Integrating Fine-Tuned LLM with Acoustic Features for Enhanced Detection of Alzheimer's Disease

Filippo Casu, Andrea Lagorio, Pietro Ruiu, Giuseppe A. Trunfio and Enrico Grosso

***Abstract*— Dementia represents a global public health concern, with the early detection of Alzheimer's disease, the most prevalent form of dementia, being of paramount importance. Given the limited availability of suitable biomarkers, research has shown that early cognitive impairment can be identified through patients' spoken language. This paper presents a multi-modal system for automatic Alzheimer's disease detection using speech. The system has been trained on spoken recordings of healthy individuals and Alzheimer's patients describing an image, a task requiring linguistic and cognitive skills. Built on fine-tuned advanced Large Language Models, audio feature extractors, and classifiers, the system, after an extensive comparison of single and multi-modal architectures, achieves optimal results with the combination of Mistral-7B, VGGish, and Support Vector Classifier, outperforming previous methods on the ADReSSo 2021 test set.**

***Index Terms*— Alzheimer's Disease, Fine-Tuning, Large Language Model, Machine Learning, Natural Language Processing.**

I. Introduction

Dementia related diseases represent a significant global public health challenge. Indeed, according to the World Health Organization [1], in 2023 more than 55 million people suffer from dementia. Furthermore, around 10 millions of new cases are diagnosed every year with an expected doubling of the total number of affected people every 20 years, in relation to the population pyramid. Alzheimer's Disease (AD) is the most common form of dementia and covers 70% of cases. It is currently the seventh leading cause of death and one of the major causes of disability and dependency among older people globally. Despite substantial efforts, dementia cost in 2019 economies globally reached 1.3 trillion US dollars. Around half of these costs are borne by family and friends providing an estimated care workload of about 5 hours per day. The problem related to undetected cases is cause of concern: recent reports show how more than 50% of the cases remain undetected (75% of all dementia cases go undiagnosed across the globe, up to 90% in low-middle-income-countries). In that sense, the importance of recognizing early signs is crucial and it is often linked to the restricted access to biomarkers, such as magnetic resonance imaging, or the limited number of clinicians and professionals involved. Several recent researches demonstrate how initial symptoms of cognitive impairment can be identified in patients' spoken language. The latter, jointly with other related features, such as semantic capabilities and vocal functionality, outlines biomarkers that are able to evaluate, through different dimensions, patients' cognitive capabilities [2]. Signs of cognitive impairment can be reflected into the voice and they can be evaluated by measuring some parameters of the acoustic waveform. This means that it is important to evaluate the capacity of the patients to control the use of the vocal cords. Memory dysfunction can affect the language of the patients and it converges to low coherence or low information density. It has been observed that Natural Language Processing (NLP) techniques are able to measure these cues of memory impairment [3]. Variations in emotion expressions and cognitive deterioration can occur simultaneously. Psycholinguistic cues of the speech are affected by these type of alternations in emotional communication, that, in many cases, derive from AD. In order to model these types of signals, nonverbal vocalization and semantic evaluations are employed [4]. Measurements of parameters deriving from acoustic waveform are an effective method to model modification in nonverbal vocalization. Whereas, semantic cues can be identified by using NLP techniques specifically developed to describe psycolinguistic part of the speech.

In this article, we present a novel framework for the automatic detection of AD by exploiting patients' speech. The input data consists of voice recordings from various subjects, both healthy and affected by Alzheimer's disease, who were asked to describe a scene depicted in an image. The task of describing the image helps to identify cognitive deficits, such as those caused by dementia or AD, as it requires the integration of various linguistic and cognitive skills. These recordings, both in their transcribed form and as audio signals, have been utilized for training the automatic detection system. Following an extensive comparison of different Large Language Models (LLMs), various vocal feature extractors, and multiple classifiers, it turns out that the optimal results are achieved using Mistral-7B, VGGish, and Support Vector Classifier (SVC). This configuration yelds an accuracy of 93%, which, to the best of our knowledge, is the highest performance currently available in the field.

The main contributions of this work are as follows:

- Development of a novel multi-modal AD detection sys-

This paragraph of the first footnote will contain the date on which you submitted your paper for review.

F.Casu, A. Lagorio, P. Ruiu, G. A. Trunfio, E. Grosso are with the Department of Biomedical Sciences at University of Sassari, Italy (e-mail: (fcasu1, lagorio, pruiu, gtrunfio, grosso)@uniss.it).

tem that integrates text and audio inputs leveraging fine-tuned LLMs as backbone.
- Comprehensive statistical analysis of various configurations of the proposed architecture, combining different LLMs, audio feature extractors, and classifiers.
- Outperformance of the State-of-the-Art (SotA) of previous AI methods on ADReSSo 2021 test set.

The paper is organized as follows: Section II provides an overview of the SotA in automatic AD detection through the speech of the patients; Section III details the dataset employed, the methodology adopted and the main aspects of the implemented architecture; Section IV presents the results obtained; Section V conducts an in-depth analysis of the findings, providing an exhaustive comparison with the most representative SotA works; finally, Section VI concludes the paper summarizing key results.

II. Related works

Several works attempted to automatically identify AD by analysing patients' data, [5], [6] and in particular speech. In this context, the challenges proposed by Interspeech, with ADReSS [7] and ADReSSo [8], represent an interesting benchmark for comparison. Furthermore, many studies have used the datasets provided by ADReSSo even in the years following the challenges. The challenge consisted of three tasks related to the assessment of AD; among them, of particular interest to this study is AD classification. This task requires developing models capable of predicting AD or non-AD labels for short speech sessions in which patients are instructed by specialized personnel to describe the Cookie-Theft picture (Figure 1). Depending on the focus of these researches, different techniques have been developed, which can be categorized into: acoustic only, purely linguistic/psycholinguistic oriented, and the joint combination. For the purposes of this work, we are particularly interested in the latter, which can, in turn, be subdivided in two sub groups. A first group is characterized by techniques where the joint combination of acoustic and linguistic/psycholinguistic features produces an actual improvement over the techniques used alone. Pappagari et al. [9] modeled the acoustic part of the speech by employing x-vectors [10], the linguistic part of the speech using the pre-trained language Bidirectional Encoder Representation from Transformers BERT [11], and the best results were achieved by jointly combining both models. Pompili et al. [12] employed a similar approach for data representation and a Bi-Long Short-Term Memory (Bi-LSTM) and Support Vector Machine (SVM) as classifiers, again achieving the best accuracy score by joint combination of the models. Martic et al. in [13] trained several Machine Learning (ML) architectures, and the best results were achieved by Logistic Regression, with acoustic embeddings extracted by using GeMAPS [14], and linguistic embeddings extracted from NLP-based techniques (e.g. TF-IDF). Chen et al., in [15], reached the best results for the acoustic part, with embeddings extracted with ComPare-2013 and GeMAPS, whereas, for the linguistic part, with embeddings extracting with BERT. Combining the two resulted in significant improvement. In the work of Rohanian et al. [16], an ML architecture with Bi-LSTM has been trained with features generated by COVAREP for the acoustic part of the speech, and Glove for the linguistic part of the speech. Pérez-Toro et al. [17], proposed the use of x-vectors for the acoustic part of the speech, word-embeddings from BERT and ELECTRA models for the linguistic part of the speech; a Radial Basis Function (RBF)-SVM was used as a classifier. In Zhu et al. [18], for the acoustic part the best results in terms of accuracy are achieved with Speech BERT, and, for the linguistic part, with Longformer. The joint combination of Speech BERT and Longformer again achieved the best accuracy.

For a second group of multimodal techniques, studies reported no improvement and, in some cases, a deterioration, typically depending on overfitting. Among those, the work of Koo et al. [19] who employed VGGish [20] and XLNet (an extended version of BERT), and trained CNN and Bi-LSTM networks. Through a trained SVM, Syed et al. [21] achieved best accuracy by extracting the acoustic features with VGGish and BERT for the linguistic part. The joint combination led to a degradation in terms of accuracy. In [22] Shah et al. modeled the acoustic part of the speech with several open-acess tools, serveral NLP techniques for the for the linguistic part, and an SVM algorithm as classifier. [23] Pappagari et al. employed x-vectors and prosody embeddings to model the acoustic part of the speech. Moreover, BERT-based models were trained on the linguistic features of speech using Automatic Speech Recognition (ASR) methods. Zolnoori et al. [4] combined the DistilBERT's word embeddings with linguistic and acoustic domain-related features and used Joint Mutual Information Maximization (JMIM) to identify and select the best speech parameters. The best results was reached by using an SVM classifier with RBF Kernel.

In some cases, works do not compare monomodal and multimodal approaches and just show the scores for the multimodal system. Two works of Balagopalan et al. [24] [25], combined different acoustic (e.g. of frequency and spectral domain) and linguistic features (lexical richness, syntactic and semantic features). By using semantic and non-semantic information of the speech, Zhu et al. in [26] proposed an architecture based on Wav2vec [27] and BERT, enabling the fine-tuning of WavBERT, and achieving the highest accuracy using ADReSSo dataset.

Finally, some studies have exclusively focused on utilizing the linguistic component of speech. In [28] Qiao et al. combined linguistic complexity and (dis)fluency features with pretrained language models, such as BERT and ERNIE models, and tested it on ADReSSo set. Roshamir et al. [29] obtained the best accuracy score with BERT jointly with a logistic regression classifier. Gómez-Zaragozá et al. [30] achieved good results by manual transcriptions. Wang et al. [31] presented an evolution of the baseline Convolutional-augmented Transformer (Conformer) that uses Learning Hidden Unit Contribution (LHUC) reaching one of the best results in terms of accuracy. Casu et al. [32] utilized three distinct pre-trained LLMs, fine-tuned over the Whisper [33] automatic transcription of audio-recorded interviews. Kong et al. [34] reached their best results by employing a Hierarchical

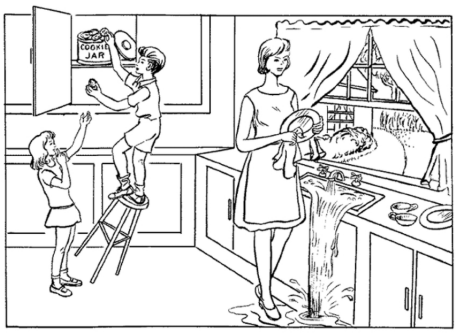

Fig. 1. The "Cookie-Theft Picture", used in the medical field to assess the cognitive and linguistic abilities of patients.

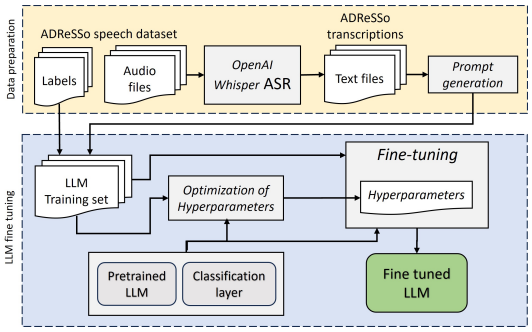

Fig. 2. The devised process for obtaining a fine-tuned LLM given the training set of labeled audio files.

Attention Network (HAN) and the joint combination of the patients's age to the model. Bertini et al. [35] only oriented the research on the acoustic component of speech, by employing an autoencoder to produce a fixed-dimensional feature vector exploiting the Log-Mel spectogram of the audio-record. The authors trained a multilayer perceptron enhaced by SpecAugmented suite.

According to the above overview, while there has been substantial progress in leveraging linguistic and acoustic features for detecting AD, current research presents notable gaps. In particular, although several studies have demonstrated the potential of integrating these features, the full capabilities of modern LLMs remain underexplored in this context, with the exception of some preliminary work presented in [32].

This article addresses these deficiencies by introducing a novel approach that deeply integrates modern LLMs, fine-tuned for the AD detection task, with comprehensive acoustic features, derived from advanced extraction techniques. To ensure a fair assessment, results obtained are carefully compared to those achieved by the most representative SotA works on the same dataset, as discussed in Sec. IV-D.

III. Data, Methodology and Implementation

We developed two methods for automatic AD detection, denoted as LLM-X and LLM-A-X, where LLM refers to the Large Language Model utilized, 'A' to the acoustic feature extractor, and 'X' to the classifier model chosen. The first approach utilizes only the linguistic features obtained through fine-tuning a pre-trained LLM, while the second approach integrates both linguistic and acoustic features extracted from the original audio files. The embeddings used to obtain linguistic and acoustic features from transformers were extracted from the final hidden layer of the selected model. This section provides a detailed description of the adopted approaches.

A. Dataset Used for Experiments

For this study, we decided to use the ADReSSo challenge dataset [8], as it contains more samples compared to the ADReSS dataset. ADReSSo includes three difficult automatic prediction problems of detecting cognitive decline using speech only. We focused on the task related to AD classification, for which a specific dataset is made available. It is an audio-recorded speech dataset from the English Pitt DementiaBank [36] and consists of recordings of picture descriptions produced by cognitively normal subjects (non-AD) and patients with an Alzheimer's disease diagnosis (AD). The subjects were asked to describe the Cookie-Theft picture (shown in Figure 1) from the Boston Diagnostic Aphasia Examination [37]. The dataset includes 166 audio files in the training set, comprising 87 from AD patients and 79 from non-AD patients. The test set consists of 71 audio files, equally divided between AD and non-AD patients (see Table I). According to the "Cookie-Theft" protocol, participants must be at least 44 years old. Table I displays information indicating the AD group participants. Notably, since the ADReSSo dataset [8] already ensures gender balance and a consistent age range among AD and non-AD participants, gender and age were excluded as predictive features.

B. Linguistic Feature Method (LLM-X)

In this method, illustrated in Figure 2, given a training set of files from the ADReSSo speech datasets, we first obtain the corresponding transcriptions into text files using Whisper [33], the OpenAI ASR system. The prompt generation process combines transcribed patient responses, provided by Whisper, an instructional text, and a detailed image description to standardize the task. It takes the following form: " *Carefully analyze the following interview with a person describing the Cookie Theft picture from the Boston Diagnostic Aphasia Examination, in which [PICTURE DESCRIPTION]. Based on your analysis, reply with 'YES' if you believe the person has signs of cognitive impairment, otherwise reply with 'NO'": [INTERVIEW].*" The result of this stage is a fine-tuned LLM used for feature vector extraction, as shown in Figure 2.

For fine-tuning, the LoRA method [38], which involves freezing the original weights and exclusively training a limited set of parameters, has been adopted in this study. For imple-

TABLE I
ADReSSo 2021 DATASET

	Train			Test		
	AD	Non-AD	Tot	AD	Non-AD	Tot
Female	58	52	110	21	23	44
Male	29	27	56	14	13	27
Total	87	79	**166**	35	36	**71**
Age Mean	69.72	66.04		68.51	66.11	

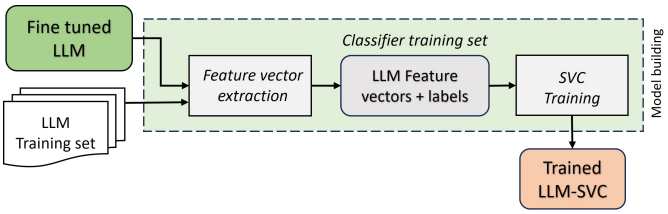

Fig. 3. Schema of an automatic AD detector based on a fine-tuned LLM and SVC (LLM-SVC).

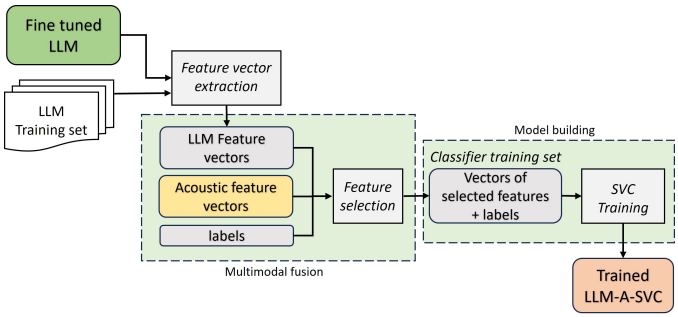

Fig. 4. Process of combining LLM feature vectors with acoustic feature vectors for classifier training in the LLM-A-SVC approach.

mentation the Optuna [39] framework has been used, which uses Bayesian optimization over a fixed number of trials (50 in this study). During this stage, the training-validation split for each fold involved 80% for training and 20% for validation.

The optimized hyperparameters include LoRA α and LoRA r, which are part of the LoRA algorithm. LoRA α is a scaling factor for the low-rank matrices, regulating their impact on model weight adaptation, while LoRA r defines the rank of these matrices. Other optimized hyperparameters include the LoRA dropout rate and the number of training epochs. After hyperparameter determination, fine-tuning of the LLMs proceeded using the LoRA algorithm based on the current training set. The fine-tuned LLM, obtained as described above, enables the production of feature vectors used to train various classifiers. Figure 3 shows an example of an automatic AD detector based on a fine-tuned LLM and SVC.

In this study we investigated the following pre-trained LLMs: Llama2-7B, Llama2-13B [40], Mistral-7B [41], and OpenHermes [42] a 7B LLM based on Mistral. Note that the 'B' following the model names indicates the billions of parameters. All the selected LLM are based on the Transformer technologies [43]. Below is a brief description of each of the LLMs used in this work:

- **Llama2**: consists of several pretrained and fine-tuned generative text models ranging in scale from 7B to 70B parameters developed and released as open source LLM by Meta. The models were pretrained on 2 trillion tokens of data from publicly available sources. Llama2-family models incorporate an adaptive attention span that has demonstrated cutting-edge results in benchmarks that necessitate profound contextual comprehension.
- **Mistral**: is the first LLM released by mistral.ai. It surpasses Llama2-13B on all benchmarks, leveraging Grouped-Query Attention for faster inference and Sliding Window Attention [44] for efficient handling of longer sequences, in which each layer attends to the previous 4,096 hidden states.
- **OpenHermes**: is a Mistral 7B fine-tuned with fully open datasets. OpenHermes was trained on 900,000 entries predominantly generated by GPT-4, sourced from open datasets spanning the AI landscape.

C. Linguistic and Acoustic Feature Method (LLM-A-X)

This method is an extension of LLM-X, which incorporates acoustic-based features extracted from the original audio files.

As shown in Figure 4, the process starts with the extraction of linguistic features using the fine-tuned LLM as described above. Additionally, acoustic features extracted from the audio files are concatenated to the linguistic features. A process of feature selection is then applied to the combined feature vectors after standardization. In fact, after the merge, selecting the most informative features helps in improving the model's performance since irrelevant or redundant features can introduce noise into the model, potentially leading to overfitting or reduced generalization capability on unseen data. The selected features of the training vectors, along with the labels, are then used for building a classifier (e.g., an SVC in Figure 4).

For the acoustic-based features, we investigated the following approaches of automatically extraction from the audio files:

- **OpenSmile**. It is an open-source toolkit for audio-speech analysis, processing, and classification [45]. It works on the classification and evaluation of specific frequencies and energies of the human voice. In this work we concentrated on the implementation of the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [14] for Voice Research. One of the lightest versions of GeMAPS was selected, generating 88-dimensional features for each audio file. These features represent relevant phonetic motor planning and model vocal psycholinguistic cues and are significantly associated with AD detection [4].
- **Wav2Vec**. This is a framework for self-supervised learning of speech representations [27]. It is pre-trained on large amounts of unlabeled audio data and fine-tuned for specific tasks like speech recognition. Wav2Vec learns representations from raw audio waveforms, using convolutional layers to process the input signal into meaningful features. The input of the model is the whole audio file. These embeddings, capture nuanced speech characteristics, including phonetic and prosodic features, which are crucial for distinguishing between different cognitive states.
- **WavLM-Base-Plus**. The WavLM-Base-Plus, a specific version of the pre-trained WavLM model for speech processing [46], demonstrates strong performance in speaker verification and speech recognition tasks. The model is composed by a convolutional feature encoder and a Transformer encoder which receive in input the whole audio file. For the aims of this work, it is potentially interesting due to its high flexibility and noteworthy results achieved in the paralinguistic Emotion Recognition task,

as variations in emotion in the patient can be considered a cue of cognitive impairments.

- **VGGish**. VGGish is a variant for audio processing of the well-known VGG image model [20]. It generates 128-dimensional embeddings from an audio window with a length of 0.96 seconds. The generation of features is based on the Log-Mel spectrogram, from the application of a Short-Time Fourier Transform (STFT) to the audio signal.

  To generate a unique dimension-fixed feature for each patient's audio interview using VGGish features, we devised a method based on a supervised autoencoder with a Gated Recurrent Unit (GRU) architecture [47]. GRUs are well known for being able to effectively capture and encode the temporal dynamics of sequential data, making it well-suited for processing audio features over time. The method is outlined in Figure 5. First, the VGGish features are fed into a GRU encoder, which processes the sequential data and encodes it into a hidden state. This hidden state encapsulates the temporal dynamics of the features, providing a compact representation of the audio sequence. Subsequently, the encoded features are passed into a GRU decoder, which attempts to reconstruct the original VGGish features. The reconstruction process ensures that the encoded features retain the essential information of the input data. The hidden state from the GRU encoder, a fixed-size vector, is then used as the acoustic feature vector.

  To utilize these acoustic feature vectors for classification, they are passed through a classification layer designed to predict the labels. The classifier is trained alongside the autoencoder, ensuring that the encoded features are not only representative of the original audio but also useful for the classification task.

  During the autoencoder training phase, two types of losses are minimized: the reconstruction loss and the classification loss. The reconstruction loss measures how accurately the decoder reconstructs the original VGGish features from the encoded representations. The classification loss assesses the performance of the classification layer in predicting the correct labels. By jointly minimizing these losses, the model learns to produce encoded features that are both representative of the input audio and effective for classification. The resulting fixed-size vectors provide a robust and compact representation of the audio data, to be used in a LLM-A-X algorithm as that in Figure 4.

D. Multimodal Fusion and dimensionality reduction

As shown in Figure 4, in order to retain only the most informative and relevant dimensions for the classification task, the multimodal fusion includes a dimensionality reduction step, implemented in terms of feature selection. For the latter process, after evaluating several alternative approaches, we employed the LinearSVC from Scikit-learn library, with an L1 penalty, to rank the features based on their importance. After training the LinearSVC, we utilized the absolute values of

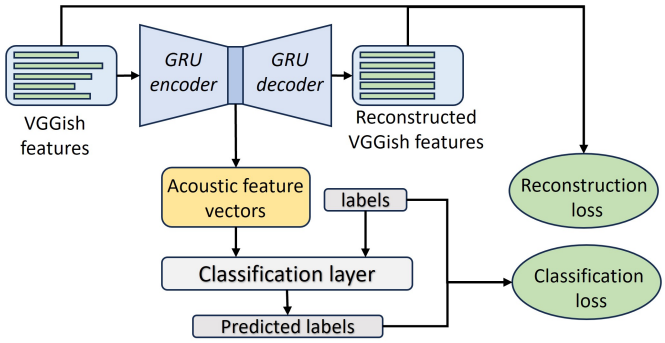

Fig. 5. The adopted method based on a supervised GRU-based autoencoder with a GRU to generate fixed-dimension feature vectors.

the learned coefficients, which are accessible via the model's `coef_` attribute, to assign a relative importance score to each feature. These coefficients reflect each feature's weight in the linear decision boundary and its influence on classification. The coefficients were then normalized to a range of $[0, 1]$ to standardize the importance scores across features. To determine the number of important features to retain, we calculate the cumulative importance of the ranked features. In particular, we select the smallest subset of features that together account for a threshold of 95% of the total importance. Based on this threshold, we identify the top features and use them to create reduced training and testing datasets.

E. Classification Layer

In both approaches (i.e. LLM-X and LLM-A-X) a classifier, indicated by X, is used to distinguish AD from non-AD patients. In the following we describe main characteristics of the classifiers used for the analysis.

- **Artificial Neural Network** (ANN): The ANN employed featured a fully connected architecture with three layers. The first layer mapped the LLM hidden state, of size 4096, to 1024 units, followed by batch normalization and a dropout of 0.01. The second layer reduced the dimension to 128 units, incorporating another dropout of 0.01. The output layer consisted of a single unit. ReLU activation functions were used throughout, and training was conducted using the BCEWithLogitsLoss with an Adam optimizer (learning rate of 1.0E-05). Early stopping was implemented with a patience of 10 epochs and a delta of 0.001.
- **Logistic Regression** (LR): This model was configured with a maximum iteration setting of 1000 and L2 regularization with a strength (C) of 1.0.
- **Random Forest** (RF): Comprising 100 decision trees, the RF model utilized the default settings for other parameters.
- **Gradient Boosting** (GB): Implemented with 50 boosting stages.
- **XGBoost** (XGB): Configured with 10 boosting rounds and the evaluation metric set to 'logloss'. The model used the default hyperparameters [48].
- **SVC**: The SVC employed a RBF kernel and was configured to output probability estimates.

- **Stacking Classifier** (SC): This classifier integrated an LR meta-model with a base ensemble comprising RF, GB, a RBF kernel SVC, and XGBoost. Each base classifier was set with 50 estimators. The ensemble utilized a 5-fold CV strategy to train the meta-model.

IV. Experimental Design and Results

In this section, we present and analyze the results obtained from various experiments conducted to evaluate the efficacy of different configurations of LLM-X and LLM-A-X.

A. Investigating LLM-X automatic AD detectors

A first set of experiments was aimed at evaluating the different pre-trained LLMs described in section III-B, for obtaining LLM-X types of automatic AD detectors based solely on linguistic features.

In the comparison, we utilized the training set delineated in Section III-A within a five-fold CV setting. For each fold, classifiers were constructed using the LLMs and assessed on the corresponding test set through various metrics.

Table II shows the results of the optimization of hyperparameters, in terms of LoRA α, LoRA r, LoRA dropout rate and the number of training epochs. Additional relevant parameters were the per-device batch size and the gradient accumulation steps, set to 2 and 8, respectively, to facilitate fine-tuning on consumer-grade GPUs. The optimizer chosen was "paged lion," a variant known for its efficiency in optimizing neural network parameters across page-sized mini-batches.

After hyperparameter determination, fine-tuning of the LLMs proceeded using the LoRA algorithm as described in Section III-B.

Then, the normalized feature vectors, of size 4096, extracted from each fine-tuned LLM served to train a variety of classifiers, each distinguished by architecture, parameters, and training configuration, as detailed in sec. III-E.

Model effectiveness was measured using metrics like accuracy, precision, recall, F1, and ROC-AUC on five-fold CV test sets, ensuring a robust assessment of generalizability across the dataset.

For each classifier, metrics were computed 25 times, each initialization using a different random state, with the results subsequently averaged. This procedure was repeated for each of the seven classifiers across five training and corresponding test sets. Consequently, for each metric and LLM, we obtained results from 35 separate experiments (seven classifiers times five folds), enabling comprehensive statistical comparisons.

Figure 6 summarizes the results of the LLM-X automatic AD detectors across all classifiers and CV folds using box plots. This graph illustrate the distribution of the performance metrics for each pre-trained LLM, providing a clear visual representation of the performance variability and robustness of each model. From Figure 6, it is evident that the models exhibited varying degrees of performance across different metrics. The median values and interquartile ranges offer insights into the central performance and variability of each LLM. Notably, Mistral-7B performed consistently well in accuracy, precision, and particularly in F1 score and ROC-AUC, demonstrating a robustness in model performance. In contrast, Mistral-7B's median recall appears slightly lower compared to other metrics.

TABLE II
Best hyperparameters found with Optuna for LLM models

Parameter	L-7	L-13	M	O
LoRA α	60	40	40	48
LoRA r	48	16	20	32
LoRA dropout	0.05	0.01	0.01	0.02
Epochs	5	6	5	5

To statistically evaluate the LLMs' performance, we applied a rank-based approach as recommended by [49]. Specifically, for each metric, we ranked the results of each LLM within the 35 experiments, assigning a rank of 1 for the best performance and 4 for the worst. The ranks for each LLM were then averaged to determine the overall performance. We utilized the Friedman test [50], a non-parametric statistical test designed to identify differences across multiple experimental conditions. The test evaluates the null hypothesis that all LLMs perform equivalently across the multiple sets of experiments, with any observed variations in ranks attributed to random variations rather than systematic differences. This rigorous statistical approach ensures that our conclusions about LLM performance are robust and supported by empirical evidence.

Upon rejecting the null hypothesis with the Friedman test at a p-value threshold of 0.05, indicating significant differences among the models, a post-hoc analysis was performed to further explore these differences. We employed the Nemenyi test [51], a common post-hoc procedure following the Friedman test, which compares all pairs of LLMs. This test assesses whether the observed rank differences between any two models are statistically significant beyond the predetermined confidence level of 0.05. The critical difference (CD) diagram was used to visually represent these statistical disparities. In the CD diagrams showed in Figure 7, models connected by a line are not significantly different from each other, while those that are unconnected demonstrate statistically significant differences. This visual tool effectively highlights which models are statistically comparable and which are distinct, providing clear guidance on the relative performance of each LLM based on the tested metrics.

According to the results in Figure 7, distinct performance patterns are evident across different metrics for the LLMs under comparison. In terms of Accuracy, Mistral-7B consistently outperformed other models, particularly Llama2-7B (L-7) and OpenHermes-7B (O), across various classifiers, indicating its superior capability in accurately classifying cases under diverse scenarios. However, its performance was statistically similar to Llama2-13B. In Precision, Mistral-7B ranked the highest, significantly surpassing all other models. Differences in Recall were less pronounced, with no model showing statistically significant superiority. For the F1 Score, both Mistral-7B and Llama2-13B demonstrated high efficiency, achieving scores of 1.8 and 2.3, respectively, and were statistically similar, suggesting a balanced trade-off between precision and recall. Similarly, in ROC-AUC, these two models exhibited superior discrimination between classes at varying threshold

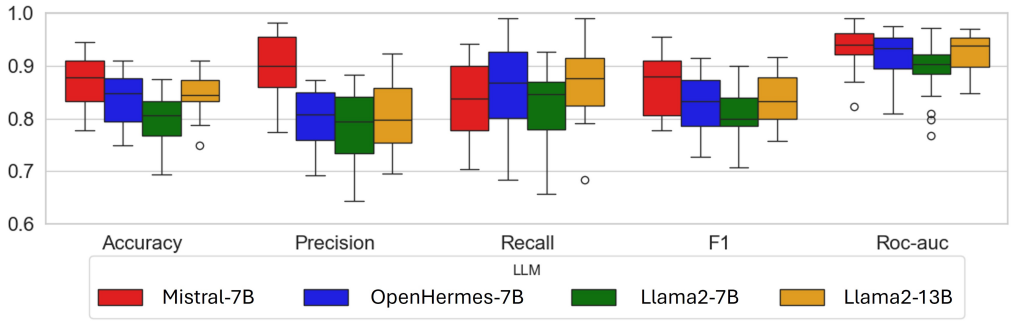

Fig. 6. Box plots summarizing all the results obtained by the LLM-X method for all considered classifiers and folds of the CV.

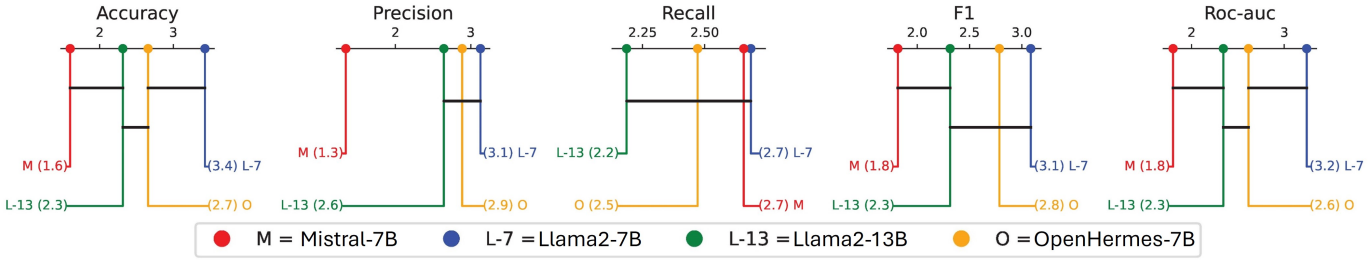

Fig. 7. Representation of the critical distances from the Nemenyi test between the average ranks of the pre-trained LLMs employed in the considered LLM-X method. LLMs connected by a horizontal line are not significantly different from each other, while those that are unconnected demonstrate statistically significant differences.

TABLE III
RESULTS OF CLASSIFICATION MODELS USING THE LLMS UNDER COMPARISON IN A 5-FOLD CV SETUP.

Classifier	LLM Features	Accuracy	Precision	Recall	F1 Score	ROC-AUC
LLM-LR	Mistral-7B	**0.898** ± 0.038	**0.913** ± 0.029	0.869 ± 0.068	**0.894** ± 0.044	**0.949** ± 0.014
	OpenHermes-7B	0.857 ± 0.042	0.835 ± 0.034	**0.878** ± 0.059	0.856 ± 0.041	0.936 ± 0.026
	Llama2-13B	0.865 ± 0.026	0.852 ± 0.044	0.871 ± 0.030	0.861 ± 0.028	0.938 ± 0.023
	Llama2-7B	0.821 ± 0.048	0.817 ± 0.067	0.835 ± 0.050	0.824 ± 0.049	0.911 ± 0.027
LLM-ANN	Mistral-7B	**0.891** ± 0.045	**0.916** ± 0.020	0.850 ± 0.076	**0.885** ± 0.052	**0.949** ± 0.013
	OpenHermes-7B	0.857 ± 0.052	0.836 ± 0.043	**0.877** ± 0.072	0.855 ± 0.054	0.935 ± 0.022
	Llama2-13B	0.868 ± 0.034	0.856 ± 0.046	0.871 ± 0.038	0.863 ± 0.039	0.938 ± 0.026
	Llama2-7B	0.823 ± 0.038	0.826 ± 0.062	0.831 ± 0.044	0.827 ± 0.039	0.916 ± 0.022
LLM-RF	Mistral-7B	**0.881** ± 0.032	**0.902** ± 0.038	0.857 ± 0.067	**0.878** ± 0.041	**0.946** ± 0.015
	OpenHermes-7B	0.853 ± 0.048	0.844 ± 0.042	0.853 ± 0.068	0.848 ± 0.051	0.930 ± 0.026
	Llama2-13B	0.864 ± 0.044	0.852 ± 0.061	**0.867** ± 0.040	0.859 ± 0.048	0.934 ± 0.032
	Llama2-7B	0.797 ± 0.029	0.806 ± 0.068	0.802 ± 0.039	0.801 ± 0.029	0.889 ± 0.014
LLM-GB	Mistral-7B	**0.877** ± 0.024	**0.896** ± 0.047	**0.857** ± 0.063	**0.874** ± 0.038	**0.887** ± 0.027
	openhermes-7B	0.799 ± 0.034	0.801 ± 0.036	0.790 ± 0.072	0.792 ± 0.021	0.839 ± 0.052
	Llama2-13B	0.834 ± 0.040	0.860 ± 0.049	0.788 ± 0.066	0.820 ± 0.042	0.879 ± 0.036
	Llama2-7B	0.805 ± 0.043	0.814 ± 0.070	0.802 ± 0.041	0.807 ± 0.048	0.839 ± 0.019
LLM-XGB	Mistral-7B	**0.869** ± 0.032	**0.880** ± 0.053	**0.860** ± 0.062	**0.868** ± 0.042	**0.944** ± 0.021
	OpenHermes-7B	0.832 ± 0.044	0.821 ± 0.029	0.838 ± 0.059	0.829 ± 0.037	0.918 ± 0.029
	Llama2-13B	0.857 ± 0.042	0.850 ± 0.070	0.852 ± 0.049	0.849 ± 0.051	0.921 ± 0.033
	Llama2-7B	0.817 ± 0.033	0.831 ± 0.067	0.816 ± 0.045	0.820 ± 0.031	0.876 ± 0.021
LLM-SVC	Mistral-7B	**0.898** ± 0.045	**0.913** ± 0.028	0.861 ± 0.077	**0.895** ± 0.049	**0.951** ± 0.014
	OpenHermes-7B	0.861 ± 0.045	0.836 ± 0.035	**0.886** ± 0.070	0.859 ± 0.046	0.935 ± 0.024
	Llama2-13B	0.868 ± 0.028	0.853 ± 0.045	0.877 ± 0.033	0.864 ± 0.032	0.939 ± 0.021
	Llama2-7B	0.825 ± 0.044	0.824 ± 0.060	0.836 ± 0.034	0.829 ± 0.041	0.911 ± 0.027
LLM-SC	Mistral-7B	**0.890** ± 0.031	**0.905** ± 0.044	**0.873** ± 0.059	**0.888** ± 0.042	**0.946** ± 0.018
	OpenHermes-7B	0.851 ± 0.026	0.838 ± 0.038	0.854 ± 0.052	0.845 ± 0.035	0.932 ± 0.026
	Llama2-13B	0.870 ± 0.044	0.870 ± 0.068	0.858 ± 0.031	0.864 ± 0.048	0.936 ± 0.029
	Llama2-7B	0.820 ± 0.027	0.834 ± 0.057	0.817 ± 0.043	0.823 ± 0.025	0.893 ± 0.017

levels, again performing equivalently.

Low p-values across accuracy, precision, and ROC-AUC confirm Mistral-7B's robust performance. While Llama2-13B occasionally approaches Mistral-7B's performance, it does not consistently excel across all metrics and classifiers. Therefore, Mistral-7B was selected for the remainder of the study due

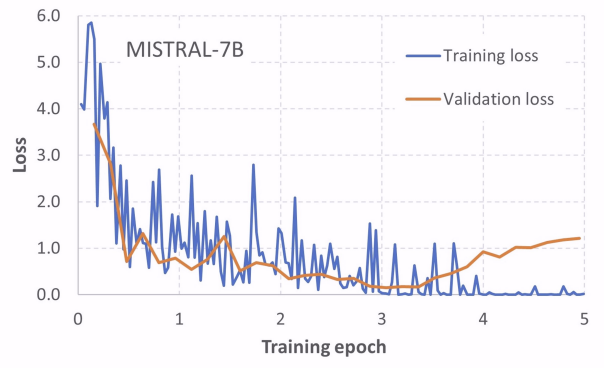

Fig. 8. Convergence plot obtained during one of the fine-tuning processes for Mistral-7B. The optimization was repeated for each fold.

to its overall superior performance metrics, demonstrated stability across different tests, and significant statistical outperformance of the alternatives in crucial aspects of classifier evaluation.

To complement the statistical analysis, we present Table III, which details the average performance metrics along with standard deviations for each LLM used in various classifiers, calculated across five folds.

In LR classifier, Mistral-7B achieved high scores with an accuracy of 0.887 ± 0.053 and precision of 0.913 ± 0.064. In the SVC, Mistral-7B led with an accuracy of 0.893 ± 0.046 and precision of 0.919 ± 0.038, indicating its effectiveness in minimizing false positives while maintaining high correct classifications. The ANN and RF classifiers showed less variation among the models. For the GB and XGB classifiers, while Mistral-7B maintained reasonable performance, with an accuracy of 0.851 ± 0.044 and precision of 0.875 ± 0.064, Llama2-13B and OpenHermes-7B provided comparable results. For SVC, Mistral-7B exhibited strong performance across all metrics, particularly excelling in precision and recall. By comparison, in the ANN classifier, while Mistral-7B showed similar precision to the LR model, it demonstrated higher accuracy and precision than in the RF classifier, indicating its capability to handle more complex patterns and dependencies in the data effectively.

A noteworthy observation is the slightly superior performance of SVC over ANN in terms of accuracy, recall, and F1 score, which could be attributed to SVC's strength in optimal hyperplane determination in high-dimensional space. The Stacked Classifier (SC), leveraging the strengths of multiple models, provided robust performance, closely matching the more complex ANN model, especially in terms of precision and ROC-AUC, despite being slightly less effective.

B. Assessment of the effect of LLM fine-tuning

To investigate the two-step training strategy, wherein a pre-trained LLM undergoes initial fine-tuning before the extraction of the last hidden state for classifier input, we conducted an additional analysis aimed to quantify the fine-tuning impact on AD detection with the Mistral-7B.Our objective was to ascertain whether the performance gains justified the additional computational costs incurred by the fine-tuning process. To give an idea of the computational burden, on a workstation furnished with an Nvidia GPU RTX 4090, fine-tuning the Mistral-7B, using a LoRA-based approach within the 5-fold CV framework, required approximately 15.27 minutes per fold. Moreover, the preliminary hyperparameter optimization phase, coupled with the final fine-tuning for each of the five training sets, cumulatively amounted to around 13 hours.

Figure 8 displays the convergence plot from fine-tuning Mistral-7B. Fine-tuning was conducted over a small number of epochs, as the validation loss plateaued early and then escalated, suggesting overfitting. This may be due to the high dimensionality of the parameter space relative to the dataset size, despite using LoRA to mitigate this. Decoupling the tuning of LLMs from classifier training helps manage this complexity and overfitting risk. Fine-tuning the LLM separately prevents biases and over-specialization, allowing for efficient classifier adjustments. This strategy preserves the fine-tuned state of the LLM, focusing computational resources on optimizing the classifier. This separation simplifies the integration of diverse inputs and facilitates the exploration of different classifier architectures. It enables a thorough evaluation of how variations in input and classifier design impact overall performance. This modular approach ensures that AD detection improvements result from both fine-tuning and the optimal combination of inputs and classifiers.

We conducted experiments using both zero-shot and fine-tuned LLM embeddings, initializing each classifier with different random states 25 times per metric. This produced 35 experimental outcomes per metric for each LLM configuration. We used the Wilcoxon signed-rank test [49], [51] to compare these paired samples, evaluating if fine-tuning significantly improved classifier performance. The p-values from the Wilcoxon test determined if observed performance differences were significant rather than random. For each metric, we calculated the mean and standard deviation of the performance scores, providing insights into the consistency and variability of performance gains from fine-tuning.

The results, shown in Table IV, reveal significant differences across all evaluated metrics, as indicated by the extremely low p-values obtained from the Wilcoxon signed-rank tests for accuracy, precision, recall, F1 score, and ROC-AUC. In terms of *Accuracy*, classifiers using fine-tuned embeddings achieved a higher average value (0.879) compared to those using zero-shot embeddings (0.700). The standard deviation for the fine-tuned models (0.038) is also lower than that for the zero-shot models (0.068), indicating more consistent performance across different runs. Similarly, *Precision* showed

TABLE IV
COMPARISON OF AVERAGE PERFORMANCE METRICS WITH AND WITHOUT FINE-TUNING OF MISTRAL-7B LLM, BASED ON A 5-FOLD CV.

Metric	Zero Shot	Fine-Tuned	p-value
Accuracy	0.700 ± 0.068	0.879 ± 0.038	$< 10^{-11}$
Precision	0.687 ± 0.081	0.901 ± 0.047	$< 10^{-11}$
Recall	0.697 ± 0.112	0.855 ± 0.068	$< 10^{-11}$
F1 Score	0.688 ± 0.082	0.876 ± 0.045	$< 10^{-11}$
ROC-AUC	0.794 ± 0.065	0.931 ± 0.028	$< 10^{-11}$

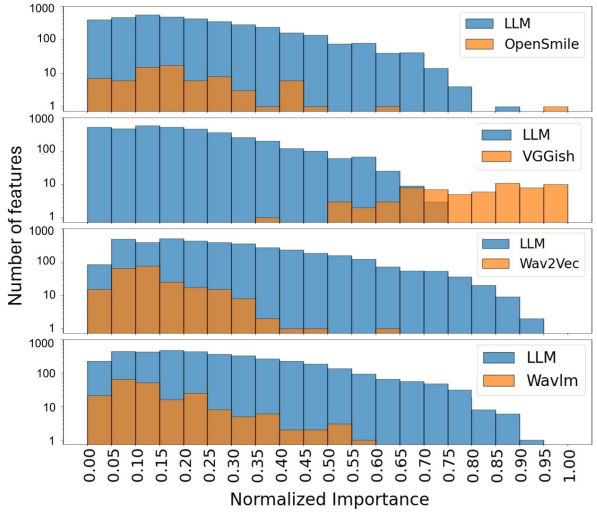

Fig. 9. Normalized feature importance for the different type of features. The number of features is reported in logarithmic scale.

a marked improvement, with fine-tuned classifiers averaging 0.901 compared to 0.687 for zero-shot. The *Recall* metric also improved with fine-tuning, from 0.697 to 0.855. This suggests that fine-tuned models are better at capturing relevant cases without being misled by noise. The F1 score, which balances the precision and recall, was significantly higher for fine-tuned classifiers (0.876) compared to zero-shot classifiers (0.688). These results, given the substantial gains in performance metrics, suggest that the computational and time investments in fine-tuning are justified, particularly in scenarios where accuracy, reliability, and robustness are critical.

C. Investigating LLM-A-X automatic AD detectors

In this section we study the effectiveness of using mixed linguistic-acoustic feature vectors for automatic AD detection, based on the LLM-A-X algorithm described in Section III-C. Using the fine tuned Mistral-7B LLM as linguistic feature vector extractor, we compared the performance of several automatic detectors of type LLM-A-SVC, where the acoustic features were those described in Section III and labelled as OpenSmile, VGGish, Wav2Vec and WavLM. In a first step of our analysis we used the different acoustic features jointly with several variations of SVC since, according to Table III, it provides a satisfactory balance between the different metrics. We tested SVC with linear kernel, RBF kernel and sigmoid kernel, each with regularization parameter selected from the set $\{0.1, 1.0, 100\}$. Moreover, to collect enough evidence on the algorithm performance we adopted the five-fold CV setting described above, repeating the training for each of the combination across five training and corresponding test sets. Consequently, for each metric and type of acoustic feature vector, we obtained results from 45 separate experiments (nine variations of SVC times five folds), enabling comprehensive statistical comparisons.

To obtain the final vector, the combined linguistic and acoustic features has been reduced according to the process explained in section III-D, which involves evaluating the importance of each single feature. To provide some insights into each feature type's contribution to the predictive capability, Figure 9 summarizes the results for one of the CV folds, with the number of selected features color-coded to distinguish between the two types. Note that the original size of LLM feature vector is 4096, while the sizes of vectors from OpenSmile, VGGish, Wav2Vec and WavLM, are 88, 64, 512 and 512, respectively. By applying the feature selection to the linguistic features only, we obtained a feature vector composed of 3767 values. In the mixed models, as can be seen in Figure 9, a substantial majority of the selected features are linguistic, with counts varying slightly depending on the acoustic feature combined. The acoustic features selected also varied, with each set showing a different impact on the model's effectiveness. In particular, in the case of LLM-OpenSmile, 3760 of the 4096 linguistic features and 72 of the 88 acoustic features were retained. In the case of LLM-VGGish, the algorithm selected 3738 linguistic features and all the 64 acoustic features. For LLM-Wav2Vec, 3825 LLM features and 224 acoustic features were selected, while for LLM-WavLM the final feature vector was composed of 3828 LLM features and 204 acoustic features.

In Figure 10, we present a comprehensive visual analysis of the performance across various metrics for different configurations of the LLM-A-X model, including an LLM-only version. The boxplots indicate that the medians of accuracy metrics are relatively clustered for all configurations, demonstrating a uniformly high baseline performance. This clustering suggests that the LLM model, both in its pure form and when integrated with various acoustic features, achieves robust effectiveness. Notably, the configuration involving LLM-VGGish showed less variability in accuracy, implying a more consistent performance across different test conditions. Regarding precision, the configurations that incorporate Wav2Vec exhibited higher median values. However, the box representing LLM-VGGish is smaller and the interquartile values are higher. The recall metric shows greater variability among the configurations, with LLM-VGGish consistently presenting a higher median value, suggesting its efficacy in retrieving a larger proportion of actual positive cases. This configuration also achieved a high median F1 score and the lowest variability, indicating a robust balance of precision and recall. Lastly, the ROC-AUC scores remained high across all configurations, with LLM-VGGish once again showing superior results and smaller variability.

We carried out the same type of statistical tests described in Section IV-A. Specifically, for each metric, we ranked the results of each LLM-A-SVC, including the version without acoustic features, within the 45 experiments, assigning a rank of 1 for the best performance and 5 for the worst. The CD diagram in Figure 11, based on the p-values from the post-hoc Nemenyi test, was used to visually represent the statistical differences using a confidence level of 0.05.

According to Figure 11, LLM-VGGish demonstrated superior performance in terms of accuracy, precision and F1, achieving the lowest rank of 2.3, 2.4 and 2.3, respectively. The CD plot shows that the performances of all the other methods were statistically equivalent. For recall, LLM-VGGish also exhibited the most robust performance, providing the lowest

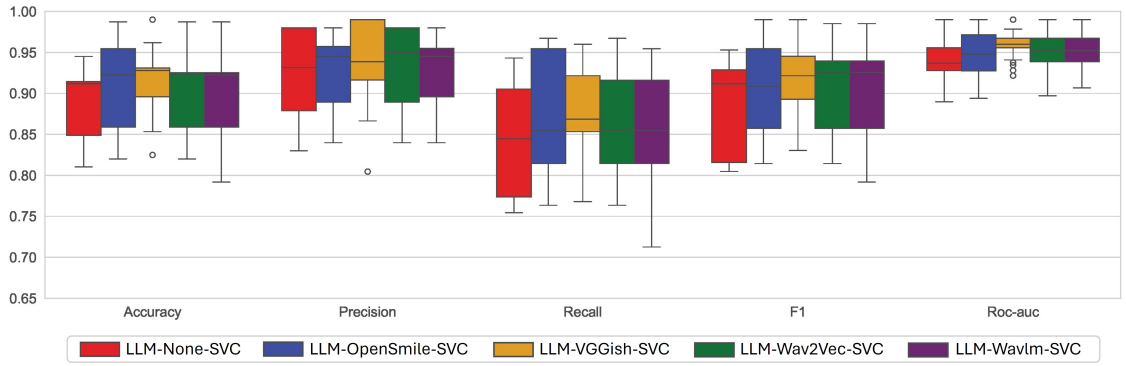

Fig. 10. Box plots summarizing all the results obtained by the LLM-A-SVC for all considered acoustic features and folds of the CV.

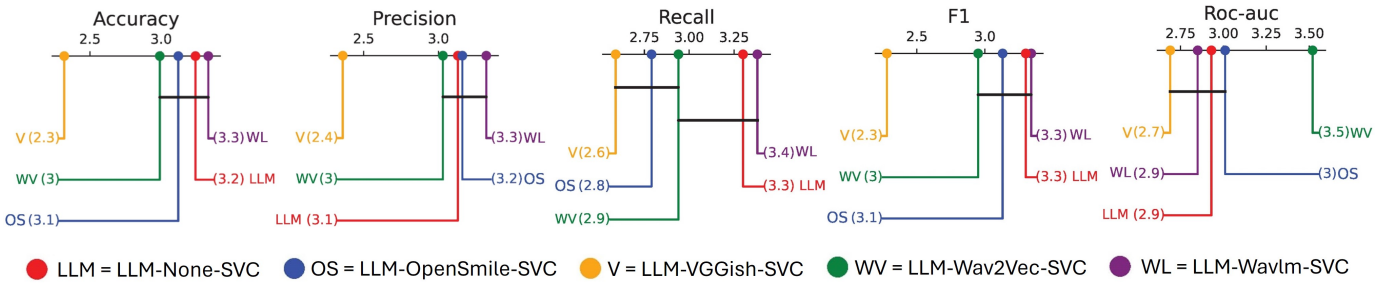

Fig. 11. Representations of the critical distances from the Nemenyi test comparing the average ranks of classifiers using different acoustic features combined with LLM embeddings. Classifiers connected by a horizontal line do not differ significantly, while those without a connecting line show statistically significant differences.

TABLE V
COMPARISON OF 5-FOLD CV AVERAGE PERFORMANCE METRICS WITH (FS) AND WITHOUT FEATURE SELECTION (NOFS) OF LLM-A-SVC USING DIFFERENT INPUT FEATURES AND A SVC WITH LINEAR KERNEL

	Features (linguistic, acoustic)	FS	NOFS	p_value
Accuracy	Mistral-7B, None	0.892 ± 0.048	0.891 ± 0.049	0.534
	Mistral-7B, OpenSmile	0.907 ± 0.051	0.891 ± 0.059	0.012
	Mistral-7B, VGGish	**0.918** ± 0.044	0.899 ± 0.051	0.003
	Mistral-7B, Wav2Vec	0.906 ± 0.050	0.888 ± 0.054	0.001
	Mistral-7B, WavLM	0.901 ± 0.048	0.882 ± 0.066	0.004
Precision	Mistral-7B, None	0.922 ± 0.049	0.914 ± 0.070	0.866
	Mistral-7B, OpenSmile	0.924 ± 0.048	0.897 ± 0.082	0.030
	Mistral-7B, VGGish	**0.946** ± 0.046	0.906 ± 0.070	0.003
	Mistral-7B, Wav2Vec	0.927 ± 0.051	0.900 ± 0.083	0.009
	Mistral-7B, WavLM	0.930 ± 0.044	0.904 ± 0.089	0.015
Recall	Mistral-7B, None	0.851 ± 0.067	0.860 ± 0.067	0.985
	Mistral-7B, OpenSmile	0.869 ± 0.072	0.835 ± 0.077	0.000
	Mistral-7B, VGGish	**0.882** ± 0.051	0.873 ± 0.068	0.009
	Mistral-7B, Wav2Vec	0.866 ± 0.068	0.835 ± 0.076	0.000
	Mistral-7B, WavLM	0.857 ± 0.069	0.829 ± 0.081	0.001
F1-Score	Mistral-7B, None	0.889 ± 0.053	0.896 ± 0.053	0.226
	Mistral-7B, OpenSmile	0.902 ± 0.058	0.886 ± 0.062	0.009
	Mistral-7B, VGGish	**0.918** ± 0.045	0.887 ± 0.052	0.000
	Mistral-7B, Wav2Vec	0.903 ± 0.057	0.889 ± 0.058	0.002
	Mistral-7B, WavLM	0.898 ± 0.054	0.888 ± 0.067	0.020
ROC-AUC	Mistral-7B, None	0.943 ± 0.028	0.948 ± 0.027	0.046
	Mistral-7B, OpenSmile	0.952 ± 0.028	0.938 ± 0.029	0.000
	Mistral-7B, VGGish	**0.959** ± 0.019	0.941 ± 0.024	0.000
	Mistral-7B, Wav2Vec	0.952 ± 0.026	0.938 ± 0.029	0.000
	Mistral-7B, WavLM	0.953 ± 0.025	0.940 ± 0.027	0.000

rank of 2.6. However, LLM-OpenSmile and LLM-Wav2Vec were statistically equivalent to the best. In terms of ROC-AUC, again LLM-VGGish outperformed the other configurations with the lowest rank of 2.7, resulting statistically superior only to LLM-Wav2Vec.

Further insights are provided by Table V, presenting the average performance metrics from the 5-fold CV and comparing the various configurations of the LLM-A-SVC based on different acoustic features, selected with the SVC method, alongside linguistic features extracted by the Mistral-7B model (column FS). Note that feature selection is applied independently within each fold of the cross-validation pipeline as part of the multimodal fusion process. This ensures that the selected dimensions are optimized for the specific training data of each fold, reflecting the natural variation across folds. Performance metrics are averaged across folds. The LLM-VGGish configuration achieves the highest average accuracy (0.918 ± 0.044), confirming that VGGish features are particularly effective in enhancing the model's ability to correctly classify AD cases. Moreover, LLM-VGGish provided the highest average precision (0.946 ± 0.046), showing the effectiveness of this feature set in identifying positive cases. Precision for other acoustic feature integrations (OpenSmile, Wav2Vec, WavLM) also exceeds that of the linguistic-only setup, reinforcing the benefit of multimodal data integration. All configurations with acoustic features exhibit a similar recall, slightly higher than the LLM baseline, with LLM-VGGish showing a slight advantage. LLM-VGGish also leads in average F1. All models based on acoustic features showed ROC-AUC scores slightly higher than LLM alone.

TABLE VI
RESULTS ON THE ADRESSO 2021 TEST SET USING BOTH THE LLM-X AND LLM-A-X ALGORITHMS, WITH MISTRAL-7B LLM FOR LINGUISTIC FEATURES AND VGGISH FOR ACOUSTIC FEATURES.

Classifier	Accuracy	Precision	Recall	F1 Score	ROC AUC
LLM-LR	0.873	0.882	0.857	0.870	0.958
LLM-ANN	0.901	0.868	0.943	0.904	0.964
LLM-RF	0.859	0.821	0.914	0.865	0.961
LLM-GB	0.873	0.861	0.886	0.873	0.873
LLM-XGB	0.845	0.816	0.886	0.849	0.846
LLM-SVC	0.915	0.914	0.914	0.914	0.976
LLM-SC	0.873	0.861	0.886	0.873	0.967
LLM-A-LR	0.887	0.909	0.857	0.882	0.950
LLM-A-ANN	0.887	0.909	0.857	0.882	0.950
LLM-A-RF	0.873	0.842	0.914	0.877	0.956
LLM-A-GB	0.873	0.861	0.886	0.873	0.873
LLM-A-XGB	0.845	0.816	0.886	0.849	0.846
LLM-A-SVC	**0.930**	**0.941**	**0.914**	**0.928**	**0.979**
LLM-A-SC	0.873	0.861	0.886	0.873	0.958

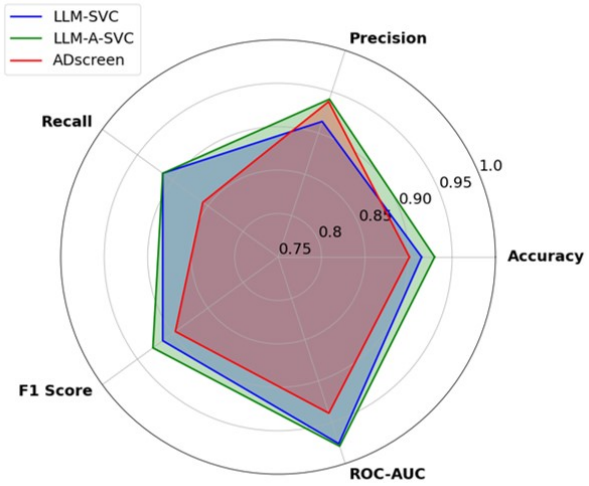

Fig. 12. Comparison between LLM-SVC, LLM-A-SVC and the best model from Zolnoori et al. (ADScreen).

Furthermore, Table V presents the results of experiments conducted with multimodal fusion without feature selection (column NOFS), enabling a comparison using the Wilcoxon signed-rank test [49], [51]. As shown, the results achieved using the proposed dimensionality reduction strategy (column FS) consistently outperformed the multimodal approaches without feature selection. The differences were statistically significant, with p-values below the significance level of 0.05. This confirms that combining linguistic and audio features, selected through appropriate techniques, enhances the model's performance.

Overall, the results discussed above, clearly illustrate the advantages of integrating acoustic features with linguistic data in the LLM-A-X models. The data also suggest that while all tested acoustic features improve model performance, the choice of specific features can be optimized based on the desired outcome, such as higher precision or better overall accuracy.

D. Final experiments and comparison with the literature

Finally, we tested the two approaches labelled as LLM-X and LLM-A-X, using the linguistic features based on the fine tuned Mistral-7B and the VGGish acoustic features. In addition, we experimented the variety of classifiers already described in Section IV-A. All the remaining settings and parameters are those described in previous experiments.

The results presented in Table VI, confirm that SVC, endowed with RBF kernel, was the classifier that better balanced all the performance metrics. Among all the classifiers, LLM-A-SVC achieved remarkable scores, with an accuracy of 0.930, a precision of 0.941 and a F1 score of 0.928.

Comparing LLM-SVC with LLM-A-SVC, we observe that the latter achieved better accuracy, precision, and F1 score. This indicates the added value of integrating acoustic features, which enhance the model's sensitivity to aspects of speech that purely linguistic models might miss. However, this integration did not significantly improve recall and ROC-AUC scores.

When comparing our results with the best reported in the recent literature, the proposed frameworks, particularly the LLM-A-SVC, exhibited superior performance metrics. For instance, the best performing model in [4] achieved an accuracy of 0.901 with a precision of 0.938, as reported in Table VII. In contrast, LLM-A-SVC model surpasses these metrics and provides a greater reliability as evidenced by its higher recall and F1 scores, which are important for clinical applications where missing diagnosis can have severe implications.

It is also interesting to highlight that the model from [4] demonstrated higher precision than our LLM-X configuration. However, LLM-X is better in terms of the remaining metrics. Figure 12 summarize the comparison from our best AD classifiers and the best proposed in [4].

As can be seen from Tables VI and VII, the proposed LLM-SVC, while achieving a lower recall compared to [32], significantly outperforms it in terms of the remaining metrics, demonstrating a more balanced and overall effective approach for AD detection. It is interesting to note that, although both methods employ fine-tuned advanced LLMs, without acoustic features, the algorithm reported in [32] is based on simultaneously training the classification layer and fine-tuning the LLM trough LoRA, which may enhance recall by better capturing subtle patterns directly related to AD features. In contrast, our approach uses a separate SVC for the final classification, which, while negatively affecting recall, contributes to improved overall performance and stability. In Figure 13, we show a comprehensive performance evaluation of the LLM-A-SVC algorithm through various diagnostic metrics.

The ROC curve demonstrates the algorithm's capability to distinguish between patients with and without AD across different threshold settings, achieving an AUC value of 0.98. This indicates, at least for the test set considered, a robust discriminatory power, which is essential for reliable medical diagnostics.

The Precision-Recall curve shows the trade-off between precision and recall. The curve initially maintains high precision with increasing recall, indicating robustness in identifying true positives without accumulating false positives. As expected,

TABLE VII
COMPARISON WITH SOTA RESULTS ON THE ADRESSO 2021 TEST SET. (ORDERED BY ACCURACY)

Method	Features	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Pérez-Toro et al. [17]	Acoustic and linguistic	0.803	n.a.	0.889	0.800	n.a.
Chen et al. [15]	Acoustic and linguistic	0.817	n.a.	n.a.	0.889	n.a.
Qiao et al. [28]	Linguistic	0.830	0.850	0.840	0.840	n.a.
Zhu et al. [26]	Acoustic and linguistic	0.831	0.871	0.771	0.818	n.a.
Rohanian et al. [52]	Acoustic and linguistic	0.840	n.a.	n.a.	n.a.	n.a.
Pappagari et al. [23]	Acoustic and linguistic	0.845	0.920	0.740	0.830	n.a.
Casu et al. [32]	Linguistic	0.901	0.854	**0.972**	0.909	n.a.
Zolnoori et al. [4]	Acoustic and linguistic	0.901	0.938	0.857	0.896	0.939
LLM-SVC (proposed)	Linguistic	0.915	0.914	0.914	0.914	0.976
LLM-A-SVC (proposed)	Acoustic and linguistic	**0.930**	**0.941**	0.914	**0.928**	**0.979**

there was a slight decline in precision at very high recall levels, reflecting the trade-off where the model begins to incorrectly classify some negatives as positives in its effort to include all possible positive cases.

The Cumulative Gains curve highlights the model's effectiveness in identifying a significant proportion of positive cases by targeting a smaller segment of the sorted sample. As can be seen, the curve sharply rises at the lower percentages of the sample, demonstrating that the model efficiently identifies a large proportion of positive cases among a small fraction of the population. This effectiveness is crucial for practical applications where prioritizing higher-risk individuals for further testing is beneficial.

The Positive Predictive Value curve, displayed across percentiles of predicted probabilities, shows a certain decline as the threshold decreased. This highlights that as the model attempts to capture more positive cases by lowering the threshold, the number of false positives increases, reducing the overall precision of the model.

The Sensitivity curve demonstrates a clear upward trend as the percentile of predicted probability increases. This indicates that as the model raises its threshold for predicting positive cases, it successfully identifies a higher number of true positives, thereby increasing sensitivity. This trend is significant in medical diagnostics where minimizing false negatives (missed true cases) is crucial. The curve illustrates that at higher thresholds, the model achieves near-perfect sensitivity, highlighting its robustness in capturing almost all positive cases of AD.

Finally, the Prediction Density curve, illustrating the probability distribution of both classes at various thresholds, showed minimal overlap between the areas.

According to the above comparison and analysis, the proposed approach, particularly the LLM-A-SVC, exhibited robust performance characteristics in terms of AD detection. In particular, the comprehensive evaluation using multiple diagnostic metrics showed the model's capability to operate efficiently across different operational thresholds, maintaining a balance between sensitivity and specificity that is crucial for clinical usage. Moreover, LLM-A-SVC model overcomes the performance of current SotA models reported in the literature, offering enhanced reliability.

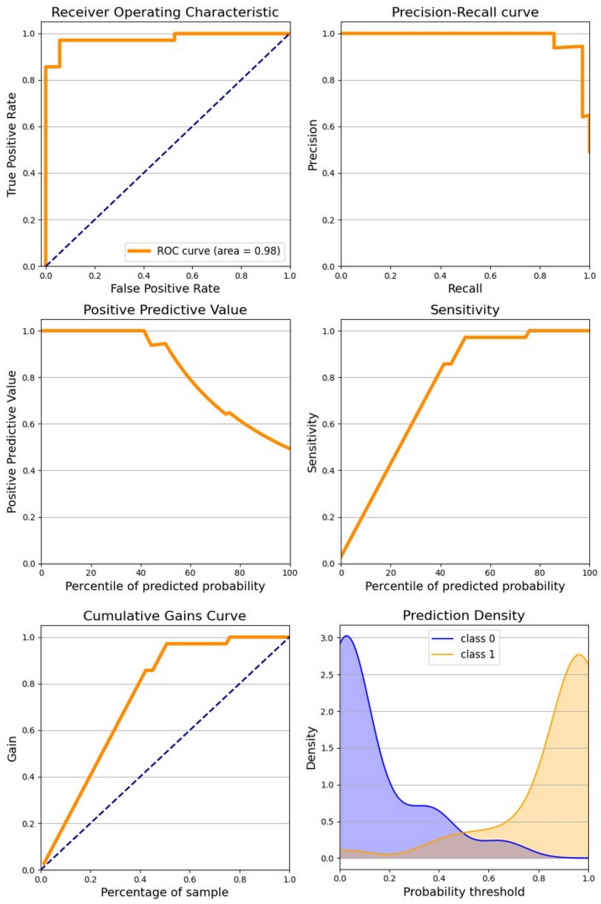

Fig. 13. Comprehensive Performance Evaluation of the LLM-A-SVC Algorithm for Alzheimer's Disease Detection.

V. DISCUSSION

The exploration of different pre-trained modern LLMs, coupled with various classifiers, provided extensive insights into their suitability for AD detection.

The considered transformer-based LLMs varied in performance, with some better capturing linguistic markers of cognitive decline. The classifiers, ranging from LR to more complex ensemble methods like RF and GB, provided useful insights on the different ML approaches that can enhance the advantages of using LLM outputs in clinical settings. Among the tested LLMs, Mistral-7B provided the best feature vectors for the AD classification task. Moreover, among the several classifiers, SVC feed by Mistral-7B led to one of the best balances between the different performance metrics.

We found that fine-tuning the LLMs on AD-specific data is

a critical factor in enhancing model performance. In fact, fine-tuned models significantly outperformed the zero-shot counterparts, exhibiting improvements in all performance metrics. This improvement highlights the adaptive capability of LLMs when they are trained on targeted datasets, allowing them to improve the ability to capture the linguistic features associated with cognitive impairments. It is reasonable to assume that the fine-tuning process can enhance the sensitivity of the LLM to the early signs of AD, which are often missed by more generalized models.

Although classifiers based on modern LLMs alone proved very effective in the AD detection task, the integration of multimodal data, combining linguistic features from LLMs with acoustic data, demonstrated superior performance over unimodal approaches. Among the considered acoustic feature vectors, the approach based on VGGish data, preprocessed by a specifically devised GRU-based supervised autoencoder, proved the most effective. The multimodal framework, particularly the LLM-A-SVC, achieved the highest marks in distinguishing between AD and non-AD cases, indicating a certain advantage of considering speech pattern analysis along with traditional linguistic analysis. This is in agreement with existing literature, which suggests that acoustic features, capturing nuances in tone, pitch, and cadence, can provide critical diagnostic information that text-based data alone cannot reveal.

The performance of both LLM-SVC and LLM-A-SVC was benchmarked against existing solutions as documented in recent literature, particularly on the ADReSSo 2021 test set. The LLM-A-SVC model not only outperformed LLM-SVC, as expected given the CV analysis, but also exhibited superior metrics compared to the best-reported models in the literature. This comparison underlines the effectiveness of the proposed approach and validate the improvements our framework brings over current methodologies, offering potential for real-world application where reliable, early detection of AD can significantly affect patient care and management.

VI. Conclusion

This study systematically investigated the potential of finely-tuned, multimodal AI systems for the enhanced detection and diagnosis of AD. By leveraging both linguistic and acoustic data, our models significantly improved the sensitivity and specificity of diagnostic tools, addressing the complex symptomatology of AD more effectively than previous AI-based methods.

Our research used pre-trained LLMs fine-tuned on AD-specific datasets, combined with advanced acoustic feature extraction. Integrating linguistic features from models like Mistral-7B with acoustic features from VGGish yielded superior performance, achieving higher accuracy, precision, recall, F1, and ROC-AUC metrics compared to unimodal approaches and literature benchmarks.

Among the key findings of our research, we outperformed the best-reported results in the literature on the ADReSSo 2021 test set. Moreover, our experiments confirmed that the multimodal approach can provide a significant increase in diagnostic accuracy, reducing the rate of undetected cases and misdiagnoses. Furthermore, we found that different ML techniques ensured robust performance, with SVC achieving the best balance of all evaluated metrics.

Future studies should integrate additional modalities like visual or behavioral data to enhance detection. Investigating the potential of integrating real-time monitoring and longitudinal data could also provide deeper insights into the progression of AD and other cognitive impairments.

The continued innovation in this field is essential for developing AI-driven diagnostic tools that can effectively support clinicians. These tools not only have the potential to improve patient outcomes through early and accurate diagnosis but also to alleviate some of the burdens on healthcare systems and caregivers. Future exploration of the ADReSSo regression task would require efforts similar to the classification task but could significantly enhance the study by providing deeper insights into cognitive decline and disease progression, complementing the detection of clinical conditions.

In conclusion, our study highlights the high potential of multimodal AI systems in the early detection and diagnosis of AD. By advancing the integration of diverse data types, we can move towards more robust and effective diagnostic methodologies, ultimately contributing to better management and treatment of AD.

Acknowledgment

This work has been developed within the framework of the project e.INS-Ecosystem of Innovation for Next Generation Sardinia (cod. ECS 00000038) funded by the Italian Ministry for Research and Education (MUR) under the National Recovery and Resilience Plan (NRRP) - MISSION 4 COMPONENT 2, "From research to business" INVESTMENT 1.5, "Creation and strengthening of Ecosystems of innovation" and construction of "Territorial R&D Leaders" - CUPJ83C21000320007.

References

[1] "World health organization," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] I. Martínez-Nicolás, T. E. Llorente, F. Martínez-Sánchez, and J. J. G. Meilán, "Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: a systematic review article," *Frontiers in Psychology*, vol. 12, 2021.

[3] Z. Liu, E. J. Paek, S. O. Yoon, D. Casenhiser, W. Zhou, and X. Zhao, "Detecting alzheimer's disease using natural language processing of referential communication task transcripts," *Journal of Alzheimer's Disease*, vol. 86, no. 3, pp. 1385–1398, 2022.

[4] M. Zolnoori, A. Zolnour, and M. Topaz, "Adscreen: A speech processing-based screening system for automatic identification of patients with alzheimer's disease and related dementia," *Artificial Intelligence in Medicine*, vol. 143, p. 102624, 2023.

[5] M. Tanveer, A. H. Rashid, M. Ganaie, M. Reza, I. Razzak, and K.-L. Hua, "Classification of alzheimer's disease using ensemble of deep neural networks trained through transfer learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1453–1463, 2021.

[6] S. Bringas, R. Duque, C. Lage, and J. L. Montaña, "Cladsi: Deep continual learning for alzheimer's disease stage identification using accelerometer data," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 6, pp. 3401–3410, 2024.

[7] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," in *Proceedings of Interspeech 2020*. ISCA, 2020, pp. 2172–2176.

[8] ——, "Detecting cognitive decline using speech only: The adresso challenge," in *Proceedings of Interspeech 2021*. Brno, Czechia: ISCA, 2021, pp. 3780–3784.

[9] R. Pappagari, J. Cho, L. Moro-Velazquez, and N. Dehak, "Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity." in *Interspeech*, 2020, pp. 2177–2181.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

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

[12] A. Pompili, T. Rolland, and A. Abad, "The inesc-id multi-modal system for the adress 2020 challenge," *arXiv:2005.14646*, 2020.

[13] M. Martinc and S. Pollak, "Tackling the adress challenge: A multimodal approach to the automated recognition of alzheimer's dementia." in *Interspeech*, 2020, pp. 2157–2161.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[15] J. Chen, J. Ye, F. Tang, and J. Zhou, "Automatic detection of alzheimer's disease using spontaneous speech only," in *Interspeech*, vol. 2021. NIH Public Access, 2021, p. 3830.

[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," *arXiv preprint arXiv:2106.09668*, 2021.

[17] P. A. Pérez-Toro, S. P. Bayerl, T. Arias-Vergara, J. C. Vásquez-Correa, P. Klumpp, M. Schuster, E. Nöth, J. R. Orozco-Arroyave, and K. Riedhammer, "Influence of the interviewer on the automatic assessment of alzheimer's disease in the context of the adresso challenge." in *Interspeech*, 2021, pp. 3785–3789.

[18] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, "Exploring deep transfer learning techniques for alzheimer's dementia detection," *Frontiers in computer science*, vol. 3, p. 624683, 2021.

[19] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, "Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition," *arXiv:2009.04070*, 2020.

[20] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," pp. 131–135, 2017.

[21] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech." in *Interspeech*, vol. 2020, 2020, pp. 2222–6.

[22] Z. Shah, J. Sawalha, M. Tasnim, S.-a. Qi, E. Stroulia, and R. Greiner, "Learning language and acoustic models for identifying alzheimer's dementia from speech," *Frontiers in Computer Science*, vol. 3, p. 624659, 2021.

[23] R. Pappagari, J. Cho, S. Joshi, L. Moro-Velázquez, P. Żelasko, J. Villalba, and N. Dehak, "Automatic Detection and Assessment of Alzheimer Disease Using Speech and Language Technologies in Low-Resource Scenarios," in *Proc. Interspeech 2021*, 2021, pp. 3825–3829.

[24] 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," *arXiv:2008.01551*, 2020.

[25] A. Balagopalan, B. Eyre, J. Robin, F. Rudzicz, and J. Novikova, "Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech," *Frontiers in aging neuroscience*, vol. 13, p. 635945, 2021.

[26] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, "WavBERT: Exploiting Semantic and Non-Semantic Speech Using Wav2vec and BERT for Dementia Detection," in *Proc. Interspeech 2021*, 2021, pp. 3790–3794.

[27] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," in *Interspeech 2019*, 2019, pp. 3465–3469.

[28] Y. Qiao, X. Yin, D. Wiechmann, and E. Kerz, "Alzheimer's Disease Detection from Spontaneous Speech Through Combining Linguistic Complexity and (Dis)Fluency Features with Pretrained Language Models," in *Proc. Interspeech 2021*, 2021, pp. 3805–3809.

[29] A. Roshanzamir, H. Aghajan, and M. Soleymani Baghshah, "Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech," *BMC Medical Informatics and Decision Making*, vol. 21, pp. 1–14, 2021.

[30] L. Gómez-Zaragozá, S. Wills, C. Tejedor-Garcia, J. Marín-Morales, M. Alcañiz, and H. Strik, "Alzheimer disease classification through asr-based transcriptions: Exploring the impact of punctuation and pauses," *arXiv:2306.03443*, 2023.

[31] T. Wang, J. Deng, M. Geng, Z. Ye, S. Hu, Y. Wang, M. Cui, Z. Jin, X. Liu, and H. Meng, "Conformer based elderly speech recognition system for alzheimer's disease detection," *arXiv:2206.13232*, 2022.

[32] F. Casu, E. Grosso, A. Lagorio, and G. A. Trunfio, "Optimizing and evaluating pre-trained large language models for alzheimer's disease detection," in *2024 32nd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. IEEE, 2024, pp. 277–284.

[33] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[34] W. Kong, "Exploring neural models for predicting dementia from language," Ph.D. dissertation, University of British Columbia, 2019.

[35] F. Bertini, D. Allevi, G. Lutero, L. Calzà, and D. Montesi, "An automatic alzheimer's disease classifier based on spontaneous spoken english," *Computer Speech & Language*, vol. 72, p. 101298, 2022.

[36] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[37] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[38] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv*, 2021. [Online]. Available: https://ar5iv.org/abs/2106.09685

[39] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," 2019.

[40] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[41] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[42] Teknium, "Openhermes-7b," 2023. [Online]. Available: https://huggingface.co/meta-llama/Llama-2-7b

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

[44] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv:1904.10509*, 2019.

[45] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[46] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, p. 1505–1518, Oct. 2022.

[47] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[48] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794.

[49] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced non-parametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, 2010.

[50] M. Friedman, "A Comparison of Alternative Tests of Significance for the Problem of m Rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[51] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[52] M. Rohanian, J. Hough, and M. Purver, "Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs," *arXiv preprint arXiv:2106.15684*, 2021.