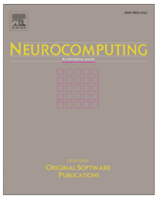
Neurocomputing 548 (2023) 126413

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



A Deep Learning-Based Multimodal Architecture to predict Signs of Dementia

David Ortiz-Perez [a,*], Pablo Ruiz-Ponce [a], David Tomás [b], Jose Garcia-Rodriguez [a], M. Flores Vizcaya-Moreno [c], Marco Leo [d]

[a] Department of Computer Science and Technology, University of Alicante, Carr. de San Vicente del Raspeig, San Vicente del Raspeig, 03690 Alicante, Spain
[b] Department of Software and Computing systems, University of Alicante, Carr. de San Vicente del Raspeig, San Vicente del Raspeig, 03690 Alicante, Spain
[c] Unit of Clinical Nursing Research, Faculty of Health Sciences, University of Alicante, Carr. de San Vicente del Raspeig, San Vicente del Raspeig, 03690 Alicante, Spain
[d] Institute of Applied Sciences and Intelligent Systems (National Research Council of Italy), 73100 Lecce, Italy

ARTICLE INFO

Article history:
Received 1 February 2023
Revised 5 May 2023
Accepted 29 May 2023
Available online 5 June 2023
Communicated by Zidong Wang

Keywords:
Multimodal
Deep learning
Transformers
Dementia prediction

ABSTRACT

This paper proposes a multimodal deep learning architecture combining text and audio information to predict dementia, a disease which affects around 55 million people all over the world and makes them in some cases dependent people. The system was evaluated on the DementiaBank Pitt Corpus dataset, which includes audio recordings as well as their transcriptions for healthy people and people with dementia. Different models have been used and tested, including Convolutional Neural Networks (CNN) for audio classification, Transformers for text classification, and a combination of both in a multimodal ensemble. These models have been evaluated on a test set, obtaining the best results by using the text modality, achieving 90.36% accuracy on the task of detecting dementia. Additionally, an analysis of the corpus has been conducted for the sake of explainability, aiming to obtain more information about how the models generate their predictions and identify patterns in the data.

1. Introduction

Nowadays, around 55 million people in the world have dementia, which is more commonly seen in older people but can also affect younger ones. Dementia is a syndrome which affects normal cognitive functions. The most common form of dementia is Alzheimer's disease, which represents 60–70% of the cases [37]. This syndrome can affect each patient in different ways and has three different stages: early, middle, and late stage. Each stage can have different symptoms that can vary from losing track of time in the early stage, forgetting recent events or becoming confused at home in the middle stage, and finally having difficulties recognising relatives or friends in a late stage. The late stage of dementia limits the autonomous life of patients, so they will need a relative or a professional to take care of them. Since dementia is usually related to older people, the number of patients who have dementia is expected to grow in the following years. This is a problem which will be more and more present in our society. Consequently, early detection is critical to limit this disease.

The contribution of this work is the creation of an architecture composed of different deep learning modules, which process text transcriptions and audio recordings, to predict symptoms of dementia, especially in the early stages of the disease. Even though at the present time there is no cure for dementia, there are treatments with or without medicines, such as therapies, that can help with the symptoms of the patients. For this reason, the early detection of this syndrome is important, since this detection and treatment can improve the quality of life of patients, their relatives, and friends.

The present work continues and expands on the study conducted by Ortiz-Perez, Ruiz-Ponce, Tomás and Garcia-Rodriguez [26]. On one hand, a research was carried out to identify suitable datasets for this task that include patients who suffer or may suffer from dementia in the future. On the other hand, the most promising deep learning techniques for text and audio classification were analysed and tested. The chosen dataset for the experiments proposed was the DementiaBank Pitt Corpus [5], which contains data in different modalities, including text and audio. A deep learning implementation using both modalities, separately and combined, was tested. Finally, in view of the good results obtained using only textual information, an analysis was carried out on the textual part of the dataset for the sake of explainability.

* Corresponding author.
E-mail addresses: dortiz@dtic.ua.es (D. Ortiz-Perez), pruiz@dtic.ua.es (P. Ruiz-Ponce), dtomas@dlsi.ua.es (D. Tomás), jgarcia@dtic.ua.es (J. Garcia-Rodriguez), flores.vizcaya@ua.es (M.F. Vizcaya-Moreno), marco.leo@cnr.it (M. Leo).

https://doi.org/10.1016/j.neucom.2023.126413
0925-2312/© 2023 The Author(s). Published by Elsevier B.V.

The remaining of the paper is organized as follows: Section 2 introduces related work regarding dementia datasets and multimodal approaches; Section 3 describes in detail the dataset used to test the present work; in Section 4 the different approaches are presented, validating them on the DementiaBank Pitt Corpus dataset and showing their performance in Section 5; Section 6 analyses the textual part of the dataset and Section 7 gives additional information on how the model makes predictions; finally, Section 8 summarises conclusions and proposes future work.

2. Related work

This section presents relevant work carried out on the dementia prediction tasks and the datasets available to this end. Additionally, a review of recent approaches dealing with multimodal information is included.

2.1. Datasets

As dementia is a highly sensitive topic, the number of datasets available for public use is limited for privacy reasons. In the medical domain, there are varied datasets on medical data, including blood test results [19] and Magnetic Resonance Imaging [38], but the focus of this section is exclusively on dementia pathology.

The following datasets were found in this area: DementiaBank Pitt Corpus [5], DemCare [13], and Praxis Gesture [24]. The DementiaBank dataset consists of a set of audio recordings in which patients having dementia or not are asked to describe a clinical image. The DemCare dataset is another multimodal dataset that contains information in video and audio modalities, including also information from physiological sensors. In this case, elderly patients (healthy or with dementia) were recorded while performing activities from their daily life, such as reading an article, watering plants, or preparing a drink. Finally, the PRAXIS Gesture dataset consists of a set of videos where actors are recorded from a front view making simple gestures, repeating them until they do them correctly.

As shown in Table 1, where the main features of each dataset are exposed, the PRAXIS Gesture dataset is the only one that does not include multimodal data. The DementiaBank dataset provides audio and their transcription. Finally, the DemCare dataset includes video, audio and physiological sensors data. All these datasets include patients who suffer from dementia and healthy patients, but they perform different tasks in each dataset, from simple ones like gestures to activities of their daily life or descriptions of an image.

Table 1
Dementia datasets identified, including modalities and activities represented.

Dataset	Modalities	Activity
DementiaBank Pitt Corpus	Audio and text	Describe an image
DemCare	Video, audio, and physiological	Activities of daily life
PRAXIS Gesture	Video	Basic gestures

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

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

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

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

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

2.2. Multimodality

In their daily lives, people perceive the world with more than one sense, for example, by combining visual and auditory stimuli. This is the basic idea behind multimodal approaches that deal with data in different modalities. An example of multimodal data is an image and a text describing it, combining visual and textual information in this way. This combination can improve the results in understanding the scene present in the image, since modalities can complement each other, giving a more complex and detailed perspective of the situation.

There are different approaches to combining modalities. One ways is the early fusion approach, combining modalities that are similar into a single vector to fit later a unified model for both modalities [34]. Another approach is the late fusion approach, in which both modalities are processed by different and independent models for each one. A straightforward approach to combining modalities is to have two different models that classify data and make a weighted sum of both to obtain the final classification. Another direct way is to obtain feature vectors for each modality from their respective models, concatenating them into a single one, and finally making another model that performs the final task using this single vector.

There are relevant implementations of these types of multimodal models, such as MMF [33], CLIP [25], and VATT [2]. MMF is a modular framework developed by FacebookAI used for vision and language multimodal research. This framework contains implementations of state-of-the-art vision and language models such as VisualBERT [16] and VilBERT Lu, Batra, Parikh and Lee [17], as well as different datasets to work on like VQA [4]. CLIP is another multimodal model which relates text and images. CLIP is pretrained using a set of images and a set of textual content to establish relationships between them. The VATT model is an approach whose main aim is to analyse video, audio, and text at the same time. Each modality has a Transformer encoder to process the input and finally a projection head to get the similarity between all those modalities using contrastive losses. This approach can be used later for downstream tasks in a variety of fields, such as video action recognition and audio event classification.

3. DementiaBank Pitt Corpus dataset

This section describes in detail the DementiaBank Pitt Corpus, which is the dataset chosen to evaluate the multimodal approach proposed in this work. This dataset contains the audio of the recording as well as a transcription of the dialogue between the interviewer and the patients. The range of age of the patients goes from 46 to 90 years, including both male and female patients. The statistics about the number of healthy (control group) and dementia patients, including the number of samples (i.e., audio recordings, since a patient can have multiple recordings) is shown in Table 2.

In order to gather this dataset, subjects were asked to describe an image. Specifically, the image used was the cookie theft picture shown in Fig. 1. This image has been used in clinical and experimental research, specifically in the field of mental and cognitive impairments.

This experiment was designed to detect some of the signs of dementia, such as having difficulties choosing the right words, choosing the wrong ones, using related or substitute words, or even not finding a word at all. Other signs shown include using words with no meaning or not related to the conversation [3].

Among recent studies working over this dataset, it is worth mentioning the work by Warnita, Inoue and Shinoda [36]. In this work, the authors used only audio data in the corpus, and the model used was a Gated CNN, achieving 73.6% accuracy. Another work that uses only audio data from the corpus is the one presented by Chakraborty, Pandharipande, Bhat and Kopparapu [6]. The authors proposed a model that analyses the audio clips in order to obtain audio biomarkers for the detection of dementia.

There are also works working only on textual modality, as in the case of [14]. In this work, the best results obtained were achieved by using CNNs combined with Recurrent Neural Networks (RNN) and the POS-tagging (morphological information) transcriptions of the utterances. The best results to date were obtained in this work, achieving an accuracy of 91.1%. The data used was downsampled because not every utterance had an accompanying POS-tagged transcription.

There are also different approaches where both modalities were combined. Such is the case of the work by Mittal, Sahoo, Datar, Kadiwala, Shalu and Mathew [23]. In this study, the authors used both modalities and two different models to weight the probability of dementia. For audio processing, a Mel Spectrogram [32] combined with an audio based model were used. For the textual part, different combinations of segment transcriptions and the full transcription were used. By using this model, the authors obtained an accuracy of 85.3%. For this multimodal project developed by Roberts [32], the authors implemented two different models to process audio and text making a prediction of whether the patient has dementia or not. After both predictions, a ponderated sum of weights was made in order to achieve a final multimodal prediction.

Te DementiaBank Pitt Corpus has a smaller subset which has been balanced in terms of age and gender, called ADReSS challenge [18]. Different works have also proposed solutions on it [22,8,21]. In the work done by Mahajan and Baths [21], the authors propose various multimodal models that combine different modalities, using raw data as text, using the original data, or extracting specific features. In contrast to the work in [32], this work extracts features from various modalities and later combines them. After combining the features through concatenation, a standard classification is performed using dense layers.

This is not the only corpus available from the DementiaBank dataset, this dataset contains other corpuses with different experiments and even in other languages. Among the available languages, we can find English, German, Mandarin, Spanish, and Taiwanese. But, Pitt Corpus and its subset, ADReSS, are the ones with more data and the most used ones in this area. The aforementioned corpus in other languages are rarely used in academic research.

Table 2
Statistics of the DementiaBank Pitt Corpus Statistics for dementia and healthy (control) patients.

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

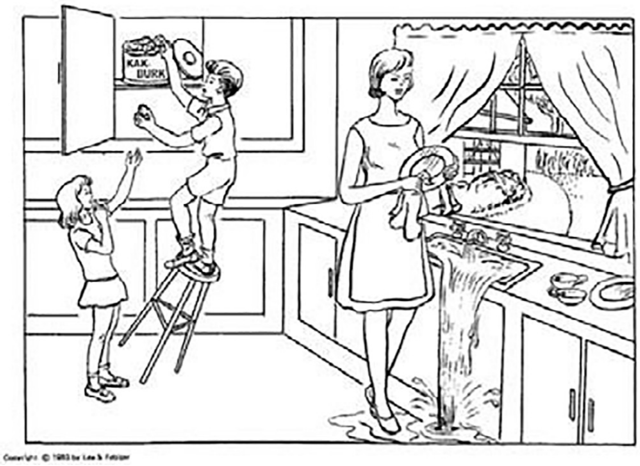


Fig. 1. The Cookie Theft Picture [15].

4. Approach

The initial dataset was split into two different subsets: the training set, with a size of 469 samples (representing 85% of the dataset) and a test set, comprising 83 samples (representing the remaining 15%). As commented before, this dataset contains data in two different modalities: the audio of the recordings and their transcriptions. To address this situation, different models were developed for each modality in order to evaluate their individual performance, finally combining them into a multimodal model trying to improve the results of the single modality approach. The remainder of this section describes the individual modalities and the combination of them in more detail.

4.1. Audio

This section describes the approach to audio analysis. Fig. 2 presents the architecture of the model implemented.

In this approach, each audio file is converted to its waveform (a graphical representation of the signals over time) and then converted to a Mel Spectrogram [32], which is a spectrogram in the Mel Scale. This scale is inspired by the way humans perceive sounds, differentiating the low frequency sounds more easily than the high frequency sounds. In the Mel Scale, two equally distanced sounds in the pitch sound equally to a listener. A CNN was used to process the spectrogram obtained. This implementation based on CNN and Mel Spectrogram of an audio achieved the state-of-the-art in previous audio classification tasks [27,10]. For this reason, it was chosen as the architecture for the present work.

The CNN model handles the spectrogram as an image. This is because these neural network architectures are used basically for

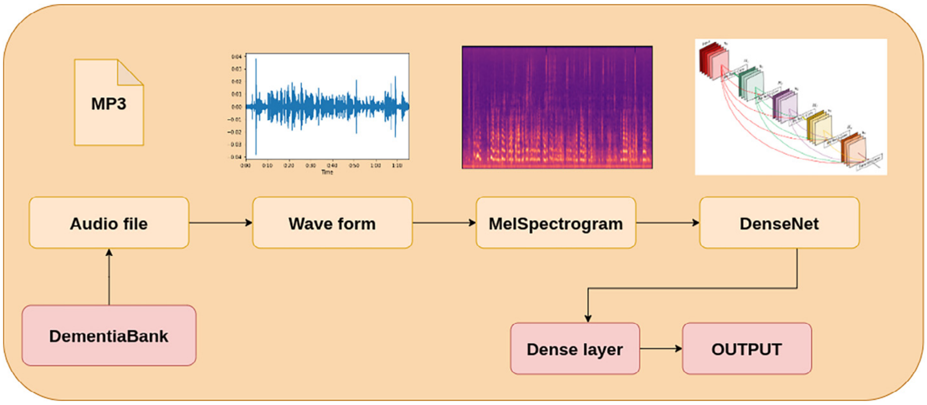


Fig. 2. Architecture of the model for audio analysis.

image tasks, such as image classification. They apply different convolutions/kernels in order to extract features from the image. In this proposal, different pre-trained CNNs have been tested: MobileNet [11], DenseNet [12], and ResNet He, Zhang, Ren and Sun [9]. The final step of this audio model is a dense layer with the outputs of the CNN in order to get as a final output the prediction of the model: dementia or healthy patient.

The best results in these experiments were obtained with the DenseNet model. In Section 5, the results obtained from the audio model are those obtained using DenseNet. The results of MobileNet and ResNet were discarded.

4.2. Text

The textual information in the corpus was encoded in CHAT format [20], a format used by TalkBank[1] in their corpora, including the DementiaBank Pitt Corpus. This format not only includes the transcription of the patients but also the transcription of the interviewer, as well as personal information from the patients and special flags representing pauses or mistaken words. For this reason, the original transcription files have been pre-processed in order to obtain a clean text transcription.

Fig. 3 shows the architecture of the model proposed for text analysis.

As shown in the architecture, the BERT model [7] is used for text analysis, a model that has achieved state-of-the-art results in many natural language processing (NLP) tasks. This model is based on the Transformers architecture [35], stacking different encoders that extract features from the text. The most interesting aspect of this architecture is the use of attention layers in order to establish relations between different words in the sentence.

In order to use the BERT model for this task, it was previously fine-tuned. After a text tokenization process, the model receives encoded words as input and returns word embeddings, which are different for each word and have a size of 768 dimensions per embedding. These embeddings are a way of representing a word as dense numerical vectors, enabling it to establish similar representations for words with similar meanings.

One way of fine-tuning BERT is to use the embeddings associated with the `[CLS]` token in a final dense layer for classification. This `[CLS]` embedding represents the whole text. However, in the experiments, this was not the only method tested for text representation. Another model used these embeddings to fit a bidirectional Long Short-Term Memory (LSTM) network. LSTMs have proven to produce good results for tasks using sequences of text, and prior to the introduction of Transformers, they were the state-of-the-art in many NLP tasks. LSTMs use the output of the words or embeddings of a layer as the input of the following one, retaining information from previous steps. They also have a mechanism to forget irrelevant data from previous segments and retain important ones. Finally, the output of the LSTM is used as input for a final dense layer to obtain the model's output. Section 5 shows the comparison of the results obtained by BERT and LSTM.

4.3. Multimodal

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

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

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

Fig. 4 shows the architecture of the model proposed for the multimodal analysis.

4.4. Other approaches

As mentioned before, the best results on this dataset were obtained in [14] by combining text features with the POS-tagging of the text. For this reason, this information was also included in the proposed model to test if it improved the results. After obtaining the POS-tagging of the text by using spaCy[2], the features were introduced in a word embedding and a bidirectional LSTM. This implementation has been done by using the embedding and LSTM lay-

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

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

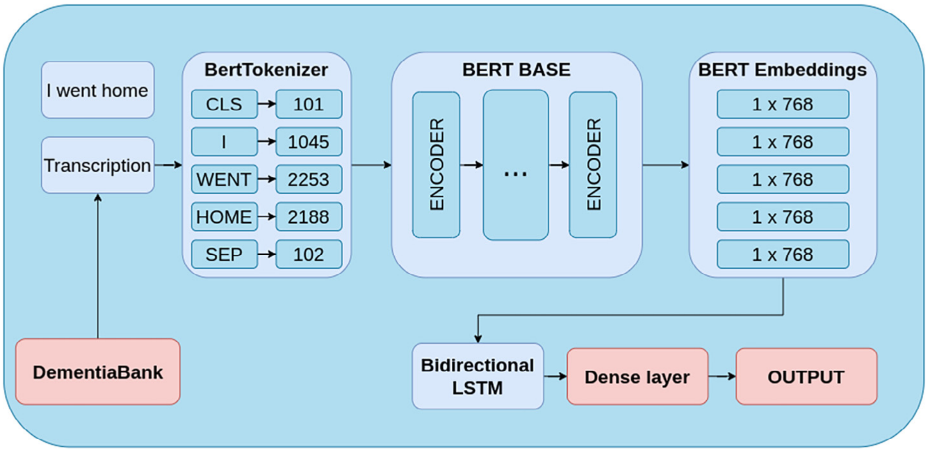


Fig. 3. Architecture of the model for text analysis.

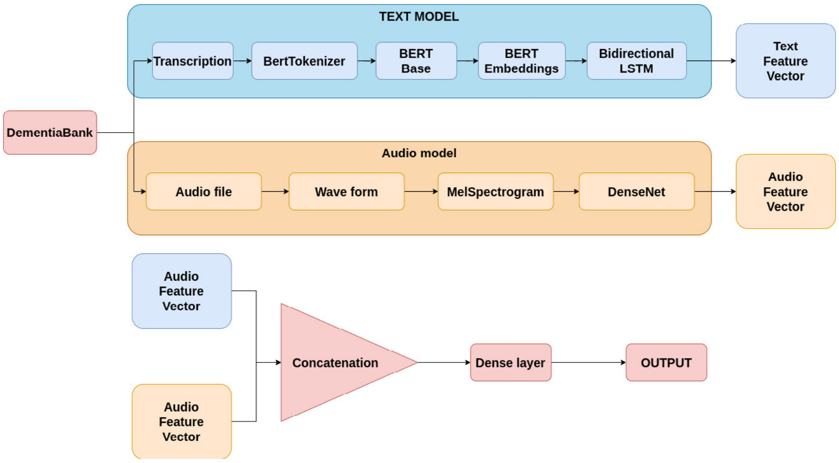


Fig. 4. Architecture of the model for multimodal analysis.

ers available in the PyTorch library [28].The embeddings with POS-tagged features are generated and then fed into the LSTM layer. After the LSTM, a dense layer for a final classification was used.

As mentioned above, the CHAT format of the transcriptions has a lot of information and not only the plain text. There are special flags that represent, for example, a pause in the patient response or a mistaken word. As some symptoms of dementia include having difficulty finding certain words, this can lead to pauses in speech or using incorrect words. These special flags may provide valuable information for this task. The occurrence of these special tokens was counted and compared between control and dementia patients. Fig. 5 displays how these flags are present in the different transcriptions of both healthy individuals and those with dementia.

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

5. Evaluation

The accuracy measure on the test set was used to evaluate the results of the models. Accuracy represents the percentage of correct predictions among all the predictions made. The obtained results for each model, excluding those mentioned in Section 4.4 (the POS-tagging and the special flags models), can be seen in Table 3.

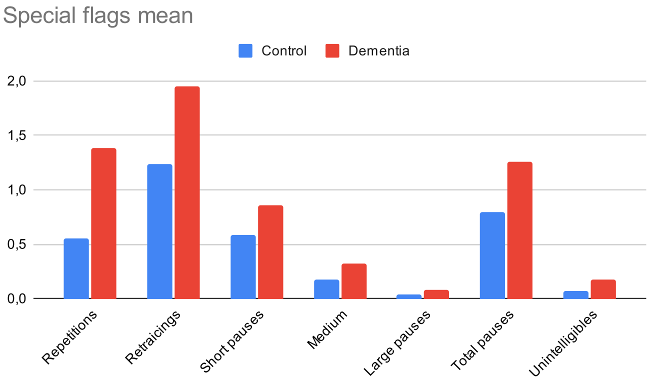


Fig. 5. Mean occurrence of special flags in dementia and control patients.

These models were discarded as they did not provide satisfactory results in our experiments. Both models failed to learn from the training set, resulting in random predictions by the model, with an accuracy of around 55%. Although there is a notable difference in the mean *repetitions*, *retracing*, *pauses*, and *unintelligible* flags between dementia and control patients (see Fig. 5), these features did not provide accurate predictions. The POS-tagging model also obtained similar results and was therefore discarded prior to the combination experiments.

Table 3
Comparison of the models evaluated.

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

The best results were obtained by the text model which used BERT embeddings along with a bidirectional LSTM (model *Text 2*). The use of this bidirectional LSTM improved the results obtained by using only the embedding of the `[CLS]` token for this task. It is worth noting that the textual data provided significantly better results than the audio content. The multimodal models proposed (*Multimodal 1* and *Multimodal 2*) did not improve the results obtained using only textual data.

Finally, Fig. 6 provides a more detailed view of how the best model (*Text 2*) performed on dementia and healthy cases separately.

Using this information, other metrics such as precision, recall, and F1-score can be calculated, which are relevant in measuring the quality of the model's predictions for dementia cases. For the *Text 2* model, all three metrics achieved the same result of 91.11%, which is a slight improvement over the accuracy obtained. This improvement is due to the fact that these metrics only consider the correct predictions of dementia cases, and the model had a higher percentage of correctly predicting dementia cases due to the larger number of dementia samples in the corpus.

6. Corpus analysis

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

First of all, the length (number of words) of the texts provided by healthy patients and patients with dementia were analysed. The result of this analysis is shown in Table 4.

The table shows that, on average, the conversations uttered by dementia patients are 10% shorter than those uttered by healthy patients (450.94 words and 503.90 words, respectively). There is also more homogeneity in terms of the number of words used in patients with dementia, as they show a lower standard deviation. Looking at the median, which is less sensitive to the presence of outliers, it also indicates the presence of shorter texts in the case of patients with dementia (404 words compared to 432 words of healthy patients). Fig. 7 and Fig. 8 show the distribution of the length of the conversations in both dementia and control patients in more detail.

6.1. N-gram frequency count

Before performing additional analysis it was necessary to carry out a series of preprocessing tasks to clean the data. To this end, the NLTK[3] library was used to lowercase all the texts, remove punctuation marks, and remove *stopwords*, that is, commonly used words in English language that do not provide useful information (e.g. "the", "a", "and").

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

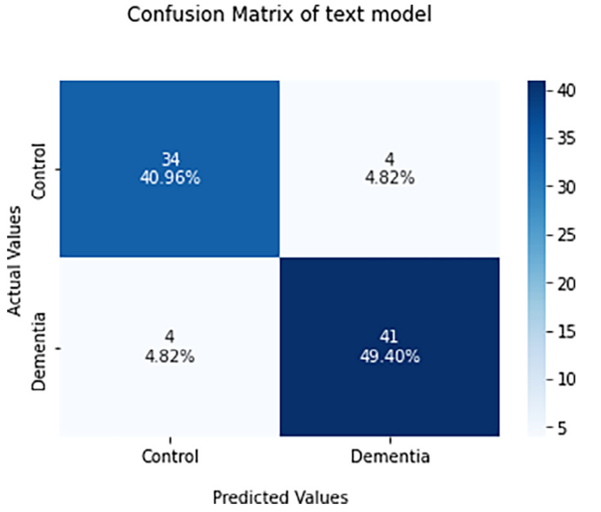


Fig. 6. Confusion matrix of the best text model over the test set.

Table 4
Central tendency measures for the length (number of words) of the conversations of dementia and control patients.

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

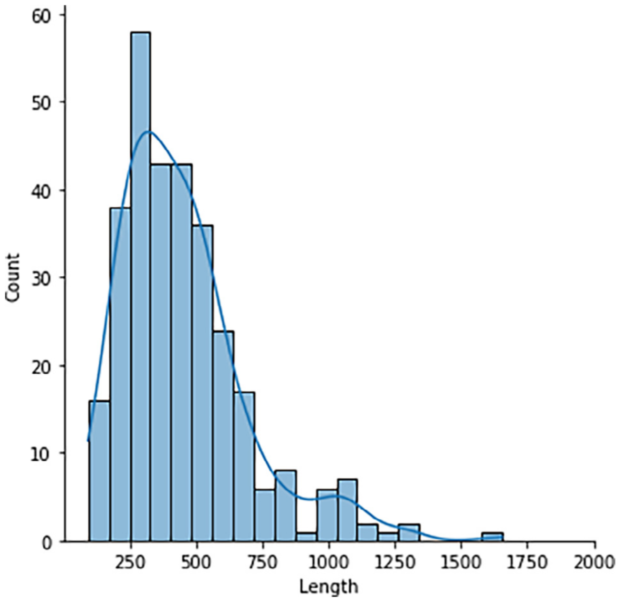


Fig. 7. Histogram of the length (number of words) of conversations by dementia patients.

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

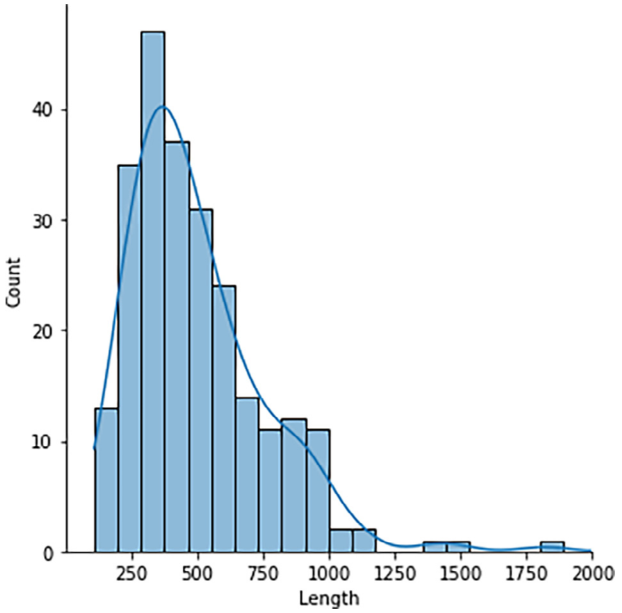


Fig. 8. Histogram of the length (number of words) of conversations by control (healthy) patients.

Table 5
List of 50 most frequent n-grams for dementia and control (healthy) patients.

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

out. Table 5 shows the 50 most frequent unigrams and bigrams for dementia and healthy individuals.

Taking a closer look at these lists, many words are common between the two different classes, but there are differences in some terms. As analyzed in Section 7, one token that can be observed in both classes but is much more common in the dementia class is the word well. Another remarkable token to distinguish between the classes is the word something, which is present in the list of unigrams of dementia patients but not in the control group. This may arise from the difficulty some dementia patients experience in properly recognizing items in the image, causing them to use a generic word instead.

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

6.2. Polarized Weirdness Index

In addition to the n-gram frequency count, an analysis was carried out by computing the Polarized Weirdness Index (PWI) [31] of the unigrams and bigrams in both dementia and healthy texts in order to extract the most characteristic words of each one. The PWI is a variant of the Weirdness Index (WI) [1], which is a metric to retrieve words characteristics of a special language with respect to their common use in general language. The intuition behind WI is that a word is highly weird in a specific corpus if it occurs significantly more often in that context than in a general language corpus. Given a specialist and a general corpus, the metric can be described as the ratio of its relative frequencies in the respective corpora. In the case of PWI, the metric compares the relative frequencies of a word as it occurs in the subset of a labeled corpus by one value of the label against its complement. In the present work, the PWI is used to compare the prevalence of words in dementia and healthy utterances.

"Table 6 displays the top 20 unigrams and bigrams extracted from the samples of dementia and healthy participants based on the PWI metric. As mentioned in the previous subsection, the bigrams of dementia patients contain the word something, while it is absent in the healthy patients' list.

6.3. Feature selection

In addition to frequency count, a feature selection procedure using χ^2 [29] was applied to identify what unigrams were considered as most relevant in order to differentiate between dementia and healthy texts. Before applying χ^2 it is necessary to transform every post into a numerical vector. The TF-IDF weighting schema was used to obtain a number representing the frequency of the token in the post (TF) and its prevalence in the dataset (IDF). The number of dimensions of each post vector is equal to the length of the vocabulary of the corpus, i.e., each dimension corresponds to one token. The value of the dimension is the TF-IDF weight if the token exists in the post or 0 otherwise. Texts were preprocessed in advance as in the previous analysis.

Table 7 shows the 50 best unigrams in order to differentiate dementia from healthy texts according to χ^2. This list shows some tokens that were appreciated with the previously obtained n-grams and also additional ones that are not as commonly used in the corpus but result in a good key to differentiate dementia.

Table 6
List of 20 most relevant unigrams based on PWI for the dementia and control (healthy) samples in the dataset.

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

7. Model explainability

Following the premise of the previous section, with the aim of identifying why textual models work so well in the multimodal corpus, the *Transformer Interpreter* software [30] was used to obtain more information about how BERT makes its predictions.

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

After analysing the tokens influence on the test set, the most significant tokens identified were those used when the patient starts to describe the image. The token Well at the beginning of a sentence influences positively when the model predicts dementia. This in turn means that the use of that token influences negatively when the model predicts a healthy patient. In Fig. 9 and Fig. 10, there are two examples of how the word Well influences both decisions made by the model. In that representation, the words highlighted in red will influence negatively the decision and those highlighted in green will influence it positively. The more intense the color, the more important the word has in the final decision.

Similarly, the use of other tokens at the beginning of a sentence positively influences the prediction of dementia, as shown in the case of Okay, which can be observed in Fig. 11 and Fig. 12. These tokens are used to introduce the sentence before describing the image, and the model gives great importance to them for the final decision of predicting dementia or not. This behavior can be observed in several other words used at the beginning of a sentence, such as So.

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

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

The trend of hesitation and uncertainty in speech can be reflected by analyzing the length of the audio samples in the datasets. Recordings of people suffering from dementia usually are longer than healthy ones, having around 20% more duration in the files. This can be visually appreciated in Fig. 13, where a histogram of the lengths is displayed comparing both control and dementia classes. In the figure, even though there are fewer samples of healthy patients, there are more samples in the leftmost part, representing less time of audio. And in contrast, in the rightmost part, the dementia classes predominate over the other.

One remarkable point is that the same word used in different contexts can have different influences. This is an important factor because not only are the words important, but also the way they are used in each situation. Additionally, the influence of a token can also vary from one sample to another, resulting in different scores between patients. Another remarkable point is that although there may be cases where a word has a very negative influence, the model can still predict the other class. This is exemplified by the use of Well for a healthy patient prediction.

8. Conclusions and future work

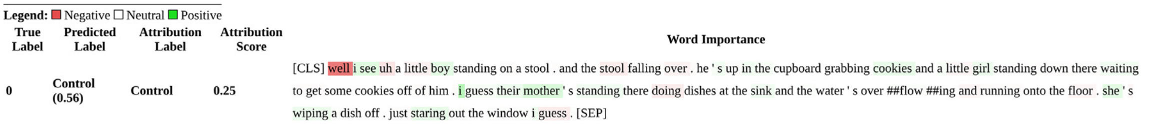
In this work, different models and approaches have been tested on the task of multimodal dementia detection. The detection of dementia, especially in the early stages, can help patients in order to improve their quality of life through different treatments.

In order to obtain the architecture of the different proposed models, a research on the state-of-the-art approaches to multimodal classification was carried out. Among the different approaches, the following one was implemented and tested: the use of CNNs for audio classification and the use of Transformers (BERT model) for text classification, including the combination of both. The model that best worked in this task was based on BERT, using only textual information. In order to explain this result, an analysis of the textual part of the dataset was carried out, including an explainability approach to determine the influence of specific words in determining the nature of the patient (dementia or healthy).

For future work, the trained model will be applied to other mental diseases that have similarities with dementia, such as aphasia disease. Another interesting future work is the addition of video modality, as well as the analysis of emotions shown in the different modalities. The aim of this is to try to identify relations between the expressed emotions in detecting symptoms of dementia. For this purpose, and with the addition of a video modality, an in-depth analysis of the facial expressions of the patients is planned, with the aim of identifying patterns and obtaining useful information, such as emotions. Facial expressions are not the only

Table 7
List of 50 most relevant unigrams according to χ^2.

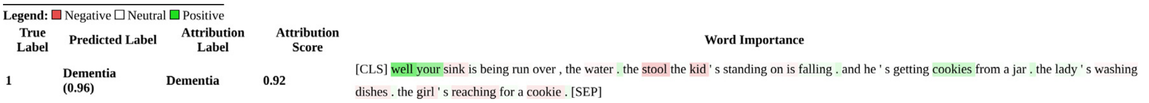
here, is, blowing, open, this, overflowing, laughs, down, window, reaching, wind, out_of, quiet, finger, action, while, moving, mouth, who, spilled, stepping, gonna, run, mother, the, um, curtains, nose, be, something, her, faucet, thing, breeze, about, they, growing, are, counter, well, get, hm, hand,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



Legend: 🟥 Negative □ Neutral 🟩 Positive

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

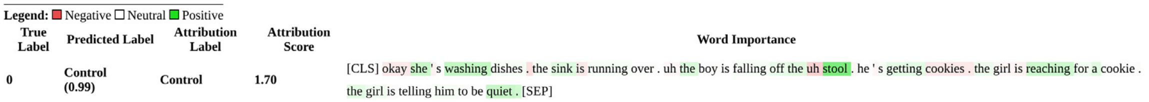
Fig. 9. Influence of the word `Well` in the prediction of a healthy patient.



Legend: 🟥 Negative □ Neutral 🟩 Positive

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

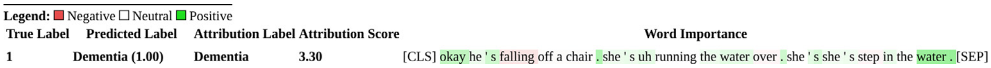
Fig. 10. Influence of the word `Well` in the prediction of a dementia patient.



Legend: 🟥 Negative □ Neutral 🟩 Positive

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

Fig. 11. Influence of the word `Okay` in the prediction of a healthy patient.



Legend: 🟥 Negative □ Neutral 🟩 Positive

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

Fig. 12. Influence of the word `Okay` in the prediction of a dementia patient.

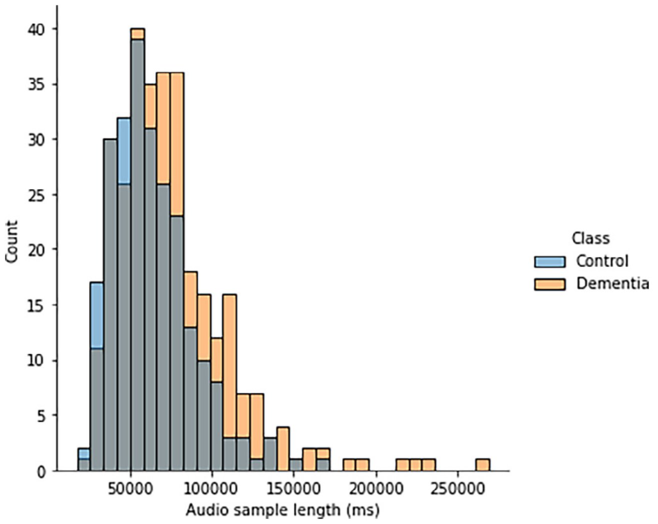


Fig. 13. Comparison of lengths of healthy and dementia patients records.

feature to be analyzed, as the evolution of the patient's pose and movements during a task or conversation can also provide insights. Cognitive diseases such as dementia can negatively affect these movements. With all this information, the work can be extended for a more complex treatment of the data.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

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

References

[1] K. Ahmad, L. Gillam, L. Tostevin, University of surrey participation in TREC8: weirdness indexing for logical document extrapolation and retrieval (WILDER), in: Voorhees, E.M., Harman, D.K. (Eds.), Proceedings of The Eighth Text REtrieval Conference, TREC, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA, 1999. pp. 1–8. URL: http://trec.nist.gov/pubs/trec8/papers/surrey2.pdf.

[2] H. Akbari, L. Yuan, R. Qian, W. Chuang, S. Chang, Y. Cui, B. Gong, VATT: transformers for multimodal self-supervised learning from raw video, audio and text, 2021. CoRR abs/2104.11178. URL: https://arxiv.org/abs/2104.11178, arXiv:2104.11178.

[3] Alzheimer, 2022. Dementia and language. https://www.alzheimers.org.uk/about-dementia/symptoms-and-diagnosis/symptoms/dementia-and-language.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, VQA: Visual Question Answering, in: International Conference on Computer Vision (ICCV), 2015, pp. 1–25.

[5] J.T. Becker, F. Boiler, O.L. Lopez, J. Saxton, K.L. McGonigle, The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis, Arch. Neurol. 51 (1994) 585–594.

[6] R. Chakraborty, M. Pandharipande, C. Bhat, S.K. Kopparapu, Identification of dementia using audio biomarkers, 2020. URL: https://arxiv.org/abs/2002.12788, 10.48550/ARXIV.2002.12788.

[7] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, 2018. CoRR abs/1810.04805. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

[8] R. Haulcy, J. Glass, Classifying alzheimer's disease using audio and text-based representations of speech, Front. Psychol. 11 (2021), https://doi.org/10.3389/fpsyg.2020.624137, https://www.frontiersin.org/article/10.3389/fpsyg.2020.624137.

[9] He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. CoRR abs/1512.03385. URL: http://arxiv.org/abs/1512.03385, arXiv:1512.03385.

[10] Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K.

W., 2016. CNN architectures for large-scale audio classification. CoRR abs/1609.09430. URL: http://arxiv.org/abs/1609.09430, arXiv:1609.09430.

[11] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861. URL: http://arxiv.org/abs/1704.04861, arXiv:1704.04861.

[12] Huang, G., Liu, Z., Weinberger, K.Q., 2016. Densely connected convolutional networks. CoRR abs/1608.06993. URL: http://arxiv.org/abs/1608.06993, arXiv:1608.06993.

[13] Karakostas, A., Briassouli, A., Avgerinakis, K., Kompatsiaris, I., Tsolaki, M., 2017. The dem@care experiments and datasets: a technical report. CoRR abs/1701.01142. URL: http://arxiv.org/abs/1701.01142, arXiv:1701.01142.

[14] Karlekar, S., Niu, T., Bansal, M., 2018. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. CoRR abs/1804.06440. URL: http://arxiv.org/abs/1804.06440, arXiv:1804.06440.

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

[16] Li, L.H., Yatskar, M., Yin, D., Hsieh, C., Chang, K., 2019. Visualbert: A simple and performant baseline for vision and language. CoRR abs/1908.03557.

[17] Lu, J., Batra, D., Parikh, D., Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Advances in Neural Information Processing Systems, Curran Associates, Inc. pp. 1–13.

[18] Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B., 2020. Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge, in: Proceedings of INTERSPEECH 2020, Shanghai, China. URL: https://arxiv.org/abs/2004.06833.

[19] I. Ma, M. Guo, C.K. Lau, V. Kandalam, C. Naugler, Dataset of test volume and tests repeated for complete blood count and electrolyte panels from hospitals in a canadian province in 2018, Data Brief 29 (2020).

[20] B. Macwhinney, The childes project: tools for analyzing talk, Child Language Teaching and Therapy 8 (2000), https://doi.org/10.1177/026565909200800211.

[21] P. Mahajan, V. Baths, Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech, Front. Aging Neurosci. 13 (2021), https://doi.org/10.3389/fnagi.2021.623607.

[22] Martinc, M., Pollak, S., 2020. Tackling the adress challenge: A multimodal approach to the automated recognition of alzheimer's dementia. 10.21437/Interspeech.2020-2202.

[23] Mittal, A., Sahoo, S., Datar, A., Kadiwala, J., Shalu, H., Mathew, J., 2020. Multi-modal detection of alzheimer's disease from speech and text. CoRR abs/2012.00096. URL: https://arxiv.org/abs/2012.00096, arXiv:2012.00096.

[24] F. Negin, P. Rodriguez, M. Koperski, A. Kerboua, J. Gonzàlez, J. Bourgeois, E. Chapoulie, P. Robert, F. Bremond, Praxis: Towards automatic cognitive assessment using gesture recognition, Expert Systems with Applications (2018).

[25] OpenAI, 2021. CLIP: Connecting Text and Images. URL: https://openai.com/blog/clip/.

[26] Ortiz-Perez, D., Ruiz-Ponce, P., Tomás, D., Garcia-Rodriguez, J., 2023. Deep learning-based dementia prediction using multimodal data, in: 17th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2022), Springer Nature Switzerland. pp. 260–269.

[27] Palanisamy, K., Singhania, D., Yao, A., 2020. Rethinking CNN models for audio classification. CoRR abs/2007.11154. URL: https://arxiv.org/abs/2007.11154, arXiv:2007.11154.

[28] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32. Curran Associates Inc, pp. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[29] Pearson, K., 1992. On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. Springer New York, New York, NY. pp. 11–28. 10.1007/978-1-4612-4380-9_2.

[30] Pierse, C., 2021. Transformers Interpret. URL: https://github.com/cdpierse/transformers-interpret.

[31] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resour. Eval. 55 (2021) 477–523, https://doi.org/10.1007/s10579-020-09502-8.

[32] Roberts, L., 2020. Understanding the mel spectrogram. https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53.

[33] Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., Rohrbach, M., Batra, D., Parikh, D., 2020. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf.

[34] C.G.M. Snoek, M. Worring, A.W.M. Smeulders, Early versus late fusion in semantic video analysis, in: Proceedings of the 13th Annual ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2005, pp. 399–402, https://doi.org/10.1145/1101149.1101236, DOI: 10.1145/1101149.1101236.

[35] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. arXiv:1706.03762.

[36] Warnita, T., Inoue, N., Shinoda, K., 2018. Detecting alzheimer's disease using gated convolutional neural network from audio data, pp. 1706–1710. 10.21437/Interspeech. 2018–1713.

[37] World Health Organization, 2023. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia.

[38] Zbontar, J., Knoll, F., Sriram, A., Muckley, M.J., Bruno, M., Defazio, A., Parente, M., Geras, K.J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzal, M., Romero, A., Rabbat, M.G., Vincent, P., Pinkerton, J., Wang, D., Yakubova, N., Owens, E., Zitnick, C.L., Recht, M.P., Sodickson, D.K., Lui, Y.W., 2018. fastmri: An open dataset and benchmarks for accelerated MRI. CoRR abs/1811.08839. URL: http://arxiv.org/abs/1811.08839, arXiv:1811.08839.