Escuela Técnica Superior de Ingeniería Informática

Universidad Nacional de Educación a Distancia (UNED)

An Artificial Intelligence Approach for Generalizability of Cognitive Impairment Recognition in Language

Autor: Mónica González Machorro

Director: Rafael Martínez Tomás

Master's Thesis (TFM)

February 2022

Máster Universitario en Investigación en Inteligencia Artificial

Abstract

Introduction: Language provides valuable information in early dementia recognition, as language impairment is a common characteristic of early dementia.

Objectives: One current limitation is a disconnection between results in previous research and clinical applications. In this paper, we propose artificial intelligence methods that address this challenge: the generalizability of studies.

Methods: We analyze language in two separate modalities: speech signal and linguistic information referring to grammar and semantics. For the first modality, we employ audio files, and for the second one, transcripts extracted from audio files. We employ a subset of the Pitt Corpus that contains early Alzheimer's disease and Mild Cognitive Impairment patients. Our proposed methods include exploring deep transfer learning models and explainable feature sets that are language-independent; applying ASR tools to obtain automatic transcripts; fine-tuning end-to-end ASR models for our task; analyzing the smallest speech unit: phonemes; and validating the most promising models on conversational speech data from other data sources.

Results: Results show that speech deep transfer learning approaches provide a content-independent solution and outperform ASR transcripts-based methods. We also demonstrate that linguistic-based deep learning methods are more robust than speech-based approaches in external validation procedures.

Conclusion: This research highlights the need of suitable ASR-tools for dementia and exploration of conversational speech data. Our main contribution is to successfully fine-tuning end-to-end and transfer learning methods in dementia recognition.

Keywords:

automatic speech recognition, artificial intelligence, early Alzheimer's recognition, language impairment, mild cognitive impairment, natural language processing.

Resumen

Introducción: Trastornos en el lenguaje se considera uno de los primeros signos del deterioro cognitivo.

Objetivos: Un reto, sin embargo, es la desconexión entre los resultados obtenidos en previas investigaciones y su aplicación en contextos clínicos. Esto se debe, en gran parte, a la falta de estandarización y de datos en este campo. La propuesta de este trabajo es emplear técnicas de inteligencia artificial para abordar este reto: la generalización.

Metodología: En este trabajo estudiamos el lenguaje en dos modalidades: el habla, que se refiere a la manifestación acústica del lenguaje y la información lingüística entendida como la gramática. Para la primera modalidad empleamos grabaciones y para la segunda transcripciones de las grabaciones. El conjunto de datos empleado es un subconjunto del Corpus Pitt que contiene pacientes con deterioro cognitivo leve y Alzheimer. Nuestra propuesta incluye explorar métodos de aprendizaje transferido y *end-to-end* tales como wav2vec, HuBERT, BERT y RoBERTa; aplicar herramientas de ASR para obtener transcripciones automáticas, explorar variables que sean independientes de la lengua y del contenido; analizar la unidad del habla más pequeñas: los fonemas; y por último, evaluar los métodos más prometedores en un conjunto de datos externo.

Resultados: Los resultados demostraron que, en el caso de métodos de aprendizaje de transferencia, la modalidad acústica no solo proporciona una solución independiente del contenido lingüístico, sino que también obtiene un mayor rendimiento que aquellos métodos basados en transcripciones producidas mediante herramientas de ASR. Los resultados también demuestran que los métodos de la modalidad lingüística son más robustos que los de la modalidad acústica.

Conclusión: Este trabajo destaca la necesidad de una herramienta ASR adecuada para la transcripción de demencia y de explorar el habla espontánea. La mayor aportación es la aplicación exitosa de modelos *end-to-end* y de aprendizaje transferido en la detección de demencia.

Palabras clave:

lenguaje, inteligencia artificial, aprendizaje transferido, enfermedad del Alzheimer, deterioro cognitivo leve.

Table of Contents

1. Introduction

1.1. Motivation

Dementia disease is a general term for degenerative diseases of the brain that interfere in the ability to remember, think, comprehend, calculate, and communicate [1], [2]. This disease is considered incurable and has a progressive nature. According to [2], "currently there are over 55 million people with dementia worldwide and there are nearly 10 million new cases every year" [2]. Moreover, "dementia is the seventh leading cause of death among all diseases and one of the major causes of disability among elderly people worldwide" [2].

There are many types of dementia, of which Alzheimer's disease (AD) accounts for 60-70% cases of dementia [2]. Other types of dementia are vascular dementia (VD), Lewy body dementia, frontotemporal dementia, among others. "Boundaries between different forms of dementia are indistinct and mixed forms often co-exist" [2].

Advanced stages of dementia are characterized by the patient's almost total dependency or inactivity and difficulty to recognize friends and relatives [2]. Behavioral changes such as aggression may also present themselves [2]. In addition, dementia disease not only has an impact on patients and their relatives but also carries social and economic implications in terms of medical care [2]. According to [2], "in 2019, the estimated total global societal cost of dementia was US $1.3 trillion, and these costs are expected to surpass US $2.8 trillion by 2030" [2].

Due to the nature of this disease, an early diagnosis is considered to be key to delay the progress of the disease [3]. Nonetheless, this is challenging since patients usually do not visit the doctor until the disease is already in an advanced state [3], even though, according to [4], dementia cognitive deficits could be detected up to ten years before its onset. This is possible due to an intermediate state between the changes in cognitive natural aging and dementia disease [5] known as Mild Cognitive Impairment (MCI). Researchers consider this preclinical period a window of opportunity for early dementia detection [5], as "approximately, 32-38% of individuals with MIC will progress to a major form of dementia within 5 years" [6].

In spite of the importance of early dementia diagnosis, the current techniques for recognizing it are "time-consuming screening tests or tests that cannot diagnose preclinical states" [3]. Common methods to diagnose dementia are the Mini-Mental Status Examination, the Clock-Drawing Test and the Alzheimer's Disease Assessment

Scale-Cognitive Subscale. According to [7], these tests have proved insufficient to accurately recognize MCI [7], [8]. Consequently, it becomes necessary to verify the tests results with diagnostic tools like Magnetic Resonance Imaging, cerebrospinal fluid and computed tomography. This, in turn, requires more resources and time which difficults early dementia recognition furthermore.

Given the severity of the disease, it is imperative to find and provide early dementia diagnosis and treatment to improve patients' living conditions. These methods should be practical, reliable and straightforward. A promising solution is language analysis, in which recent artificial intelligence growth has allowed researchers to find encouraging results [9], [10]. However, there are still key challenges to employ language as biomarker for dementia [10]. One of them is the potential applicability of a research in clinical scenarios, namely how to generalize approaches as much as possible.

1.2. Objective

As mentioned above, language analysis presents a potential improvement for early dementia recognition. In spite of recent developments in artificial intelligence, one key challenge remains: the generalizability of approaches. To tackle this issue, this paper's main objective is to propose machine learning (ML) methods that are transcription-free, content and language-independent and that do not require a big amount of data. The best-performing methods are evaluated on conversational data as well. The paper focuses on AD and MCI identification. Accordingly, we set the following objectives:

- customize a balanced subset of the Pitt Corpus to study MCI, probable AD and control patients. Pitt Corpus is the most studied corpus on dementia in speech [10], however, previous research has mostly focused on only AD detection;
- obtain non-human generated transcripts. Thus, we apply an Automatic Speech Recognition (ASR) tool to obtain transcripts;
- identify relevant explainable acoustic and linguistic features to distinguish patients with early AD and MCI from healthy control (HC) group in English;
- analyze the role of phonetic units in dementia detection;
- evaluate feature-based and deep transfer learning-based methods for AD and MCI detection in language;
- fine-tune end-to-end pre-trained ASR models to recognize MCI and AD;

- evaluate and compare methods employing hand-generated transcripts and ASR transcripts;
- and after the initial evaluation, we propose to evaluate the deep learning methods on an external data set containing conversational speech.

2. Related work

2.1. Dementia in language

In the past decades, researchers have focused on developing diagnostic tools for early dementia detection. One of the most investigated domains is the aspect of memory due to the high impact the disease has on it [10]. For instance, two common symptoms of dementia are to lose track of time and to forget old memories. Studies have proved that the presence of memory deficits is a consistent factor in the profile of dementia patients [11].

Closely related to memory, the language domain has gained interest in the research community since early types of dementia are characterized by language impairment, an abnormality understood as "deficits in the cognitive process of forming language with structure and meaning" [12]. For instance, dementia patients have trouble finding a word (anomia), word meanings, hypofluency, hyperfluency, repetition and naming, among others [8], [12], [13]. They also present longer speech pauses. Especially in the early stage of the disease, it is common to call language impairment "empty speech", since the speech is correctly structured in terms of grammar but misses some words or contains unclear meanings [12], [13]. Later stages of dementia are characterized by a lack of verbal fluency and comprehension [13]. These symptoms make linguistic and speech analysis useful tools to identify early stages of dementia.

Other advantages of using speech and linguistic information as biomarkers for dementia are: the simplicity and non-invasiveness of data collection [10] and "the fact that speech problems may be manifest at different stages of the disease, making it a life-course assessment that has value unlimited by disease stage" [10]. Moreover, computerized analysis of speech and ML techniques have developed increasingly in the past years, which has attracted increasing interest in the field of dementia recognition in language [10], [14], [15]. Specifically, in 2020 and 2021 the INTERSPEECH conference presented two special sessions: Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) [15] and Detecting Cognitive Decline through Speech Only

(ADReSSo) [14], which “aim to develop methods that can assist in the automated prediction of AD from speech data” [15], [16]. These sessions have obtained high performance using state-of-the-art architectures both in the speech and linguistic domain.

In this paper, we analyze language in two separate modalities: speech and linguistic information (such as grammar). Collecting dementia speech and linguistic data, however, is not an easy task in comparison to other language identification tasks [10], [14], due to several constraints involved such as recruiting elderly people. Generally, dementia in language is investigated through two main data types: spontaneous speech and elicited speech [14], [17]. In the first one, speakers have spontaneous conversations or monologues covering broad topics. Examples are dialogue data, in which the patient has a conversation with the physician. This data type represents spontaneous conversation that is neither limited by time nor by topic. One drawback of this type of data is that recordings may highly vary in length [17], leading up to data sets with imbalanced amounts of data from particular speakers.

The second data type, elicited speech, is one of the most common techniques employed in dementia speech and linguistic analysis [10]. It provides a sample of discourse, where topic and length can be controlled by the interviewer [17]. A common approach consists of the patient describing a scene, such as the famous Cookie Theft Picture from the Boston Diagnostic Aphasia Examination [18]. In this Picture Description task, the participant sees an image of a kitchen in black and white and is asked to describe the scene (figure 1). This is a quick, objective and non-invasive assessment of an individual’s cognitive status [19]. Nevertheless, the task is unnatural and “limited richness and length” [17], [20]. In addition, the picture used is “an outdated depiction of domestic life” [17], [20] and the ability to describe it highly depends on cultural circumstances.

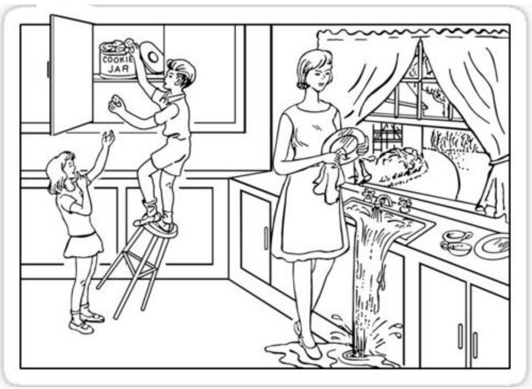


***Figure 1**. The standardized Cookie Theft Picture. Image obtained from [18].*

Nonetheless, given its simplicity and standardization [14], [15], it is one of the most used tasks in dementia linguistic and speech analysis [10], [14], [15]. For instance, studies have shown an accuracy of 81% using this task to distinguish between AD and healthy patients [21]. Deep learning models have achieved similar accuracy rates for this task. [19] fine-tuned a BERT-based text sequence classification that reached 83.3% accuracy [19].

Other approaches are to narrate a learned story and recall tests, in which patients are asked to recall and describe different routines they carry out [10] [17], for example, “the procedure for making a cup of tea” [17]. However, recall tests may result in a simplified speech as this type of discourse rarely occurs in everyday conversation [17]. In the two following sections, we explain in detail why speech and linguistic analysis are useful in dementia recognition and what previous research has reported.

2.1.1. Speech analysis

Speech corresponds to phonetics, i.e., the human voice production [22], and it can be understood as what has been said (paralinguistic). Due to speech requiring precise and complex movements controlled by brain functions [23], [24], it is well documented that speech can carry information about each speaker, for example, his cognitive and emotional state [23]. For instance, it has been found that in a patient with Parkinson’s disease the excessive loss of dopaminergic neurons leads to muscular rigidity, tremor and disturbances of gait and posture, which in turn leads to a more monotone speech [22], [25]. Another example is stuttering (speech disorder related to the flow of speech) which is associated to alterations in dopaminergic function [26]. It has been found that a person who stutters usually presents higher amounts of a dopamine precursor [22]. Regarding dementia, [26] found that “speech and orofacial apraxis are nearly always present in AD, regardless of disease stage” [26]. Apraxis (or apraxia) is a disorder that hinders a person’s ability to verbally communicate due to damage in the frontal lobe [26]. This disorder also worsens as the disease progresses [26].

Speech analyses have mostly focused on the study of utterances and conversations [27]. However, in order to acquire a deeper understanding of how dementia affects speech, it is important to analyze specific speech units, i.e., phonemes. It has been proven that not these units do not pose the same relevance detecting cognitive diseases in speech [24], “since each one devices from a different narrowing, articulation and configuration of the vocal tract" [24]. [28] stated that "AD presents articulatory difficulties at early stages of the disease" [28], [29]. Moreover, [29] found that "phonetic representations in transcripts can be understood as analogous to data augmentation"[29] and reported an accuracy of 77% by applying a FastText classifier [29]. Given that this research's main objective is the generalizability of studies and most of the early dementia detection work was conducted in English, we aim to further study phonetic representations in dementia since phonemes are the smallest processing unit and, therefore, they "are more generalizable across languages" [10].

The variance of speech is captured in speech features, also known as acoustic or paralinguistic features. These features are well-studied [30] since they are robust and generalizable due to minimal pre-processing [31]. Nonetheless, a challenge of speech feature analysis remains as to "the majority of state-of-the-art speech features do not generalize between languages and therefore are probably not modeling language impairment in dementia as a neurocognitive function" [12].

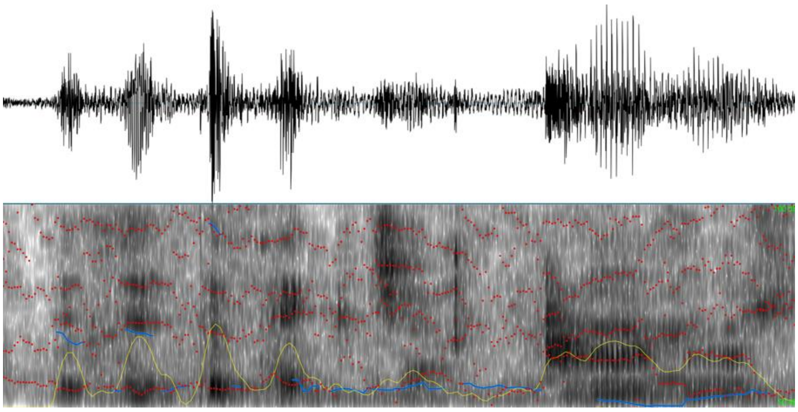


***Figure 2**. Example of a spectro-temporal representation of speech from which acoustic features are computed. The red dots represent the formants of the speech signal.*

In general terms, one can identify two types of analyses of acoustic features: low-level descriptors (LLDs) and high-level descriptors (HLDs). LLDs are closely related to the signal itself, such as fundamental frequency (F0), which refers to the frequency of voiced speech signals; jitter; shimmer; spectral and formants, "features known to model spoken content represented by position, amplitude and bandwidth" [32] (figure 2); and duration, "energy and intensity based on the amplitude in different intervals" [32]. The final result of LLDs is a value over a short window of time.

HLDs refer to statistical functionals of how the LLD have been extracted, such as the mean, median, standard deviance, variance, skewness, percentiles, quartiles ranges, ratio, linear and quadratic regression coefficient, ranges, zero-crossing rate and centroid [32]. The final result of the HLD calculation is an "aggregated information over an entire utterance" [33].

There are multiple examples of successful studies employing acoustic features to distinguish between dementia speech and healthy speech [10], [12], [14]. For instance, [9] used LLDs to train supervised ML architectures such as Random Forest (RF), Support Vector Machine (SVM) and Bayes Network, in which the best-performing algorithm obtained 82% of accuracy in classifying early AD [9]. [3] applied the number of silent and filled pauses, the length of utterances and the speech tempo on a SVM and achieved a 74% accuracy classifying HC from patients with AD and MCI [3].

Recently, also deep neural networks methods have shown potential in speech classification tasks among dementia recognition [10], [34]. These methods can be broken down into two broad categories: sequential and convolutional. Convolutional neural networks (CNNs) "operate on the basis of LLD as they can exploit the local correlations of the speech signals in both time and frequency dimensions" [35]. The focus lies on log-Mel and Mel-Frequency Cepstral Coefficients (MFCCs) spectrograms, "as they transform speech signals to feature maps as in computer vision applications" [35]. For example, [36] employed pre-trained models such as Inception v3, ResNet50, VGG16, AlexNet and Vision Transformer to classify AD and non-AD patients. The best-performing model obtained an accuracy of 65% [36]. [35] also investigated the use of computer vision architectures to detect dementia in speech data by employing a pre-trained ResNet18 model with 62% accuracy [35].

Finally, methods that utilize raw waveform on CNNs have also shown promising results [37]. These methods allow capturing information related to the speech production directly from raw audio files. [37] trained raw audio-based CNNs and achieved a 74% accuracy [37]. More recent approaches consist of adapting ASR models to dementia detection. [38], for example, used wav2vec 2.0 to "embed the raw audio into representations using the last hidden state of the pre-trained model and classify audio samples using these embeddings as input to classifiers" [38] like SVM and linear regression (LR). This resulted in a 67% accuracy [38]. [39] explored wav2vec 2.0 on dementia classification for Mandarin Chinese and reported a 77% accuracy [39].

Paper	Dataset	Language	Task	Method	Acc.
[3]	MCI-mAD database	Hungarian	MCI and AD	SVM	0.74
[17]	private dataset	English	MCI and AD	SVM	0.76
[35]	ADReSS	English	AD	ResNet-LSTM	0.625
[36]	ADReSS	English	AD	Vision Transformer	0.653
[37]	ADReSS	English	AD	CNN	0.74
[39]	2021 NCMMSC dataset	Mandarin Chinese	MCI and AD	wav2vec	0.798
[38]	ADReSS	English	AD	SVM	0.67
[9]	Pitt Corpus	English	MCI and AD	RF	0.65
[40]	ADReSS	English	AD	CNN + gating mechanism	0.64
[41]	private dataset	French	MCI and AD	SVM	0.79

***Table 1**. Summary of the latest previous work for speech-based approaches. LSTM stands for long-short term memory.*

2.1.2. Linguistic analysis

As mentioned above, dementia not only affects how things are said (speech) but also what has been said (linguistic information), i.e., the system of words and symbols used in language, specifically the domains of semantics, syntaxis and morphology. The gradual degeneration of the various cortical areas caused by dementia have an impact on how to structure sentences, naming objects, recalling past experiences and word meanings, among others [10]. This linguistic analysis – in contrast to speech analysis – is based on transcripts either made by specialists or using ASR tools.

The semantic field is concerned with the meaning of words. It is generally accepted that one of the first signs of impairment occur in the semantic domain [12], [13]. Common aspects in the study of dementia are fillers and hesitations [3] since dementia patients usually present more hesitations than healthy patients [3]. Also, word frequency, word length and word repetition are among the most studied semantic features [12]. Other features are frequency of keywords, frequency of unique words, type-to-token ratio (calculated by dividing all unique words by the total word count) and information units, which computes the number of unique entities that a person mentions. For example, [12] found 11 semantic features relevant to identify dementia in English and French such as percentage of keywords, number of unique keywords and number of unique information units [12].

Syntax and morphology represent the structure of language and word forms. This domain evaluates the sentence complexity and the use of parts-of-speech [12]. An analysis based on it is highly language-dependent [10], however, [12] found that some syntactic features can “represent syntactic impairment that overlap between languages” [12]. Syntax and morphology are also closely related to semantics since they depend on each other to form intelligible sentences, therefore, it is common that studies combine these domains [10]. As for dementia, [12] found that dementia patients use less nouns and adpositions [12]. [3] proved that patients with MCI often “substitute content words with fillers or indefinite pronouns, and they also appear to use a lot of paraphrases indicating uncertainty” [3].

Furthermore, researchers used word embeddings to better capture linguistic information where words that have the same meaning have a similar representation. This technique has been proven useful in many natural language processing (NLP) tasks, among them dementia detection. For example, [42] trained a SVM to classify AD from non-AD using linguistic features and reported a 71% accuracy [42]. [43] used word embeddings to successfully identify MCI patients from English speech transcripts and obtained an accuracy of 60%. Other studies [44] measured lexical diversity to compare dementia versus AD [44].

Parallel to the speech modality, this field has gained interest due to deep learning method development. For example, [16] extracted lexical information from transcripts to train a LSTM and achieved a 70% accuracy [16]. Other studies like [45] investigated text-based CNNs and reported a 88% accuracy. [37] trained a bi-directional Hierarchical Attention Network on transcripts and found an 81% accuracy.

Among these methods, it is important to mention transformers-architectures-based pre-trained models, particularly BERT (Bidirectional Encoder Representations from Transformer) and RoBERTa (Robustly Optimized BERT Pretraining Approach) models [46], [47]. These pre-trained models allow researchers to “fine-tune on specific tasks by adding just one additional output layer” [46], [48]. Previous works have demonstrated promising results exploring these models. For example, [48] adapted a pre-trained BERT model to detect AD from non-AD and achieved an 82.1% accuracy [48]. [49] reported an 87% accuracy on the ADReSS dataset using RoBERTa [49]. Another example is [50], who combined wav2vec and BERT to detect dementia in speech. They extracted semantic information from speech data adapting wav2vec and then classified this information using BERT. They reported an 83% accuracy. [51] used BERT embeddings to train a RF classifier and obtained an accuracy of 85.4% [51].

This domain has demonstrated better results than speech-based models [10], [29]. These results, however, are dependent on whether the transcripts are generated by a human observer or by an ASR system. As expected, the highest performance is obtained when the results are human-generated. However, this approach is time-consuming and requires domain-knowledge to achieve high accuracy. Unfortunately, previous work has mostly focused on human-generated transcripts (table 3).

Paper	Dataset	Transcripts	Language	Task	Method	Acc.
[3]	MCI-mAD database	human	Hungarian	MCI and AD	SVM	0.82
[49]	ADReSS	human	English	AD	RoBERTa	0.87
[50]	ADReSS	ASR	English	AD	BERT	0.83
[48]	ADReSS	human	English	AD	BERT	0.821
[37]	ADReSS	human	English	AD	LSTM	0.84
[35]	ADReSS	human	English	AD	FastText	0.83
[51]	ADReSS	human	English	AD	RF	0.85
[29]	ADReSS	human	English	AD	FastText	0.77
[39]	NCMMSC dataset	human	Mandarin Chinese	AD	RoBERTa	0.75
[39]	NCMMSC	ASR	Mandarin Chinese	AD	RoBERTa	0.53
[45]	Pitt Corpus	human	English	various types of dementia	CNN	0.88

***Table 2**. Summary of latest previous work for linguistic-based approaches.*

2.1.3. Linguistic fusion

As seen in the previous two sections, speech and linguistic analysis are closely related and equally important in human communication. Recently, studies have addressed the fusion of these two main domains aiming to imitate how humans process language as accurately as possible. In other words, “in the real word, people often associate information from multiple forms, i.e., speech and linguistic modalities” [52]. This multimodal language analysis offers advantages such as “language disambiguation and language sparsity issues” [52].

According to [33], “the dominant computational paradigm for fusing different modalities has been that of shallow fusion” [33]. This paradigm is generally divided into feature-level or early fusion and decision-level or late fusion [33], [53]. “Some researchers have also followed a hybrid approach by performing fusion at the feature as well as the decision level” [53].

The early fusion method consists of processing each modality independently and to merge them before training a single ML model, which could also, be understood as unimodal classification. In other words, features from each modality are extracted from the input and combined [53], [54]. This results in a high dimensionality feature vector that can be reduced through feature selection methods. One advantage of this approach is that “it can utilize the correlation between multiple features from different modalities at an early stage which helps in better task accomplishment”[53]. Disadvantages are that this approach “has to deal with many heterogeneous features” [54] and “it is hard to represent the time synchronization between multimodal features” [53].

Late fusion, on the other hand, “corresponds to training separate models for each modality and combining their independent decisions” [33]. It is also known as rule-based fusion method because it “includes statistical rule-based methods such as majority voting and linear weighted fusion” [53]. These schemes “generally perform well if the quality of temporal alignment between different modalities is good” [53].

Likewise, deep fusion modalities have gained interest. These procedures employ deep learning methods to learn multidimensional representations [33]. It is similar to shallow fusion with the difference that deep fusion “consists of several differential modules, some of them unimodal and others multimodal, which are jointly trained” [33].

In recent years, both feature-based and deep fusion modalities have shown competitive performance in dementia detection in language. Especially ADReSS and ADReSSo [14], [15] have predominantly focused on multimodal analyses of acoustic and language data. [42] normalized scores obtained from SVM models trained on text and audio modalities and merged the scores by the average, achieving an overall accuracy of 75% [42]. [19] found an 81% accuracy training a SVM model with acoustic and language features [19]. Another example is [17], who applied an early fusion approach by extracting various linguistic and speech features on the Picture Description task to train a SVM and reported a 76% accuracy [17].

2.2. Current challenges

Even though dementia in speech and language has been thoroughly investigated, there are still important challenges to tackle. The first main challenge is data collection, which involves several constraints concerning participant recruitment and data type. This results in data scarcity – a common aspect in studies of pathological disorders – which in turn leads to many data sets being imbalanced, that is, the number of participants for each class is not evenly distributed across the different categories [10], [24]. Moreover, demographic variables such as sex, age and education have shown impact on dementia diagnosis and development [10]. Therefore, these variables should also be considered when collecting dementia data [9], [10], [19], [38] since “conclusions drawn from imbalanced data are subject to a greater probability of bias, especially in small datasets” [10].

This problem is related to the evaluation techniques reported when manipulating imbalanced data sets since accuracy is not always appropriate for the task as it is not robust for imbalanced classes [10]. Nevertheless, it is the most common performance metric. According to [10], evaluation metrics like F1, precision and recall are more fitted for clinical studies because they “summarize the rates of false positives and false negatives” [10]. Another relevant metric is AUC (Area Under the Curve) [10].

Another main aspect to take into consideration when analyzing dementia in language is the variability and multi-faceted character of language [12]. As mentioned above, language depends on cognitive processes therefore changes in any language modality do not necessarily indicate an early stage of dementia but might also be caused by any other process involved in language, such as task performance. To avoid this, it is important to aim for robust feature sets that, ideally, are not language-dependent or task-dependent.

The third challenge is generalizability, which “is the degree to which a research approach may be attempted with different data, different settings or real practice” [10]. This is an essential part in dementia recognition in language “as it shows how translatable research is” [10]. [10] found that the majority of previous studies present low generalizability. Reasons for this include dependency on content, such as either manual annotations or word content, and targeting only one linguistic aspect, “particularly if this aspect is language-dependent” [10]. Moreover, most studies neither validate the results on external data sets or collect the data in real life settings nor study conversational speech, which is “more representative of real speech” [10].

In this paper, we aim to address these challenges by proposing ML methods that are transcription-free, content and language-independent and that do not require a large amount of data. We analyze phonetic representations to classify AD and MCI from HC since the smaller the processing unit is the more generalizable the results are across languages [10]. Furthermore, the best performing models are validated on an external data set containing conversational speech.

3. Materials and methods

3.1. Methodology

To obtain our objectives, we define the following steps:

- Data processing, where a data set is customized based on class labels. This data set contains audio files and transcripts obtained from recordings. The acoustic and transcript data is processed for further analysis. We also define the external validation set using conversational speech data.
- Baseline approach, where we define a baseline method based on demographic variables: age and sex.
- Feature extraction, in which we extract two types of features: acoustic and linguistic features. We focus on analyzing general feature sets that are less content-dependent and that have shown competitive results in other speech and linguistic tasks.
- Feature selection, where two methods are used: a correlation-based filter method and a feature selection technique based on importance of weights.
- Feature-based approach. We apply a SVM, a RF and a Gaussian Mixture Model (GMM) to both the acoustic and linguistic modality separately.
- Multimodal fusion, where we explore early and late feature-based fusion.
- Deep learning-based approach. The acoustic and linguistic modality are explored separately. For the acoustic modality, we propose three pre-trained computer vision CNNs and two self-supervised pre-trained ASR models: wav2vec XLSR and HuBERT. For the linguistic modality, we explore LSTM networks, FastText model and self-supervised pre-trained BERT model and RoBERTa model.

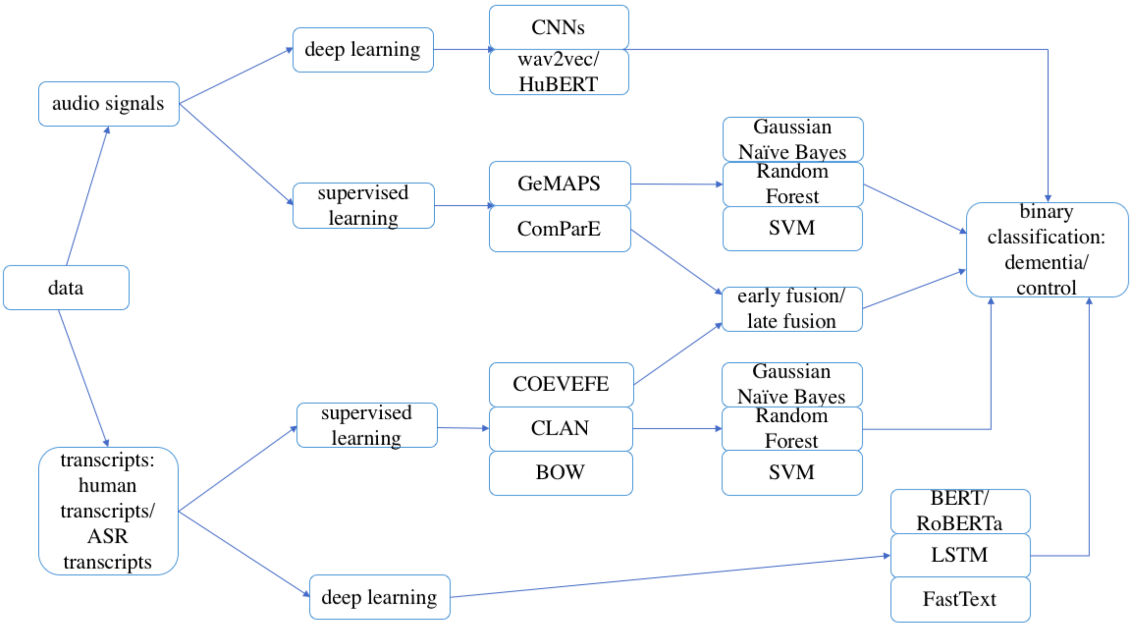


Figure 3. *Overview of the paper's methodology for dementia detection in language.*

3.2. Data set

The data set used in this project is based on the Picture Description task explained in section 2.1.: the English Pitt Corpus [55] collected at University of Pittsburgh School of Medicine and available at DementiaBank (https://dementia.talkbank.org/). This longitudinal study was conducted over the span of four years. It contains 551 speech samples and hand-made transcripts, of which 309 samples belong to the dementia group and 243 to the HC group. HC consists of patients that present no signs of neurocognitive decline. There are 398 individual participants: 208 dementia patients, 104 HC and 85 patients with unknown diagnosis. The dementia group contains probable AD, MCI and VD. According to [10], this is the "largest data set available for dementia recognition" [10].

The audio and transcript data consists of three sections: responses to the Cookie Theft Picture task, responses to the Word Fluency task, and responses to a recall task [55]. The recordings were collected in interviews where two speakers can be identified: the interviewer and the participant. Speech samples have been enhanced using an implementation of spectral subtraction performed for the ADReSS Challenge [15].

The ADReSS 2020 and ADReSSo 2021 challenges utilized a balanced subset of the Pitt Corpus containing only probable AD and HC samples. Nonetheless, since this paper's focus is early dementia detection, we consider MCI and AD groups as a broad group ("dementia"). Our motivation to jointly study MCI and AD samples is that, according to [10], "studies aiming for a three way classification have reported inconclusive results and present biases related to the quality of data set" [10]. In addition, classification accuracy tends to decrease in comparison to HC versus AD when more than one stage of dementia is considered [9]. This is, up to a great extent, due to the few available samples for MCI, an aspect that it is "consistent with the fact that distinguishing early stages of dementia from healthy speech is challenging even for trained experts" [9]. We consider this challenge makes the study of MCI an even more pressing matter.

We process the original data set to ensure a balanced one following the next steps: first, we drop all missing values and select patients with probable AD and MCI from the dementia group. This gives us a subset containing 389 samples: 214 samples with probable AD, 27 with MCI and the rest 148 with HC. We include AD and MCI in the same class: dementia. This subset is further broken down into train and test sets. Each set is under-sampled to contain the same number of classes. The final training set contains 118 control and 118 dementia samples while the testing set is composed of 30 control and 30 dementia samples. For reproducibility purposes, the original audio samples and transcripts are available at DementiaBank upon request, while the exact specification of the subset is available at: https://github.com/monicagoma/masters_thesis_dementia.

Regarding demographics, 62% of the samples correspond to females and 38% to males on the training and testing set both for dementia and HC samples. This means that sex is not balanced for class. However, as mentioned in [56], "it is estimated that worldwide, 61% of the people with dementia are women and 39% are men" [56]. Therefore, our data set's distribution (figure 4) is proximate to a realistic representation. In both the train and the test set, the average age is 68 years for the two classes, the minimum age for HC is 47 years and 49 years for dementia, and the maximum age for HC is 80 years and 90 years for dementia. Age and sex are not balanced for class. We consider balancing these variables produce an even smaller data set. To compensate this statistical bias, we provide thoughtful evaluation metrics explained in section 3.6.

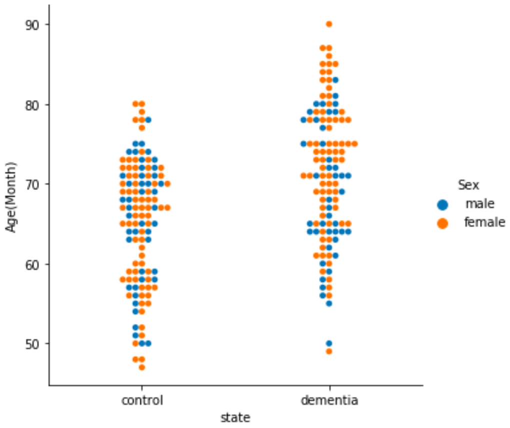


Figure 4*. Shows the age and sex distribution for the training set and held-out test set.*

3.3. External validation procedure

As mentioned in section 1.1., one important challenge in dementia recognition in language is its generalizability. However, the majority of previous studies "present neither external validation procedures nor a system design that involves them" [10]. This is a constraint in potential applications since promising results may transfer poorly to real-life scenarios. A straightforward solution is to evaluate the proposed methods on other databases. Nonetheless, there are few available data sets regarding dementia in language, of which the majority can be found at DementiaBank. As for the English language, there is no data set big enough to be used as an external validation set apart from the Pitt Corpus.

Therefore, we customize a validation set by employing various data sets in English available at DementiaBank. Specifically, we use the Wisconsin Longitudinal Study (WLS) [57], the Lanzi data set [6] and the Kempler data set [58]. We select these data sets for the following reasons: 1. None of them is sufficiently large to be used as an external validation set. 2. They present different protocols for data collection in contrast to the defined train and test sets, which focus on the Picture Description task. 3. They use American English analogous to the Pitt Corpus. 4. For the three data sets, there are transcripts and audio files available. 5. These transcripts are human-generated and comply with CHAT transcription requirements [59], an aspect that eases data preprocessing and feature extraction.

For the dementia group, we employ the Kempler and Lanzi data set. The Kempler data set contains seven conversations between investigators and participants diagnosed with AD. The range of age is between 65 and 87 years old. The Lanzi data set is composed of six files from semi-structured interviews. The interview is carried out using 11 openended questions. Participants present MCI [6]. Range of age is between 65 and 88. These two data sets use conversational speech data, which is considered “the most desirable data type in dementia detection” [10] since it is representative of spontaneous speech [10]. Up to our knowledge, previous work has not attempted to validate its results on external conversational speech data.

For the HC group, we use the WLS data set, which is an extended study of a sample of 10,317 students graduating from high school in Wisconsin in 1957 [57]. There, “participants were interviewed up to six times across 60-years between 1957 and 2011” [48]. For this research, we only focus on the last collected samples, i.e., those from 2011. Participants performed several language tasks, among them the Cookie Theft Picture description task. Since “the metadata of WLS do not provide dementia related diagnoses” [48] but only a “limited set of cognitive test scores” [48], we follow [48] methodology to select a group of HC comparable to the HC group in the Pitt Corpus. We use “the verbal fluency scores and an answer of ‘yes’ to the question ‘Have you ever been diagnosed with mental illness?’ as inclusion/exclusion criteria” [48]. Those who reported no mental illnesses and obtained verbal fluency scores above the normative threshold are considered the control group. From the resulting files, we select 13 participants: seven females and six males with a range of age between 60 and 72 in the year 2011. This external validation set is balanced for class but not for age and sex (figure 5).

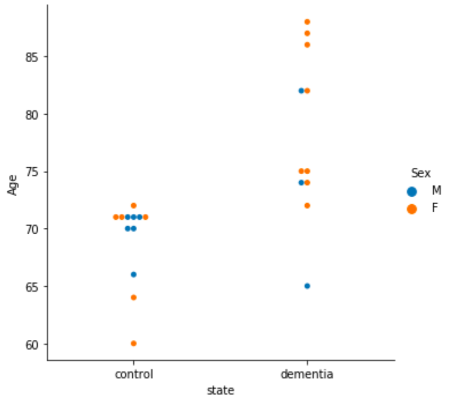


Figure 5*. Shows the age and sex distribution for the external validation set.*

3.4. Data processing

The first processing step is to obtain a uniformed sampling rate of all acoustic files by either up-sampling or down-sampling. The Pitt Corpus has a sampling rate of 44.1kHz, while the external validation set contains various sampling rates. We define

16kHz as the uniformed sampling rate since this is the sampling rate employed to pre-train the ASR models, HuBERT and wav2vec.

As mentioned above, the main data set and the external validation data set are extracted from interviews. Therefore, we carry out an acoustic segmentation based on speaker, also known as speaker diarization. Speaker diarization is the process of automatically identifying audio segments from different speakers. In other words, it is "answering the question: who spoke when?" [60]. This presents a challenging task especially due to background noises, overlapping utterances and the fact that speakers are rarely balanced during a conversation. In the case of the Pitt Corpus there are two participants: the interviewer and the patient. The first one asks the initial question but mostly remains silent as the patients describes the picture. This procedure, nonetheless, varies highly from interview to interview.

Speaker diarization is carried out using pyAudioAnalysis [60], a Python package that applies speaker diarization by employing a clustering approach. The steps are as following: "MFCCs in a short-term basis and its corresponding means and standard deviation are extracted" [60] while keeping a window size of 50 milliseconds [60]. Then, a k-means clustering method is carried out. The user has to provide the number of clusters, which refers to the number of speakers. If it is unknown, the model will apply the process on a wide range of clusters [60]. Finally, a smoothing step is performed "combining a median filtering on the extracted cluster and a Viterbi Smoothing step" [60]. In this paper, given the interview setting, we set the number of clusters as 2.

Moreover, to achieve full automation in AD and MCI detection pipeline, it is important to employ an ASR system for the transcriptions. Therefore, we propose two types of input: human-generated transcripts and ASR-generated transcripts. The former ones are available at DementiaBank database following CHAT transcription requirements [59]. We use the program CLAN version 10 to preserve the text information solely from the participant prescinding from the interviewer's transcripts. These human-generated transcripts already contain all relevant information to identify speakers. They also provide morphological and syntactical cues to simplify language analyses.

The latter use Google Cloud Speech-To-Text (STT) (https://cloud.google.com/speech-to-text). This is an open-domain ASR tool version 3.8.1. We choose this tool because it claims to reach a word error rate of 4.9% [61], which means a high performance considering "human transcripts report an average a word error rate of 4%" [61]. The aim of ASR-generated transcripts is to further automate the process of dementia recognition. We expect, nonetheless, that human-generated transcripts outperform ASR transcripts because elderly speech is difficult to transcribe by ASR tools [62], since “they are optimized on non-domain data” [62]. Additionally, a characteristic of the speech of dementia patients is “an increased amount of agrammatical sentences and incorrect word inflections” [62] that are not properly recognized by the ASR tool.

Given the small data set and due to memory constraints, we also perform a simple data augmentation method that consists of splitting the audio samples into small segments. This is carried out for each of the sets as implemented by [39]. Accordingly, the audio samples are cropped to a maximum duration of 8 seconds. As stated by [39], however, this method does not benefit neither acoustic feature-based nor transcript-based approaches. Thus, we decide to only use these samples for the acoustic deep learning methods and to obtain the ASR transcripts. Another reason to apply this method is that the proposed end-to-end ASR models not only are computationally expensive but also by using smaller segments performance may increase since it provides less dimensionality to the models [39].

3.5. Evaluation techniques

As we are employing a data set balanced for class, the proposed models are assessed in terms of accuracy, recall, precision and F1. We select these metrics since, according to [10], they are adequate to report the classification performance and clinically relevant to reduce potential bias [10]. Besides, these metrics have been reported in previous related research [14], [15].

Accuracy (Acc.) is the most reported metric in classification tasks because its understanding is intuitive. This metric is calculated using the following formula: $\frac{True\ Positives + True\ Negatives}{all\ samples}$.

Precision shows the proportion of positive identifications that are correct. It refers to “the ratio of the number of true positives to the number of instances classified” [63]. This metric is reported per class: dementia and control. Mathematically, it is expressed as: $\frac{True\ Positive}{True\ Positive + False\ Positive}$.

Recall is the proportion of true positives given the total number of instances. This metric is also known as sensitivity. It is reported per class as well. The formula to calculate it is: $\frac{True\ Positive}{True\ Positive + False\ Negative}$. F1 score “is calculated by weighting precision and recall equally and building an harmonic mean" [64]. The formula to calculate this metric is: $2 * \frac{Precision*Recall}{Precision+Recall}$. It is also reported per class. For the best-performing approaches, confusion matrices are available in Appendix. These figures illustrate the correct and incorrect predictions per class.

Moreover, to further report the rate of true positives and false positives, we report AUC and plot the Received Operating Characteristic Curve (ROC). This metric is only explored in the deep learning approaches since this is where the focus of this research lies. In addition, it measures a binary classifier's ability by considering different thresholds. AUC "measures the entire two-dimensional area underneath the entire ROC curve" [65]. ROC curves are plotted for the best-performing deep learning methods.

To be consistent with the literature, each feature-based approach is evaluated using 5-fold Cross-validation (CV) as an attempt to "mitigate the potential bias caused by using a small dataset" [10]. The final 5-fold CV performance is the average result obtained in each fold. The final model's performance is evaluated on the held-out test set that represents 20% of the subset.

To prevent overfitting, deep learning approaches are also evaluated using 5-fold CV except for the models: BERT, HuBERT, wav2vec and RoBERTa and FastText. This decision is due to constraints of GPU memory since these models have been pre-trained on big data. This choice is also in line with previous literature employing these architectures [39], [64]. Deep learning models are evaluated on the held-out test set and on an external data set, which contains conversational speech (section 3.3). To account for stochastic initialization [48], values are averaged across three runs.

3.6. Baseline model

To compare our proposed methods, we present a baseline model. This model is implemented by training a SVM classifier on two demographic variables – sex and age – without a feature selection technique. We select SVM because, according to [10], this is not only the most common model used in prior work but also the supervised learning algorithm with the highest performance in previous studies [10], [34]. Parameter optimization resulted in: a linear kernel and C value equal to 1. The final model is evaluated on the held-out test set (table 3).

Approach	**Acc.**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
Baseline 5-fold CV	.63	.70 / .55	.65 / .51	.63 / .63
Baseline held-out test set	.66	.69 / .64	.65 / .69	.73 / .60

Table 3*. Baseline results on 5-fold CV and held-out test set.*

3.7. Supervised learning

3.7.1. Feature extraction and selection

To obtain acoustic features, we focus on the temporal and prosodic characteristics of speech. Specifically, we automatically extract functionals of speech (section 2.1.1.) using the feature sets ComParE 2016 [66] containing 6,373 features and GeMAPS [67] composed of 62 functionals from the toolkit openSMILE 3.0 [30] (open-source Speech and Music Interpretation by Large-Space Extraction). These feature sets have been broadly studied and have shown competitive results in different speech classification tasks [10]. Examples for these features are MFCCs, F0, length of unvoiced segments, loudness, jitter, shimmer, among others.

In the case of linguistic features, we extract features using the FLUCALC toolkit from the CLAN program [59], which tracks the frequencies of the various fluency indicators available in the transcripts [59]. This toolkit extracts a total of 62 features such as word repetitions, blockings, retraces, monosyllabic repetitions, percentage of phrase repetitions and typical disfluencies [59]. Initially, this toolkit was created for stutter (i.e. diffluent speech) language analysis, which makes it especially suitable for dementia analysis since stuttering is usually present in early dementia [68].

Additionally, we extract linguistic features using the Core Variable Feature Extraction Feature Extractor (COVFEFE) tool [69]. This tool was introduced in [68] as a feature extraction tool for dementia recognition in speech and language. [17] applied COVFEFE's features on a SVM classification model and reported a 76% accuracy on the Picture Description task. The toolkit provides a total of 472 features. Features include lexical richness, percentage of parts-of-speech, readability, grammatical constituents of the syntax tree, etc. [68].

In addition to CLAN and COVEFE features, we apply Bag-Of-Words (BOW), a simple sentence representation technique, using the count vectorizer function from the scikit-learn library version 1.0.1, which analyzes the frequency of words. To do so, we remove punctuations, drop stopwords and tokenize each sentence. We select this method since it has been thoroughly investigated in various text classification tasks. Regarding dementia, [70] trained an artificial neural network with BOW and reported a 0.91 AUC [70]. One drawback of this method is that it only counts the word frequencies without taking the word order and sentence structure into account [71].

Given the large dimensionality of the features [72], we apply two feature selection methods: a filter method based on a statistical index of feature differences between classes, and a wrapper method based on weights. In the first method, we identify relevant features using the Analysis of Variance (ANOVA) ranking scheme, which is a common statistical method that analyses variables under various conditions. The threshold to select relevant features is to filter those features, of which the p-value is higher than 0.05 [73]. In the second method, we find the most relevant features according to the coefficients from fitting a Logistic Regression model defined with a maximum number of iterations of 1,000 [74]. The threshold for selecting the relevant features is a scaling factor: '1.25*mean' [74]. This intends to select those features whose importance is greater or equal than the threshold [74]. We select these two methods because they are simple feature selection techniques and have been widely investigated [75].

3.7.2. Proposed ML classifiers

We believe that, even though feature-based classifiers have not shown a performance as competitive as deep learning architectures in dementia in language (table 1 and 2), they are worth exploring since it is possible to determine each feature importance in the classification task. This means that these methods are explainable. That is relevant to the generalizability of studies since it refers to the capacity with which parameters can justify the results. Therefore, we propose three supervised learning models: SVM, RF and Gaussian Naïve Bayes (GNB). SVM and RF have shown the highest performance for supervised learning (table 1 and 2) and [38] reported high accuracy using GNB [38]. These approaches are developed using scikit-learn version 1.0.1 in Python 3.7.12. Each model is optimized using hyperparameters selected via grid search 5-fold CV on the training set. The two feature selection methods optimize the number of features for each model, except for GNB, which employs the feature selection technique that obtains the most promising results in SVM and RF. For reproducibility purposes, Appendix presents a detailed list of all the hyperparameters obtained from the grid search 5-fold CV for each model.

We also explore two multimodal approaches: early fusion and late fusion explained in section 2.1.3. To do so, we only focus on the best unimodal performing models. Parameter optimization is also performed via grid search 5-fold CV on the training set. In the case of late fusion, we apply weighted and unweighted majority voting. Unweighted majority voting is a technique in which "the final decision is the one where the majority of the classifiers reach a similar decision" [53]. In the case of weighted majority voting, the decision is reached by weighting each classifier's decision with the obtained accuracy. Majority voting approach is a well-known and broadly used technique, of which there are numerous examples of its successful application. For instance, [76] employed multiple classifiers for speaker identification and the "output scores of all classifiers were fused in a late integration approach to obtain the majority decision regarding the identity of the unknown speaker" [53], [76]. One advantage of this method is that "it works at the decision level by combining the prediction scores available for each modality" [54].

3.8. Deep learning

Feature-based approaches, nonetheless, usually "do not generalize well and often demand some level of domain expertise" [36]. Therefore, we explore deep learning models with special emphasis on transfer learning, which "refers to the use of existing knowledge in a different domain" [64]. This approach has allowed the development of many tasks where data collection is scarce, e.g., in health related ones. Given that this paper's main objective is to propose generalizable methods with potential applications, we believe transfer learning can provide a solution for this challenge.

In the case of the speech modality, we propose two methods: pre-trained CNNs of the computer vision domain and fine-tuning end-to-end ASR architectures. For the linguistic modality, we propose two types of input: human-generated transcripts and transcripts obtained from an ASR system. The reason is that, even though we expect human-generated transcripts to achieve a better performance, it is essential to explore ASR systems in order to reach a complete automation [10]. The two types of transcripts are applied to the following architectures: LSTM neural network, FastText and self-supervised pre-trained models BERT and RoBERTa. In the case of FastText method, which is the only method that does not support transfer learning, we explore the relevance of phonetic representations.

Deep learning models are developed using Pytorch version 1.10.0 in Python 3.7.12. Experiments are conducted on Google Colab Pro using a 25GB NVIDIA Tesla

K80 GPU. All models expect ASR architectures and BERT, RoBERTa and FastText are evaluated using 5-fold CV. These approaches are also evaluated on a held-out test and on an external validation set. To account for stochastic initialization, overall performance is averaged across three runs.

3.8.1. Acoustic modality

According to [36], for dementia detection in speech “few works have employed pre-trained models of the computer vision domain such as ResNet or MobileNet” [36]. These methods depend on spectrograms, which are obtained from slicing raw audio files into “overlapping windows of small-time frames and applying Fourier Transform to each window” [77]. “Fourier Transform decomposes a signal into its constituent frequencies” [77]. The result is an image (figure 7), which allows us to apply computer vision architectures to audio downstream tasks. The most common type of spectrogram is Mel-spectrogram, in which the amplitude from a linear scale is transformed to Mel scale [77]. “Mel scale aims to mimic the non-linear human ear perception of sound” [77]. To obtain the spectrograms, we use librosa library version 0.8.0. Spectrograms are extracted with 128 Mel bands, a hop length equal to 512 and a length of the Fast Fourier Transform (FFT) window of 2,048.

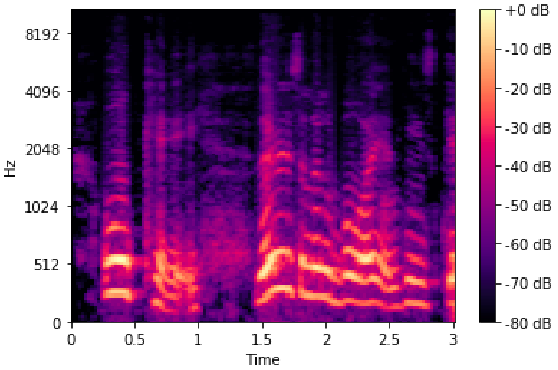


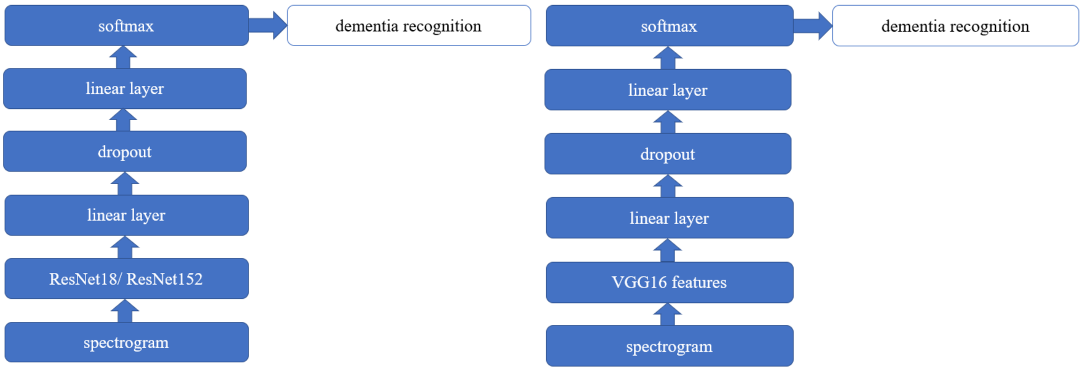
Figure 6. *Example of Mel-spectrogram used to feed pre-trained CNNs.*

One advantage of pre-trained CNNs on speech tasks is that these architectures “do not necessitate domain knowledge for their design” [78]. Consequently, this could lead to more successful models due to less assumptions taken regarding the task [78]. Drawbacks, nevertheless, are that this technique assumes that spectrograms are images instead of speech representations [78] and that during the spectrogram transformation some information from the wav file may be lost [78].

The pre-trained CNNs we propose are: VGG16, ResNet18 and ResNet152, which have been widely used in the computer vision domain [79]. VGG16 is characterized by adding more filters to increase the model's depth [79]. We choose this model because [36] employed it to detect dementia using Log-Mel spectrograms from the ADReSS speech data set. [36] reported a 56.65% accuracy. We aim to improve this performance by using Mel-spectrograms besides exploring MCI and AD instead of only early AD as done by the ADReSS data set [15].

ResNet18 and ResNet152 are residual neural networks introduced in [80]. These networks were trained on the ImageNet database – a data set that contains more than 100,000 images [80]. We select ResNet18 because it has shown as promising results as deeper neural networks such as DenseNet201 while being less computationally demanding [81]. We choose ResNet152 to compare results obtained with ResNet18. Furthermore, up to our knowledge, these two networks have not yet been investigated in speech classification tasks for AD and MCI.

In comparison to VGG neural networks like VGG16, ResNet has fewer filters and lower complexity [82]. It also uses less filters, which in turn makes them computationally faster than VGG [82]. These neural networks have many variants according to the number of layers. ResNet18 contains 18 deep layers and ResNet152 contains 152 deep layers.



***Figures** 7 **and** 8*. *Figure 7 (left) shows the architecture defined for ResNet18 and ResNet152 model, figure 8 (right) for VGG16.*

To train the CNNs, we keep the original structures and re-estimate the two last hidden layers to comply with our classification task (figures 7 and 8). After the second to last layer, a dropout layer is adopted with a rate of 0.3. We also set softmax as the activation function in the output layer. We define cross entropy loss as criterion since it is usually used to measure the difference between two likelihood distributions [79]. As optimizer we use Adam with an empirically defined learning rate of 0.001. The batch size to load the images is set to 32. The number of epochs is defined empirically: for ResNet152 is 35, for ResNet18 is 70 and VGG16 is 22.

In addition, given that our objective is to propose generalizable methods, we believe that the best practice to detect dementia in speech is by employing raw audio files without preprocessing steps like spectrogram transformation. This means using an end-to-end methodology. Hence, we apply end-to-end self-supervised ASR frameworks, wav2vec 2.0 and HuBERT, which, up to our knowledge, have not yet been studied in our task except for wav2vec 2.0, which was investigated for AD and MCI classification for the Mandarin Chinese language [39] (table 1). None of these models have been previously applied on the English language.

In the past two years, "speech representation learning has been increasingly dominated by self-supervised frameworks using contrastive loss" [83]. Among several, wav2vec 2.0 is the most common framework since it has shown a performance improvement and a decrease in training time [84]. This framework was initially developed as an ASR tool in order to tackle the lack of transcribed data. However, by fine-tuning pre-trained wav2vec 2.0, the model can be applied to downstream tasks.

Wav2vec 2.0 framework, as illustrated in figure 9, "contains three modules: feature encoder, quantization module and contextualized encoder. The first module consists of multiple convolution layers represented by the blue trapezoids (figure 9) and self-attention layers" [85]. The input is a one-dimensional raw waveform. The convolutional layers output latent speech representations. In the second module, "latent speech representations are learned in an unsupervised way; as a consequence, one has to build a ground truth to calculate a loss to optimize the Transformer in training" [64]. This ground truth is "created by quantization that transforms the continuous latent speech representation into a discrete vector called quantized representation" [64].

In the last module, some quantized representations are masked and fed into a Transformer. Transformers are built upon decoders and encoders, and they employ the concept of self-attention [64]. The model aims to predict the masked quantized representations using the contrastive loss function. The model's goal is to "optimize the contrastive loss, which enforces the model to identify the true quantized latent speech representations" [86].

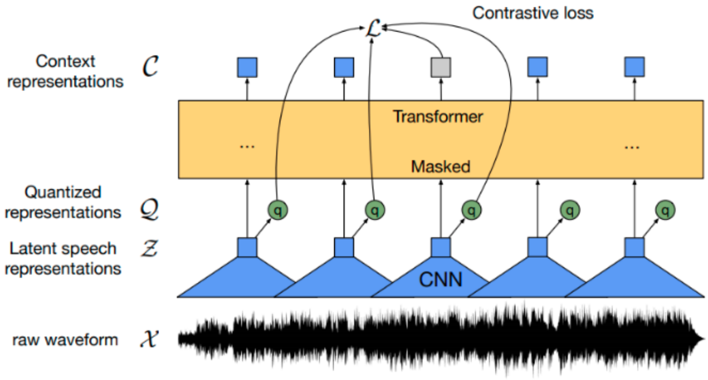


Figure 9. *Illustrates wav2vec 2.0 framework. Taken from* https://huggingface.co/.

Building upon wav2vec 2.0, Facebook released wav2vec 2.0 XLSR, which presents a key difference compared to wav2vec 2.0: it can learn speech representations across multiple languages. This is possible by pre-training the model using multi-lingual data sets. [64] states that "the most significant benefit of sharing these representations is the possibility to use language features from another language without pre-training the model again on a new language" [64], [87]. In this paper, we focus on wav2vec 2.0 XLSR. The reason is that this model could be applied to other languages, even though our data sets employ American English. Specifically, we fine-tune 'wav2vec2-large-xlsr-53', which is, until now, the biggest model released by Facebook. This model is available in the HuggingFace's Transformers library.

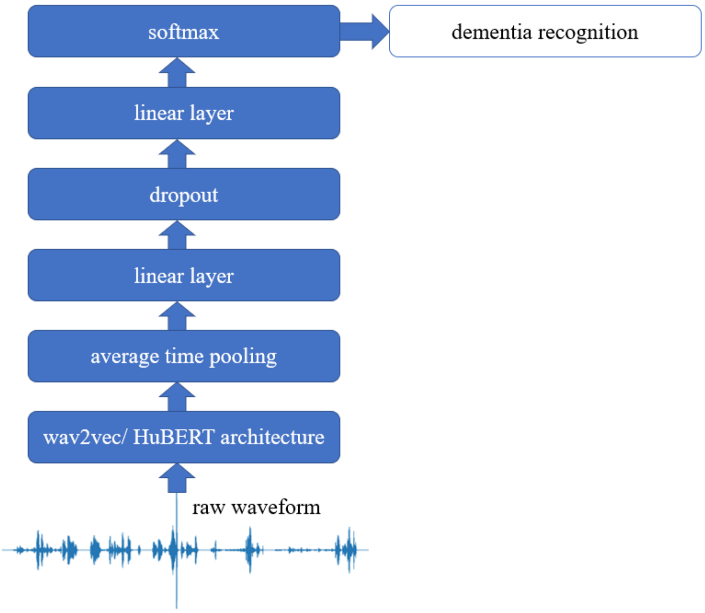


Figure 10. *Steps followed to fine tune wav2vec and HuBERT for dementia recognition.*

For fine-tuning, the batch size set is 1 primarily due to memory constraints. To be consistent with previous work fine-tuning wav2vec, we employ mean as the pooling mode strategy applied to merge 3D representations from an audio into 2D representations [85]. On top of the model, two linear layers are added to map the representations of wav2vec into our classification task. Softmax activation function is the last layer defined. The learning rate is set to $1e - 4$ and number of training epochs is empirically set to 10. We set a 0.3 dropout rate.

The second self-supervised ASR model applied in this paper is HuBERT (Hidden Unit BERT), which is a recent ASR tool released in June 2021 [88]. This model has key differences compared to the wav2vec 2.0 architecture: HuBERT uses "an offline clustering step to generate noisy labels for Masked Language Model pre-training" (figure 11) [86], [88]. This clustering is a "k-means clustering algorithm applied on 39-dimensional MFCC features" [88]. Another difference is that it employs cross-entropy loss as criterion whereas wav2vec uses contrastive loss function, and it "re-uses embeddings from BERT encoder to improve targets" [89]. According to [88], "the HuBERT model is forced to learn both acoustic and language models from continuous inputs" [88].

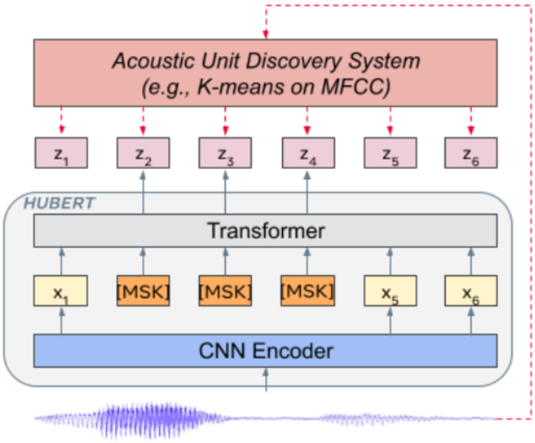


***Figure 11**. HuBERT framework. It "predicts hidden cluster assignments of the masked frames generated by one or more iterations of k-means clustering" [88]. Image taken from [88].*

To employ HuBERT, we follow the same steps as in figure 10. We apply mean as the pooling mode strategy and two linear layers on top of the HuBERT representations. We set a 0.3 dropout rate and softmax as activation function. The pre-trained model used is 'hubert-large-ls960-ft' obtained from HuggingFace's Transformers library. This model has been previously fine-tuned on 960 hours of Librispeech using the model 'hubert-large-ll60k'. Audio files are loaded using a batch size of 1 and we empirically select the number of epochs for fine-tuning as 7. The final optimization steps are 880. We expect wav2vec 2.0 XLSR and HuBERT architectures to outperform pre-trained CNNs from the computer vision domain. The reason is that these two systems have been pre-trained specifically on audio.

3.8.2. Linguistic modality

As mentioned above, for the linguistic modality we propose two types of input: human-generated transcripts and ASR-generated transcripts. These two data types are applied on the same architectures. We hypothesize that the human-generated transcripts will obtain the highest performance and the ASR-generated transcripts will perform better than the baseline model. Our proposed architectures are: LSTM, FastText, BERT and RoBERTa.

We explore LSTM neural network because this model has shown promising results in the task of dementia recognition, for instance, [37] obtained an 84% accuracy applying a bidirectional LSTM [37]. This is a popular recurrent neural network architecture, which is characterized by "adding an additional weight to the network to create cycles and maintain an internal state" [90]. This property enables us to capture long dependencies [90]. "One disadvantage of this network is that it struggles with long-range dependencies" [90].

To train LSTM, we use a pre-trained GloVe model vector (Global Vectors for word representations) [91], specifically 'glove.6B.100d', to extract lexical representations. This model vector contains 6B tokens, 400K vocabulary words and 100 dimensional vectors [91]. GloVe is "an unsupervised learning algorithm for obtaining word embeddings" [91]. The reason for selection this model vector is that it trains "on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space" [91].

The model is initialized with the GloVe word embeddings, setting the size of the vocabulary with a maximum value of 725. The batch size is set to 30. The number of hidden nodes is 32 and the number of layers is 2. Bi-direction is set to true. In total, the model has 131,957 trainable parameters. The most suitable number of epochs is empirically defined as 40. The criterion is cross entropy loss and the optimizer is Adam, as performed in [37]. Dropout is defined with a rate of 0.5 and softmax activation function is defined as the last layer.

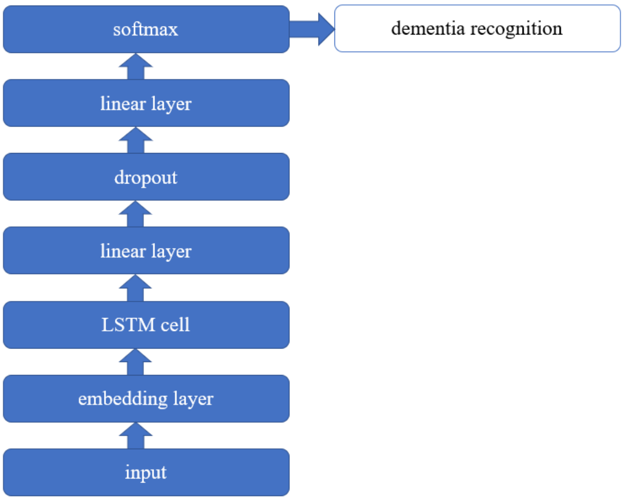


Figure 12. *LSTM architecture*

The second proposed model, FastText, has previously obtained competitive performance in dementia detection [35]. Even though this model is content-dependent since it depends on n-grams, [29] found that the model is able to successfully learn phoneme embeddings [29]. We select this architecture because this finding suggests that FastText could be a useful architecture for potential clinical applications since the smaller the processing unit is, the more generalizable a study is across languages [10]. This architecture contains two layers: an embedding layer, in which the word embeddings are calculated, and a linear layer [35]. This model can train and evaluate faster than many other deep models [92].

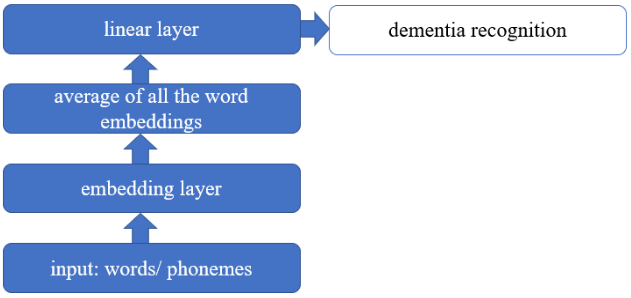


Figure 13. *FastText architecture*

To train FastText, we use trigrams. We also conducted experiments for bigrams and 4-grams without an improvement in performance. As mentioned in [35], “this bag of trigrams acts as additional features to capture some information about the local word order” [35]. To be consistent with previous studies [29], we set the number of epochs to 300 and a learning rate of 0.05. The embedding dimension is 100. To analyze the role of phonetic representations, we apply two types of inputs: text and phonemes following [29]'s methodology. An example for the first one is: 'the scene is in the kitchen'. The second one becomes: 'DH AH0 IS IY1 N IH1 Z IH0 N D AH0 K IH1 CH AH0 N'. This phonetic transcription is carried out using the CMUDict library. We keep the vowel stress in the pronunciation, which is represented by the numbers appended to the phonemes [29]. The reason is that [29] found that removing the stress reduced performance [29].

The third approach is the state-of-the-art BERT architecture. Recently, self-supervised pre-trained models using language model (LM) or masked-language model (MLM) have achieved promising results on a large variety of NLP tasks. Among them, BERT is the most successful structure, which was initially released by Google in late 2018 [46]. "BERT is built on Transformer networks to pre-train bidirectional representations of text by conditioning on both left and right contexts jointly in all layers" [19]. It is capable of learning dependencies beyond the fixed-length limitations [64].

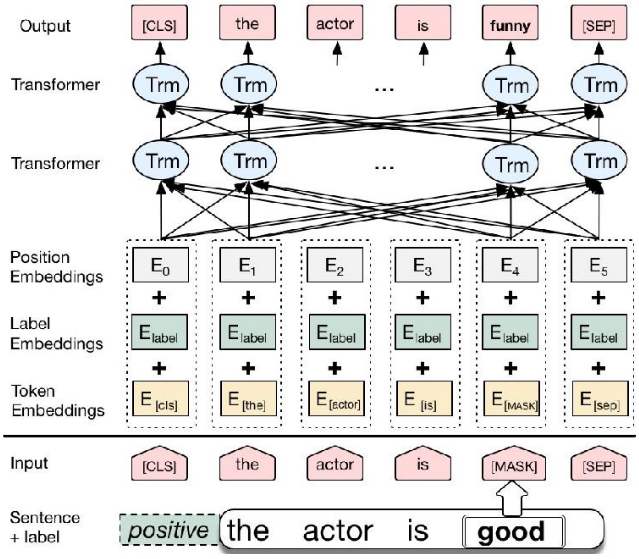


***Figure 14**. Example of BERT architecture. Figure taken from* [93].

BERT can take multiple sentences as input. "A special token, CLS, is prepended to the first segment, which can be used in an extra classification objective in addition to the MLM objective" [94]. "MLM enables BERT to learn to predict words within a sentence" without having to predict every next token [64]. Another special token, SEP, is inserted at the end of each segment to indicate segment boundaries [93]. Then, BERT masks 15% of the words [64] (figure 14). The aim is to predict the masked word and to "check if there is a word missing or an incorrect word in the sentence" [64].

One important difference between BERT and traditional neural networks is that it has access to the entire input whereas previous architectures such as LSTM usually have access to the input one by one. This means that BERT considers the words' context rather than learning sequentially. This is possible due to the use of attention modules, "which take into account other words in a unit of text when generating a word representation during pre-training and subsequent tasks" [48].

This model can be used in two different ways: as a high quality linguistic feature extractor or as a mean to fine-tune the model on a downstream task. In this paper, we focus on the second application because it allows us to explore a relevant transfer learning approach for linguistic data. BERT has also been proven successful for dementia detection, for example, [48] fine-tuned BERT and obtained an 81.9% accuracy. [50] combined wav2vec and BERT to detect dementia in speech. They reported an 83% accuracy. These studies have shown higher performance than those employing other types of word embeddings or handcrafted features.

In this research, we use the model 'bert-based-uncased' to initialize the classifier, which has been pre-trained on English and introduced in [46]. This model consists of 12 layers and was trained on BooksCorpus containing around 800 million words [46]. We add a classification layer at the end to map the BERT output to our task [19]. The maximum sequence length of the input is set to 512 tokens. Empirically, dropout is defined as 0.1. Batch size is set to 16 because previous work has defined this value as optimal for BERT [19]. We use Adam as optimizer and linear scheduling as learning rate. Theses parameters are chosen based on previous studies fine-tuning BERT, such as [19], [48]. During fine-tuning, we optimize the number of epochs to 5 within a range of 1 to 10. According to prior studies, BERT does not need a lot of epochs but rather has shown competitive performance by employing few epochs even on small data sets [19]. Label probabilities are computed using softmax.

Our fifth proposed method is RoBERTa architecture, which is built based on BERT's language masking strategy. RoBERTa, however, presents a key difference: it "removes BERT's next-sequence pre-trained objective and trains with much larger mini-batches and learning rates" [47]. We select this model because it has shown competitive performance in studies using human-generated transcript [49] and we believe it is necessary to address this architecture using ASR-generated transcripts in order to ensure the generalizability of results.

The model is initialized with the pre-trained ‘roberta-base’ model from the HuggingFace’s Transformers Library. This model was pre-trained on five large corpora of English [47]. We add a two-dimensional linear layer with softmax activation function. Dropout is empirically defined as 0.3. We use cross entropy loss as the loss function and Adam as optimizer with a learning rate of $1e-05$. Batch size is defined as 16 and the maximum input length is 512, following BERT implementation details. After trying different number of epochs, we set the final value to 7. We expect that BERT and RoBERTa will obtain the most promising results not only on the human transcripts but also on the ASR transcripts.

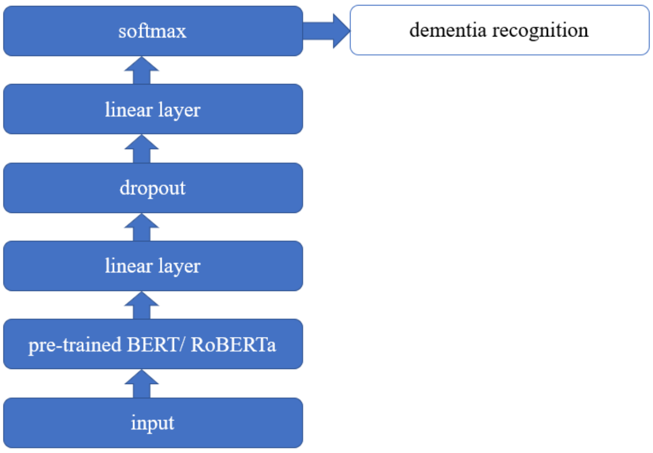


***Figure 15**. BERT and RoBERTa fine-tuning steps.*

4. Results

4.1. Feature-based approach

Appendix contains all parameters on which every model is trained on after parameter optimization and confusion matrixes for the best-performing models.

4.1.1. Acoustic-feature-based approach

Results on the 5-fold CV (table 4) and on the held-out set (table 5) show that the two feature sets, ComParE and GeMAPS, are successful at detecting dementia in speech since the two obtain results in line with previous work (table 1). The best-performing model – RF – employs the ComParE feature set. Moreover, results (table 5) suggest that the importance of weights is more effective than ANOVA. It is relevant to notice, nonetheless, that half of the models do not outperform the baseline model which obtained an accuracy of .66 (table 3).

Approach: **5-fold CV**	**FS**	**Acc.**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
SVM GeMAPS	weights	.67	.67 / .65	.67 / .67	.45 / .85
SVM GeMAPS	ANOVA	.62	.61 / .63	.63 / .62	.59 / .65
SVM ComParE	weights	.72	.62 / .78	1 / .64	.45 / 1
SVM ComParE	ANOVA	.67	.70 / .66	.66 / .71	.76 / .61
RF GeMAPS	weights	.69	.70 / .67	.67 / .70	.71 / .65
RF GeMAPS	ANOVA	.65	.67 / .64	.64 / .67	.69 / .61
RF ComParE	**weights**	**.72**	**.72 / .73**	**.73 / .72**	**.71 / .73**
RF ComParE	ANOVA	.70	.70 / .70	.69 / .71	.58 / .79
GNB ComParE	weights	.64	.59 / .67	.67 / .61	.53 / .73
GNB GeMAPS	weights	.65	.64 / .66	.66 / .65	.63 / .67

***Table 4**. Five-fold cross-validation results on training set. FS stands for feature selection. The best-performing model is displayed in bold.*

Approach: **held-out set**	**FS**	**Acc.**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
SVM GeMAPS	weights	.73	.75 / .71	.80 / .65	.77 / .70
SVM GeMAPS	ANOVA	.63	.61 / .66	.57 / .70	.65 / .62
SVM ComParE	weights	.73	.78 / .57	.69 / .78	.91 / .45
SVM ComParE	ANOVA	.68	.42 / .42	.60 / .74	.92 / .30
RF GeMAPS	weights	.63	.68 / .58	.69 / .57	.67 / .60
RF GeMAPS	ANOVA	.61	.69 / .54	.66 / .57	.72 / .51
RF ComParE	**weights**	**.74**	**.79 / .67**	**.75 / .74**	**.79 / .67**
RF ComParE	ANOVA	.63	.68 / .56	.68 / .57	.69 / .55
GNB ComParE	weights	.65	.72 / .54	.67 / .61	.77 / .49
GNB GeMAPS	weights	.67	.71 / .61	.71 / .60	.70 / .62

***Table 5**. Results obtained on the held-out test set for the acoustic modality.*

4.1.2. Linguistic-feature-based approach

Tables 6 and 7 demonstrate that most linguistic-feature-based models outperform the acoustic-feature-based models (table 5). From the three feature sets extracted, BOW obtains the highest performance outperforming the baseline model by a large margin (plus .12 accuracy). These findings suggest that the frequency of words is relevant to distinguish dementia language from control. Nonetheless, compared to previous work (table 2), performance is poorer. The reason is that most of the previous work has focused on deep learning techniques, for which we present results in section 4.2.2.

Approach: **5-fold CV**	**FS**	**Acc.**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
SVM COEVEFE	**weights**	**.79**	**.78 / .80**	**.83 / .78**	**.75 / .83**
SVM COEVEFE	ANOVA	.73	.71 / .74	.75 / .71	.68 / .77
SVM CLAN	weights	.65	.65 / .65	.66 / .64	.60 / .69
SVM CLAN	ANOVA	.70	.67 / .72	.65 / .67	.62 / .78
SVM BOW	weights	.69	.73 / .71	.72 / .77	.76 / .69
SVM BOW	ANOVA	.71	.57 / .69	.71 / .76	.77 / .65
RF COEVEFE	weights	.69	.81 / .75	.73 / .88	.91 / .56
RF COEVEFE	ANOVA	.76	.75 / .76	.78 / .74	.73 / .79
RF CLAN	weights	.64	.63 / .63	.65 / .67	.65 / .62
RF CLAN	ANOVA	.63	.59 / .64	.64 / .64	.62 / .64
RF BOW	weights	.78	.77 / .79	.72 / .77	.75 / .80
RF BOW	ANOVA	.74	.73 / .75	.74 / .75	.71 / .76
GNB COEVEFE	weights	.70	.68 / .72	.75 / .68	.64 / .77
GNB CLAN	weights	.74	.74 / .74	.74 / .74	.74 / .74
GNB BOW	weights	.56	.55 / .57	.57 / .55	.53 / .59

***Table 6**. Five-fold cross-validation results on training set.*

Approach: **held-out test**	**FS**	**Acc.**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
SVM COEVEFE	weights	.62	.62 / .61	.62 / .61	.61 / .62
SVM COEVEFE	ANOVA	.68	.72 / .55	.60 / .77	.88 / .43
SVM CLAN	weights	.73	.75 / .71	.71 / .77	.80 / .67
SVM CLAN	ANOVA	.75	.78 / .72	.70 / .83	.87 / .63
SVM BOW	weights	.73	.73 / .74	.74 / .73	.71 / .75
SVM BOW	ANOVA	.71	.59 / .76	1 / .62	.42 / 1
RF COEVEFE	weights	.61	.64 / .59	.60 / .63	.68 / .55
RF COEVEFE	ANOVA	.63	.67 / .59	.61 / .67	.75 / .52
RF CLAN	weights	.67	.66 / .68	.68 / .66	.63 / .70
RF CLAN	ANOVA	.74	.75 / .58	.62 / .87	.93 / .43
RF BOW	weights	.74	.71 / .77	.79 / .71	.65 / .83
RF BOW	**ANOVA**	**.78**	**.79 / .78**	**.80 / .77**	**.77 / .79**
GNB COEVEFE	weights	.57	.58 / .56	.57 / .57	.59 / .55
GNB CLAN	weights	.72	.74 / .69	.69 / .76	.80 / .63
GNB BOW	weights	.52	.54 / .49	.53 / .50	.55 / .48

***Table 7**. Results obtained on the held-out test set for the linguistic modality.*

4.1.3. Multimodal feature-based approaches

Given the results obtained in each individual modality, we explore two fusion methods: early fusion and late fusion. The idea is to boost performance by combining acoustic and linguistic information.

4.1.3.1.Early fusion

Early fusion focuses on the best-performing algorithms (tables 8 and 9): SVM and RF algorithms. We exclude GNB since it obtains the lowest performance in unimodal classification. The highest performance (.78) is achieved by SVM implementing BOW and ComParE feature sets using importance of weights as feature selection technique. The lowest accuracy (.65), however, is reached by the same architecture but using ANOVA as feature selection technique. Performance does not improve compared to unimodal linguistic classification (table 7). This finding encourages further exploration of acoustic feature sets.

Approach:	**FS**	**Acc.**	**F1**	**Precision**	**Recall**
5-fold CV			dementia / control	dementia / control	dementia / control
SVM CLAN + GeMAPS	ANOVA	.70	.73 / .65	.66 / .76	.83 / .57
SVM CLAN + GeMAPS	weights	.66	.60 / .70	.71 / .63	.52 / .79
SVM CLAN + ComParE	ANOVA	.72	.73 / .72	.72 / .73	.75 / .70
SVM CLAN + ComParE	weights	.74	67 / .79	.92 / .67	.52 / .96
SVM BOW + ComParE	ANOVA	.77	.74 / .78	.80 / .74	.70 / .83
SVM BOW + ComParE	weights	.77	.78 / .76	.73 / .81	.83 / .71
RF CLAN + GeMAPS	ANOVA	.60	.63 / .61	.60 / .64	.65 / .58
RF CLAN + GeMAPS	weights	.72	.75 / .70	.70 / .75	.79 / .65
RF BOW + ComParE	ANOVA	.74	.68 / .79	.87 / .69	.57 / .92
RF BOW + ComParE	**weights**	**.89**	**.89 / .90**	**.95 / .85**	**.83 / .96**
RF CLAN + ComParE	ANOVA	.68	.53 / .72	.82 / .60	.39 / .91
RF CLAN + ComParE	weights	.81	.82 / .80	.78 / .86	.88 / .75

Table 8. *Five-fold cross-validation results on training set.*

Approach:	**FS**	**Acc.**	**F1**	**Precision**	**Recall**
held-out set			dementia / control	dementia / control	dementia / control
SVM CLAN + GeMAPS	ANOVA	.72	.70 / .73	.74 / .70	.67 / .77
SVM CLAN + GeMAPS	weights	.68	.74 / .64	.80 / .58	.70 / .70
SVM CLAN + ComParE	ANOVA	.70	.71 / .69	.69 / .71	.73 / .67
SVM CLAN + ComParE	weights	.73	.81 / .59	.75 / .71	.88 / .50
SVM BOW + ComParE	ANOVA	.65	.67 / .63	.64 / .67	.70 / .60
SVM BOW + ComParE	**weights**	**.78**	**.79 / .78**	**.80 / .77**	**.77 / .79**
RF CLAN + GeMAPS	ANOVA	.66	.65 / .67	.68 / .64	.62 / .70
RF CLAN + GeMAPS	weights	.75	.78 / .71	.86 / .63	.72 / .80
RF BOW + ComParE	ANOVA	.74	.77 / .73	74 / .77	.81 / .69
RF BOW + ComParE	weights	.72	.71 / .73	.73 / .72	.70 / .75
RF CLAN + ComParE	ANOVA	.68	.73 / .61	.63 / .79	.87 / .39
RF CLAN + ComParE	weights	.67	.82 / .51	.72 / .85	.96 / .37

Table 9. *Results from the early fusion approach.*

4.1.3.2. Late fusion

Late fusion employs the predictions of the best-performing models on the held-out test set from each individual modality: RF using ComParE and BOW (tables 5 and 7). The late fusion approaches are unweighted majority voting and weighted majority voting explained in section 2.13. Results (table 10) show an improvement in performance by .01 compared to unimodal linguistic classification (table 7) applying weighted majority voting. Unweighted majority voting does not increase performance. This finding further stresses the exploration of other acoustic feature sets.

Approach	**Acc.**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
unweighted majority voting	.76	.80 / .76	.85 / .62	.75 / .76
weighted majority voting	**.79**	**.82 / .72**	**.87 / .67**	**.80 / .79**

Table 10. *Results from weighted and unweighted majority voting.*

4.2. Deep learning approach

4.2.1. Acoustic modality

From the five approaches for the acoustic modality only three (VGG16, ResNet152 and HuBERT) outperform the baseline model when evaluated on the held-out test set (table 14) and neither of the models outperform the baseline model when evaluated on external conversational speech data (table 15), in which the highest accuracy is .65. Results from the ResNet152 and HuBERT methods are in accord with the current state-of-art results (table 1). Table 13 suggests that VGG16 is overfitting. Results on the external validation set suggest that our proposed models are not robust enough to generalize on other types of data like conversational speech, except for ResNet152, which maintains a similar performance (table 13). Figure 16 displays the trade-off between true and false positives for HuBERT on the held-out and external validation set.

Approach: **5-fold CV**	**Acc.**	**AUC**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
CNNs (VGG16)	.61	.61	.38 / .72	.57 / 1	1 / .23
CNNs (ResNet18)	.54	.53	.14 / .68	1 / .52	.08 / 1
CNNs (ResNet152)	**.71**	**.74**	**.64 / .79**	**1 / .65**	**.47 / 1**

Table 13. *5-fold CV results on the training set.*

Approach: **held-out test set**	**Acc.**	**AUC**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
CNNs (VGG16)	.68	.68	.76 / .54	.61 / 1	1 / .37
CNNs (ResNet18)	.51	.51	.12 / .67	.67 / .51	.07 / .97
CNNs (ResNet152)	.68	.72	.76 / .61	.64 / .88	.93 / .61
wav2vec XLSR	.62	.63	.66 / .59	.66 / .59	.66 / .60
HuBERT	**.74**	**.78**	**.77 / .72**	**.77 / .72**	**.77 / .71**

Table 14. *Results for the acoustic modality on the held-out test.*

Approach: **external validation set**	**Acc.**	**AUC**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
CNNs (VGG16)	.65	.66	.53 / .73	.83 / .60	.38 / .92
CNNs (ResNet18)	.50	.50	0 / .63	0 / .50	0 / 1
CNNs (ResNet152)	**.69**	**.69**	**.73 / .64**	**.65 / .78**	**.85 / .54**
wav2vec XLSR	.47	.47	.26 / .16	.85 / .09	.16 / .76
HuBERT	.53	.58	.68 / .14	.52 / 1	.08 / 1

Table 15. *Results for the acoustic modality on the external validation set.*

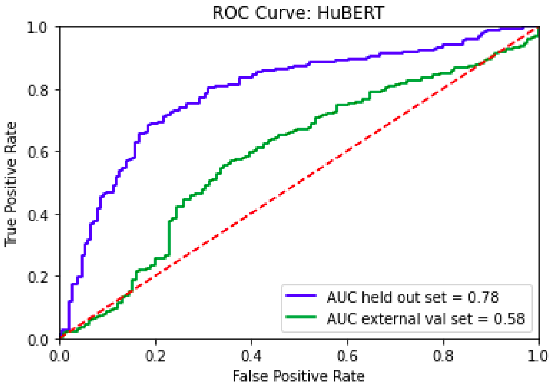


***Figure 16**. Rate of true positives and false positives for the HuBERT model on the held-out test set and external validation set.*

4.2.2 Linguistic modality

The linguistic modality explores two types of inputs: human-generated transcripts and ASR-generated transcripts. In the first case, results (tables 16 and 17) demonstrate accuracies are in line with previous work (table 2), except for LSTM GloVe and phoneme-based FastText. BERT (acc. .81) and RoBERTa (acc. .81) both achieve similar results.

Performance on the external validation set (table 18) does not decrease but rather increases on all methods apart from phoneme-based FastText. This finding demonstrates that these models generalize well on conversational data. Recall results on the external validation set (table 18) display high results in the dementia class, which indicates that it is easier for the models to identify true positives in conversational speech data. Figure 17 shows a better performance of BERT on the external validation set compared to the held-out test set.

Approach: **5-fold CV**	**Acc.**	**AUC**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
LSTM GloVe	.65	.73	.71 / .59	.88 / .39	.42 / .85

***Table 16**. Results for the linguistic modality on 5-fold CV train set. Transcripts are human-made.*

Approach: **held-out set**	**Acc.**	**AUC**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
LSTM GloVe	.56	.65	.53 / .60	.58/ .55	.42 / .69
BERT	**.81**	**.82**	**.79 / .81**	**.85 / .76**	**.73 / .87**
RoBERTa	.80	.80	.78 / .82	.88 / .75	.70 / .90
FastText	.75	.75	.75 / .73	.74 / .75	.76 / .73
FastText phonemes	.66	.61	.65 / .67	.67 / .65	.63 / .70

***Table 17**. Results for the linguistic modality on held-out test set. Transcripts are human-made.*

Approach: **external validation set**	**Acc.**	**AUC**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
LSTM GloVe	.72	.70	.76 / .66	.61 / 1	1 / .47
BERT	**.85**	**.88**	**.86 / .83**	**.76 / 1**	**1 / .71**
RoBERTa	.88	.82	.89 / .87	.80 / 1	1 / .70
FastText	.76	.76	.80 / .70	.66 / 1	1 / .53
FastText phonemes	.65	.63	.72 / .47	.57 / 1	1 / .30

Table 18*. Results for the linguistic modality on the external validation set. Transcripts are human-made.*

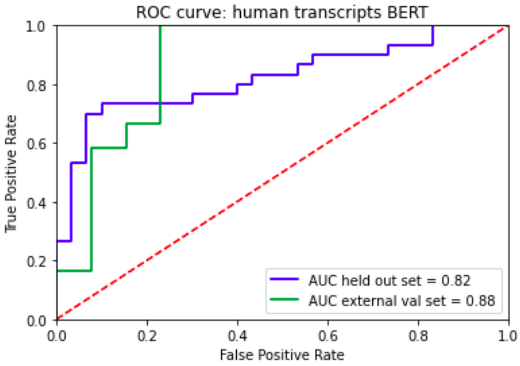


Figure 17*. ROC curves for BERT on the held-out test set and external validation set.*

Regarding the ASR-generated transcripts, none of the methods (tables 19, 20 and 21) outperform the baseline model. RoBERTa is the best-performing method with an acc. of .64 on the held-out test and .63 on the external validation. Compared to the human-generated transcript results (tables 17 and 18), there is a decrease in performance by a large margin. For example, a comparison between BERT (table 17) and RoBERTa (table 20) shows a drop of .17 acc. Figure 18 illustrates the low performance for the two sets. We assume this performance is due to the ASR tool used to obtain the transcripts. It is important to notice that the FastText model exploring phonemes as input achieves the lowest performance, which suggests that phonemes are not as effective as words embeddings to detect dementia.

Approach: **5-fold CV**	**Acc.**	**AUC**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
LSTM GloVe	.62	.63	.60 / .76	.51 / .78	.69 / .54

Table 19*. 5-fold CV on the training set. Transcripts are ASR-generated.*

Approach: **held-out set**	**Acc.**	**AUC**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
LSTM GloVe	.61	.66	.54/ .68	.61/ .66	.63/ .59
BERT	.63	.63	.68 / .57	.65/ .61	.72 / .53
RoBERTa	**.64**	**.56**	**.69 / .59**	**.67 / .63**	**.73 / .56**
FastText	.62	.62	.65 / .58	.65 / .58	.64 / .59
FastText phonemes	.60	.53	.63 / .57	.64 / .56	.62 / .58

Table 20*. Results for the linguistic modality on the held-out set. Transcripts are ASR-generated.*

Approach: **external validation set**	**Acc.**	**AUC**	**F1** dementia / control	**Precision** dementia / control	**Recall** dementia / control
LSTM GloVe	.57	.56	.60/ .20	.56/ .58	.64/ .50
BERT	.58	.61	.86 / .26	0.92 / .19	.80 / .41
RoBERTa	**.63**	**.60**	**.93 / .34**	**.93 / .37**	**.94 / .32**
FastText	.58	.57	.81 / .22	.91 / .14	.72 / .43
FastText phonemes	.57	.56	.74 / .21	.92 / .13	.61 / .52

Table 21*. Results for the linguistic modality on the external validation set. Transcripts are ASR-generated.*

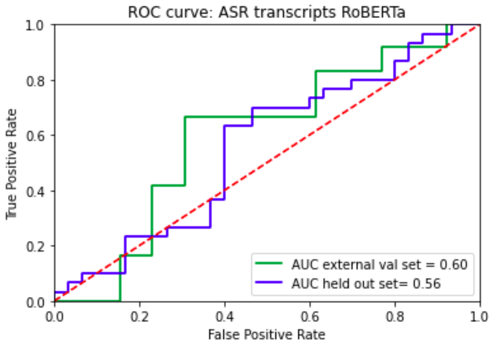


Figure 18. *ROC curves for RoBERTa on the held-out test set and the external validation set. Transcripts are ASR-generated.*

5. Discussion and limitations

5.1. Dementia in language

As shown in this paper, language analysis provides a non-invasive and accessible method for dementia recognition. However, in this analysis there are some factors to take into consideration. The first aspect is the complexity of language processing. “Speech and linguistic processes do not work in isolation and are usually intertwined with other cognitive and physical processes” [12]. It means that a language analysis for dementia not exclusively concerns the cognitive state of the speaker but, among others, his age, culture, sex and emotional state. Particularly, the age of the participants should be taken into consideration. As mentioned in [95], “a person’s age can be predicted with fair accuracy by speech characteristics including voice tremor, pitch, speaking rate, loudness and fluency” [95]. This aspect becomes of special interest when working with imbalanced data sets, since risk of statistical bias arises. If dementia patients are significantly older than the control ones, models might only detect the participants’ age.

This is also related to the Picture Description Task itself. As mentioned above, the Pitt Corpus employs this task to elicit speech samples from patients with an early type of dementia and no signs of neurocognitive decline. This task does not exclusively display language impairment but "cognitive performance of multiple neurocognitive functions" [12]. Therefore, a person with AD or MCI may have problems performing the task by not recalling common words or stuttering not necessarily due to language deficits but concentration issues [12]. Moreover, the task is limited by its scope and setting, especially since the Cookie Theft Picture is "an outdated depiction of domestic life" [17], [20] and the ability to describe it highly depends on cultural circumstances. The best solution to avoid this limitation is to employ conversational speech [10], since it represents a more natural and open type of conversation. By that, it tackles the mentioned disadvantages of elicited speech.

A further factor is model interpretability, which is essential for potential clinical applications as "clinicians' ability to interpret the model's decisions is crucial for adopting these technologies" [10]. Due to the scope of this paper, we have not addressed this issue. Future work should focus on inherently interpretable models such as decision trees and k-Nearest Neighbors [10].

Additionally, language in dementia presents various constraints in data collection. For example, recordings ought to be carried out in a medical context using adequate equipment and explicit consent from patients or relatives. These requirements, among other factors, cause the majority of available data sets to be relatively small compared to related fields such as emotion recognition [3], [8]. This limitation is distinctly noticeable in our external validation set, since it contains few samples and two different data types. This is important to take into consideration when drawing conclusions from the external validation results: They could have been influenced by the possibility of conversational speech being easier to identify than elicited speech. In addition, a small data set presents high variance. This constraint also causes incomplete data since participants might not finish the data collection process. The Pitt Corpus, for instance, has some participants with only one sample and some with more than one. This may add an additional bias to the obtained results. Moreover, our subset does not consider the education variable, which, according to [10], is a relevant factor in dementia. Future work should aim to consider years of education and ensure that participants are represented with a maximum of one sample in each the training set and the testing set.

5.2. Data preprocessing and baseline model

As mentioned in section 3.4., acoustic data is segmented based on each speaker using the pyAudioAnalysis library. This segmentation is challenging due to factors such as background noises, speaker overlapping and imbalance [60]. Therefore, it is common to notice overlapping speech between patient and interviewer even after speaker diarization. We consider this a possible bias in our acoustic models. In fact, in the case of the Picture Description task, the interviewer's speech contribution has shown an improvement in the dementia classification task [96]. This may be because "interviewers intuitively adapt their behavior to better communicate or interact with the patient" [96]. This influence of the interviewer's contribution should be further investigated by future work.

Our baseline model reaches a high accuracy (.66). This suggests that in the subset of the Pitt Corpus, even though the class is balanced, age and sex provide a relevant statistical bias to the data. This aspect could be avoided by either using a fully balanced data set like the ADReSS data set [15] or customizing a subset that is not exclusively balanced for class but also for education, age and sex. Due to the already small subset, we believe the second option is not feasible. To tackle this problem, we provide a thoughtful choice of evaluation methods (section 3.5.).

5.3. Feature-Based models

Acoustic feature sets, ComParE and GeMAPs, developed to detect emotions in speech [66], [67], are language-independent and do not rely on word content. Therefore, these feature sets offer a generalizable approach for dementia recognition. However, it is important to notice that acoustic features do not exclusively reflect the neurocognitive state of a person but the speaker's education, age, sex and emotional state. In other words, not all relevant features can be considered speech features for AD and MCI but there could be many other factors involved. Our results are in line with previous studies applying the ADReSS data set, which, as mentioned in section 2.1., is a subset of the Pitt Corpus focusing solely on AD detection. This finding suggests that MCI and AD can be jointly explored.

For the linguistic modality, the best-performing feature set is content-dependent (table 7). This suggests results would likely differ if applied to other data types (e.g., spontaneous speech) and languages or if transcripts were generated using an ASR system.

The CLAN feature set, which depends on fluency indicators, achieves a performance poorer than related work. Early and late fusion – against previous expectations – do not outperform the unimodal linguistic classification. This finding encourages further exploration of acoustic feature sets.

Concerning generalizability, one advantage that feature-based models offer is explainability, since they allow to determine each feature's importance in the classification tasks. Therefore, there is need for more curated features since our results are not promising enough, especially regarding acoustic features.

5.4. Deep Learning models

Our proposed deep learning approaches mostly focus on transfer learning since this method does not require a large amount of training data. Concerning the acoustic modality, the proposed methods show promising results on the held-out test set but a drop in performance when evaluated on external data except for ResNet152. This suggests that these three deep transfer learning models do not generalize well on conversational data. Moreover, VGG16 show signs of overfitting, this suggests other CV techniques should be considered.

As the results show, wav2vec XLSR does not outperform our proposed baseline model. HuBERT model reaches the highest accuracy on the held-out test set but presents a drop in performance on the external set. This is in line with [86]: "efforts on applying wav2vec and HuBERT to speech classification tasks have not yet proved a higher performance than traditional models" [86]. We hypothesize HuBERT outperforms wav2vec due to a key difference: the loss function. As mentioned in section 3.8.1., HuBERT employs cross-entropy loss whereas wav2vec uses contrastive loss, which is a more complex, but less stable loss function [88], [97]. We believe further research in this direction is necessary.

Another potential explanation for these results is that wav2vec and HuBERT have been trained on non-domain data. In order to properly fine-tune these two models on dementia detection, one would need to train them on dementia speech samples. This task is currently unfeasible since there are no big data sets for dementia available.

A further possible cause is the potential external noise created by the down-sampling and up-sampling techniques. These sampling techniques align the sampling rate of an audio file by reducing or increasing it. Wav2vec XLSR and HuBERT have been trained on raw audio files with a sampling rate of 16k Hz whereas the Pitt Corpus was recorded with a sampling rate of 41k Hz. Therefore, a down-sampling process is carried out (section 3.4.). This process can add external noise, which can interfere in the pre-trained ASR model implementation.

Furthermore, the lack of CV may lead to overfitting. Unfortunately, given memory constraints, CV techniques exceed our resources. Finally, ASR-based models might perform better with longer speech samples. In this paper, we employ a data augmentation method that consists of splitting the audio samples in 8-second segments. These models, however, were originally trained on segments which are around 1 minute long [88]. This suggests that wav2vec XLSR and HuBERT require segments longer than 8 seconds to learn speech representations.

It is important to highlight that neither wav2vec nor HuBERT models have been previously investigated in dementia detection in the English language. On that account, we cannot contextualize our results with previous studies. We believe that further research is needed since these methods offer a solution that does not require a lot of fine-tuning data, is end-to-end, content-independent and – in the case of wav2vec XLSR – language-independent.

For the linguistic modality, BERT and RoBERTa (using human-generated transcripts) achieve the best results on the held-out test set and external validation set. They outperform by a large margin, for example, BERT AUC outperforms LSTM GloVe by .17 on the held-out test set. These findings are in line with previous work and suggest that both models are robust enough to generalize on other data types. Similar to wav2vec XLSR and HuBERT, the lack of CV in BERT and RoBERT may lead to overfitting.

ASR-generated transcripts, on the other hand, present poor performance compared to human-generated transcripts. A possible explanation is that the Google Cloud STT ASR-tool may not be suitable to transcribe dementia speech since it is characterized by filled pauses, hesitations and "empty speech" (section 2.1) [62]. A solution on how ASR tools can improve the transcription is by adding speech pauses and punctuation [62]. These results highlight the relevance of either employing a transcription-free methodology – an acoustic-based approach – or developing ASR tools specifically for dementia speech recognition.

FastText using phoneme embeddings does not improve performance in comparison to the use of word embeddings. This finding suggests that phonetic representations do not augment dementia recognition. This is not line with what [29] reported. Therefore, we believe that future work should further explore the role of phonemes in dementia.

It is important to notice as well that, when evaluated on the external validation set, all proposed models are better at correctly classifying MCI and AD samples than those from the HC group. This could be a consequence from the subset used since the dementia group is defined using conversational speech whereas the control group contains only Picture Description task samples.

5.5. Future work and conclusion

This paper's objective is to propose ML methods that address the generalizability of results. Our study demonstrates that:

- AD and MCI can be jointly analyzed in comparison to HC. This joint analysis obtains performance in line with previous work based on AD.
- Transfer learning is a promising solution for the generalizability of results in dementia detection. Especially pre-trained CNNs from the computer vision domain (ResNet152) and pre-trained BERT and RoBERTa models.
- Linguistic-based approaches generalize well on external validation procedures using conversational speech. This methodology has not been previously applied in dementia recognition.
- The widely used Google Cloud STT ASR tool might not be suitable for automatically transcribing dementia speech.
- Fine-tuning end-to-end pre-trained HuBERT model shows potential to be a huge contribution to improve generalizability since it does neither depend on transcripts nor on content. This architecture has not been previously explored in dementia detection.
- Phonetic representations do not augment dementia recognition.
- Explainable linguistic features outperform explainable acoustic features.
- Acoustic features do not improve performance in an early and late multimodal fusion.

To support the obtained results, future work should replicate this study with balanced data sets. Our plan for future projects is to employ BERT and HuBERT embeddings in a deep multimodal fusion approach on our proposed subset of the Pitt Corpus and on the ADReSS data set. The aim is to combine the best-performing models in order to increase performance and tackle the main challenges identified in the two modalities: poor generalizability in the HuBERT model and a drop in performance of the BERT model when using ASR-transcripts. Additionally, to address these challenges, future work should also focus exclusively on conversational speech data and investigate other languages than English. BERT and wav2vec XLSR already provide a technique to apply the same models to different languages. Follow-up work should also focus on improving ASR transcriptions for dementia speech.

6. Bibliography

[1] “*What Is Dementia?”*. Accessed on: Oct. 10, 2021. [Online]. Available: https://www.cdc.gov/aging/dementia/index.html

[2] World Health Organization, *Dementia*. September, 2021. Accessed on: Oct. 10, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[3] G. Gosztolya *et al*, "Identifying Mild Cognitive Impairment and mild Alzheimer’s disease based on spontaneous speech using ASR and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181-197, 2019. Available: http://dx.doi.org/10.1016/j.csl.2018.07.007. DOI: 10.1016/j.csl.2018.07.007.

[4] K. B. Rajan, R. S. Wilson, J. Weuve, L. L. Barnes and D. Evans, "Cognitive impairment 18 years before clinical diagnosis of Alzheimer disease dementia," *Neurology*, vol. 85, *(10)*, pp. 898-904, 2015. Available: https://www.ncbi.nlm.nih.gov/pubmed/26109713. DOI: 10.1212/WNL.0000000000001774.

[5] D. Beltrami *et al*, "Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline?" *Frontiers Aging Neuroscience*, vol. 10, 2018. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2018.00369/full. DOI: 10.3389/fnagi.2018.00369.

[6] A. Lanzi, M. Bourgeois and S. Wallace, "GROUP EXTERNAL MEMORY AID TREATMENT FOR MILD COGNITIVE IMPAIRMENT," *Alzheimer's & Dementia*, vol. 13, *(7)*, pp. 257, 2017. Available: https://dx.doi.org/10.1016/j.jalz.2017.06.121. DOI: 10.1016/j.jalz.2017.06.121.

[7] A. T. Patocskai *et al*, "Is there any difference between the findings of Clock Drawing Tests if the clocks show different times?" *Journal of Alzheimer's Disease*, vol. 39, *(4)*, pp. 749-757, 2014. Available: https://www.ncbi.nlm.nih.gov/pubmed/24270210. DOI: 10.3233/JAD-131313.

[8] G. Szatloczki, I. Hoffman, V. Vincze, J. Kalman and M. Pakaski, "Speaking in Alzheimer’s Disease, is That an Early Sign? Importance of Changes in Language Abilities in Alzheimer’s Disease," *Frontiers Aging Neuroscience*, vol. 7, 2015. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2015.00195/full. DOI: 10.3389/fnagi.2015.00195.

[9] R. Chakraborty, M. Pandharipande, C. Bhat, S. K. Kopparapu, "Identification of Dementia Using Audio Biomarkers," 2020. Available: https://arxiv.org/abs/2002.12788

[10] S. de la Fuente Garcia, C. W. Ritchie and S. Luz, "Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review," *Journal of Alzheimer's Disease*, vol. 78, *(4)*, pp. 1547-1574, 2020. Available: https://www.ncbi.nlm.nih.gov/pubmed/33185605. DOI: 10.3233/JAD-200888.

[11] P. Pastoriza-Domínguez *et al*, "Speech pause distribution as an early marker for Alzheimer’s disease," *Speech Communication*, vol. 136, pp. 107-117. 2022. Available: https://www.sciencedirect.com/science/article/pii/S0167639321001333.DOI:10.1101/2020. 12.28.20248875.

[12] H. Lindsay, J. Tröger and A. König, "Language Impairment in Alzheimer’s Disease—Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through

Multilingual Machine Learning," *Frontiers in Aging Neuroscience*, vol. 13, *(11)*, 2021. Available: https://search.proquest.com/docview/2529003568. DOI: 10.3389/fnagi.2021.642033.

[13] B. Klimova and K. Kuca, "Speech and language impairments in dementia," *Journal of Applied Biomedicine*, vol. 14, *(2)*, pp. 97-103, 2016. Available: https://dx.doi.org/10.1016/j.jab.2016.02.002. DOI: 10.1016/j.jab.2016.02.002.

[14] S. Luz, F. Haider, S. de la Fuente, D. Fromm, B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo Challenge," 2021. Available: https://arxiv.org/abs/2104.09356. DOI: 10.1101/2021.03.24.21254263.

[15] S. Luz, F. Haider, S. de la Fuente, D. Fromm, B. MacWhinney, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," *Proc. Interspeech 2020*, pp. 2172-2176, 2020. DOI: 10.21437/Interspeech.2020-2571.

[16] M. Rohanian, J. Hough and M. Purver, "Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech," *Interspeech 2020*, pp. 2187-91, 2020. Available: https://www.isca-speech.org/archive/interspeech_2020/rohanian20_interspeech.html. DOI: 10.21437/interspeech.2020-2721.

[17] N. Clarke, T. R. Barrick and P. Garrard, "A Comparison of Connected Speech Tasks for Detecting Early Alzheimer's Disease and Mild Cognitive Impairment Using Natural Language Processing and Machine Learning," *Frontiers in Computer Science (Lausanne)*, vol. 3, 2021. Available: https://doaj.org/article/ef7ae92f93544eefacafacbab4dfa2cd. DOI: 10.3389/fcomp.2021.634360.

[18] H. Goodglass and E. Kaplan, "Boston Diagnostic Aphasia Examination Booklet," *PA: Lea & Febiger.*, 1983.

[19] A. Balagopalan, B. Eyre, F. Rudzicz and J. Novikova, "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection," *Proc. Interspeech 2020*, 2020. Available: https://arxiv.org/abs/2008.01551. DOI: 10.21437/Interspeech.2020-2557.

[20] S. Berube *et al*, "Stealing Cookies in the Twenty-First Century: Measures of Spoken Narrative in Healthy Versus Speakers With Aphasia," *American Journal of Speech-Language Pathology*, vol. 28, *(1S)*, pp. 321-329, 2019. Available: https://www.ncbi.nlm.nih.gov/pubmed/30242341. DOI: 10.1044/2018_AJSLP-17-0131.

[21] K. C. Fraser, J. A. Meltzer and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, *(2)*, pp. 407-422, 2016. Available: https://www.ncbi.nlm.nih.gov/pubmed/26484921. DOI: 10.3233/JAD-150520.

[22] K. Simonyan, B. Horwitz and E. D. Jarvis, "Dopamine regulation of human speech and bird song: A critical review," *Brain and Language*, vol. 122, *(3)*, pp. 142-150, 2012. Available: https://dx.doi.org/10.1016/j.bandl.2011.12.009. DOI: 10.1016/j.bandl.2011.12.009.

[23] T. Johnstone, "The effect of emotion on voice production and speech acoustics," Ph.D dissertation, Psychology Department, Univ. of Western Australia, 2017.

[24] L. Moro-Velazquez *et al*, "Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's Disease," *Scientific Reports*, vol. 9, *(1)*, pp. 19066-16, 2019. Available: https://www.ncbi.nlm.nih.gov/pubmed/31836744. DOI: 10.1038/s41598-019-55271-y.

[25] O. Hornykiewicz, "Chemical neuroanatomy of the basal ganglia — normal and in Parkinson's disease," *Journal of Chemical Neuroanatomy*, vol. 22, *(1)*, pp. 3-12, 2001. Available: https://dx.doi.org/10.1016/S0891-0618(01)00100-4. DOI: 10.1016/S0891-0618(01)00100-4.

[26] M. L. Cera, K.S. Ortiz, P. H. Ferreira and T. S. Minett, "Speech and orofacial apraxias in Alzheimer's disease," *International Psychogeriatrics*, vol. 25, *(10)*, pp. 1679-1685, 2013. Available: https://dx.doi.org/10.1017/S1041610213000781. DOI: 10.1017/S1041610213000781.

[27] W. S. Horton, "Theories and approaches to the study of conversation and interactive discourse," in *The Routledge Handbook of Discourse Processes*, 2nd ed., Routledge, 2018, pp. 22-68.

[28] K. Croot, J. Hodges, J. Xuereb and K. Patterson, "Phonological and Articulatory Impairment in Alzheimer's Disease: A Case Series," *Brain and Language*, vol. 75, *(2)*, pp. 277-309, 2000. Available: https://dx.doi.org/10.1006/brln.2000.2357. DOI: 10.1006/brln.2000.2357.

[29] E. Edwards *et al*, "Multiscale System for Alzheimer's Dementia Recognition Through Spontaneous Speech," *Proc. Interspeech 2020*, 2020. Available: https://www.isca-

speech.org/archive/interspeech_2020/edwards20_interspeech.html. DOI: 10.21437/interspeech.2020-2781.

[30] Eyben Florian, Wöllmer Martin, Schuller Björn, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," *Proc.ACM Multimedia (MM)*, pp. 1459-1462, 2010.

[31] A. Satt *et al*, "Speech-Based Automatic and Robust Detection of Very Early Dementia," *Proc. Interspeech 2014*, 2014. Available: https://www.isca-speech.org/archive/interspeech_2014/satt14_interspeech.html. DOI: 10.21437/Interspeech.2014-544.

[32] B. Schuller *et al*, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," *Proc. Interspeech 2007*, 2007. Available: https://www.semanticscholar.org/paper/The-relevance-of-feature-type-for-the-automatic-of-Schuller-Batliner/af089c9240ad42586d0be0e12d1e5ab324703853.

[33] A. Triantafyllopoulos *et al*, "Multistage linguistic conditioning of convolutional layers for speech emotion recognition," 2021. Available: https://arxiv.org/abs/2110.06650.

[34] I. Vigo, L. Coelho and S. Reis, "Speech- and Language-Based Classification of Alzheimer's Disease: A Systematic Review," *Bioengineering (Basel)*, vol. 9, *(1)*, pp. 27, 2022. Available: https://www.ncbi.nlm.nih.gov/pubmed/35049736. DOI: 10.3390/bioengineering9010027.

[35] A. Meghanani, C. S. Anoop and A. G. Ramakrishnan, "An exploration of log-mel spectrogram and MFCC features for alzheimer's dementia recognition from spontaneous speech," *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 670-677, 2021. DOI: 10.1109/SLT48900.2021.9383491.

[36] L. Ilias, D. Askounis and J. Psarras, "Detecting Dementia from Speech and Transcripts using Transformers," 2021. Available: https://arxiv.org/abs/2110.14769.

[37] N. Cummins *et al*, "A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition," *Proc. Interspeech 2020*, pp. 2182-86, 2020. Available: https://www.isca-speech.org/archive/interspeech_2020/cummins20_interspeech.html. DOI: 10.21437/interspeech.2020-2635.

[38] A. Balagopalan and J. Novikova, "Comparing Acoustic-based Approaches for Alzheimer's Disease Detection," *Proc. Interspeech 2021*. 2021. Available: https://www.isca-speech.org/archive/interspeech_2021/balagopalan21_interspeech.html. DOI: 10.21437/Interspeech.2021-759.

[39] Y. Qin *et al*, "Exploiting Pre-Trained ASR Models for Alzheimer's Disease Recognition Through Spontaneous Speech," *Proc. NCMMSC2021*, 2021. Available: https://arxiv.org/abs/2110.01493.

[40] P. Mahajan and V. Baths, "Acoustic and Language Based Deep Learning Approaches for Alzheimer's Dementia Detection From Spontaneous Speech," *Frontiers in Aging Neuroscience*, vol. 13, 2021. Available: https://www.ncbi.nlm.nih.gov/pubmed/33613269. DOI: 10.3389/fnagi.2021.623607.

[41] A. König *et al*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, *(1)*, pp. 112-124, 2015. Available: https://pubmed.ncbi.nlm.nih.gov/27239498/ . DOI: 10.1016/j.dadm.2014.11.012.

[42] E. L. Campbell *et al*, "Alzheimer's Dementia Detection from Audio and Text Modalities," 2020. Available: https://arxiv.org/abs/2008.04617.

[43] L. B. d. Santos *et al*, "Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts," *ACL*, 2017. Available: https://www.semanticscholar.org/paper/Enriching-Complex-Networks-with-Word-Embeddings-for-Santos-J%C3%BAnior/c21a76d1f48fc7163a8f532a78fab183ca168fe6.

[44] S. A. Sajjadi *et al*, "Abnormalities of connected speech in semantic dementia vs Alzheimer's disease," *Aphasiology*, vol. 26, *(6)*, pp. 847-866, 2012. Available: http://www.tandfonline.com/doi/abs/10.1080/02687038.2012.654933. DOI: 10.1080/02687038.2012.654933.

[45] F. Di Palo and N. Parde, "Enriching Neural Models with Targeted Features for Dementia Detection," ACL, 2019. Available: https://aclanthology.org/P19-2042/.

[46] J. Devlin *et al*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018. Available: https://arxiv.org/abs/1810.04805.

[47] Y. Liu *et al*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019. Available: https://arxiv.org/abs/1907.11692.

[48] Y. Guo, C. Li, C. Roan, S. Pakhomov, T. Cohen, "Crossing the “Cookie Theft” Corpus Chasm: Applying What BERT Learns From Outside Data to the ADReSS Challenge Dementia Detection Task," *Frontiers in Computer Science (Lausanne)*, vol. 3, 2021. Available: https://doaj.org/article/417d2905f8ed446884c6ff7f860e4453. DOI: 10.3389/fcomp.2021.642517.

[49] M. S. S. Syed, Z. S. Syed, M. Lech and E. Pigorova, "Automated Screening for Alzheimer’s Dementia Through Spontaneous Speech," *Proc. Interspeech 2020*, 2020. Available: https://www.isca-speech.org/archive/interspeech_2020/syed20_interspeech.html. DOI: 10.21437/interspeech.2020-3158.

[50] Y. Zhu *et al*, "WavBERT: Exploiting Semantic and Non-semantic Speech using Wav2vec and BERT for Dementia Detection," *Proc. Interspeech 2021*, pp. 3790-94, 2021. Available: https://www.isca-speech.org/archive/pdfs/interspeech_2021/zhu21e_interspeech.pdf. DOI: 10.21437/Interspeech.2021-332.

[51] R. Haulcy and J. Glass, "Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech," *Frontiers in Psychology*, vol. 11, pp. 624137, 2020. Available: https://www.ncbi.nlm.nih.gov/pubmed/33519651. DOI: 10.3389/fpsyg.2020.624137.

[52] F. Chen *et al*, "Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis," 2019. Available: https://arxiv.org/abs/1904.08138.

[53] P. K. Atrey, M. A. Hossain, A. E. Saddik and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, *(6)*, pp. 345-379, 2010. Available: https://link.springer.com/article/10.1007/s00530-010-0182-0. DOI: 10.1007/s00530-010-0182-0.

[54] E. Morvant, A. Habrard and S. Ayache, "Majority vote of diverse classifiers for late fusion," in *Structural, Syntactic, and Statistical Pattern Recognition*, Springer, Berlin Heidelberg, pp. 153-162, 2014.

[55] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, J. and K. L. McGonigle, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, *(6)*, pp. 585-594, 1994. Available: https://dementia.talkbank.org/access/0docs/Becker1994.pdf.

[56] *Women and dementia a marginalised study*, 2015. Accessed on: Oct. 10, 2021. [Online]. Available:https://www.alzheimersresearchuk.org/about-us/our-influence/policy-work/reports/women-dementia/

[57] P. Herd, D. Carr and C. Roan, "Cohort Profile: Wisconsin longitudinal study (WLS)," *International Journal of Epidemiology*, vol. 43, *(1)*, pp. 34-41, 2014. Available: https://www.ncbi.nlm.nih.gov/pubmed/24585852. DOI: 10.1093/ije/dys194.

[58] D. Kempler, S. Curtiss and C. Jackson, "Syntactic Preservation in Alzheimer's Disease," *Journal of Speech and Hearing Research*, vol. 30, *(3)*, pp. 343-350, 1987. Available: http://jslhr.asha.org/cgi/content/abstract/30/3/343. DOI: 10.1044/jshr.3003.343.

[59] B. MacWhinney, *The Childes Project.* Taylor and Francis, 2014.

[60] T. Giannakopoulos, "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis," *PloS One*, vol. 10, *(12)*, 2015. Available: https://www.ncbi.nlm.nih.gov/pubmed/26656189. DOI: 10.1371/journal.pone.0144610.

[61] H. Chen, "Does Word Error Rate Matter?," 2021. Accessed on: Jan. 23, 2022. [Online]. Available: https://www.smartaction.ai/blog/does-word-error-rate-matter/

[62] L. Toth *et al*, "A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech," *Current Alzheimer Research*, vol. 15, *(2)*, pp. 130-8, 2018. Available: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1567-2050&volume=15&issue=2&spage=130. DOI: 10.2174/1567205014666171121114930.

[63] S. Luz, S. de la Fuente and P. Albert, "A Method for Analysis of Patient Speech in Dialogue for Dementia Detection," 2018. Available: https://arxiv.org/abs/1811.09919.

[64] P. Fivian and D. Reise, "Speech Classification using Wav2vec 2.0.," B.S Thesis, Zürcher Hochschule für Angewandte Wissenchaften, Zurich, 2021.

[65] *Classification: ROC Curve and AUC*. Accessed on: Jan. 13, 2022. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es_419

[66] B. Schuller *et al*, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," *Proc. Interspeech 2013*, 2013. Available:

https://www.isca-speech.org/archive/interspeech_2013/schuller13_interspeech.html. DOI: 10.21437/Interspeech.2013-56.

[67] F. Eyben *et al*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *T-Affc*, vol. 7, *(2)*, pp. 190-202, 2016. Available: https://ieeexplore.ieee.org/document/7160715. DOI: 10.1109/TAFFC.2015.2457417.

[68] J. O. de Lira, T. S. Minnet, P. H. Ferreira and K. Z. Ortiz, "Analysis of word number and content in discourse of patients with mild to moderate Alzheimer's disease," *Dementia & Neuropsychology*, vol. 8, *(3)*, pp. 260-265, 2014. Available: https://www.ncbi.nlm.nih.gov/pubmed/29213912. DOI: 10.1590/S1980-57642014DN83000010.

[69] M. Komeili *et al*, "Talk2Me: Automated linguistic data collection for personal assessment," *PloS One*, vol. 14, *(3)*, pp. e0212342, 2019. Available: https://www.ncbi.nlm.nih.gov/pubmed/30917120. DOI: 10.1371/journal.pone.0212342.

[70] P. Klumpp, J. Fritsch and E. Noeth, "ANN-based alzheimer's disease classification from bag of words," *ITG-Symposium 282 - Speech Communication*, pp. 1-4, 2018. Available: https://ieeexplore.ieee.org/document/8578051.

[71]J. Blanco, *Hacking Scikit-Learn's Vectorizers*. Accessed on Jan. 13, 2022. Available: https://towardsdatascience.com/hacking-scikit-learns-vectorizers-9ef26a7170af.

[72] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, *(1)*, pp. 273-324, 1997. Available: https://dx.doi.org/10.1016/S0004-3702(97)00043-X. DOI: 10.1016/S0004-3702(97)00043-X.

[73] S. Hu, "Complete Feature Selection Techniques 4-1 Statistical Test & Analysis,". Accessed on Dec. 05, 2021. Available: https://summer-hu-92978.medium.com/complete-feature-selection-techniques-4-1-statistical-test-analysis-611ede242fa0

[74] Scikit-Learn, *sklearn.feature_selection.SelectFromModel*. Accessed on Dec. 29, 2021. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html#sklearn.feature_selection.SelectFromModel.

[75] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Mach. Learn. Res., pp. 1157–82, 2003. Available: https://dl.acm.org/doi/10.5555/944919.944968.

[76] V. Radova and J. Psutka, "An approach to speaker identification using multiple classifiers," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1135-38, 1997. Available: https://ieeexplore.ieee.org/document/596142. DOI:10.1109/ICASSP.1997.596142.

[77] H. Sura, *Audio Classification using Librosa and Pytorch*, 2020. Accessed on Dec. 29, 2021. Available: https://medium.com/@hasithsura/audio-classification-d37a82d6715

[78] J. Pons. *Why do spectrogram-based VGGs rock?*. 2019. Accessed on Dec. 12, 2021. Available: https://towardsdatascience.com/why-do-spectrogram-based-vggs-rock-6c533ec0235c.

[79] E.J. Velo-Fuentes, "Introducción a los métodos Deep Learning basados en Redes Neuronales," M.S Thesis, Universidad de Santiago de Compostela, 2020. Available: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1654.pdf.

[80] K. He, X. Zhang, S. Ren, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016. Available: https://ieeexplore.ieee.org/document/7780459 . DOI: 10.1109/CVPR.2016.90.

[81] K. K. Bressem *et al*, "Comparing different deep learning architectures for classification of chest radiographs," *Scientific Reports*, vol. 10, *(1)*, pp. 13590, 2020. Available: https://www.ncbi.nlm.nih.gov/pubmed/32788602. DOI: 10.1038/s41598-020-70479-z.

[82] G. Boesch, *Deep Residual Networks (ResNet, ResNet50) – Guide in 2022*. Accessed on Jan. 22, 2021. Available: https://viso.ai/deep-learning/resnet-residual-neural-network/.

[83] J. Weston, R. Lenain, U. Meepegama and E. Fristed, "Learning De-identified Representations of Prosody from Raw Audio," *Proceedings of the 38th International Conference on Machine Learning*, 2021. Available: https://arxiv.org/abs/2107.08248.

[84] S. Schneider *et al*, "wav2vec: Unsupervised Pre-training for Speech Recognition," 2019. Available: https://arxiv.org/abs/1904.05862.

[85] J. Yuan *et al*, "The Role of Phonetic Units in Speech Emotion Recognition," 2021. Available: https://arxiv.org/abs/2108.01132.

[86] Y. Wang, A. Boumadane and A. Heba, "A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding," 2021. Available: https://arxiv.org/abs/2111.02735.

[87] A. Conneau, A. Baevski, R. Collobert, A. Mohamed and M. Auli, "Unsupervised Cross-lingual Representation Learning for Speech Recognition," *Proc. Interspeech 2021*, 2021. Available: https://www.isca-speech.org/archive/interspeech_2021/conneau21_interspeech.html.DOI: 10.21437/Interspeech.2021-329.

[88] W. Hsu *et al*, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *Taslp*, vol. 29, pp. 3451-3460, 2021. Available: https://ieeexplore.ieee.org/document/9585401. DOI: 10.1109/TASLP.2021.3122291.

[89] J. Bgn, "HuBERT: How to Apply BERT to Speech, Visually Explained," 2021. Available: https://jonathanbgn.com/2021/10/30/hubert-visually-explained.html#:~:text=HuBERT%20builds%20targets%20via%20a,process%20using%20Gumbel%2Dsoftmax

[90] S. Minaee *et al*, "Deep Learning--based Text Classification," *ACM Computing Surveys*, vol. 54, *(3)*, pp. 1-40, 2021. Available: https://dl.acm.org/doi/10.1145/3439726. DOI: 10.1145/3439726.

[91] J. Pennington, R. Socher and C. Manning, "GloVe: Global Vectors for Word Representation," 2014. Available: https://nlp.stanford.edu/projects/glove/.

[92] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "Bag of Tricks for Efficient Text Classification," *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427- 31, 2017. Available: https://aclanthology.org/E17-2068/.

[93] X. Wu *et al*, "Conditional BERT Contextual Augmentation," *Computational Science – ICCS 2019. ICCS 2019*, pp. 84-95, 2019. Available: https://link.springer.com/chapter/10.1007/978-3-030-22747-0_7 . DOI: 10.1007/978-3-030-22747-0_7.

[94] Y. Jia, H. Zen, J. Shen, Y. Zhang and Y. Wu, "PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS," *Proc. Interspeech 2021*, pp. 151-55, 2021. Available: https://www.isca-speech.org/archive/interspeech_2021/jia21_interspeech.html. DOI: 10.21437/Interspeech.2021-1757.

[95] K. Yorkston, M. Bourgeois and C. Baylor, "Communication and Aging," *Phys Med Rehabil Clin N Am*, 2010. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074568/.DOI: 10.1016/j.pmr.2009.12.011.

[96] P. A. Pérez-Toro *et al*, "Influence of the Interviewer on the Automatic Assessment of Alzheimer’s Disease in the Context of the ADReSSo Challenge," Proc. Interspeech 2021, pp. 3758-89, 2021. Available: https://www.isca-speech.org/archive/interspeech_2021/pereztoro21_interspeech.html. DOI: 10.21437/Interspeech.2021-1589.

[97] M. Moradshahi, H. Paangi, M. Lam, P. Smolensky and J. Gao, "HUBERT Untangles BERT to Improve Transfer across NLP Tasks," 2019. Available: https://arxiv.org/abs/1910.12647.

7. Appendix

A.

Hyper-parameters are tuned using 5-fold grid search cross validation on the customized training set (section 3.2.). The SVM has the following parameters: a radial basis function (rbf) or linear kernel and a C value: 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 1.5, 2, 2.5, 3, 10, 1000, 100. All other parameters are set to default. The GNB classifier is fit with variance smoothing coefficients of $1\,e-10, 1\,e-9, 1e-8$. RF parameters are: number of estimators as 50, 100, 300, 500, 800 and 1000; criterion either gini or entropy with a minimum samples per split of either 2, 3 or 4. Bootstrap was set to False and True. All other parameters are set to default.

Approach	FS	Hyperparameters

SVM GeMAPS	weights	C = 0.05
		Kernel = linear

SVM GeMAPS	ANOVA	C = 2
		Kernel = linear

SVM ComParE	weights	C = 1
		Kernel = rbf

SVM ComParE	ANOVA	C = 0.005
		Kernel = linear

RF GeMAPS	weights	Criterion = Gini
		Minimum samples per split = 2
		Number of estimators = 500
		Bootstrap = False

RF GeMAPS	ANOVA	Criterion = Entropy
		Minimum samples per split = 2
		Numbers of estimators = 500
		Bootstrap = False

RF ComParE	weights	Criterion = Gini
		Minimum samples per split = 3
		Numbers of estimators = 3000
		Bootstrap = False

RF ComParE	ANOVA	Criterion = Entropy
		Minimum samples per split = 3
		Numbers of estimators = 500
		Bootstrap = False

GNB ComParE	weights	variance smoothing coefficient = 1 e-10
GNB GeMAPS	weights	variance smoothing coefficient = 1 e-9

Table 22. *Hyper-parameters settings for each acoustic feature-based model.*

Approach	FS	Hyperparameters

SVM COEVEFE	weights	C = 0.1
		Kernel = linear

SVM COEVEFE	ANOVA	C = 100
		Kernel = rbf

SVM CLAN	weights	C = 3
		Kernel = linear

SVM CLAN	ANOVA	C = 0.01
		Kernel = rbf

SVM BOW	weights	C = 0.001
		Kernel = linear

SVM BOW	ANOVA	C = 0.01
		Kernel = linear

RF COEVEFE	weights	Criterion = Gini
		Minimum samples per split = 2
		Numbers of estimators = 100
		Bootstrap = True

RF COEVEFE	ANOVA	Criterion = Gini
		Minimum samples per split = 2
		Number of estimators = 100
		Bootstrap = False

RF CLAN	weights	Criterion = Gini

		Minimum samples per split = 3 Number of estimators = 300 Bootstrap = True

RF CLAN	ANOVA	Criterion = Gini
		Minimum samples per split = 3
		Number of estimators = 500
		Bootstrap = True

RF BOW	weights	Criterion = Gini
		Minimum samples per split = 4
		Number of estimators = 300
		Bootstrap = True

RF BOW	ANOVA	Criterion = Gini
		Minimum samples per split = 3
		Number of estimators = 100
		Bootstrap = False

GNB COEVEFE	weights	variance smoothing coefficient = 1 e-10
GNB CLAN	weights	variance smoothing coefficient = 1 e-9
GNB BOW	weights	variance smoothing coefficient = 1 e-9

Table 23. *Hyper-parameters settings for each linguistic feature-based model.*

Approach	**FS**	**Hyperparameters**

SVM CLAN + GeMAPS	ANOVA	C = 2.5
		Kernel = linear

SVM CLAN + GeMAPS	weights	C = 1.5
		Kernel = rbf

SVM CLAN + ComParE	ANOVA	C = 0.05
		Kernel = linear

SVM CLAN + ComParE	weights	C = 3
		Kernel = rbf

SVM BOW + ComParE	ANOVA	C = 0.001
		Kernel = linear

RF CLAN + GeMAPS	ANOVA	Criterion = Entropy
		Minimum samples per split = 2
		Numbers of estimators =300
		Bootstrap = True

RF CLAN + GeMAPS	weights	Criterion = Gini
		Minimum samples per split = 3
		Numbers of estimators = 300
		Bootstrap = False

RF BOW + ComParE	weights	Criterion = Gini
		Minimum samples per split = 4
		Numbers of estimators = 50
		Bootstrap = True

RF CLAN + ComParE	weights	Criterion = Gini
		Minimum samples per split = 2
		Numbers of estimators = 50
		Bootstrap = True

RF CLAN + ComParE	ANOVA	Criterion = Entropy
		Minimum samples per split = 3
		Numbers of estimators = 50
		Bootstrap = False

Table 24. *Hyper-parameters settings for each multimodal feature-based model.*

B.

This section presents confusion matrices for the best-performing models in each approach. It compliments results reported in section 4.

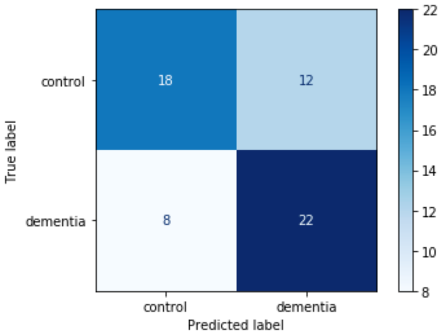


Figure 19*. Confusion matrix obtained from the baseline model evaluated on the held-out test set.*

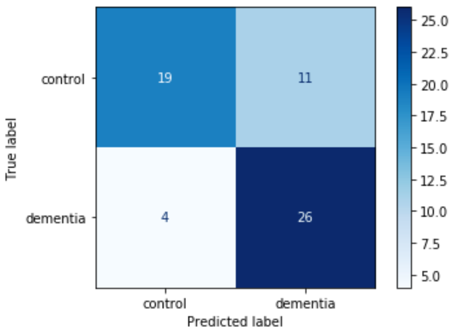


Figure 20*. Confusion matrix for RF with feature set ComParE and weights as FS method.*

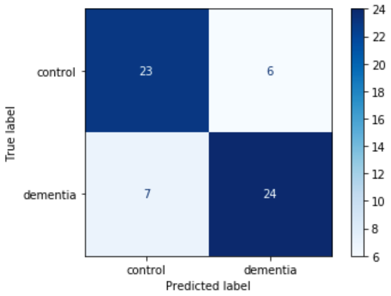


Figure 21*. Shows the confusion matrix given by RF using BOW and ANOVA as FS method.*

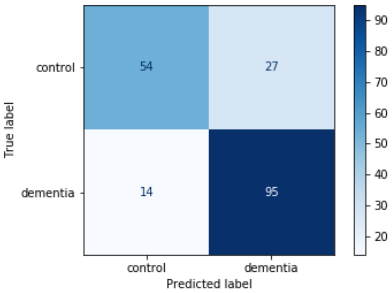


Figure 22. *Confusion matrix given by weighted majority voting.*

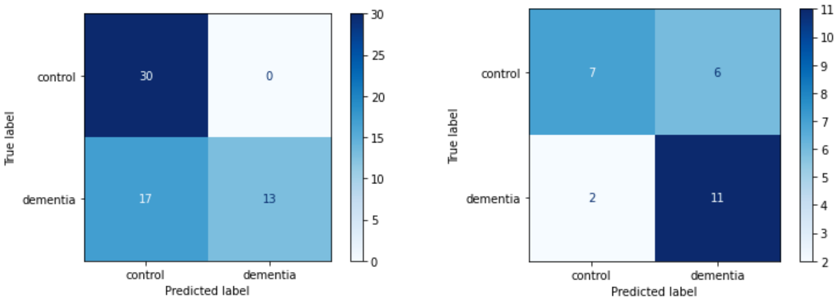
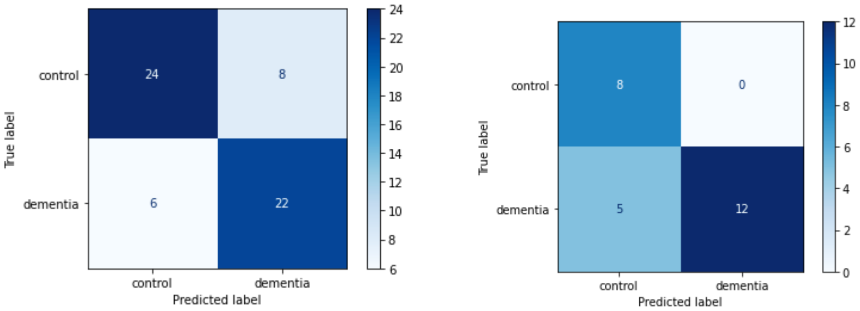


Figure 23 and 24. *Figure 23 (left) is the confusion matrix given by ResNet152 on the held-out test set, figure 24 (right) is given by ResNet152 on the external validation set.*



Figures 25 and 26. *Figure 25 (left) is the confusion matrix given by the BERT model evaluated on the held-out test, figure 26 (right) is the confusion matrix given by the BERT model evaluated on the external validation set. Transcripts are human-made.*