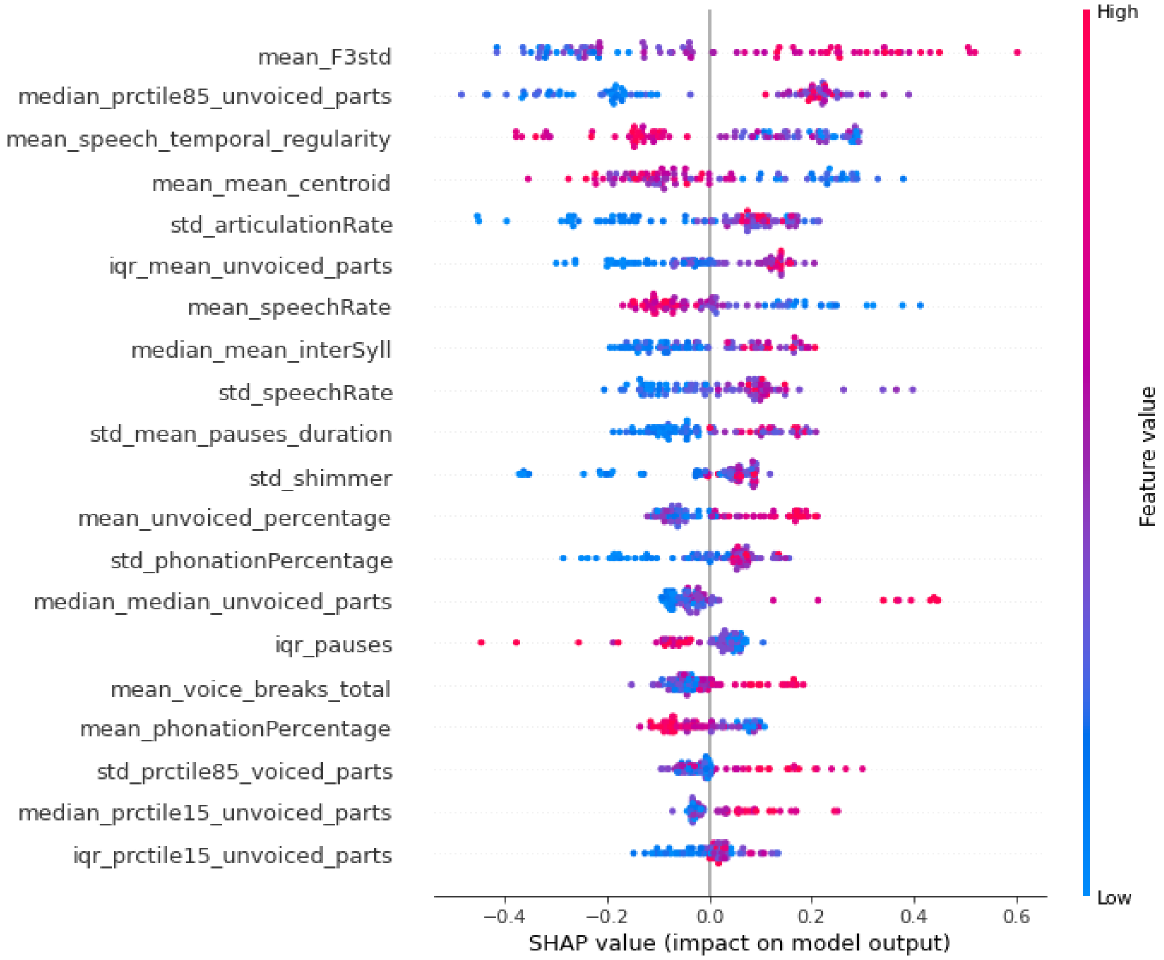

Figure 4.2: *Classification - Binary settings - Feature ranking - Italy*

The confusion matrix of the best model - CatBoostClassifier - with features computed on 15s segments- shown in Figure 4.1 confirms the good performances stated by the metrics.

From the feature ranking in Figure 4.2, it can be seen that the most discriminating features are the standard deviation of the 3^{rd} formant (the tonal modulation of the voice), the number of voiced parts, temporal regularity, and the the speech and articulation rate. In particular, lower values of F3 and articulation rate suggest a subject with no cognitive impairment. On the contrary, low values of speech temporal regularity suggest that the subject has mild cognitive impairment, which is in line with the literature [42].

4.1.2. Multiclass classification

The performance of multiclass classification models in discriminating subjects in the 3 groups is shown in Tables 4.4 and 4.5. The models were applied on datasets of features computed at 5-10 and 15s and then on the dataset with the features computed at the different time scales altogether. The comparison was carried out in terms of F1-score on the validation sets and when the performances were the same, the simplest model was chosen as the best one. CatBoost overall obtains the best results for all the time scales. The model with all the features computed at 5-10-15s altogether is the one that performs better with an F1-score of 64% on the validation set.

Table 4.4: *Classification - Multiclass settings - Only Acoustic Features - Italy*

	Time segment	**Accuracy**	**Recall**	**F1-score**
CATBOOST	5	**0,66** $(\pm 0,02)$	0,66 $(\pm 0,02)$	0,63 $(\pm 0,02)$
	10	**0,66** $(\pm 0,02)$	**0,67** $(\pm 0,02)$	0,63 $(\pm 0,02)$
	15	0,65 $(\pm 0,02)$	0,65 $(\pm 0,02)$	0,62 $(\pm 0,02)$
	5-10-15	0,65 $(\pm 0,03)$	0,65 $(\pm 0,03)$	**0,64** $(\pm 0,03)$
SVM	5	0,53 $(\pm 0,03)$	0,53 $(\pm 0,02)$	0,50 $(\pm 0,03)$
	10	0,52 $(\pm 0,05)$	0,52 $(\pm 0,05)$	0,49 $(\pm 0,04)$
	15	0,52 $(\pm 0,03)$	0,52 $(\pm 0,03)$	0,48 $(\pm 0,02)$
	5-10-15	0,52 $(\pm 0,03)$	0,52 $(\pm 0,03)$	0,50 $(\pm 0,05)$
LR	5	0,54 $(\pm 0,03)$	0,54 $(\pm 0,03)$	0,51 $(\pm 0,03)$
	10	0,54 $(\pm 0,03)$	0,54 $(\pm 0,03)$	0,51 $(\pm 0,03)$
	15	0,55 $(\pm 0,02)$	0,55 $(\pm 0,02)$	0,51 $(\pm 0,02)$
	5-10-15	0,54 $(\pm 0,02)$	0,54 $(\pm 0,02)$	0,52 $(\pm 0,04)$

Table 4.5: *Classification - Multiclass settings - Acoustic Features with Demographic Information - Italy*

	Time segment	Accuracy	Recall	F1-score
CATBOOST	5	**0,69**($\pm 0,02$)	0,69 ($\pm 0,02$)	0,67 ($\pm 0,01$)
	10	0,66 ($\pm 0,02$)	0,67 ($\pm 0,02$)	0,63 ($\pm 0,02$)
	15	0,65 ($\pm 0,02$)	0,65 ($\pm 0,02$)	0,62 ($\pm 0,02$)
	5-10-15	0,68 ($\pm 0,03$)	0,69 ($\pm 0,03$)	0,66 ($\pm 0,03$)
SVM	5	0,59 ($\pm 0,02$)	0,59 ($\pm 0,02$)	0,56 ($\pm 0,03$)
	10	0,66 ($\pm 0,04$)	0,56 ($\pm 0,04$)	0,53 ($\pm 0,04$)
	15	0,60 ($\pm 0,03$)	0,60 ($\pm 0,03$)	0,56 ($\pm 0,03$)
	5-10-15	0,55 ($\pm 0,03$)	0,55 ($\pm 0,03$)	0,53 ($\pm 0,05$)
LR	5	0,59 ($\pm 0,03$)	0,59 ($\pm 0,03$)	0,56 ($\pm 0,03$)
	10	0,69 ($\pm 0,03$)	**0,70** ($\pm 0,03$)	**0,68** ($\pm 0,03$)
	15	0,65 ($\pm 0,02$)	0,65 ($\pm 0,02$)	0,62 ($\pm 0,02$)
	5-10-15	0,55 ($\pm 0,03$)	0,55 ($\pm 0,03$)	0,53 ($\pm 0,04$)

From Table 4.5, it can be seen that the information about age, sex, and years of education indeed improves the model performances, mainly increasing the metrics of the simpler models, obtaining high recall (70%) and F1-score even with Logistic Regression.

Table 4.6: *Classification - Multiclass settings - Performance on test set - Italy*

Time segment	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
5-10-15	CatBoost	0,66	0,68	0,66	0,63

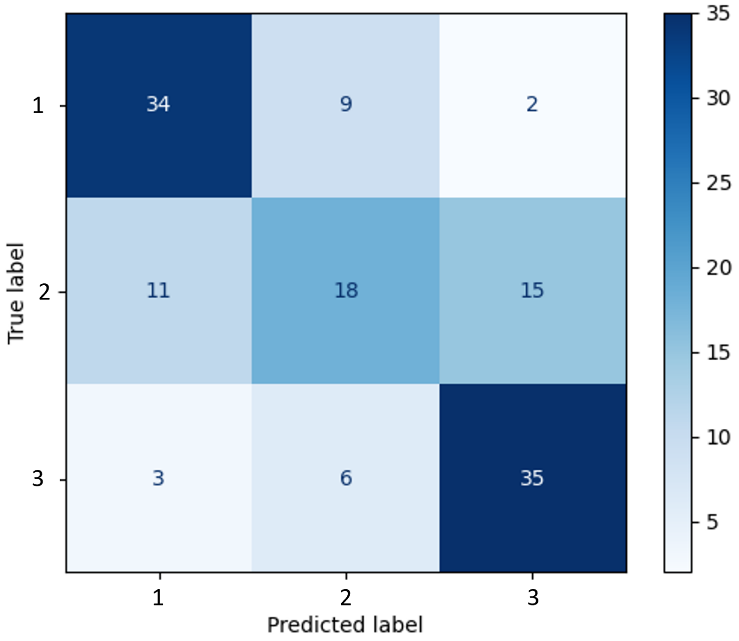

Figure 4.3: *Classification - Multiclass settings - Confusion Matrix on test set - Italy*
Class 1 represents Group 1 subjects (MMSE>26), Class 2 represents Group 2 subjects (20≤MMSE≤26) and Class 3 represents Group 3 subjects (MMSE>26).

In Table 4.6, all the metrics - i.e. accuracy, precision, recall, and F1-score - of the performance on the test set of the best model are shown. From validation to test set, metrics do not change much suggesting that there is no overfitting, with a good generalization power, although the large number of features. On the test set, considering that the prediction was between 3 classes (hence if a model predicted the outcomes in a random way 33% in accuracy would be expected), performances were good obtaining 66% in accuracy. From the confusion matrix in Figure 4.3 it can be seen that healthy patients and subjects with a severe cognitive impairment can be well distinguished, whereas the group with mild dementia tends to be confused, but overall they are correctly predicted.

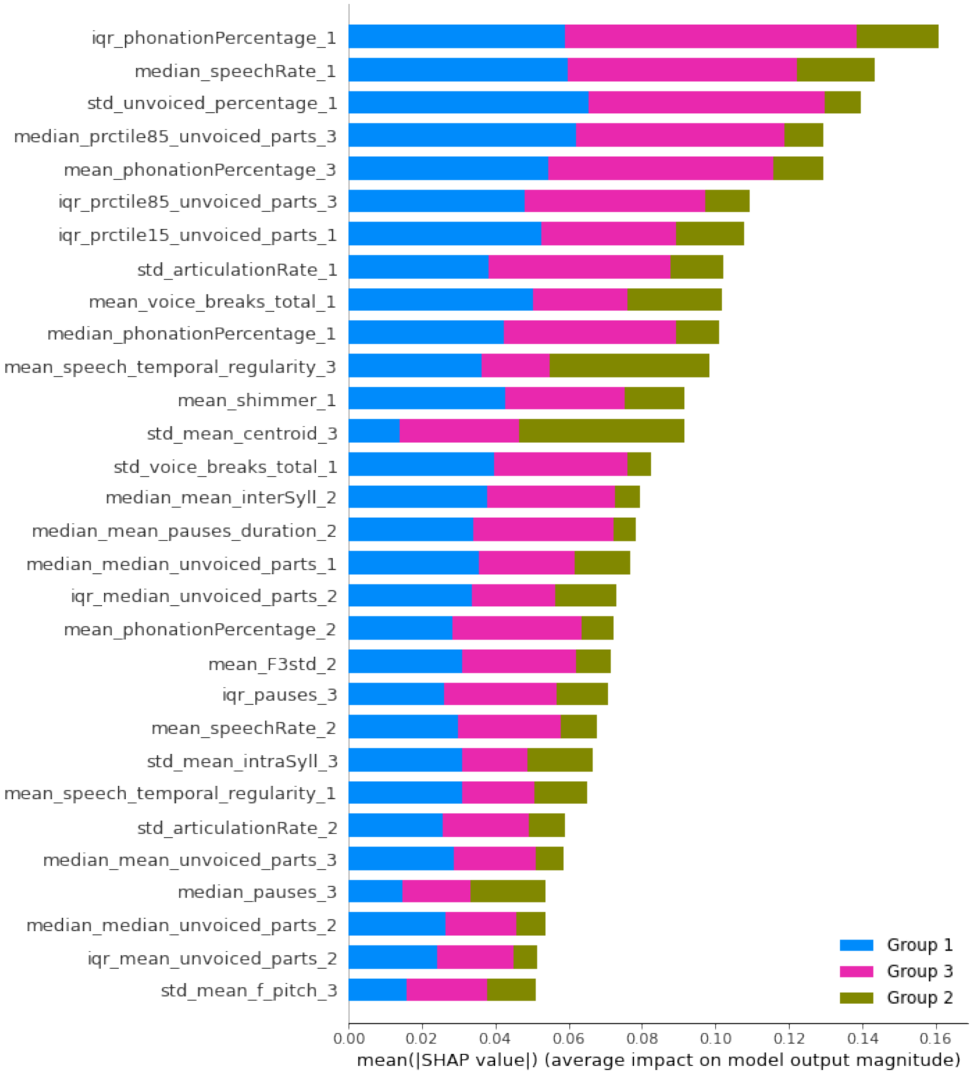

Figure 4.4: *Classification - Multiclass settings - Feature ranking - Italy*

The ranking in Figure 4.4 shows that the most significant features are mainly estimated on 5 and 15s segments. In this case, the most discerning ones are related to the speed of speech, in particular, the variation of the percentage of syllables in the time window and the mean number of syllables per second, which have a great contribution to predicting group 1 (healthy) subjects and group 3 (high cognitive impairment) subjects. Instead, the most informing feature for group 2 (mild cognitive impairment) subjects is the speech temporal regularity computed on segments of 15s.

4.1.3. Regression analysis

The performance of regression models for predicting MMSE scores among the Italian-speaking participants is shown in Table 4.7. Dataset 1 corresponds to the dataset with only acoustic features; dataset 2 instead represents the dataset with the addition of age, sex, and years of education. The models were applied on datasets of features computed at 5-10 and 15s and then on the dataset with the features at the different time scales altogether. The comparison was carried out in terms of mean absolute error on the validation sets, in order to reduce the error between the predicted value and the real target. When the performances were the same, the simplest model was chosen as the best one. The errors obtained were similar overall, slightly worse with Support Vector Regressor (SVR) with respect to the other two models. However, Ridge Regression (LR in Table, since it is a Linear Regression) with features computed at 15s proves to be the best model, achieving a mean absolute error of 3,46 in the validation set. In this case, a simpler model performs not only in a comparable way, but better than a much more complex model such as CatBoost.

Table 4.7: *Regression - MMSE prediction - Italy*

	Time segment	**CATBOOST**	**SVR**	**LR**
Dataset 1	5	3,52	3,75	3,65
	10	3,51	3,68	3,66
	15	3,47	3,59	**3,46**
	5-10-15	3,48	3,92	3,62
Dataset 2	5	3,40	3,77	3,49
	10	3,39	3,80	3,57
	15	3,33	3,52	3,37
	5-10-15	3,42	4,01	3,51

Mean absolute error in predicting MMSE score (range 0-30)

Table 4.8: *Regression - MMSE prediction - Performance on test set - Italy*

Time segment	**Model**	**MAE**	**RMSE**
15	RIDGE REGRESSION	3,39	4,16

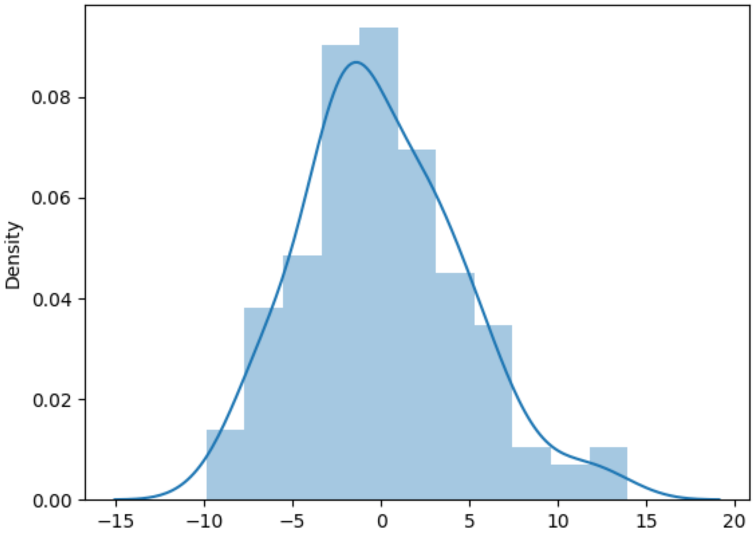

Figure 4.5: *Regression - MMSE prediction - Density plot of residuals on test set - Italy*

In Table 4.8, the mean absolute error and root mean absolute error of the performance of the best model on the test set are shown, achieving for the former 3,39 and the latter 4,16. The improvement - decrease - in the mean absolute error from validation to test shows that the model does not overfit, thus having a good generalization power.

Although in Figure 4.5 residuals have quite a normal distribution overall, the boxplot of residuals in Figure 4.7 suggests that there is a slight polarization, especially for the lower scores to be predicted that should need further investigation, which is confirmed by the density plot in Figure 4.6.

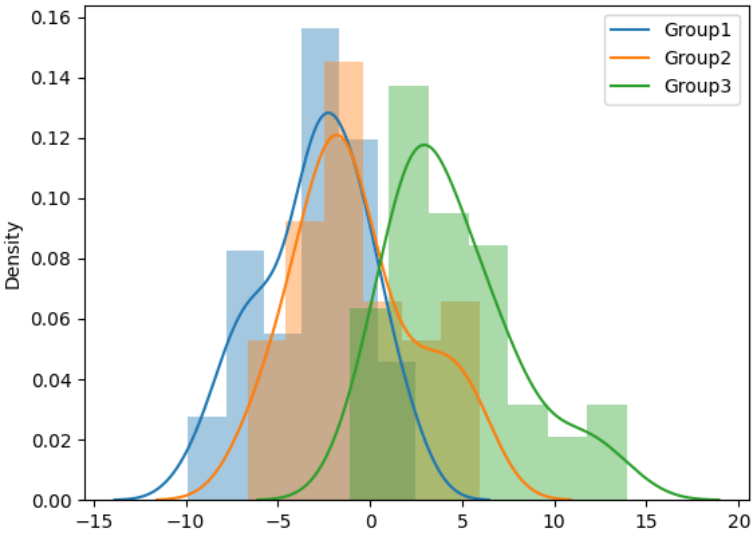

Figure 4.6: *Regression - MMSE prediction - Density plot of residuals per class on test set - Italy*
Group 3 corresponds to the plot of residuals in predicting subjects with MMSE<20, Group 2 to residuals obtained when predicting scores of the subjects with 20≤MMSE≤26, and Group 1 to the distribution of residuals when predicting subjects with MMSE>26

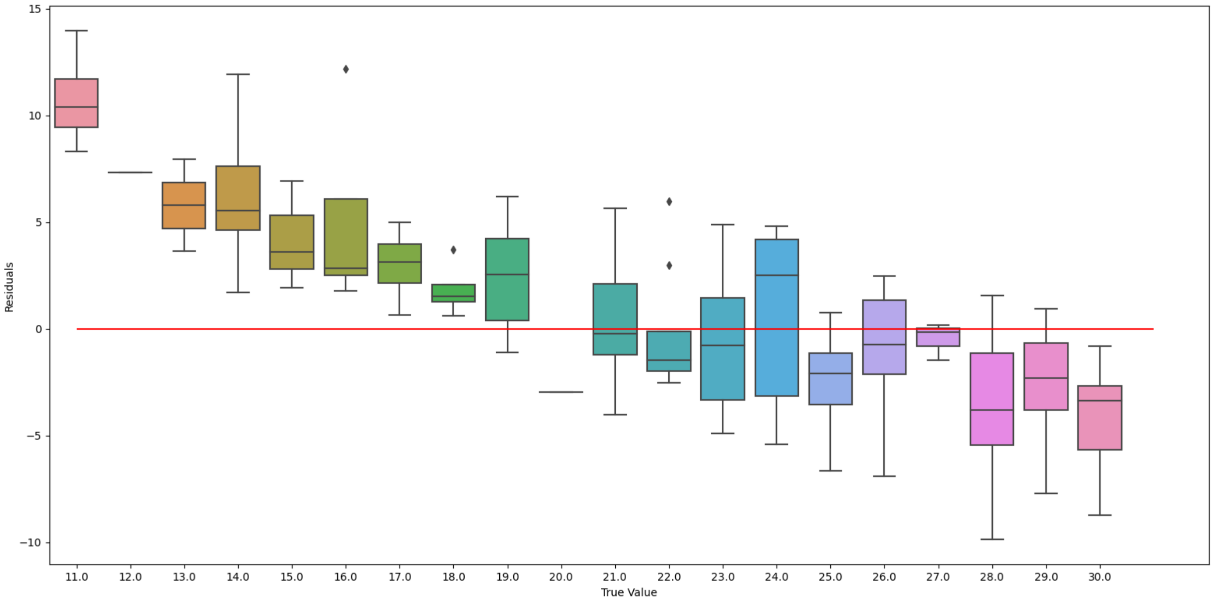

Figure 4.7: *Regression - MMSE prediction - Box-plot of residuals on test set - Italy*

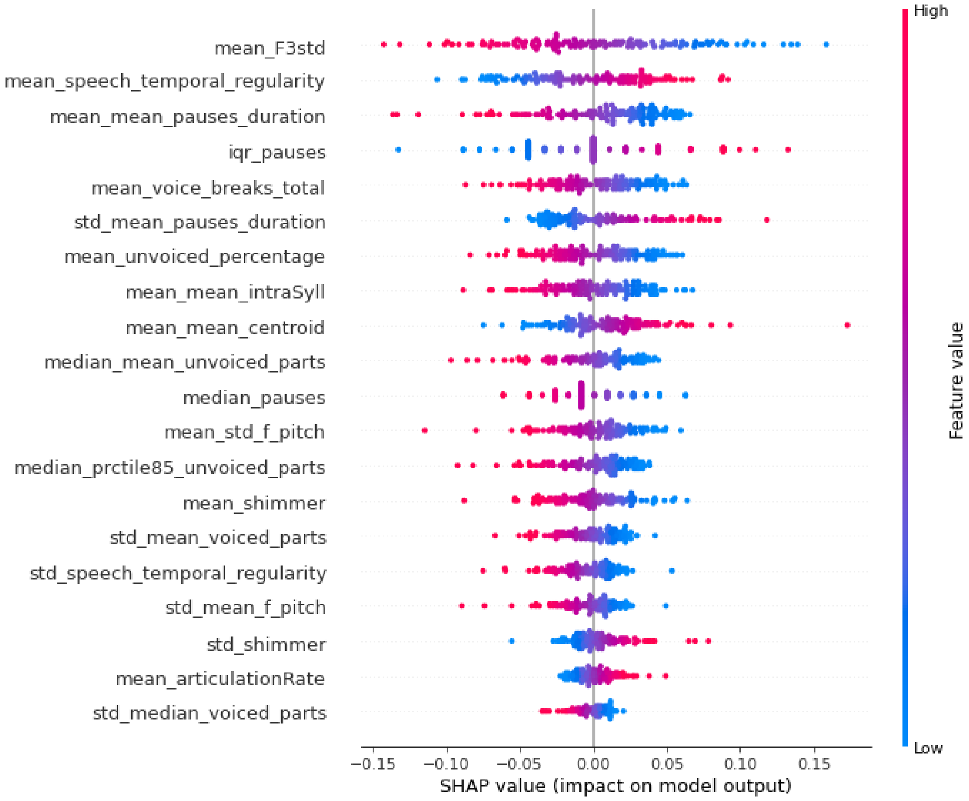

Figure 4.8: *Regression - MMSE prediction - Feature ranking - Italy*

As for binary classification, in Figure 4.8 it can be seen that the most informant features are the 3^{rd} formant and speech temporal regularity, as well as information about the number of pauses and their mean duration. In particular, as for the classification, lower values of 3^{rd} formant yield higher MMSE scores, hence no cognitive impairment, whereas lower values of temporal regularity suggest lower scores, therefore cognitive impairment.

4.2. Spanish Dataset

4.2.1. Binary classification

The performance of binary classification models in dividing Spanish-speaking subjects between group 1 and group 2 is shown in 4.9 and 4.10.

The models were applied on datasets of features computed at 5-10 and 15s and then on the dataset with the features at the different time scales altogether. The comparison was carried out in terms of F1-score on the validation sets. Accuracy, recall, and F1-score are comparable to those obtained in binary classification with the Italian dataset. CatBoost is still the classifier that overall performs better, increasing performances by 10% with respect to Support Vector Machines and Logistic Regression. Indeed, considering the F1-score, the best model for the Spanish dataset is CatBoost with features computed on 15s lengths, which achieves an F1-score of 76% and an accuracy of 77%.

Table 4.9: *Classification - Binary settings - Only Acoustic Features - Spain*

	Time segment	**Accuracy**	**Recall**	**F1-score**
CATBOOST	5	0,76 ($\pm 0,04$)	**0,76** ($\pm 0,05$)	0,74 ($\pm 0,03$)
	10	0,76 ($\pm 0,02$)	0,72 ($\pm 0,03$)	0,74 ($\pm 0,03$)
	15	**0,77** ($\pm 0,03$)	0,74 ($\pm 0,05$)	**0,76** ($\pm 0,03$)
	5-10-15	0,75 ($\pm 0,03$)	0,69 ($\pm 0,05$)	0,72 ($\pm 0,03$)
SVM	5	0,63 ($\pm 0,04$)	0,62 ($\pm 0,06$)	0,62 ($\pm 0,04$)
	10	0,64 ($\pm 0,03$)	0,60 ($\pm 0,05$)	0,62 ($\pm 0,04$)
	15	0,58 ($\pm 0,04$)	0,54 ($\pm 0,05$)	0,55 ($\pm 0,06$)
	5-10-15	0,61 ($\pm 0,02$)	0,63 ($\pm 0,04$)	0,60 ($\pm 0,05$)
LR	5	0,65 ($\pm 0,02$)	0,67 ($\pm 0,04$)	0,66 ($\pm 0,04$)
	10	0,64 ($\pm 0,03$)	0,60 ($\pm 0,04$)	0,62 ($\pm 0,03$)
	15	0,62 ($\pm 0,04$)	0,60 ($\pm 0,05$)	0,60 ($\pm 0,05$)
	5-10-15	0,63 ($\pm 0,01$)	0,61 ($\pm 0,03$)	0,61 ($\pm 0,03$)

Table 4.10: *Classification - Binary settings - Acoustic and Demographic Features - Spain*

	Time segment	**Accuracy**	**Recall**	**F1-score**
CATBOOST	5	0,76 $(\pm 0,03)$	**0,74** $(\pm 0,04)$	**0,75** $(\pm 0,04)$
	10	**0,76** $(\pm 0,02)$	0,72 $(\pm 0,05)$	0,74 $(\pm 0,04)$
	15	0,75 $(\pm 0,03)$	0,70 $(\pm 0,04)$	0,72 $(\pm 0,03)$
	5-10-15	0,75 $(\pm 0,03)$	0,70 $(\pm 0,06)$	0,73 $(\pm 0,03)$
SVM	5	0,63 $(\pm 0,03)$	0,66 $(\pm 0,05)$	0,64 $(\pm 0,05)$
	10	0,63 $(\pm 0,02)$	0,64 $(\pm 0,03)$	0,62 $(\pm 0,03)$
	15	0,62 $(\pm 0,04)$	0,62 $(\pm 0,05)$	0,61 $(\pm 0,05)$
	5-10-15	0,61 $(\pm 0,04)$	0,62 $(\pm 0,04)$	0,61 $(\pm 0,05)$
LR	5	0,65 $(\pm 0,03)$	0,68 $(\pm 0,05)$	0,66 $(\pm 0,03)$
	10	0,66 $(\pm 0,03)$	0,63 $(\pm 0,04)$	0,64 $(\pm 0,04)$
	15	0,65 $(\pm 0,03)$	0,65 $(\pm 0,03)$	0,65 $(\pm 0,03)$
	5-10-15	0,62 $(\pm 0,02)$	0,65 $(\pm 0,04)$	0,60 $(\pm 0,03)$

Table 4.11: *Classification - Binary settings - Performance on test set - Spain*

Time segment	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
15	CATBOOST	0,69	0,72	0,70	0,69

In Table 4.11, all the indicators of the performance on the test set of the best model are shown.

On the test set performances worsen slightly achieving an accuracy of 69% and an F1-score of 69% as well, thus slightly overfitting, anyway by looking at the confusion matrix on the test set in Figure 4.9, it can be seen that the algorithm still performs well, succeeding in the discrimination between the two classes.

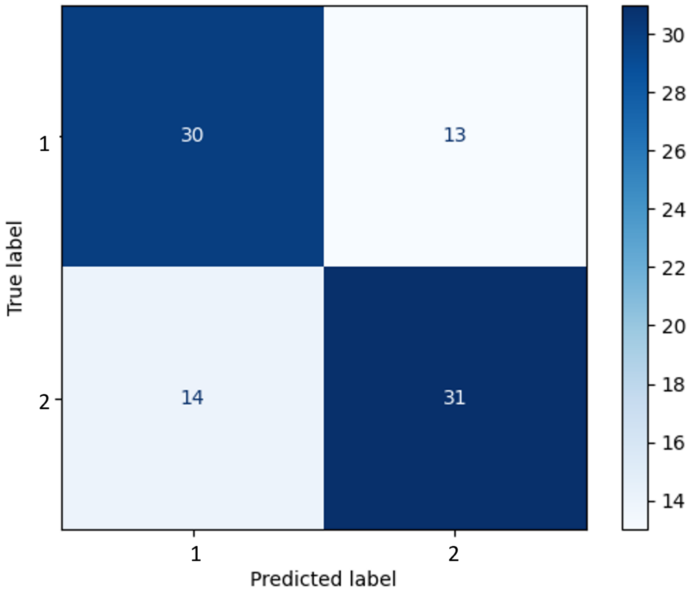

Figure 4.9: *Classification - Binary settings - Confusion Matrix on test set - Spain*
Class 1 represents Group 1 subjects (MMSE>26) and Class 2 represents Group 2 subjects (20≤MMSE≤26).

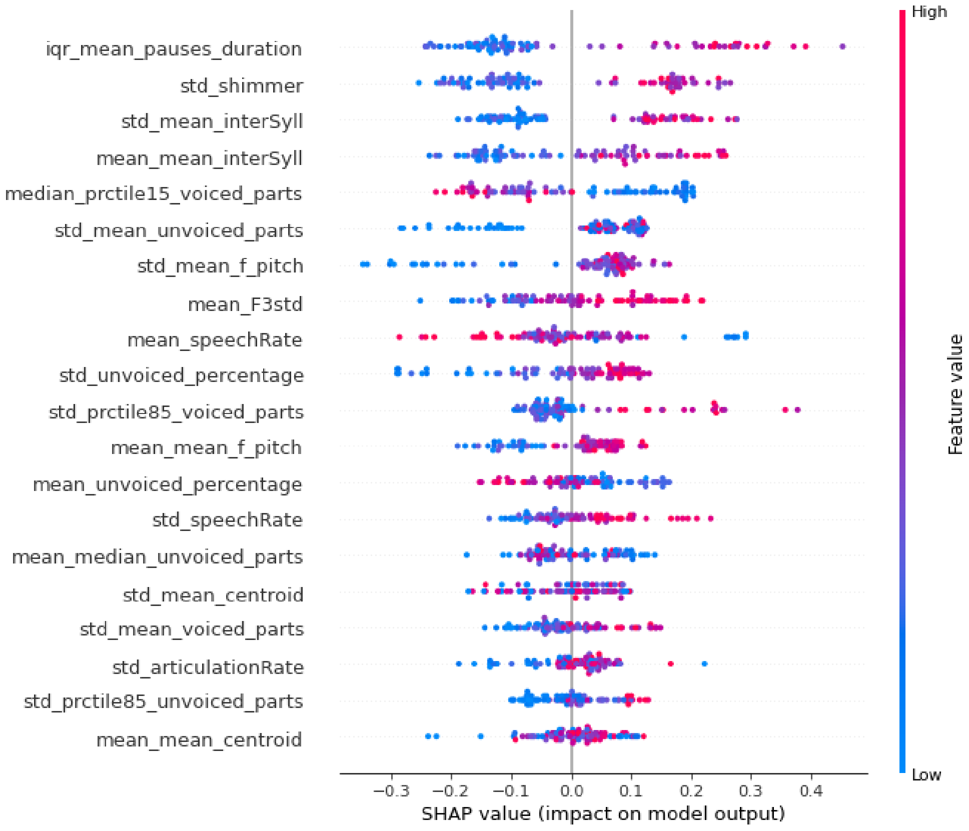

Figure 4.10: *Classification - Binary settings - Feature ranking - Spain*

In Figure 4.10, it can be seen that for the Spanish dataset, the most significant features are different than the ones from the Italian dataset, highlighting indeed there is high variability in acoustic features among languages. In this case, the most discriminating ones are the variation of the duration of pauses, the duration between syllables, and the variation of shimmer (changes in amplitude of vibration of the vocal fold). In particular, low values of these features suggest healthy subjects, whereas higher values are noticed in subjects with mild cognitive impairment.

4.2.2. Multiclass classification

The performance of multiclass classification models in discriminating Spanish-speaking subjects among the three groups is shown in Tables 4.12 and 4.13. The models were applied on datasets of features computed at 5-10 and 15s and then on the dataset with the features at the different time scales altogether. The choice of the best model was carried out in terms of F1-score on the validation sets and when the performances were the same, the simplest model was chosen as the best one. CatBoost overall has higher accuracy, recall, and F1-score with respect to the other two models. In Table 4.13, the performances of the dataset with the addition of age, sex, and years of education do not improve much. Considering only the datasets with acoustic features, in an aim of longitudinal monitoring, with an accuracy of 64% and F1-score of 62%, the best model is CatBoost with features computed on 15s segments.

Table 4.12: *Classification - Multiclass settings - Only Acoustic Features - Spain*

	Time segment	Accuracy	Recall	F1-score
CATBOOST	5	0,64 (±0,02)	**0,64** (±0,02)	0,62 (±0,02)
	10	0,62 (±0,03)	0,63 (±0,03)	0,61 (±0,03)
	15	0,64 (±0,03)	**0,64** (±0,02)	**0,63** (±0,03)
	5-10-15	**0,65** (±0,03)	0,64 (±0,03)	0,62 (±0,03)
SVM	5	0,51 (±0,03)	0,51 (±0,02)	0,50 (±0,03)
	10	0,52 (±0,03)	0,52 (±0,03)	0,50 (±0,03)
	15	0,50 (±0,03)	0,51 (±0,03)	0,49 (±0,03)
	5-10-15	0,51 (±0,02)	0,51 (±0,02)	0,49 (±0,05)
LR	5	0,54 (±0,02)	0,54 (±0,02)	0,54 (±0,02)
	10	0,53 (±0,02)	0,53 (±0,02)	0,51 (±0,02)
	15	0,52 (±0,03)	0,52 (±0,03)	0,50 (±0,02)
	5-10-15	0,51 (±0,03)	0,51 (±0,03)	0,50 (±0,05)

Table 4.13: *Classification - Multiclass settings - Acoustic Features with Demographic Information - Spain*

	Time segment	Accuracy	Recall	F1-score
CATBOOST	5	**0,65**(±0,03)	0,65 (±0,03)	**0,64** (±0,03)
	10	0,64 (±0,03)	0,64 (±0,03)	0,63 (±0,03)
	15	0,64 (±0,03)	0,64 (±0,03)	0,63 (±0,03)
	5-10-15	0,65 (±0,03)	**0,66** (±0,03)	0,64 (±0,03)
SVM	5	0,53 (±0,02)	0,54 (±0,02)	0,52 (±0,03)
	10	0,52 (±0,03)	0,52 (±0,03)	0,51 (±0,04)
	15	0,53 (±0,02)	0,53 (±0,02)	0,52 (±0,03)
	5-10-15	0,51 (±0,02)	0,51 (±0,02)	0,49 (±0,05)
LR	5	0,52 (±0,03)	0,57 (±0,03)	0,57 (±0,03)
	10	0,54 (±0,02)	0,54 (±0,02)	0,53 (±0,02)
	15	0,55 (±0,03)	0,55 (±0,03)	0,54(±0,02)
	5-10-15	0,54 (±0,02)	0,54 (±0,02)	0,52 (±0,04)

Table 4.14: *Classification - Multiclass settings - Performance on test set - Spain*

Time segment	Model	Accuracy	Precision	Recall	F1-score
15	CATBOOST	0,55	0,56	0,54	0,53

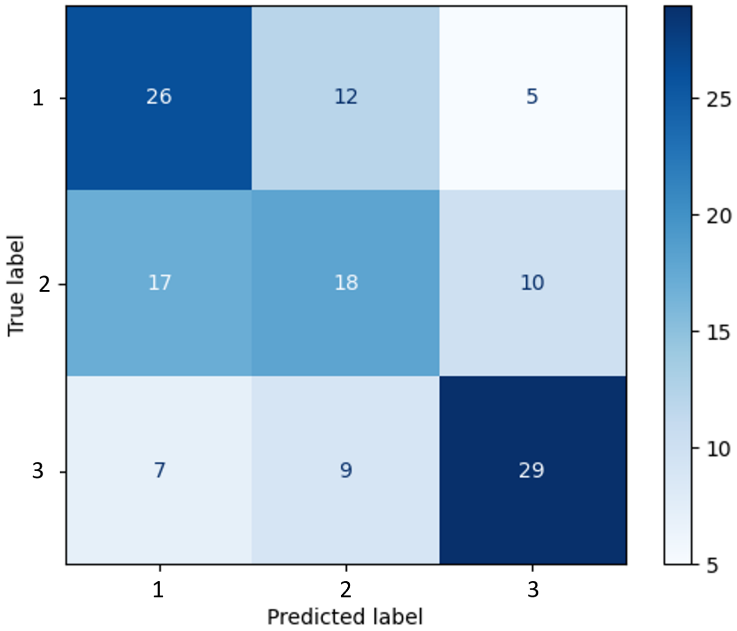

Figure 4.11: *Classification - multiclass settings - Confursion matrix of test set - Spain*
Class 1 represents Group 1 subjects (MMSE>26), Class 2 represents Group 2 subjects (20≤MMSE≤26) and Class 3 represents Group 3 subjects (MMSE>26).

In Table 4.14, all the indicators of the performance on the test set of the best model are shown. On the test set, the model yields an accuracy of 55%, thus, unfortunately, showing overfitting. By looking at the confusion matrix in Figure 4.11, it can be seen that the model mainly struggles to distinguish between group 1 (healthy subjects, class 1 in the figure) and group 2 subjects (mild cognitive impairment subjects, class 2 in the figure).

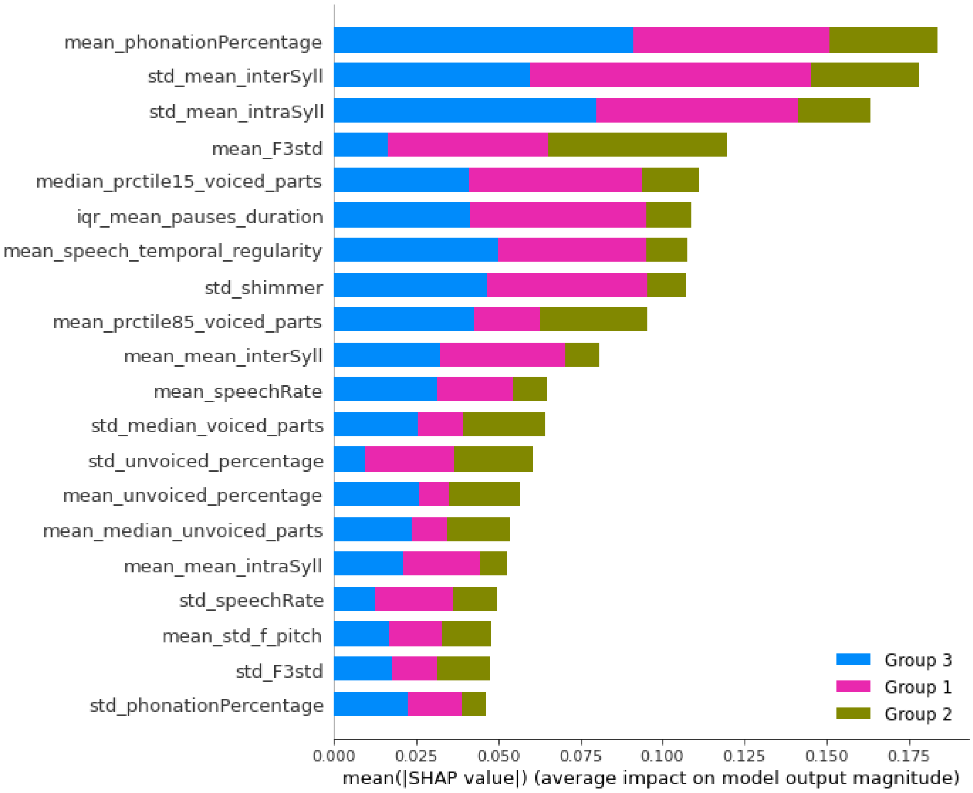

Figure 4.12: *Classification - multiclass settings - Feature ranking - Spain*

The feature ranking in Figure 4.12 shows that the most important features are the duration of syllables and the time interval between them - as for the binary classification - as well as the mean percentage of syllables among the audio segments, i.e. the phonation percentage.

4.2.3. Regression analysis

The performance of regression models for predicting MMSE scores among the Spanish-speaking participants is shown in Table 4.15. Dataset 1 corresponds to the dataset containing only acoustic features, whereas Dataset 2 is the one with the addition of age, sex, and years of education. The models were applied on datasets of features computed at 5-10 and 15s and then on the dataset with the features at the different time scales altogether. The comparison was carried out in terms of minimization of mean absolute error on the validation sets and when the performances were the same, the simplest model was chosen as the best one. Overall, CatBoost and Ridge Regression (in Table 4.15 referenced as LR, since it is a Linear Regression) performed better than Support Vector Machines. Moreover, the addition of the demographic features in Dataset 2 shows an improvement in the mean absolute error for all three regressors. Anyway, considering only the performances of the acoustic features alone, the best model is CatBoost with features computed on 5s segments, which achieves a mean absolute error of 3,57 on the validation set.

Table 4.15: *Regression - MMSE prediction - Spain*

	Time segment	**CATBOOST**	**SVR**	**LR**
Dataset 1	5	**3,57**	4,19	3,78
	10	3,67	4,20	3,76
	15	3,70	4,19	3,79
	5-10-15	3,64	4,51	3,97
Dataset 2	5	3,54	3,97	3,57
	10	3,59	3,93	3,62
	15	3,58	3,78	3,58
	5-10-15	3,55	4,03	3,68

Mean absolute error in predicting MMSE score (range 0-30)

Table 4.16: *Regression - MMSE prediction - Performance on test set - Spain*

Time segment	**Model**	**MAE**	**RMSE**
5	CATBOOST	3,70	4,50

In Table 4.16, mean absolute error and root mean squared error obtained with CatBoostRegressor on the test set are shown, respectively 3,70 and 4,50.

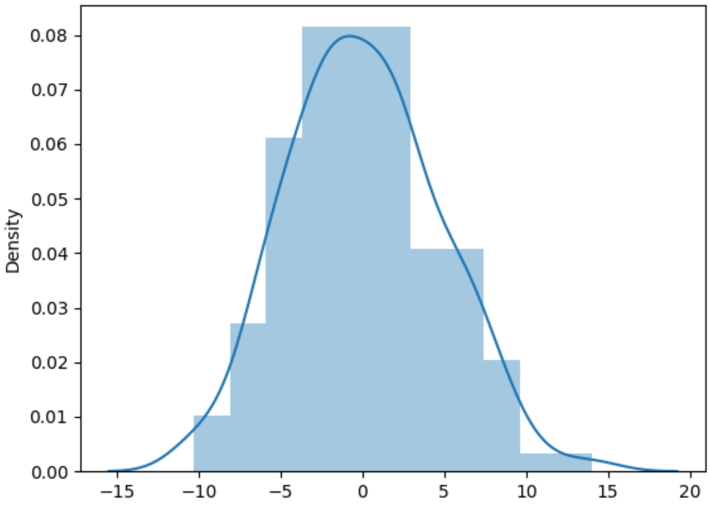

Figure 4.13: *Regression - MMSE prediction - Density plot of residuals on test set - Spain*

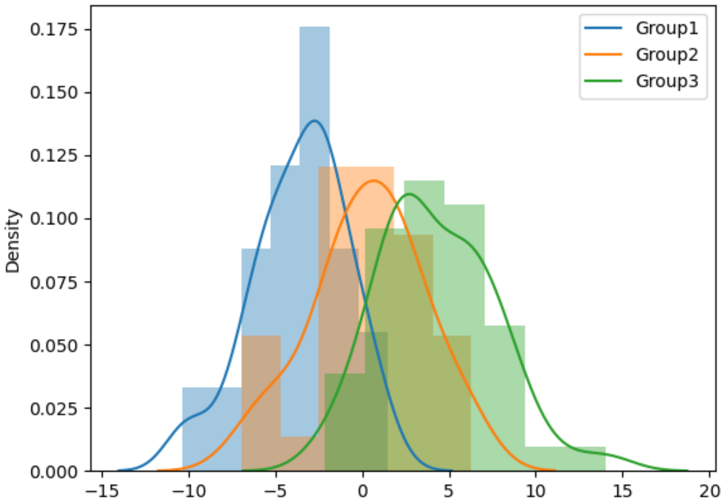

Figure 4.14: *Regression - MMSE prediction - Density plot of residuals on test set - Spain*
Group 3 corresponds to the plot of residuals in predicting subjects with MMSE<20, Group 2 to residuals obtained when predicting scores of the subjects with 20≤MMSE≤26, and Group 1 to the distribution of residuals when predicting subjects with MMSE>26

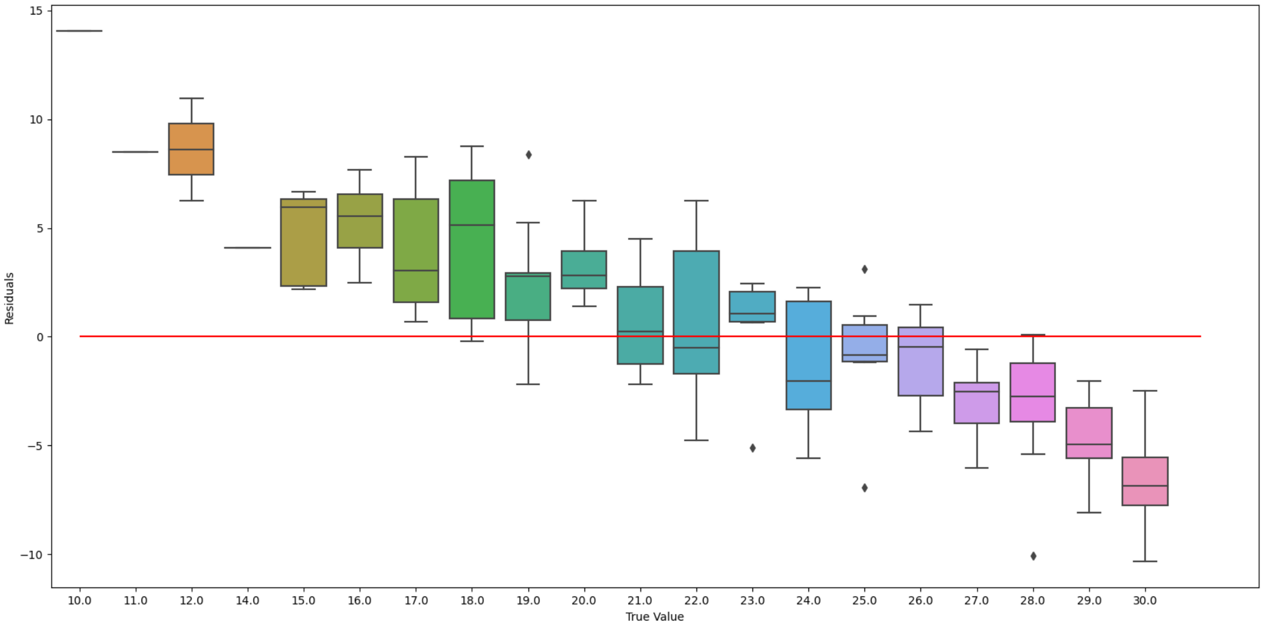

Figure 4.15: *Regression - MMSE prediction - Box-plot of residuals on test set - Spain*

By looking at the density plot of the residuals in Figure 4.13, overall residuals have a normal distribution, but more in-depth, in the density plot per class in Figure 4.13 it can be seen that the model tends to overestimate the lower scores and underestimates the higher ones. This trend is confirmed by the boxplots of the residuals per score in Figure 4.15, as well.

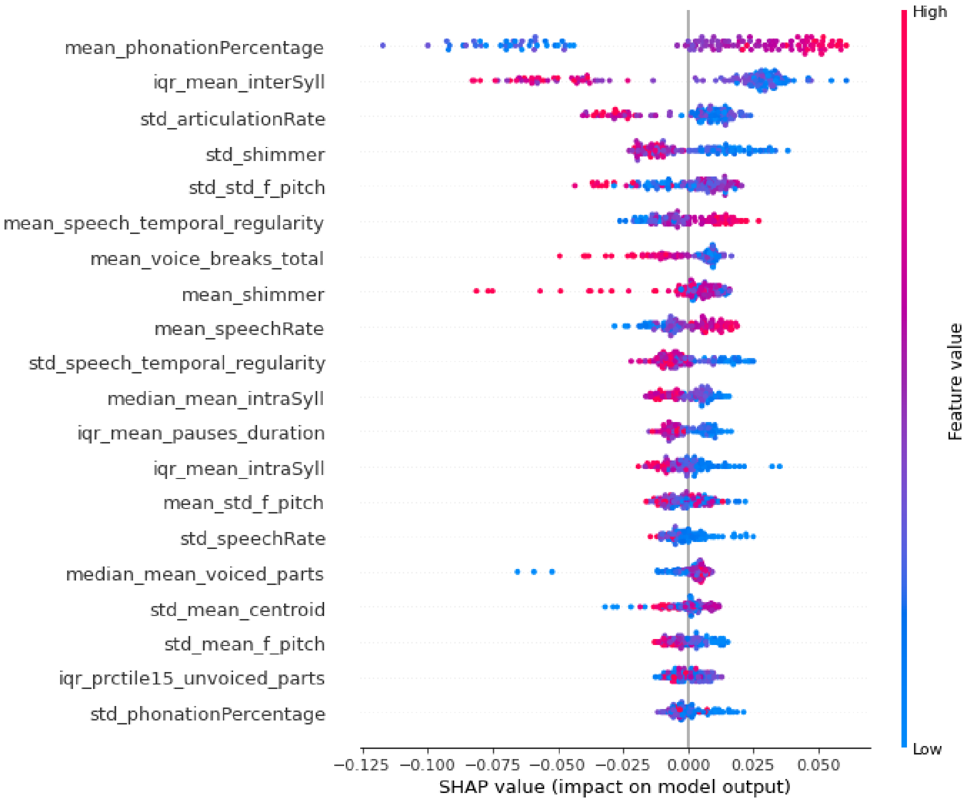

Figure 4.16: *Regression - MMSE prediction - Feature ranking - Spain*

Feature ranking in Figure 4.16 corroborates the previously shown feature rankings of the other two tasks (binary and multiclass classification) since it can be seen that the most important features are still the phonation percentage, shimmer, and the duration between the syllables. Higher values of shimmer tend to predict lower scores, whereas higher values of phonation percentage suggest higher scores, results that are consistent with the ranking seen in Figure 4.10.

4.3. Pitt Corpus

4.3.1. Binary classification

The performance of binary classification models in dividing English-speaking subjects between group 1 and group 2 is shown in Tables 4.17 and 4.18.

Table 4.17: *Classification - Binary settings - Only Acoustic Features - Pitt Corpus*

	Time segment	Accuracy	Recall	F1-score
CATBOOST	5	**0,73** (±0,02)	0,66 (±0,04)	**0,64** (±0,04)
	10	0,72 (±0,01)	**0,70** (±0,06)	0,63 (±0,04)
	15	0,72 (±0,02)	0,68 (±0,03)	0,63 (±0,04)
	5-10-15	0,73 (±0,02)	0,68 (±0,04)	0,63 (±0,03)
SVM	5	0,64 (±0,01)	0,62 (±0,07)	0,55 (±0,04)
	10	0,66 (±0,01)	0,49 (±0,08)	0,58 (±0,01)
	15	0,68 (±0,01)	0,62 (±0,05)	0,55 (±0,06)
	5-10-15	0,66 (±0,01)	0,46 (±0,05)	0,41 (±0,05)
LR	5	0,65 (±0,01)	0,51 (±0,04)	0,54 (±0,06)
	10	0,62 (±0,06)	0,45 (±0,04)	0,42 (±0,05)
	15	0,55 (±0,03)	0,55 (±0,03)	0,54(±0,02)
	5-10-15	0,67 (±0,01)	0,41 (±0,03)	0,45 (±0,04)

Table 4.18: *Classification - Binary settings - Acoustic Features with Demographic Information - Pitt Corpus*

	Time segment	Accuracy	Recall	F1-score
CATBOOST	5	0,76 (±0,03)	**0,74** (±0,04)	**0,75** (±0,04)
	10	**0,76** (±0,02)	0,72 (±0,05)	0,74 (±0,04)
	15	**0,76** (±0,02)	0,68 (±0,03)	0,67 (±0,05)
	5-10-15	0,76 (±0,02)	0,71 (±0,04)	0,68 (±0,05)
SVM	5	0,63 (±0,03)	0,66 (±0,05)	0,64 (±0,05)
	10	0,63 (±0,02)	0,64 (±0,03)	0,62 (±0,03)
	15	0,71 (±0,01)	0,40 (±0,04)	0,47 (±0,05)
	5-10-15	0,70 (±0,04)	0,43 (±0,03)	0,59 (±0,05)
LR	5	0,65 (±0,03)	0,68 (±0,05)	0,66 (±0,03)
	10	0,66 (±0,03)	0,63 (±0,04)	0,64 (±0,04)
	15	0,71 (±0,01)	0,67 (±0,05)	0,46 (±0,05)
	5-10-15	0,71 (±0,01)	0,41 (±0,01)	0,49 (±0,03)

Table 4.19: *Classification - Binary settings - Performance on test set - Pitt Corpus*

Time segment	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
5	CATBOOST	0,67	0,55	0,52	0,52

The models were applied on datasets of features computed at 5-10 and 15s and then on the dataset with the features at the different time scales altogether. The comparison was carried out in terms of F1-score on the validation sets and when the performances were the same, the simplest model was chosen as the best one. CatBoost has slightly better accuracy with respect to Support Vector and Logistic Regression, whereas in recall and F1-score it has much better scores. In the end, the model with the best performance on the validation set is indeed CatBoost with features computed on segments of 5s, providing an accuracy of 73% and F1-score of 64%.

Overall the addition of demographic information in Table 4.18 mainly improves accuracy, recall, and F1-score, although, in Support Vector Machines with features computed at 15s, it can be seen that recall and F1-score worsen, as if the addition of new features confuses the algorithm.

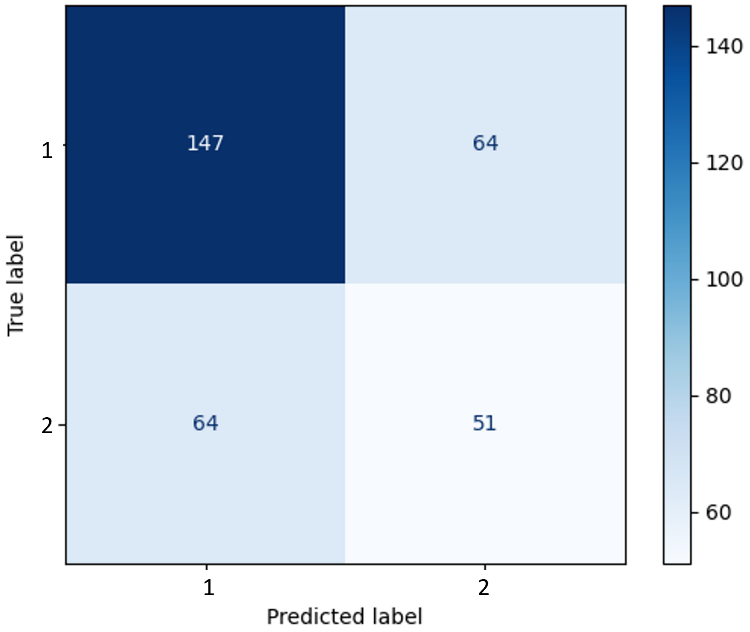

Figure 4.17: *Classification - Binary settings - Confusion Matrix on test set - Pitt Corpus*
Class 1 represents Group 1 subjects (MMSE>26) and Class 2 represents Group 2 subjects (20≤MMSE≤26).

In Table 4.19, accuracy, precision, recall, and F1-score of the performance on the test set of the best model are shown. On the test set accuracy, recall, and F1-score lower, to 67%, 52%, and 52% respectively, struggling to predict the subjects with mild cognitive impairment, as can be seen in Figure 4.17. This phenomenon may be explained by the strong imbalance between the 2 classes (211 subjects in group 1 versus 115 subjects in group 2).

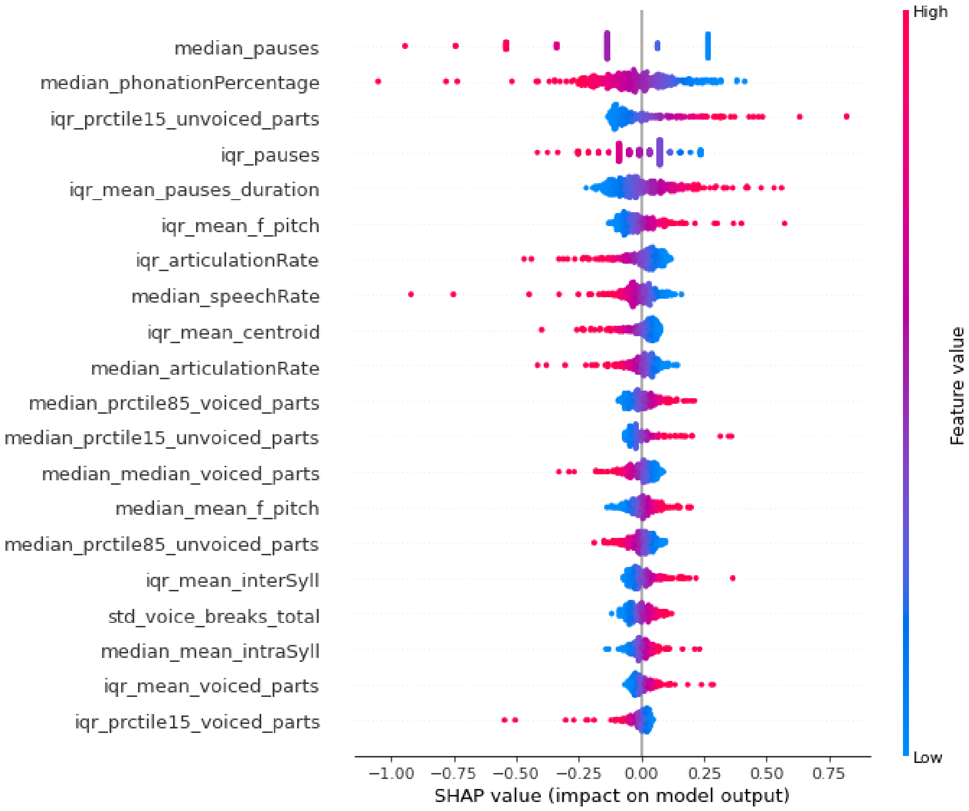

Figure 4.18: *Classification - Binary settings - Feature ranking - Pitt Corpus*

Feature ranking in Figure 4.18 shows that the most important features are those related to the number of pauses, and those related to the syllables, i.e. the phonation percentage, articulation rate, and speech rate. In particular, for these last three features, higher values (hence a higher speed in speech) encode healthy subjects, whereas lower speed suggests a mild impairment.

4.3.2. Multiclass classification

The performance of multiclass classification models in discriminating English-speaking subjects among the three groups is shown in Tables 4.20 and 4.21. The first one corresponds to the dataset with only acoustic features, whereas the latter considers the dataset with the addition of demographic information. The models were applied on datasets of features computed at 5-10 and 15s and then on the dataset with the features at the different time scales altogether. The comparison was carried out in terms of F1-score on the validation sets and when the performances were the same, the simplest model was chosen as the best one. Moreover, the best model was chosen by considering the performances obtained with acoustic features only, to evaluate the capability of the models to be applied for longitudinal monitoring. The addition of demographic information does not improve much the prediction between the three groups, confirming that it is possible to obtain good classification performances with acoustic features only, hence good performance on longitudinal monitoring. As in the binary classification, the best model was CatBoost which reached an accuracy of 64% and an F1-score of 62% on the validation set, with the dataset with features computed on segments of 5s.

Table 4.20: *Classification - Multiclass settings - Only Acoustic Features - Pitt Corpus*

	Time segment	Accuracy	Recall	F1-score
CATBOOST	5	**0,64** ($\pm 0,02$)	**0,64** ($\pm 0,02$)	0,62 ($\pm 0,02$)
	10	0,62 ($\pm 0,03$)	0,63 ($\pm 0,03$)	0,61 ($\pm 0,03$)
	15	0,64 ($\pm 0,03$)	**0,64** ($\pm 0,02$)	**0,63** ($\pm 0,03$)
	5-10-15	0,65 ($\pm 0,03$)	0,64 ($\pm 0,03$)	0,62 ($\pm 0,03$)
SVM	5	0,51 ($\pm 0,03$)	0,51 ($\pm 0,02$)	0,50 ($\pm 0,03$)
	10	0,52 ($\pm 0,03$)	0,52 ($\pm 0,03$)	0,50 ($\pm 0,03$)
	15	0,50 ($\pm 0,03$)	0,51 ($\pm 0,03$)	0,49 ($\pm 0,03$)
	5-10-15	0,49 ($\pm 0,04$)	0,49 ($\pm 0,04$)	0,48 ($\pm 0,05$)
LR	5	0,54 ($\pm 0,02$)	0,54 ($\pm 0,02$)	0,54 ($\pm 0,02$)
	10	0,53 ($\pm 0,02$)	0,53 ($\pm 0,02$)	0,51 ($\pm 0,02$)
	15	0,52 ($\pm 0,03$)	0,52 ($\pm 0,03$)	0,50 ($\pm 0,02$)
	5-10-15	0,51 ($\pm 0,03$)	0,51 ($\pm 0,03$)	0,50 ($\pm 0,05$)

Table 4.21: *Classification - Multiclass settings - Acoustic Features and Demographic Information - Pitt Corpus*

	Time segment	Accuracy	Recall	F1-score
CATBOOST	5	**0,65**(±0,03)	**0,65** (±0,03)	**0,64** (±0,03)
	10	0,64 (±0,03)	0,64 (±0,03)	0,63 (±0,03)
	15	0,64 (±0,03)	0,64 (±0,03)	0,63 (±0,03)
	5-10-15	0,64 (±0,03)	0,64 (±0,03)	0,63 (±0,03)
SVM	5	0,53 (±0,02)	0,54 (±0,02)	0,52 (±0,03)
	10	0,52 (±0,03)	0,52 (±0,03)	0,51 (±0,04)
	15	0,53 (±0,02)	0,53 (±0,02)	0,52 (±0,03)
	5-10-15	0,51 (±0,02)	0,51 (±0,02)	0,49 (±0,05)
LR	5	0,57 (±0,03)	0,57 (±0,03)	0,56 (±0,03)
	10	0,54 (±0,02)	0,54 (±0,02)	0,53 (±0,02)
	15	0,55 (±0,03)	0,55 (±0,03)	0,54(±0,02)
	5-10-15	0,54 (±0,02)	0,54 (±0,02)	0,52 (±0,04)

Table 4.22: *Classification - Multiclass settings - Performance on test set - Pitt Corpus*

Time segment	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
5	CATBOOST	0,49	0,46	0,46	0,45

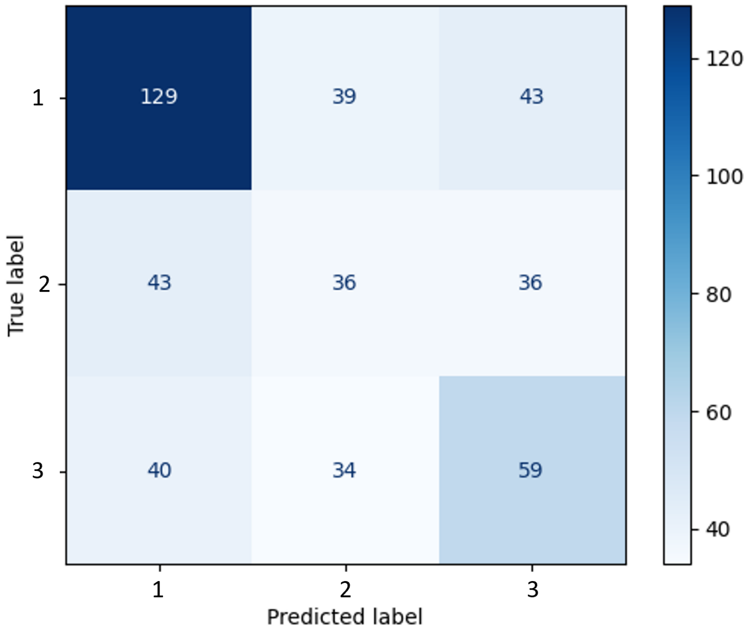

Figure 4.19: *Classification - multiclass settings - Confursion matrix of test set - Spain*
Class 1 represents Group 1 subjects (MMSE>26) and Class 2 represents Group 2 subjects (20≤MMSE≤26) and Class 3 represents Group 3 patients (MMSE<20).

In Table 4.22, all the indicators - namely accuracy, precision, recall, and F1-score - of the performance on the test set of the best model are shown, highlighting that on the test set its metrics worsen, reaching an accuracy of 49%. This value is still acceptable considering that the prediction occurs among three classes, therefore a score indicating randomness of prediction and inability to correctly classify subjects would be around 33%. From the confusion matrix, it can still be noticed that the algorithm struggles to correctly predict mildly impaired subjects, but overall it correctly discriminates between healthy subjects and severely impaired ones.

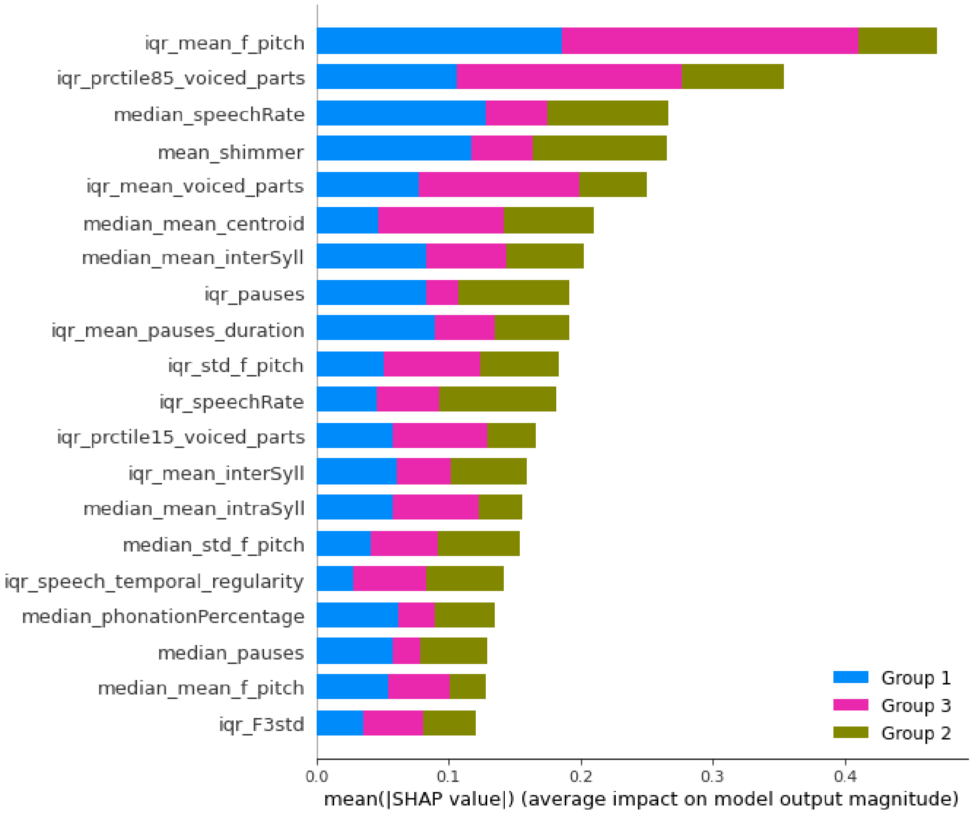

Figure 4.20: *Classification - multiclass settings - Feature ranking - Pitt Corpus*

From the feature ranking in Figure 4.20, the variation of pitch throughout the audio recording plays a paramount role in discriminating between healthy patients and severely impaired ones. Shimmer and variation of speech rate, instead, mainly contribute to the prediction of the subjects with mild cognitive decline.

4.3.3. Regression analysis

The mean absolute error obtained applying the regression models for predicting MMSE score among the English-speaking participants is reported in Table 4.23. The models were applied on datasets of features computed at 5-10 and 15s and then on the dataset with the features at the different time scales considered altogether. Dataset 1 corresponds to the dataset with only acoustic features, whereas dataset 2 corresponds to the one with the addition of information related to age, sex, and years of education. The comparison was carried out in terms of minimization of mean absolute error on the validation sets and when the performances were the same, the simplest model was chosen as the best one. In this case, MAEs are generally slightly higher with respect to those seen for the Italian and Spanish datasets, reaching overall more than 5 with dataset 1, while with dataset 2 the metric improves, proving once again that demographic information help in the evaluation of the MMSE score. Anyway, regarding the dataset with only acoustic features, CatBoost with the dataset that considers the features computed at the different scales altogether slightly improves performances, decreasing the mean absolute error up to 4,93 on the validation set.

Table 4.23: *Regression - MMSE prediction - Pitt Corpus*

	Time segment	**CATBOOST**	**SVR**	**LR**
Dataset 1	5	5,05	5,18	5,21
	10	5,06	5,29	5,29
	15	5,06	5,17	5,12
	5-10-15	**4,93**	5,53	5,18
Dataset 2	5	4,64	4,73	4,72
	10	4,63	4,73	4,71
	15	4,56	4,62	4,64
	5-10-15	4,58	4,97	4,79

Mean absolute error in predicting MMSE score (range 0-30)

Table 4.24: *Regression - MMSE prediction - Performance on test set - Pitt Corpus*

Time segment	**Model**	**MAE**	**RMSE**
5-10-15	CATBOOST	4,27	5,47

In Table 4.16, the mean absolute error and root mean squared error of the performance on the test set of the best model are shown. On the test set, the performance improves, lowering the mean absolute error (MAE) to 4,27. Although the errors may seem high considering the scoring range that has to be predicted - from 0 to 30 - results are perfectly consistent with the performances obtained with this dataset by the other participants of the challenge [38].

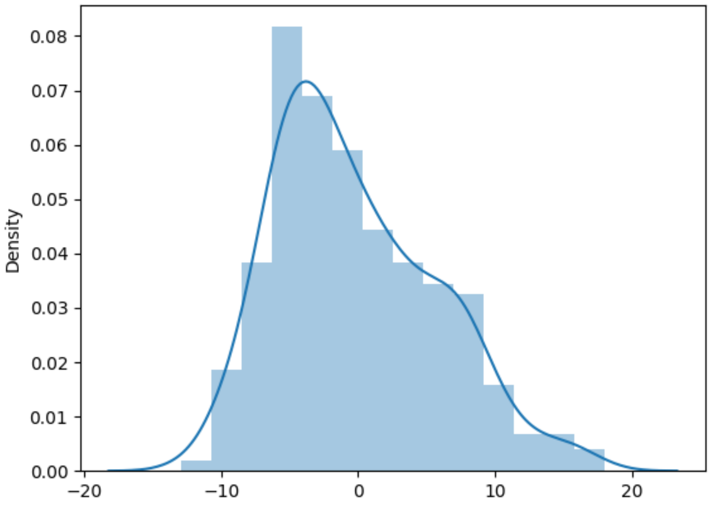

Figure 4.21: *Regression - MMSE prediction - Density plot of residuals on test set - Pitt Corpus*

The overall density plot of the residual in Figure 4.21 does not have exactly a normal distribution, suggesting some bias in the results.

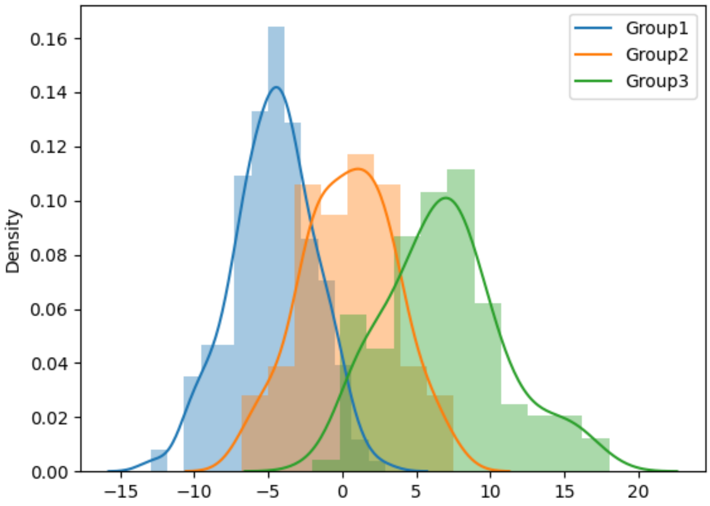

Figure 4.22: *Regression - MMSE prediction - Density plot of residuals on the test set per class - Pitt Corpus*
Group 3 corresponds to the plot of residuals in predicting subjects with MMSE<20, Group 2 to residuals obtained when predicting scores of the subjects with 20≤MMSE≤26, and Group 1 to the distribution of residuals when predicting subjects with MMSE>26

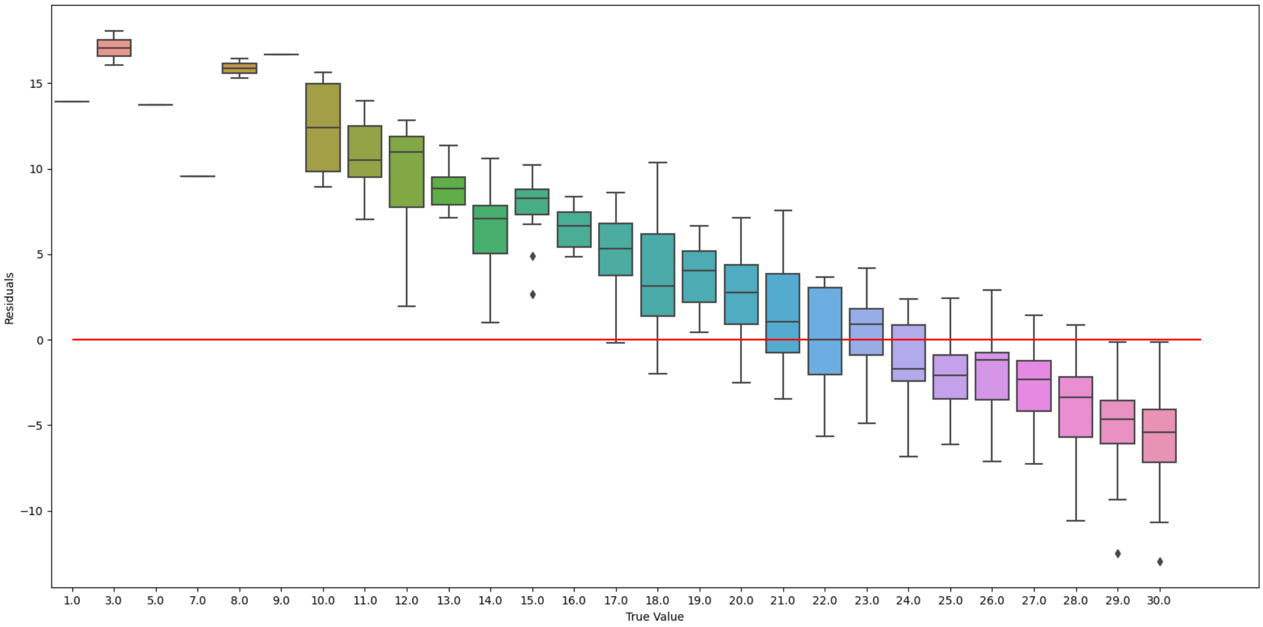

Figure 4.23: *Regression - MMSE prediction - Box-plot of residuals on test set - Pitt Corpus*

Regarding the plot of residuals in Figures 4.23, it can be seen that it still highly overestimates the scores of the class with severe impairment which is confirmed in the density plot per class of the residuals in Figure 4.22, probably due to the higher range of score in this group, and the lower numerosity for each score, but it still needs further investigation.

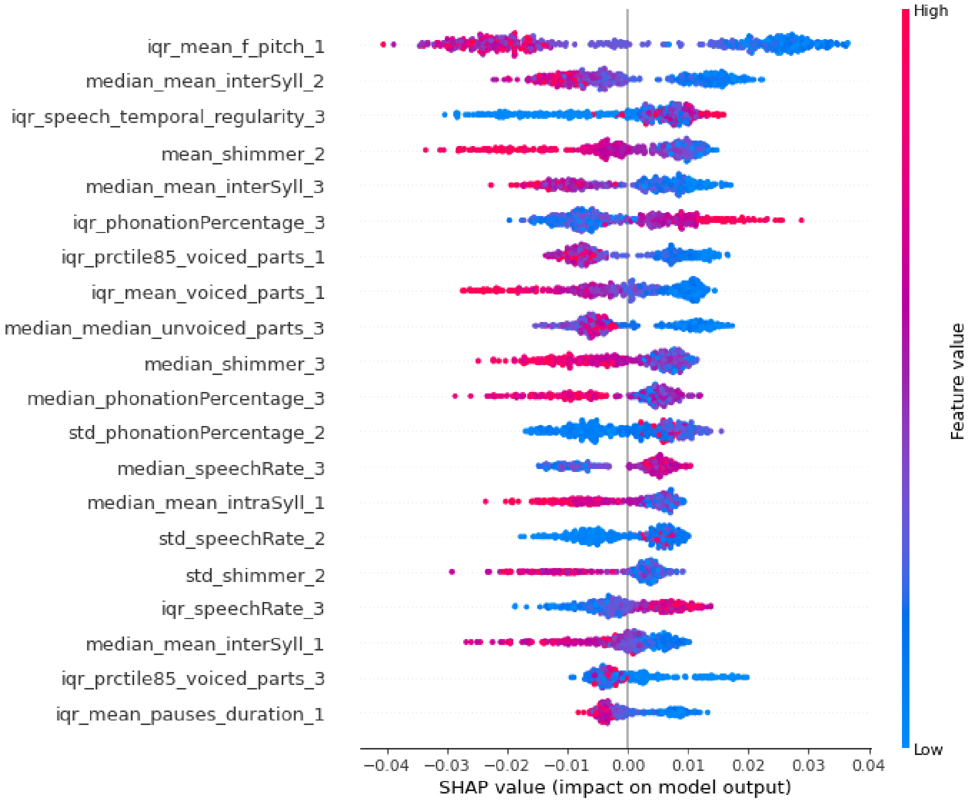

Figure 4.24: *Regression - MMSE prediction - Feature ranking - Pitt Corpus*

The feature ranking in Figure 4.24 confirms that in this case, the model finds it difficult to find a good generalization rule since there is no clear separation between feature values and their contribution to the model. Overall, higher values of pauses between syllables (*median_ mean_ interSyll_ 2*) as well as shimmer computed on the 10s segments seem to be found in subjects with lower scores.

5 | Discussion and Limitations

In the context of the worldwide increase in life expectancy, and therefore cognitive decline in elderly people, this thesis presents the analysis of acoustic features extracted from speech of different languages. Three different datasets have been employed in order to evaluate the ability of the acoustic features derived from spontaneous speech to discriminate between different levels of cognitive decline and normal aging, regardless of the language involved.

Similar results were achieved regardless of the time scale used for the computation of features. This shows that there is no need for long audio recordings and it allows to speed up computational time. Considering the features computed at the different scales altogether does not significantly improve performances, but it should need further investigation. Moreover, the addition of the demographic information of subjects does improve the prediction of cognitive impairment, but not in a significant way, and performances in the detection of cognitive decline are good even leaving out this demographic information, hence promising good results for an application in *longitudinal monitoring*.

Results highlighted that different sets of features are relevant depending on the considered idiom and on the specific task. Overall, an increase in phonation percentage, as well as speech rate in healthy subjects, is noticed in the three datasets. Shimmer is shown to be significant in Spanish-speaking subjects, highlighting that a larger variation throughout the segments is predictive of higher cognitive impairment.

The addition of new features such as *speech temporal regularity* and *mean duration of pauses* highly contribute to the detection of cognitive impairment, mainly in the Italian and Spanish datasets respectively, as can be seen in the feature ranking plots in Figures 4.2, 4.10, and 4.8.
Moreover, speech temporal regularity confirms the trends described in literature since lower values are shown in subjects with higher Mini-Mental State Examination scores, therefore healthy subjects, and higher values (hence higher values of the Mel-Frequency Cepstral Coefficients) are seen in subjects with cognitive decline [42].

Results for the Italian dataset by considering only smaller segments for feature extraction have obtained comparable performances to those in [39] that evaluated instead the whole recordings of more than two minutes. In particular, for binary classification the best model of this work yields on the validation set an accuracy of 77% without the addition of age, sex and years of education, compared to the 72% achieved in [39], whereas for multiclass classification it achieves 64% in accuracy without the demographic information with respect to 56% in [20]. Moreover, in this work, with nested 10-Fold CV, it was possible have an estimate on how the model performs on unseen data, whereas in the aforementioned work the Authors have only validated results with standard 5-fold Cross-Validation, without estimating the model performance on unseen data. With respect to [33], F1-score in binary classification was lower, but, in the case of the current work it was obtained without considering demographic features such as age and years of education, important indicators of cognitive decline, to evaluate the possibility to employ acoustic features for longitudinal monitoring.
Still, the models seem to overfit since performances on unseen data worsen, thus there is the need for further investigation. The problem may be the lack of generalization power of the model, therefore it would be useful to implement a feature selection algorithm to keep only the

most significant features.

Regarding the Spanish dataset, classification was performed on a larger dataset than the one in [30], obtaining slightly worse performances on the binary classification. Still, SHAP analysis in Figure 4.10 confirmed that fluency is an important aspect of the evaluation of cognitive decline from spontaneous speech.

Finally, the prediction of the Mini-Mental State Examination (MMSE) score from acoustic features overall obtained a mean absolute error of about 4 in all the 3 datasets. Although this error seems high with respect to the range of scores 0-30 on the test, it is perfectly in line with the few results found in the literature [15]. Furthermore, it achieves better results than both [35] and [36] with the same starting Pitt Corpus dataset. Yet, regression task seems to yield high mean absolute errors with respect to the prediction of the MMSE. Anyway, it was shown that different rating styles among clinicians that administer the MMSE and variance in test-retest scoring can lead to a within-subject inter- and intra-rater standard deviation of 3.9 to 4.8, with higher variation in low-scoring subgroups of subjects [49], therfore the MAE obtained from the three datasets in the present work is comparable to such variability.

6 | Conclusions and Future Developments

The purpose of this work has been to analyze cognitive decline based on acoustic features to improve early detection in elderly people, by designing new acoustic features and employing machine learning techniques for classification, and regression applied to datasets of different languages. First, the former algorithm for feature extraction was optimized by adding new features and allowing segmentation of the audio recordings. Then, an analysis of the duration of the segments was carried out by evaluating the performance of each model, for both the classification and the regression tasks.

The good performances obtained without considering the demographic characteristics are promising to define an application for longitudinal monitoring of cognitive decline in elderly people, e.g. the development of an application for automatic feature extraction. In particular, the results obtained from the analysis of the duration of segments suggest that it would be feasible to design for example a real-time acoustic feature extraction mobile app. Moreover, instead of pre-defined buffers on which to compute the features, it would be useful to identify and segment speeches into homogeneous regions or to compute them in different ranges.

To avoid the generally seen overfitting of the performances on the test sets, it may be useful to employ existing feature selection algorithms or implement one *ad-hoc*. For example, a recursive feature elimination algorithm could be implemented starting from the SHAP values: at each iteration, the model would compute the Shapley values and the last features in the ranking would be eliminated. In the end, the final set of features would be the one that optimizes the chosen scoring function. Moreover, future work should employ larger sets to validate and test the models and investigate other languages to verify whether the defined features are as a matter of fact idiom-independent. It would be interesting to use deep learning methods not only for the machine learning tasks but for the feature extraction as well. Indeed, other works in this field showed promising results [50].

Further improvements in the recognition of cognitive decline would be to evaluate and predict the emotional state of the subjects since previous studies have widely shown the influence of emotion on acoustic features. In this regard, in Appendix A the main findings in the literature on Speech Emotion Recognition and the assessment of emotions are outlined for completeness.

In conclusion, the results of this work show that the extracted acoustic features from spontaneous speech provide a good discrimination power between healthy subjects and those with signs of cognitive decline, as well as in predicting Mini-Mental Score from voice, regardless of the language spoken. The use of small-time segments allows to compute the features in a faster way, thus this work represents a first step to enabling the implementation of applications for large-scale, real-time monitoring the process of cognitive impairment in everyday activities, without directly involving clinical assessments and visits.

A | Speech Emotion Recognition

A.1. Alterations of acoustic profiles in emotion expression

There is evidence that emotion produces changes in respiration, phonation and articulation [51]. Banse and Scherer in [51] studied how vocal parameters varied depending on the staged emotion.

Fourteen emotions have been considered in this case, including

- Hot and cold anger
- Panic fear and anxiety
- Despair and sadness
- Happiness and contempt
- Disgust and boredom

It has been noticed that for intense emotions such as despair, hot anger, panic fear, fundamental frequency (mean F0) is the highest whereas the lowest for contempt and boredom. The remaining emotions — happiness, anxiety, sadness, disgust and cold anger — are located in the middle range.

A similar pattern is displayed for the energy, since its high correlation with fundamental frequency. The mean energy for sadness, although its F0 is in the middle range, has the lowest mean in energy.

Moreover, Sadness is characterized by a particularly low speech rate. Hot anger and panic fear display an increase in speech rate, whereas despair shows a slight decrease.

A.2. Acoustic features for emotion recognition

A literature review about acoustic features for emotion recognition has been carried out. The most informative acoustic features can be divided in 4 categories:

- Frequency related
- Vocal tract related
- Energy related
- Spectrum related
- Syllable related

A.2.1. Frequency related features

The pitch signal, also known as the glottal waveform, is produced from the vibration of the vocal folds and it contains information about emotion, since it describes the tension of the vocal folds and the subglottal air pressure. The main feature that is considered in this case is the vibration rate of the vocal folds which is called *fundamental frequency of F0* or *pitch frequency*. An example on how pitch affects emotion can be seen for harsh emotions such as anger and disgust, since they are characterized by low velocity, therefore a lower frequency [52].

A.2.2. Energy

The short-term speech energy is usually used for emotion recognition, since it represents the arousal level of emotions.

A feature of this kind is the energy related to certain frequency bands. There are many contradictions in identifying the best frequency band of the power spectrum in order to classify emotions. Anyway, the Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) provide a better representation than the frequency bands since they additionally exploit the human auditory frequency response. These parameters are derived from the short-term Fourier spectrum of the acoustic signal. MFCCs are preferred over Linear Frequency Cepstral Coefficients (LFCCs) since they allow better suppression of insignificant spectral variation in the higher frequency bands.

A.2.3. Vocal tract features

The shape of the vocal tract is modified by the emotional states.The features used to describe the shape of the vocal tract include:

- Formants
- Frequency-related coefficients

The formants are one of the quantitative characteristics of the vocal tract. They represent resonances of vocal tract. In the frequency domain, the location of vocal tract resonances depends on the shape and the physical dimensions of the vocal tract, forming the the overall spectrum, thus the definition of formants. Each formant is characterized by its center frequency and its bandwidth. In this context, subjects during stress or under depression do not articulate voiced sounds with the same effort as in the neutral emotional state [53]. The formants can be used to discriminate the articulated speech from the relaxed one. In the first case, the formant bandwidth is gradual, whereas in the latter the formant bandwidth is narrow with steep sides.

A.2.4. Spectral features

The spectral centroid is a measure used in digital signal processing to characterise a spectrum. It indicates where the center of mass of the spectrum is located. Perceptually, it has a robust connection with the impression of brightness of a sound.

It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights [54].

A.2.5. Syllable-related features

These parameters are linked to speech fluency. One of the main indicators is articulation rate which is defined as the number of syllables produced in a timed speech sample discarding all silent parts from the sample. Moreover, duration of articulation and duration of voiced segments can be used as (inverse) measures of speech rate or tempo.

A.3. Dimensional approach for emotion representation

An alternative way to emotion analysis is the dimensional approach, according to which, emotions can be represented using specific dimensions that stem from psychophysiology [55, 56]. In particular, emotions are mainly represented in a 2D or 3D space where each point corresponds to a separate emotion state [57].

The most universal model seems to be Mehrabian's Pleasure-Arousal-Dominance (PAD) model which measures emotional tendencies and affective states along three dimensions:

- Pleasure (also called valence)
- Arousal

- Dominance

Its continuous nature allows to model intermediate states of emotions that may not have an a-priori label [58].

An example of the 3D space is showed in FigureA.1.

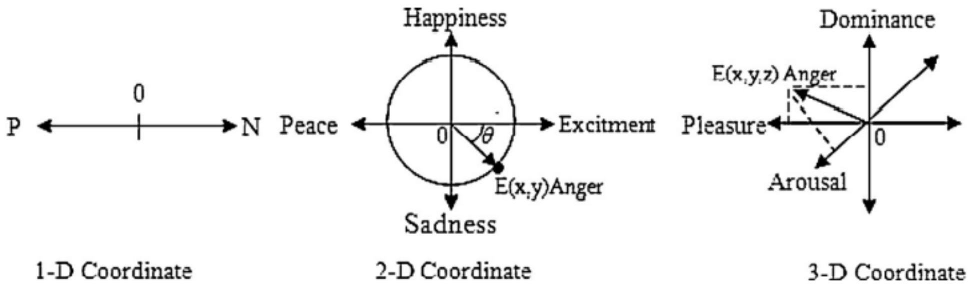

Figure A.1: *PAD space*
On the right, 1D space defined by valence; centre: 2D space defined by Valence and Arousal; left: 3D space defined by Valence, Arousal and Dominance

The three dimensions vary from -1 to 1. Valence characterizes whether the feeling showed is positive or negative. Arousal represents the intensity of the emotion ranging from calm to active. Lastly, dominance describes how much the emotion is present and important inside the subject.

Russell [59] showed that two dimensions of the PAD Model - Pleasure and Arousal - helped account for the main portion of the variance in affective states.

A.3.1. Self-Assessment Manikin (SAM)-Likert Scale

When mapping of emotions in PAD space, sometimes it is used the Self-Assessment Manikin (SAM)-Likert Scale, a pictographic scale that allows to annotate valence, arousal and dominance independently with a point from 1 to 5 [60, 61].

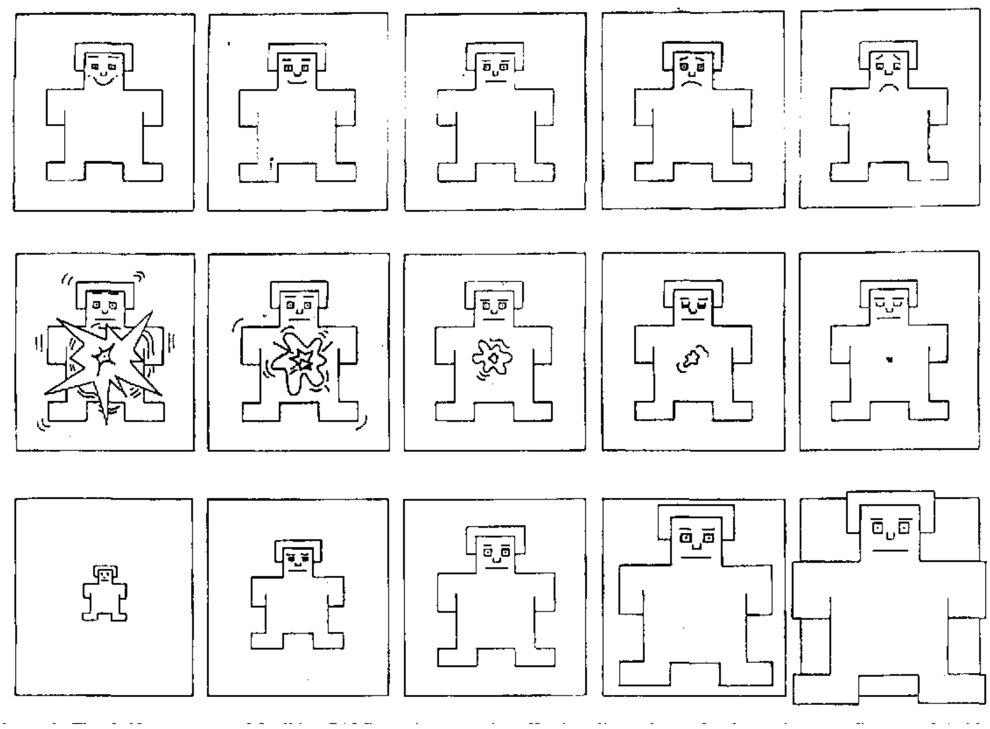

Figure A.2: *Self-Assessment Manikin Likert Scale*
First row corresponds to valence, second row to arousal and the last one to dominance.

A.4. Classification of emotional level

In article [62] feasibility of real-time emotion recognition was studied. Features were computed through openSmile software. A search problem was carried out through WEKA software, employed for machine learning, to detect the optimal set of features. The whole sentence is considered, for computing the features, even though it is real-time evaluation. In this case, emotions displayed were rated by two researchers with a psychological background and no self-assessment was carried out.

In [56], speech segments coming from vocal messages from weChat were mapped in the 3D PAD space. The considered emotions were happyness, sadness, anger, surprise, fear, and neutral whose coordinates in the space were estimated through Support Vector Regression. The recognition rate was more than 80% for each emotion.

Similar results were obtained in [63] where Multi-Layer Perceptron, Network-Based Fuzzy Inference System and Generic Self-Organizing Fuzzy Neural Network were employed to estimate coordinates in Valence-Arousal space of speech segment from Berlin Emotional Speech Database and later through K-means clustering, areas corresponding to different emotions are detected.

In [64] before emotion classification the Authors achieved a better performance reaching 90% of accuracy. In this case, as a previous step before actual classification, gender detection was performed and 2 separate algorithms were developed. Indeed, it was showed that gender-specific classifiers or regressors give higher accuracy in mapping of emotions. In both [63, 64], the Berlin emotional speech database (Emo-DB) has been used for training and testing. Emo-DB is considered the standard database for emotion recognition, indeed.

Some studies have been conducted in music emotion recognition, as well. For example, in [65], audio segments are mapped in the 2D space, defined by arousal and valence. In this case, features were extracted with PsySound and Marsyas, two open-source toolboxes for psychoacoustic feature extraction (such as loudness, pitch and dissonance), to construct a 114-dimension feature space. A further dimensionality reduction was then performed through WEKA software, obtaining an r^2 of about 58% for arousal and 20% for valence.

In [66] 6 emotions were considered: happiness, fear, sadness, neutral, surprise, and disgust. PRAAT was used for feature extraction and the study was carried out with the SAVEE database, containing audio signals of 4 male actors performing 7 different emotions, for a total of 480 British English utterances. Decision trees or random forests were the classifiers used since they allow an automatic feature selection. In [55], comparison between human and automatic annotations has been carried out for mapping scenes from 30 movies of different genres. Finally, in [67] a multiple combination of features was tried on Emo-DB and an Indian database, showing that among all the features, spectral ones are the most significant for detecting emotions.

B | Mini-Mental State Examination - MMSE

MINI MENTAL STATE EXAMINATION

(Folstein M.F., Folstein S.E., McHugh P.R. "Mini Mental State" a practical method for grading the cognitive state of patients for the clinicians. J Psychait Res 1975; 12: 189 – 198)

AREE INDAGATE	PUNTEGGIO
ORIENTAMENTO	
Il paziente sa riferire: giorno del mese, anno, mese, giorno della settimana e stagione.	**(0) (1) (2) (3) (4) (5)**
Il paziente sa riferire: luogo in cui si trova, a quale piano, città, regione, stato	**(0) (1) (2) (3) (4) (5)**
MEMORIA	
L'esaminatore pronuncia ad alta voce tre termini (casa, pane, gatto) e chiede al paziente di ripeterli immediatamente. L'esaminatore deve ripeterli fino a quando il paziente non li abbia imparati (max 6 ripetizioni). Tentativi n.______________	**(0) (1) (2) (3)**
ATTENZIONE E CALCOLO	
Partendo da 100 far contare sottraendo 7 all'indietro fermandosi dopo le prime cinque risposte. In caso di difficoltà di calcolo far scandire al contrario la parola "mondo" ("odnom").	**(0) (1) (2) (3) (4) (5)**
RICHIAMO DELLE TRE PAROLE (RICHIAMO ALLA MEMORIA)	
Richiamare i tre termini precedentemente imparati.	**(0) (1) (2) (3)**
LINGUAGGIO	
Il paziente deve riconoscere due oggetti. Come si chiama questo (matita)? Come si chiama questo (orologio)?	**(0) (1) (2)**

Il paziente deve ripetere la frase “TIGRE CONTRO TIGRE”	**(0) (1)**
Il paziente deve eseguire un compito su comando: a) prenda un foglio con la mano destra, b) lo pieghi a metà; c) e lo butti dal tavolo/metta sul pavimento.	**(0) (1) (2) (3)**
Far eseguire al paziente il comando scritto “CHIUDA GLI OCCHI”	**(0) (1)**
Far scrivere al paziente una frase di senso compiuto formata almeno da soggetto e verbo. N.B. conservare il materiale	**(0) (1)**
Far copiare al paziente un disegno (pentagoni intersecati). N.B. conservare il materiale	**(0) (1)**
Totale complessivo **Totale complessivo aggiustato**	**_____________/30** **_____________/30**

NOME E COGNOME DEL PAZIENTE (iniziali):_______________________

DATA SOMMINISTRAZIONE:______________________

C | Geriatric Depression Scales - GDS

Geriatric Depression Scale (GDS)

(Yesavage JA, Rose TL, Lum O, Huang V, et al. Development and validation of geriatric depression screening: a preliminary report. J Psychiatr Res 1983;17:37-49)

		SI	NO
1	E' soddisfatto della sua vita?	0	1
2	Ha abbondonato molte delle sue attività e dei suoi interessi?	1	0
3	Ritiene che la sua vita sia vuota?	1	0
4	si annoia spesso	1	0
5	Ha speranza nel futuro?	0	1
6	E' tormentato da pensieri che non riesce a togliersi dalla testa?	1	0
7	E' di buon unore per la maggior parte del tempo?	0	1
8	Teme che le stia per capitare qualcosa di brutto?	1	0
9	Si sente felice per la maggior parte del tempo?	0	1
10	Si sente spesso indifeso?	1	0
11	Le capita spesso di essere irrequieto e nervoso?	1	0
12	Preferisce stare a casa, piuttosto che uscire a fare cose nuove?	1	0
13	Si preoccupa frequentemente per il futuro?	1	0
14	Pensa di avere più problemi di memoria della maggior parte delle persone?	1	0
15	Pensa che sia bello stare al mondo, adesso?	0	1
16	Si sente spesso abbattuto e triste. adesso?	1	0
17	Trova che la sua condizione attuale sia indegna di essere visstuta?	1	0
18	Si tormenta molto pensando al passato?	1	0
19	Trova che la sita sia molto eccitante?	0	1
20	Le risulta difficiel iniziare ad occuparsi di nuovi progetti?	1	0
21	Si sente pieno di energia?	0	1
22	Pensa di essere in una situazione priva di speranza?	1	0
23	Pensa che la maggior parte delle persona sia in una condizione migliore della sua?	1	0
24	Le capita spesso di turbarsi per cose poco importanti?	1	0
25	Ha frequentemente voglia di piangere?	1	0
26	Ha difficoltà a concentrasi?	1	0
27	Si alza con piacere la mattina?	0	1
28	Preferisce evitare gli incontri sociali?	1	0
29	Le riesce facile prendere delle decisioni?	0	1
30	ha la mente lucida come prima?	0	1

Punteggio totale **____/30**

GDS - Geriatric Depression Scale (forma breve)

1. È fondamentalmente soddisfatto della sua vita?	Sì	No
2. Ha abbandonato molte delle sue attività e dei suoi interessi?	Sì	No
3. Sente che la sua vita è vuota?	Sì	No
4. Si annoia spesso?	Sì	No
5. È di buon umore la maggior parte del tempo?	Sì	No
6. Ha paura che qualcosa di brutto stia per succederle?	Sì	No
7. Si sente più felice nella maggior parte del tempo?	Sì	No
8. Si sente spesso impotente?	Sì	No
9. Preferisce restare a casa piuttosto che uscire e fare cose nuove?	Sì	No
10. Ritiene di avere più problemi con la memoria della maggior parte delle persone?	Sì	No
11. Pensa che la vita sia meravigliosa?	Sì	No
12. Si sente piuttosto inutile così com’è?	Sì	No
13. Si sente pieno di energie?	Sì	No
14. Ha l’impressione che la sua situazione sia disperata?	Sì	No
15. Pensa che la maggior parte delle persone sia migliore di lei?	Sì	No
Punteggio: ___/15 Un punto per “No” alle domande 1, 5, 7, 11, 13 Un punto per “Sì” alle altre domande	Normale	3 ± 2
	Lievemente depresso	7 ± 3
	Molto depresso	12 ± 2
Adattata da Sheikh JI, Yesavage JA: “Geriatric depression scale (GDS): Recent evidence and development of a shorter version,” in *Clinical Gerontology: A Guide to Assessment and Intervention*, edited by TL Brink. Binghamton, NY, Haworth Press, 1986, pp. 165-173. © By The Haworth Press, Inc.		

Bibliography

[1] European Commission, Directorate-General for Economic, and Financial Affairs. *The 2021 ageing report : underlying assumptions and projection methodologies*. Publications Office, 2021. DOI: doi/10.2765/733565.

[2] Michael D. Hurd et al. “Monetary Costs of Dementia in the United States”. In: *New England Journal of Medicine* 368.14 (2013). PMID: 23550670, pp. 1326–1334. DOI: 10.1056/NEJMsa1204629. eprint: https://doi.org/10.1056/NEJMsa1204629. URL: https://doi.org/10.1056/NEJMsa1204629.

[3] María Luisa Barragán Pulido et al. “Alzheimer’s disease and automatic speech analysis: A review”. In: *Expert Systems with Applications* 150 (2020), p. 113213. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.113213. URL: https://www.sciencedirect.com/science/article/pii/S0957417420300397.

[4] Vicki A. Freedman and Brenda C. Spillman. “Disability and Care Needs Among Older Americans”. In: *The Milbank Quarterly* 92.3 (2014), pp. 509–541. DOI: https://doi.org/10.1111/1468-0009.12076. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0009.12076. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0009.12076.

[5] Jessica Robin et al. “Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations”. In: *Digital Biomarkers* 4 (Oct. 2020), pp. 99–108. DOI: 10.1159/000510820.

[6] Jingying Wang et al. “Acoustic differences between healthy and depressed people: a cross-situation study”. In: *BMC Psychiatry* 19.1 (Oct. 2019), p. 300. ISSN: 1471-244X. DOI: 10.1186/s12888-019-2300-7. URL: https://doi.org/10.1186/s12888-019-2300-7.

[7] L. S. Low et al. “Detection of clinical depression in adolescents’ speech during family interactions”. In: *IEEE Trans Biomed Eng* 58.3 (Mar. 2011), pp. 574–586.

[8] Natália B. Mota, Mauro Copelli, and Sidarta Ribeiro. “Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance”. In: *npj Schizophrenia* 3.1 (Apr. 2017). DOI: 10.1038/s41537-017-0019-3. URL: https://doi.org/10.1038/s41537-017-0019-3.

[9] Sunghye Cho et al. “Automatic Detection of Autism Spectrum Disorder in Children Using Acoustic and Text Features from Brief Natural Conversations”. In: Sept. 2019, pp. 2513–2517. DOI: 10.21437/Interspeech.2019-1452.

[10] Meysam Asgari, Liu Chen, and Eric Fombonne. “Quantifying Voice Characteristics for Detecting Autism”. In: *Frontiers in Psychology* 12 (2021). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.665096. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2021.665096.

[11] A. Imran et al. “AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app”. In: *Inform Med Unlocked* 20 (2020), p. 100378.

[12] G. Cavallaro et al. “Acoustic voice analysis in the COVID-19 era”. In: *Acta Otorhinolaryngol Ital* 41.1 (Feb. 2021), pp. 1–5.

[13] Fasih Haider, Sofia de la Fuente, and Saturnino Luz. “An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer’s Dementia in Spontaneous Speech”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.2 (2020), pp. 272–281. DOI: 10.1109/JSTSP.2019.2955022.

[14] Jochen Weiner et al. “Investigating the Effect of Audio Duration on Dementia Detection using Acoustic Features”. In: *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Com-*

munication Association. 2018. URL: https://www.csl.uni-bremen.de/cms/images/documents/publications/Interspeech2018_WeinerEtAl.pdf.

[15] Saturnino Luz et al. “Alzheimer’s Dementia Recognition through Spontaneous Speech: The ADReSS Challenge”. English. In: *INTERSPEECH 2020.* ISCA, Oct. 2020, pp. 2172–2176. DOI: 10.21437/Interspeech.2020-2571. URL: http://www.interspeech2020.org/.

[16] Guy M. McKhann et al. “The diagnosis of dementia due to Alzheimer’s disease: Recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease”. In: *Alzheimer’s & Dementia* 7.3 (2011), pp. 263–269. DOI: https://doi.org/10.1016/j.jalz.2011.03.005. eprint: https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1016/j.jalz.2011.03.005. URL: https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2011.03.005.

[17] Florian Eyben et al. “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing”. In: *IEEE Transactions on Affective Computing* 7.2 (2016), pp. 190–202. DOI: 10.1109/TAFFC.2015.2457417.

[18] audEERING. *emobase feature set.* https://audeering.github.io/opensmile-python/api-smile.html. 2014.

[19] Björn Schuller et al. “The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load”. In: Sept. 2014.

[20] Matteo Caielli. *Automatic speech analysis for early detection of functional cognitive impairment in elderly population.* Master’s Thesis. 2018/2019.

[21] K. E. Forbes-McKay and A. Venneri. “Detecting subtle spontaneous language decline in early Alzheimer’s disease with a picture description task”. In: *Neurological Sciences* 26.4 (Oct. 2005), pp. 243–254. ISSN: 1590-3478. DOI: 10.1007/s10072-005-0467-9. URL: https://doi.org/10.1007/s10072-005-0467-9.

[22] Elaine Giles and Karalyn Patterson. “Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer’s type: Missing information”. In: *Aphasiology* 10 (May 1996), pp. 395–408. DOI: 10.1080/02687039608248419.

[23] *Mini-Mental State Examination (MMSE).* https://www.healthdirect.gov.au/mini-mental-state-examination-mmse.

[24] “Classification”. In: *Business Intelligence.* John Wiley & Sons, Ltd, 2009. Chap. 10, pp. 221–275. ISBN: 9780470753866. DOI: https://doi.org/10.1002/9780470753866.ch10. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470753866.ch10. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470753866.ch10.

[25] Anna Dorogush, Vasily Ershov, and Andrey Gulin. “CatBoost: gradient boosting with categorical features support”. In: (Oct. 2018).

[26] *What is Gradient Boosting? How is it different from Ada Boost?* https://medium.com/analytics-vidhya/what-is-gradient-boosting-how-is-it-different-from-ada-boost-2d5ff5767cb2.

[27] J. Elith, J. R. Leathwick, and T. Hastie. “A working guide to boosted regression trees”. In: *Journal of Animal Ecology* 77.4 (2008), pp. 802–813. DOI: https://doi.org/10.1111/j.1365-2656.2008.01390.x. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2656.2008.01390.x. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2656.2008.01390.x.

[28] Aaron J Masino et al. “Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data”. In: *PloS one* 14.2 (2019). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0212665. URL: https://europepmc.org/articles/PMC6386402.

[29] *SHAP Documentation.* https://shap.readthedocs.io/en/latest/index.html#..

[30] F. Martínez-Sánchez et al. “Oral reading fluency analysis in patients with Alzheimer disease and asymptomatic control subjects”. In: *Neurologia* 28.6 (2013), pp. 325–331.

[31] Alexandra König et al. "Use of Speech Analyses within a Mobile Application for the Assessment of Cognitive Impairment in Elderly People". In: *Current Alzheimer research* 14 (Aug. 2017). DOI: `10.2174/1567205014666170829111942`.

[32] László Tóth et al. "A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech". In: *Current Alzheimer Research* 14 (Nov. 2017). DOI: `10.2174/1567205014666171121114930`.

[33] Laura Calzà et al. "Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia". In: *Computer Speech & Language* 65 (2021), p. 101113. ISSN: 0885-2308. DOI: `https://doi.org/10.1016/j.csl.2020.101113`. URL: `https://www.sciencedirect.com/science/article/pii/S0885230820300462`.

[34] K. C. Fraser et al. "Predicting MCI Status From Multimodal Language Data Using Cascaded Classifiers". In: *Front Aging Neurosci* 11 (2019), p. 205.

[35] Muhammad Shehram Shah Syed et al. "Automated Screening for Alzheimer's Dementia Through Spontaneous Speech". In: *Proc. Interspeech 2020*. 2020, pp. 2222–2226. DOI: `10.21437/Interspeech.2020-3158`.

[36] Mireia Farrús and Joan Codina-Filbà. *Combining Prosodic, Voice Quality and Lexical Features to Automatically Detect Alzheimer's Disease*. 2020. DOI: `10.48550/ARXIV.2011.09272`. URL: `https://arxiv.org/abs/2011.09272`.

[37] Maria Yancheva, Kathleen Fraser, and Frank Rudzicz. "Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias". In: *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*. Dresden, Germany: Association for Computational Linguistics, Sept. 2015, pp. 134–139. DOI: `10.18653/v1/W15-5123`. URL: `https://aclanthology.org/W15-5123`.

[38] Z. Fu, F. Haider, and S. Luz. "Predicting Mini-Mental Status Examination Scores through Paralinguistic Acoustic Features of Spon-

taneous Speech". In: *Annu Int Conf IEEE Eng Med Biol Soc* 2020 (July 2020), pp. 5548–5552.

[39] E. Ambrosini et al. "Automatic speech analysis to early detect functional cognitive decline in elderly population". English. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2019*. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2019 ; Conference date: 23-07-2019 Through 27-07-2019. Institute of Electrical and Electronics Engineers Inc., July 2019, pp. 212–216. DOI: `10.1109/EMBC.2019.8856768`.

[40] Natick, Massachusetts: The MathWorks Inc. *MATLAB*. Version R2021b. 2021. URL: `https://it.mathworks.com/help/releases/R2021b/matlab/index.html`.

[41] Min Xu et al. "HMM-based audio keyword generation". English. In: *5th Pacific Rim Conference on Multimedia, Tokyo, Japan, November 30 - December 3, 2004. Proceedings, Part III*. Springer, 2004, pp. 566–574.

[42] Aharon Satt et al. "Evaluation of Speech-Based Protocol for Detection of Early-Stage Dementia". In: Aug. 2013. DOI: `10.21437/Interspeech.2013-32`.

[43] Francesca Lunardini et al. "The MOVECARE Project: Home-based Monitoring of Frailty". In: *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*. 2019, pp. 1–4. DOI: `10.1109/BHI.2019.8834482`.

[44] K. D. Mueller et al. "Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks". In: *J Clin Exp Neuropsychol* 40.9 (Nov. 2018), pp. 917–939.

[45] J. A. Yesavage et al. "Development and validation of a geriatric depression screening scale: a preliminary report". In: *J Psychiatr Res* 17.1 (1982), pp. 37–49.

[46] Theodore W Anderson, Donald A Darling, et al. "Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes". In: *The annals of mathematical statistics* 23.2 (1952), pp. 193–212.

[47] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[48] Liudmila Prokhorenkova et al. "CatBoost: Unbiased Boosting with Categorical Features". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., 2018, pp. 6639–6649.

[49] David Molloy and Timothy Standish. "A Guide to the Standardized Mini-Mental State Examination". In: *International psychogeriatrics / IPA* 9 Suppl 1 (Feb. 1997), 87–94, discussion 143. DOI: 10.1017/S1041610297004754.

[50] Mikel de Velasco, Raquel Justo, and María Inés Torres. "Automatic Identification of Emotional Information in Spanish TV Debates and Human&ndash;Machine Interactions". In: *Applied Sciences* 12.4 (2022). ISSN: 2076-3417. DOI: 10.3390/app12041902. URL: https://www.mdpi.com/2076-3417/12/4/1902.

[51] R. Banse and K. R. Scherer. "Acoustic profiles in vocal emotion expression". In: *J Pers Soc Psychol* 70.3 (Mar. 1996), pp. 614–636.

[52] Dimitrios Ververidis and C. Kotropoulos. "Emotional speech recognition: Resources, features, and methods". In: *Speech Communication* 48 (Sept. 2006), pp. 1162–1181. DOI: 10.1016/j.specom.2006.04.003.

[53] D. J. France et al. "Acoustical properties of speech as indicators of depression and suicidal risk". In: *IEEE Trans Biomed Eng* 47.7 (July 2000), pp. 829–837.

[54] Geoffroy Peeters. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. Tech. rep. Icram, 2004.

[55] Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. "A dimensional approach to emotion recognition of speech from movies". In: *2009 IEEE International Conference on Acous-*

tics, Speech and Signal Processing. 2009, pp. 65–68. DOI: 10.1109/ICASSP.2009.4959521.

[56] Weihui Dai et al. "Emotion recognition and affective computing on vocal social media". In: *Information & Management* 52.7 (2015). Novel applications of social media analytics, pp. 777–788. ISSN: 0378-7206. DOI: https://doi.org/10.1016/j.im.2015.02.003. URL: https://www.sciencedirect.com/science/article/pii/S037872061500018X.

[57] Albert Mehrabian. "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament". In: *Current Psychology* 14.4 (Dec. 1996), pp. 261–292. ISSN: 1936-4733. DOI: 10.1007/BF02686918. URL: https://doi.org/10.1007/BF02686918.

[58] Stephen W. Gilroy et al. "PAD-based multimodal affective fusion". In: *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 2009, pp. 1–8. DOI: 10.1109/ACII.2009.5349552.

[59] James Russell. "A Circumplex Model of Affect". In: *Journal of Personality and Social Psychology* 39 (Dec. 1980), pp. 1161–1178. DOI: 10.1037/h0077714.

[60] Margaret M. Bradley and Peter J. Lang. "Measuring emotion: the Self-Assessment Manikin and the Semantic Differential." In: *Journal of behavior therapy and experimental psychiatry* 25 1 (1994), pp. 49–59.

[61] Alberto Betella and Paul F. M. J. Verschure. "The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions". eng. In: *PloS one* 11.2 (Feb. 2016). PONE-D-15-40714[PII], e0148037–e0148037. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0148037. URL: https://doi.org/10.1371/journal.pone.0148037.

[62] Kiavash Bahreini, Rob Nadolski, and Wim Westera. "Data Fusion for Real-time Multimodal Emotion Recognition through Webcams and Microphones in E-Learning". In: *International Journal of Human–Computer Interaction* 32.5 (2016), pp. 415–430. DOI:

10.1080/10447318.2016.1159799. eprint: https://doi.org/10.1080/10447318.2016.1159799. URL: https://doi.org/10.1080/10447318.2016.1159799.

[63] Norhaslinda Kamaruddin and Abdul Wahab Abdul Rahman. "Valence-arousal approach for speech emotion recognition system". In: *2013 International Conference on Electronics, Computer and Computation (ICECCO)*. 2013, pp. 184–187. DOI: 10.1109/ICECCO.2013.6718259.

[64] Igor Bisio et al. "Gender-Driven Emotion Recognition Through Speech Signals For Ambient Intelligence Applications". In: *IEEE Transactions on Emerging Topics in Computing* 1.2 (2013), pp. 244–257. DOI: 10.1109/TETC.2013.2274797.

[65] Yi-Hsuan Yang et al. "A Regression Approach to Music Emotion Recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.2 (2008), pp. 448–457. DOI: 10.1109/TASL.2007.911513.

[66] Fatemeh Noroozi et al. "Vocal-based emotion recognition using random forests and decision tree". In: *International Journal of Speech Technology* 20.2 (June 2017), pp. 239–246. ISSN: 1572-8110. DOI: 10.1007/s10772-017-9396-2. URL: https://doi.org/10.1007/s10772-017-9396-2.

[67] Shashidhar Koolagudi and K. Rao. "Emotion recognition from speech using source, system, and prosodic features". In: *International Journal of Speech Technology* 15 (June 2012). DOI: 10.1007/s10772-012-9139-3.

Acknowledgments

Siamo arrivati alla fine di questo percorso accademico e vorrei ringraziare tutte le persone che mi hanno sostenuto e mi sono state accanto, chi dall'inizio chi si è aggiunto lungo la strada.

Ringrazio la Professoressa Simona Ferrante per avermi dato l'opportunità di lavorare a questo progetto, dandomi la possibilità di imparare molto più di quanto potessi mai immaginare. Grazie alla Professoressa Ambrosini per avermi seguito con così tanta costanza e pazienza e grazie a Eugenio per le sue brillanti intuizioni e consigli.

Ringrazio tutti i miei amici che ho incontrato in questo lungo percorso accademico, chi c'è da sempre chi si è aggiunto negli anni.

E infine, un enorme ringraziamento va alla mia famiglia, che mi sostiene da sempre in ogni mio interesse e passione, ognuno a modo proprio, e crede in me anche quando io non riesco.

Milano, 07/06/2022
Chiara